UNIVERSITÉ DE CAEN - NORMANDIE

ÉCOLE DOCTORALE MIIS - ED 590

Mathématiques, information, ingénierie des systèmes

MÉMOIRE

pour l'obtention de l'

Habilitation à Diriger des Recherches

Présentée et soutenue par

Alexis LECHERVY

Travaux sur l'apprentissag frugal et la fusion de modalités

préparée au GREYC dans l'équipe Image

soutenue le 11 juillet 2025

Composition du jury :

Rapporteure	:	Valérie GOUET-BRUNET	-	LASTIG (DR - Univ. Gustave Eiffel / IGN)
Rapporteure	:	Céline HUDELOT	-	MICS (PR - CentralSupélec)
Rapporteur	:	David PICARD	-	IMAGINE (DR - École Nationale des Ponts et Chaussées)
Examinateur	:	Stéphane CANU	-	LITIS (PR - INSA de Rouen)
Garant	:	Frédéric JURIE	-	GREYC (PR - Unicaen)

Table des matières

Remerciements

La rédaction de ce mémoire est l'occasion de remercier chaleureusement toutes les personnes qui ont contribué, par leur soutien et leurs échanges, à mon parcours scientifique.

Mes premiers remerciements s'adressent aux membres de mon jury. Je suis particulièrement reconnaissant envers Valérie Gouet-Brunet, Céline Hudelot et David Picard pour l'honneur qu'ils m'ont fait en acceptant d'évaluer ce manuscrit et mes travaux de recherche. Je remercie également sincèrement Stéphane Canu d'avoir accepté de faire partie de ce jury en tant qu'examinateur et d'en avoir accepté la présidence.

J'adresse ma plus profonde gratitude à Frédéric Jurie. Je le remercie pour la confiance qu'il m'a témoignée durant toutes ces années au GREYC, pour sa disponibilité, ses conseils avisés et son soutien. Sa vision scientifique, sa grande expérience et sa bienveillance m'ont été d'une grande aide.

Je tiens également à renouveler mes remerciements à mes directeurs de thèse, Philippe-Henri Gosselin et Frédéric Precioso, qui m'ont initié à la recherche et m'ont transmis leur passion et le goût de la découverte. Je remercie aussi Matthieu Cord et Nicolas Thome qui ont encadré mon post-doctorat au LIP6 et dont les conseils ont été précieux pour la suite de ma carrière.

La recherche est une aventure collective, et je tiens à remercier les personnes avec qui j'ai eu la chance d'encadrer des thèses : Frédéric Jurie, Stéphane Pateux, Luc Brun, Olivier Étard, Samia Ainouz, Hind Laghmara, Youssef Chahir, Gaël Dias, Fabrice Maurel, Abder El Moataz, Aurélien Corroyer et David Tschumperlé. J'ai énormément appris à vos côtés, tant sur le plan scientifique qu'humain et j'espère que nous continuerons ces collaborations fructueuses. Je remercie également Abdel-illah Mouaddib, Gaele Simon et Bruno Mermet pour la thèse dont nous venons de débuter l'encadrement.

Ces années n'auraient pas été les mêmes sans l'environnement stimulant du GREYC. Je remercie l'ensemble de mes collègues de l'équipe Image, permanents et non-permanents, ainsi que tous les membres du laboratoire, chercheurs, enseignants, enseignants-chercheurs, personnels administratifs et techniques, pour leur aide au quotidien, les échanges fructueux et l'ambiance conviviale qui règne au laboratoire.

Je tiens à remercier tout particulièrement les doctorants, post-doctorants, ingénieurs et étudiants que j'ai eu le privilège de co-encadrer : Binod Bhattarai, Romain Négrel, Sovann En, Valentin Vielzeuf, Dennis Conway, Clément Benoist, Darshan Venkatrayappa, Thibault Durand, Paul Dequidt, Shivang Agarwal, Mathieu Séraphim, Lucas Deregnaucourt, Youva Addad, Raphaëlle Lemaire, Azamat Kaibaldiyev, et plus récemment Jérôme Cartier, Éric Hu et Nour El Imene Ait Salem. Cette habilitation n'existerait pas sans les idées que nous avons partagées et sans le travail remarquable que vous avez accompli pour leur donner vie. Ce mémoire est nourri de vos travaux, de votre curiosité, de vos idées et de vos réussites. Mes remerciements vont également aux étudiants et doctorants avec qui j'ai eu la chance de collaborer

plus ponctuellement, notamment Ilyass Moummad, Alexandre Perrier, François Ledoyen, Noémie Moreau, Prince Jha et Chinmay Rane. Nos discussions ont alimenté mes propres réflexions et ont enrichi le contenu de ce manuscrit.

Sur un plan plus personnel, mes pensées les plus affectueuses vont à ma famille. À mes parents, mes grands-parents, mon frère et mes sœurs, et à tous mes proches, votre soutien infaillible a toujours été une force.

Un grand merci à mes enfants, Nathalie et Mathis, qui sont un rayon de soleil dans ma vie et un réconfort précieux dans les moments de doute.

Enfin, je remercie infiniment ma femme, Liang, pour me supporter malgré mes nombreux défauts et pour son soutien de chaque instant. Le bonheur de t'avoir à mes côtés est mon plus sûr repère et je suis heureux que tu m'accompagnes sur les chemins de la vie.

Première partie

Synthèse de mes activités de recherche

CHAPITRE 1

Introduction

Sommaire

1.1 Introduction générale

Le présent manuscrit est écrit en vue de mon passage de l'*Habilitation à Diriger des Recherches* (HDR). Il vise à démontrer ma capacité à initier et piloter des projets de recherche originaux, à encadrer de jeunes chercheurs (doctorants, post-doctorants, ingénieurs), et à contribuer activement à l'avancement des connaissances en apprentissage machine et vision par ordinateur.

Depuis l'obtention de mon doctorat en décembre 2012, mes activités de recherche, menées principalement au sein du laboratoire GREYC (UMR 6072) de l'Université de Caen Normandie à partir de 2013, se sont inscrites dans un contexte dynamique. La dernière décennie a vu l'émergence de modèles d'apprentissage profond aux performances remarquables, mais souvent au prix d'une demande croissante en données et en ressources de calcul. Simultanément, la nécessité de traiter des informations de natures diverses pour appréhender la complexité du monde réel s'est accrue.

Face à ces défis et opportunités, mes travaux se sont progressivement articulés autour de deux axes complémentaires et interdépendants :

— **L'apprentissage frugal** : Comment développer des modèles performants tout en étant économes en données (données limitées ou faiblement annotées) et/ou en ressources de calcul (complexité algorithmique, mémoire) ?

— **L'apprentissage multimodal** : Comment fusionner et exploiter efficacement des informations issues de sources hétérogènes (capteurs variés, texte, image...) pour améliorer la robustesse et la pertinence de nos systèmes ?

Ce manuscrit synthétise l'ensemble de mes travaux de recherche, offrant une vue d'ensemble de mes publications, projets et collaborations. En resituant mes contributions par rapport aux avancées récentes

de la littérature, je souhaite illustrer la pertinence et l'évolution de ma démarche scientifique. Il met également en lumière mon expérience acquise en matière d'encadrement scientifique et de gestion de projets. Une vision prospective de mes orientations futures est également esquissée.

Ce document se concentre sur l'analyse et le positionnement de mes travaux, sans entrer dans le détail excessif des développements mathématiques, afin de privilégier la vision d'ensemble. Pour une compréhension technique approfondie, le lecteur est invité à consulter la littérature citée ainsi que la sélection d'articles reproduite en annexe. Compte tenu de l'évolution rapide du domaine, j'ai choisi de discuter mes travaux, y compris les plus anciens, à la lumière des connaissances actuelles, quitte à ce que certains choix techniques d'alors puissent sembler aujourd'hui datés. Pour des raisons de concision et de cohérence thématique, seule une partie de mes publications est incluse ; la liste exhaustive est disponible en fin de document.

1.2 Contexte

Ce document offre une présentation synthétique de mes activités de recherche menées entre ma prise de fonction en tant que maître de conférences en 2013 et aujourd'hui, en 2025. Ces travaux ont été réalisés au sein du laboratoire Greyc (Groupe de Recherche en Informatique, Image et Instrumentation de Caen) UMR 6072. Ils sont réalisés dans la continuité de mes travaux de thèse dont une rapide synthèse est faite dans la section suivante.

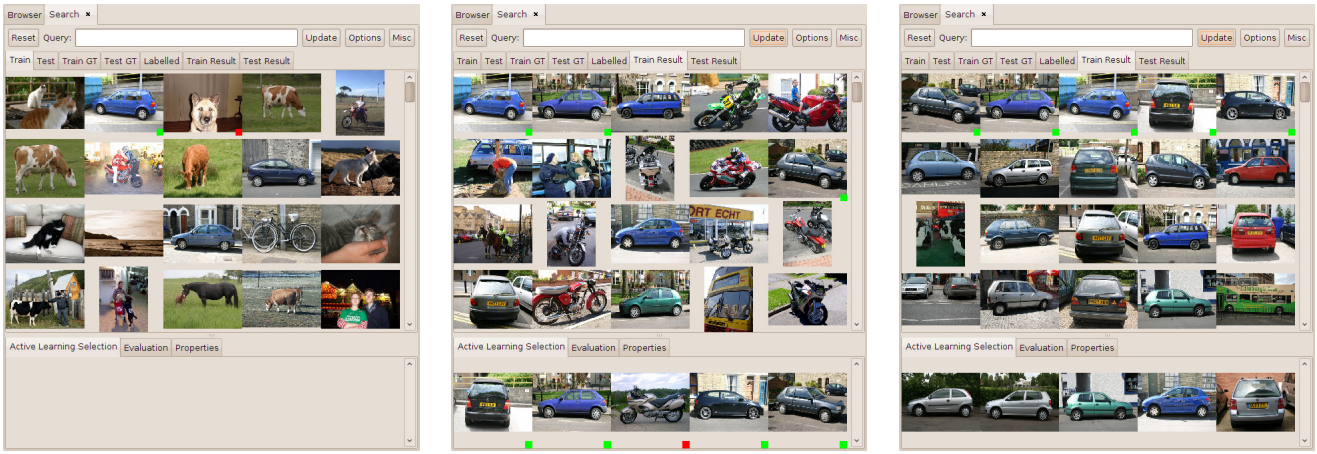
1.2.1 Thèse

Mes travaux de thèse (2009-2012), dirigés par Philippe-Henri Gosselin et Frédéric Precioso au laboratoire ETIS et intitulés "Apprentissage interactif et multi-classes pour la détection de concepts sémantiques dans des données multimédia", constituent le point de départ de mes activités de recherche. Ils exploraient déjà des questions liées à la réduction de l'effort d'annotation via l'apprentissage interactif (une forme de frugalité en données) et à l'apprentissage de fonctions noyaux via des méthodes de Boosting, jetant ainsi les bases de mes intérêts ultérieurs pour l'efficacité et la pertinence des représentations en apprentissage automatique. Les deux sections suivantes détailleront brièvement ces travaux de thèse avant de développer les recherches menées depuis 2013 dans le reste de ce manuscrit.

1.2.1.1 Méthode de boosting interactif

Références des travaux associés : [Lechervy 2010b, Lechervy 2010a]

La construction de base de données pour l'apprentissage machine est un processus long et coûteux. Il est néanmoins nécessaire pour obtenir des résultats en adéquation avec les attentes des utilisateurs. Il est en effet tout à fait possible avec un même jeu de données non labellisés de réaliser différentes labelisations cohérentes qui cependant ne regrouperait pas les mêmes exemples ensembles. La supervision



(a) (b) (c)

FIGURE 1.1 – Protocole expérimental interactif. (a) Au départ, on annote deux images. Une voiture dans la classe recherchée et un chien non recherché. (b) Le système fait un premier classement et propose sur la dernière ligne de nouvelles images à annoter. (c) Après quelques annotations, on converge vers la classe recherchée et il n'est plus nécessaire de continuer l'annotation.

est alors indispensable pour guider le choix vers une classification répondant aux besoins voulus. Afin de réduire les besoins en annotations, plusieurs stratégies ont été proposées, en particulier les méthodes d'apprentissage actif permettant d'introduire l'utilisateur dans la boucle d'apprentissage. La recherche interactive introduit un "dialogue" entre l'utilisateur et le système en vue de construire un ensemble d'apprentissage avec un nombre limité d'exemples (un exemple d'utilisation est donnée en figure Fig. 1.1). L'objectif est d'obtenir le minimum de supervision pour répondre à la tâche demandée tout en limitant l'effort d'annotation. La littérature a proposé plusieurs stratégies pour répondre à cette question, mais les méthodes de *Boosting* étaient moins explorées dans le contexte de l'apprentissage actif. Les méthodes de Boosting reposent sur la combinaison de plusieurs classifieurs (dits "faibles") pour en construire un meilleur (dit "fort"). La construction de l'ensemble des classifieurs faibles dans lequel est réalisé le choix de ces derniers, est un élément important et des techniques ont été proposées pour le construire dynamiquement, dans le cadre d'analyse de données vidéos. Notre travail a consisté à adapter ces techniques au contexte de la recherche interactive afin de proposer un nouvel algorithme de boosting interactif.

Mes travaux ont principalement contribué sur deux points : la méthode de sélection active des exemples à annoter et la construction de l'ensemble des classifieurs faibles.

Les méthodes de boosting reposent sur des classifieurs faibles qui doivent être simples et rapides à calculer. Nous proposons une construction des classifieurs faibles à partir des images sélectionnées. Ils consistent à calculer des distances entre une image d'entrée et des zones spécifiques de l'image génératrice du classifieur. Cela nous permet d'avoir des classifieurs faibles construits directement à partir des images choisies. Le choix des exemples suivants à annoter est réalisé de manière à sélectionner les exemples tels que s'ils étaient annotés apporteraient des classifieurs faibles qui maximisent le critère de sélection de l'algorithme de boosting, mesuré sur les exemples précédemment annotés pondérés de

manière équivalente. Notre méthode de choix des exemples est spécifique aux approches de boosting et entraine à la fois une augmentation des images annotées et des classifieurs faibles disponibles.

Mes travaux ont été réalisés sur une approche d'apprentissage de classement d'image inspirée de la méthode RankBoost [Freund 2003] et comparé aux méthodes d'approche active sur les SVM de l'époque illustrés sur la base VOC 2006.

1.2.1.2 Méthode de boosting pour l'apprentissage de fonctions noyaux

Références des travaux associés : [Lechervy 2012b, Lechervy 2012a, Lechervy 2014]

La notion de produit scalaire est un élément clé de beaucoup d'approche d'apprentissage machine. Leur extension via des fonctions noyaux a ouvert un nouveau champ de possibilités, permettant notamment l'apprentissage de classifieurs non-linéaires en s'appuyant sur des classifieurs par hyperplan tel que les SVM. Néanmoins, le choix de la fonction noyau est une question difficile et il peut être pertinent de vouloir injecter des étapes d'apprentissage dans sa conception. Les approches de MKL (*Multiple Kernel Learning*) proposent d'apprendre une combinaison de plusieurs fonctions noyaux. Ces approches, bien que performantes, souffrent d'une augmentation significative du coût de calcul.

Afin de réduire ces contraintes, j'ai proposé une approche de construction de fonction noyau efficace en termes de mémoire et de temps de calcul, en m'appuyant sur un algorithme de boosting. Ma méthode construit itérativement l'espace de représentation où est effectué le produit scalaire de la fonction noyau, une dimension après l'autre. Le choix des nouvelles dimensions est réalisé afin d'optimiser un critère d'alignement avec une fonction noyau cible.

1.2.2 Après thèse

À la suite de ma thèse, j'ai poursuivi ma recherche par un postdoctorat au LIP6 à Paris, sous la supervision de Matthieu Cord et en collaboration avec Nicolas Thome. En 2013, j'obtiens le poste de maitre de conférence au département de Mathématiques/Informatique de l'université de Caen et je rejoins le laboratoire de recherche en informatique Greyc en intégrant l'équipe *Image*. Mes travaux de thèse sur la réduction de l'effort d'annotation et l'apprentissage efficace de représentations ont naturellement posé les bases des deux axes majeurs développés par la suite et présentés dans ce mémoire :

— L'apprentissage frugal que je présenterai dans le chapitre 2.

— L'apprentissage multimodal qui fera l'objet du chapitre 3.

1.3 Organisation du manuscrit

Le manuscrit est structuré autour de trois grandes parties, chacune visant à offrir une vue d'ensemble complète de mes activités de recherche, de publication et d'enseignement.

Synthèse des activités de recherche et présentation des thématiques de recherche

La première partie du manuscrit présente une synthèse détaillée de mes activités de recherche. Elle commence par une introduction générale qui situe le contexte et les objectifs de mes travaux.

Pour présenter de manière cohérente mes activités de recherche, j'ai regroupé mes travaux autour de deux thématiques principales :

— **L'apprentissage frugal** (chapitre 2) : Cette section explore l'apprentissage dans des environnements où les ressources sont limitées. Cela inclut les contextes de données limitées (section 2.1) et ceux avec des ressources de calcul faibles (section 2.2).

— **L'apprentissage multimodal** (chapitre 3) : Cette section se concentre sur l'exploitation de différentes sources d'information pour améliorer la performance des modèles d'apprentissage automatique.

Pour chaque thématique, je commence par un état de l'art actuel du domaine. Ensuite, je présente les travaux que j'ai réalisés en rapport avec ces thèmes et j'en fais une analyse critique en comparaison avec les avancées actuelles. La première partie se conclut par une ouverture sur les perspectives futures de mes recherches, offrant des pistes pour de futurs développements.

Publications

La deuxième partie regroupe une sélection d'articles que j'ai publiés, couvrant l'ensemble de mes activités de recherche. Cette section est essentielle pour les lecteurs souhaitant connaître les références des publications sous-jacentes à mes travaux ainsi que les détails des méthodes utilisées.

Activités d'enseignant-chercheur

La dernière partie du manuscrit est consacrée à une description détaillée de mes activités d'enseignant-chercheur. Elle constitue un CV détaillé de mon parcours académique, incluant mes expériences professionnelles, mes enseignements, et mes responsabilités administratives. Cette section permet aux lecteurs de comprendre le contexte dans lequel mes recherches ont été menées et comment elles s'intègrent dans mon parcours professionnel.

En conclusion, ce manuscrit vise à offrir une vue d'ensemble complète et détaillée de mes activités de recherche, de publication et d'enseignement, tout en mettant en lumière les contributions significatives que j'ai apportées dans les domaines de l'apprentissage frugal et multimodal.

CHAPITRE 2

Apprentissage frugal

Sommaire

Depuis environ une décennie, les progrès dans le domaine des architectures de réseaux de neurones ont conduit à des performances exceptionnelles dans de nombreuses tâches. Ce développement a été rendu possible par la construction de structures de réseaux de neurones de plus en plus complexes, caractérisées par une profondeur croissante et un nombre accru de paramètres. Cependant, cette complexité a un coût : l'apprentissage de ces réseaux devient de plus en plus ardu, exigeant des quantités de données toujours plus massives et des puissances de calcul considérable.

Les architectures de réseaux de neurones les plus récentes, avec leurs multitudes de paramètres, exigent désormais des ensembles d'entraînement composés de millions d'exemples et mobilisent des ressources de calcul qui ne sont souvent accessibles qu'aux géants technologiques. Cette réalité soulève non seulement des questions d'accessibilité et d'équité, mais aussi des préoccupations environnementales majeures liées à la consommation énergétique de tels entraînements et déploiements. Face à ces enjeux, il devient impératif de développer des approches alternatives, plus sobres en ressources et plus simples à mettre en œuvre, tout en cherchant à préserver, voire améliorer, les performances.

Ce chapitre aborde ce défi sous l'angle de l'*apprentissage frugal*. S'inspirant de la définition du Larousse – *frugal* signifiant «Qui se nourrit de peu, qui vit d'une manière simple» –, nous distinguons deux dimensions essentielles de la frugalité en apprentissage automatique :

1. **La frugalité des données** ("qui se nourrit de peu") : Elle vise à construire des modèles efficaces à partir de peu d'exemples d'entraînement, de données faiblement labellisées ou bruitées.

2. **La frugalité des ressources** ("qui vit d'une manière simple") : Elle concerne le développement de méthodes capables d'apprendre et d'opérer avec des ressources de calcul et/ou de mémoire limitées.

Ce chapitre est structuré autour de ces deux dimensions, présentant pour chacune un état de l'art, mes contributions spécifiques, et une discussion les resituant dans le contexte actuel et futur de la recherche. La section 2.1 traitera de la frugalité des données, tandis que la section 2.2 abordera la frugalité des ressources.

2.1 Frugalité des données

L'apprentissage automatique moderne est fortement dépendant de la disponibilité de vastes corpus de données souvent annotées, dont la création est souvent longue, coûteuse et parfois éthiquement problématique. La notion de frugalité des données émerge alors comme une réponse à ce défi, cherchant à maximiser l'information extraite des données existantes tout en minimisant le besoin d'en acquérir de nouvelles. Réduire la dépendance aux données massives est crucial non seulement pour abaisser les coûts, mais aussi pour accélérer les cycles de développement et rendre l'IA applicable dans des domaines où les données sont intrinsèquement rares ou difficiles d'accès.

Cette section se focalise sur les stratégies permettant d'apprendre efficacement avec des données limitées. Après une revue de l'état de l'art (section 2.1.1), je détaillerai mes contributions portant sur deux approches spécifiques : l'apprentissage *cross-domaine* (section 2.1.2), qui vise à transférer la connaissance acquise sur un domaine source vers un domaine cible disposant de peu de données, et l'apprentissage de métriques (section 2.1.3), qui permet de construire des espaces de représentation où la similarité sémantique peut être mesurée efficacement, facilitant l'apprentissage à partir de très peu d'exemples.

L'apprentissage cross-domaine propose de pallier le manque de données dans un domaine spécifique en intégrant des informations provenant d'autres domaines. Ces méthodes permettent d'obtenir des résultats intéressants même lorsque les données disponibles dans un domaine particulier sont limitées. En exploitant les similarités et les complémentarités entre différents domaines, ces approches ouvrent des perspectives intéressantes pour répondre aux défis de données de l'apprentissage machine.

L'apprentissage de métrique introduit une fonction de similarité qui permet de comparer des exemples entre eux à un niveau sémantique. Cette fonction est particulièrement utile pour entraîner des classifieurs avec très peu de nouvelles données étiquetées, ce qui est idéal pour des contextes d'apprentissage avec peu d'exemples ("few-shot learning") ou même avec un seul exemple ("one-shot learning"). En construisant une métrique robuste, il devient possible de généraliser à partir de très peu de données, tout en maintenant une haute performance.

2.1.1 État de l'art

L'apprentissage avec peu d'exemples ("few shot learning"), que j'ai nommé frugalité de données dans ce manuscrit, correspond à un champ d'étude qui a beaucoup évolué ces dernières années. Plusieurs études de synthèse ont été publiées récemment, notamment celles de [Song 2023, Chen 2023, Antonelli 2022, Jadon 2023, Gharoun 2024, Parnami 2022, Tian 2024, Wang 2021a].

L'objectif de ces approches est de permettre un apprentissage avec un nombre limité de données en introduisant des a priori dans les modèles. Plusieurs raisons peuvent motiver cette démarche. La plus évidente est la disponibilité limitée de données. Certaines applications ne disposent pas de grandes bases de données. Cela peut être dû à la difficulté de collecter des données, aux coûts potentiels d'annotation, ou à un déséquilibre entre les classes. On peut illustrer cela avec les cas de la recherche d'événements rares ou les problèmes éthiques liés à la constitution de certaines bases de données. Par exemple, il est parfois moralement impossible de réaliser certains tests de toxicité *in vivo* sur des sujets humains pour étudier la dangerosité de certaines molécules chimiques. L'apprentissage avec peu de données peut également être motivé par des considérations de rapidité et de réduction des coûts. En effet, avoir moins de données accélère les calculs et réduit significativement les besoins d'annotation et de stockage.

L'apprentissage avec peu d'exemples est une capacité humaine [Lake 2017] qui nous permet d'apprendre de nombreuses choses avec très peu de nouvelles informations. Par exemple, nous sommes capables de comprendre et d'appliquer les règles d'un jeu de société après un nombre limité de parties. Une description textuelle des règles nous permet même généralement d'apprendre directement à jouer, ce qui s'apparente à de l'apprentissage sans exemple ("zero-shot learning").

La principale problématique à laquelle on est confronté dans l'apprentissage avec peu d'exemples est la difficulté à bien généraliser l'erreur optimisée. En effet, les modèles ont tendance à sur-apprendre dans ce contexte, et l'erreur empirique est peu fiable en présence de peu de données. Il est donc indispensable d'introduire un a priori pour compenser ce défaut.

D'après la taxonomie proposée par [Wang 2021a], la littérature identifie trois endroits où des a priori peuvent être introduits pour réduire les besoins de données d'apprentissage : les données elles même, les modèles et les algorithmes.

2.1.1.1 Ajout d'a priori dans les données

La première famille de méthodes vise à pallier le manque de données en enrichissant la base d'apprentissage à l'aide de connaissance extérieur. Ces techniques sont généralement très dépendantes de la nature des données et des problèmes étudiés. L'augmentation peut se faire en changeant les données d'entrées uniquement, en changeant les labels associés à ces données uniquement ou en changeant les deux.

Concernant le changement des données d'origine, de nombreuses techniques d'augmentation ont été proposées [Wang 2024, Zhou 2024, Nanthini 2023], reposant principalement sur des invariances dans les problèmes abordés. Par exemple, dans le cadre des images, on utilise la translation, la rotation, le

retournement, le cisaillement, la mise à l'échelle, la réflexion, le recadrage ... Un lecteur intéressé peut se référer à [Mumuni 2022, Xu 2023] pour les augmentations sur les images et à [Feng 2021] pour les textes.

Une autre façon d'augmenter le nombre de données d'apprentissage est d'utiliser d'autres bases de données faiblement supervisées ou non annotées et de produire une annotation automatique. Cela peut se faire en propageant des annotations d'exemples similaires déjà annotées [Douze 2018], ou à l'aide d'une annotation automatique des exemples non supervisés par un réseau de neurones pouvant être celui en cours d'apprentissage [Lee 2013].

Enfin il est également possible de produire à la fois des données et des labels pour augmenter la variabilité d'une base de données. Cela peut se faire en faisant des fusions de plusieurs exemples en un. Ainsi Mixup [Zhang 2018a] propose de faire une somme pondérée de deux images et de leurs deux labels pour produire un nouvel exemple annoté. AugMix [Hendrycks 2020] propose une méthode analogue avec la fusion de trois images, tandis que d'autres auteurs vont s'intéresser à proposer de meilleures techniques de fusion des images [Lee 2020, Kim 2021, Kim 2020] ou des techniques de labelisation alternative [Inoue 2018, Chou 2020]. On peut également citer la méthode CutMIx [Yun 2019] et ses variantes [Walawalkar 2020, DeVries 2017, French 2020] qui s'appuient sur une combinaison de patch pour produire de nouvelles images annotées avec une combinaison de labels.

Certains auteurs proposent d'utiliser des architectures génératives pour produire de nouvelles données. Ainsi [Bowles 2018, Frid-Adar 2018, Kaur 2021] utilise des modèles génératifs adverses (GAN [Goodfellow 2014]) pour augmenter la base de données. Les GAN peuvent également être utilisé pour améliorer les données [Ashraf 2021, Kupyn 2018, Sharma 2019] ou proposer des vues et des poses différentes [Zhang 2022]. Il est par ailleurs possible d'utiliser des modèles de transfert de style pour introduire de la diversité à partir d'une base de données existante [Gatys 2016, Li 2018b]. Dans [Ruffino 2022], les auteurs produisent des images polarimétriques à partir d'un modèle génératif prenant en entrée des images RGB. Les images produites sont ensuite utilisées pour apprendre un modèle de réseau de neurones pour l'analyse de scène routière.

Ces différentes stratégies d'enrichissement des données que nous venons de voir, visent toutes à fournir plus d'informations sur les exemples au modèle. Une approche complémentaire consiste non pas à augmenter les données, mais à contraindre l'espace des modèles possibles ou à guider leur apprentissage, comme nous allons le voir dans la section suivante.

2.1.1.2 Ajout d'a priori dans les modèles

Afin de réduire les risques de sur-apprentissage lorsque l'on utilise peu de données, on peut réduire la taille de l'espace des hypothèses dans lequel on fait la recherche des paramètres du modèle à apprendre. En effet, plus l'espace des hypothèses est petit, plus le risque de sur-apprentissage est faible, néanmoins au prix d'un risque empirique possiblement plus important. L'ajout d'un a priori qui contraint le modèle dans un sous-espace possible peut se faire à l'aide de plusieurs techniques. On peut utiliser une

construction spécifique de descripteurs, une mémoire externe, un apprentissage multi-tâche, ou encore des modèles génératifs, que je vais détailler ci-dessous.

Avant l'avènement des techniques de deep learning, on utilisait des descripteurs conçus à la main qui contenaient intrinsèquement une expertise du problème étudié. L'arrivée des modèles appris de bout en bout à changer les habitudes en ne dissociant plus la conception des descripteurs et l'apprentissage d'un modèle sur la tâche cible. Il n'était alors plus possible d'injecter des connaissances dans la représentation des données. Depuis quelques années, nous observons un retour de la dissociation de la construction d'un descripteur universel et de l'apprentissage spécifique d'un modèle sur une tâche cible. Ces descripteurs ne sont désormais plus conçus à la main, mais sont les résultats d'un apprentissage d'une représentation [Bengio 2013] ou d'une métrique [Li 2023b]. Ces nouveaux descripteurs peuvent être utilisés directement en ajoutant une simple couche linéaire que l'on apprend pour résoudre la tâche finale (on parle de *linear probing* [Kumar 2022]) ou ils peuvent être intégrés dans une architecture plus globale et servir de point de départ à un modèle apprenable de bout en bout. On peut distinguer différentes approches d'apprentissage de représentation : les méthodes supervisées, non-supervisées et auto-supervisées.

Les méthodes supervisées de construction de descripteurs tendent à disparaitre. Leur usage est restreint à des tâches que l'on aurait spécifiées et les représentations construites ne peuvent pas être appliquées dans d'autres contextes. Cependant parmi les approches nécessitant une supervision, on peut citer [Tian 2020b] qui propose de trouver via une méthode de méta-learning, la meilleure représentation des données pour apprendre par la suite un classifieur linéaire à partir de cette représentation pour différentes tâches. Les auteurs montrent qu'une bonne représentation permet d'avoir des résultats comparables à un modèle profond entièrement entrainé. On peut également citer la méthode DIABLO [Jacob 2020], qui repose sur un bloc d'attention basé sur un dictionnaire appris pour construire des descripteurs d'images. Ce dictionnaire, composé de vecteurs prototypes, permet de sélectionner et d'agréger les caractéristiques pertinentes des images selon plusieurs stratégies (*feature-wise* ou *dimension-wise*). Les labels des classes sont utilisés dans les fonctions de perte pour optimiser les distances entre les *embeddings*. Cependant, les classes utilisées pour l'entraînement sont distinctes de celles utilisées pour le test, afin d'évaluer la généralisation du modèle. La performance est mesurée à l'aide du Recall@K, qui évalue la capacité du modèle à retrouver des images de la même classe que l'image requête parmi les K-images les plus proches dans le reste de la base de test.

Les méthodes non-supervisé de construction de représentation reposent sur l'hypothèse qu'il existe un espace de représentation des données de dimension inférieure facilitant les tâches que l'on cherche à résoudre. On peut distinguer les méthodes probabilistes, des méthodes d'apprentissage de variété. Un lecteur intéressé par plus de précision sur ces approches peut se référer à [Bengio 2013].

Nous avons vu ces dernières années de nombreuses techniques de construction de représentation s'appuyant sur des mécanismes d'auto-supervision. On est parti de l'idée d'utiliser des méthodes de générations de pseudo-labels pour enrichir les bases de données pour s'orienter progressivement vers des méthodes où les pseudo-labels peuvent être construit directement avec le contenu des données elle-même sans supervision.

L'apprentissage contrastif [Jaiswal 2021, Hu 2024] est une méthode d'apprentissage auto-supervisée qui vise à apprendre des représentations en rapprochant les exemples similaires en terme de sémantique et en éloignant les exemples dissemblables dans l'espace latent construit. L'objectif est d'entraîner un modèle à distinguer les paires positives des paires négatives, en s'appuyant sur une métrique de similarité pour comparer les plongements des images. Par opposition, les approches non contrastives [Grill 2020, Chen 2021, Caron 2021, Assran 2023] se concentrent exclusivement sur le rapprochement des paires positives, évitant le besoin (parfois coûteux) d'échantillonner des négatifs. Ces deux familles connaissent une forte croissance depuis 2021.

Ces méthodes n'utilisent pas d'annotations manuelles et créent des labels directement à partir des données elles-mêmes via une *tâche prétexte*.

Les différentes méthodes proposées dans la littérature se différencient principalement autour des éléments suivants :

— Les *tâches prétextes* utilisées,
— L'utilisation ou non de pairs négatifs (contrastif ou non contrastif),
— L'architecture d'apprentissage,
— Les fonctions de coût utilisées.

Le choix de la *tâche prétexte*, qui permet de générer des pseudo-labels à partir des données, est un élément important objet de nombreuses recherches. Il existe différentes catégories de *tâches prétextes* qui dépendent beaucoup de la nature des données à traiter.

Pour les images, il est possible par exemple d'utiliser des transformations visuelles via du floutage, de la distorsion de couleurs ou de la conversion en niveaux de gris ou de l'ajout de bruit gaussien comme dans [Caron 2020, Chen 2020a]. Ces techniques sont faciles à implémenter mais capture difficilement la sémantique haut niveau nécessaire à des tâches de classification fine. Il est également possible d'utiliser des transformations géométriques tel que redimensionnement, des recadrage aléatoire, des retournements ou des rotations [Chen 2020a]. Ces méthodes cherchent à acquérir la compréhension d'invariances spatiales. Certains proposent d'effectuer des tâches sur l'agencement spatial du contenu des images. Par exemple, le puzzle d'images (Jigsaw) incite le modèle à reconstruire une image à partir de ses parties mélangées [Misra 2020], tandis que l'ordre des trames dans une séquence vidéo exige que le modèle apprenne à réordonner ces éléments [Qian 2021, van den Oord 2019, Lorre 2020]. [Ouali 2021] propose une tâche contrastive spatiale et non global reposant sur un mécanisme d'attention permettant d'aligner les représentations locales des deux images. Enfin, il est possible d'utiliser plusieurs vues d'une même scène pour rapprocher les représentations d'images capturées sous divers angles [Sermanet 2018, Tian 2020a, Bachman 2019] ou sous divers modalités [Li 2019b].

Pour le textes, les *tâches prétextes* peuvent être la prédiction de token masqué (BERT [Devlin 2019]), la prédiction du token suivant d'une séquence de manière autorégressive (GPT [Radford 2018]) ou la réorganisation de phrase qui ont été mélangées aléatoirement (BART [Lewis 2020a]). Il est également possible d'apprendre à prédire si deux phrases sont consécutives ou voisines dans un texte. Ainsi BERT [Devlin 2019] apprend à vérifier si deux phrases se suivent et *Skip-Thought Vectors* [Kiros 2015] utilise

la prédiction de phrases voisines en s'inspirant de la méthode Skip-gram pour les mots.

Les techniques génératives à base de masquage utilisées pour le texte ont également été utilisés pour l'image (I-JEPA [Assran 2023]) et la vidéo (V-JEPA [Bardes 2024a]).

La manière de construire les ensembles de pairs positifs et négatifs est un point différenciant de certains papiers. L'approche classique suit le déroulement présenté dans [Ge 2021]. Pour ce faire, on utilise une donnée de référence (l'ancre) et une version augmentée de cette donnée, qui constituent une paire positive. Les autres données du *batch* traité ou de l'ensemble d'entraînement constituent les paires négatives. [Park 2020] découpe les images en patches, les éléments à la même position dans l'ancre et dans l'image augmentée constitue les pairs positives tandis que les autres patches de la même image correspondent aux pairs négatives. [Quan 2023] construit les pairs positives en prenant des vues de la même ancre et les pairs négatives en prenant une partie de l'objet d'intérêt de l'ancre et une partie de l'arrière-plan. [Jang 2023] propose de traiter les pairs négatives pas niveau de difficulté. Certaines méthodes utilisent un nombre d'exemples négatifs faibles [Zbontar 2021] tandis que d'autres (BYOL [Grill 2020], SimSiam [Chen 2021], DINO [Caron 2021], VICReg [Bardes 2022, Bardes 2024b]...) utilisent uniquement des pairs positives et sont non contrastives. La question du non effondrement sur une solution trivial dans ces cas a été étudié par [Tian 2021].

Les fonctions de coût généralement utilisé pour l'apprentissage sont des similarités cosinus. Certaines méthodes utilisent également la *Noise Constrastive Estimation* (NCE) [Gutmann 2010] ou des variantes tel que InfoNCE [van den Oord 2019].

Les architectures proposées sont soit :

— De bout-en-bout [Chen 2020a, van den Oord 2019]. Différents encodeurs sont utilisés pour l'ancre et la donnée augmentée.

— Utilise une banque de mémoire [Wu 2018, Misra 2020]. Des représentations intermédiaires issues d'itération précédentes sont utilisées pour une des branches.

— Utilise un mécanisme de momentum [He 2020] permettant de construire la banque de mémoire à la volée en s'appuyant en partie sur les itérations précédantes d'un des encodeurs.

— Utilise un rapprochement à un centre de cluster [Caron 2020, Xie 2016].

[Bendou 2022] reprend plusieurs idées déjà présentées pour concevoir une nouvelle architecture. Les auteurs proposent d'utiliser plusieurs backbones entrainés de manière supervisée et auto-supervisée à l'aide d'augmentation par recadrage des images. Les sorties de ces backbones sont ensuite concaténés, puis centrée vis-à-vis des classes et projeté sur la sphère unité.

Outre l'apprentissage de représentations générales, une autre stratégie pour intégrer des a priori dans les modèles est l' apprentissage multi-tâches, qui exploite les synergies entre différentes tâches, en combinant des tâches génériques à des tâches spécifiques aux problèmes étudiés [Zhang 2021b, Ruder 2017, Zhang 2018d, Thung 2018]. Différentes stratégies sont possibles. On peut partager une partie extraction de caractéristique commune et spécialiser les dernières couches du réseau comme illustré dans [Zhang 2018c, Hu 2018, Motiian 2017, Benaim 2018]. Par exemple, [Zhang 2018c] utilise cette approche pour la catégorisation visuelle fine en partageant les premières couches pour l'information gé-

nérique et en utilisant des couches finales distinctes pour des sorties spécifiques. De même, [Hu 2018] applique cette méthode à deux tâches sur du traitement du langage naturel pour des textes juridiques, en partageant une fonction d'encodage commune pour la description des affaires criminelles.

Une autre approche consiste à aligner les représentations ou des couches intermédiaires entre les tâches, comme le font [Yan 2015, Luo 2017].

L'ajout de contraintes durant l'apprentissage ou dans l'architecture permet ainsi de pallier le manque de données. Selon cette logique, les *Physics-informed neural networks* (PINNs) [Raissi 2019] proposent d'ajouter des contraintes issues de modélisations physiques différentiables pour apprendre des réseaux de neurones sur des problèmes physiques pour lesquels le nombre de points réels issus d'expérimentations est limité. Les *Fourier Neural Operators* (FNO) [Li 2021b] propose un nouvel opérateur, paramétré dans l'espace de Fourier, pour résoudre des équations aux dérivées partielles (EDP). Les modèles orientés par la logique [Ledaguenel 2024] permettent d'introduire des contraintes logiques entre les sorties du réseau, ce qui permet d'intégrer des connaissances a priori dans les modèles.

Enfin une dernière famille de méthode que l'on peut citer pour apprendre avec peu de données sont les techniques d'apprentissage sans exemple [Pourpanah 2022, Chen 2023, Cao 2023]. L'idée de ces méthodes est d'utiliser une autre modalité tel que le texte pour décrit les données à récupérer et ainsi ne pas avoir à utiliser de données annotées. Il faut pour cela disposer de représentations multimodales alignées tel que des données sémantiquement identiques dans chaque modalité soit représenté de manière similaire. On peut citer la méthode CLIP [Radford 2021] pour réaliser cela ainsi que les modèles de fondations [Gu 2023, Liu 2024, Madan 2024] qui tendant à se développer actuellement.

Au-delà de l'enrichissement des données ou de l'intégration de contraintes dans l'architecture ou la représentation, un troisième levier pour l'apprentissage frugal consiste à agir sur l'algorithme d'apprentissage lui-même, ce que nous verrons dans la partie suivante.

2.1.1.3 Ajout d'a priori dans les algorithmes

Le dernier levier qui peut être utilisé lorsque l'on s'attaque à des problèmes où l'on fait face à un faible nombre de données est l'ajout d'a priori dans l'algorithmes d'apprentissage. Cela peut se faire soit en initialisant correctement les méthodes d'apprentissage pour partir d'une solution plus proche de l'optimum possible, soit en utilisant des méthodes de méta-learning pour choisir le chemin d'optimisation le plus efficace.

La technique la plus simple à mettre en œuvre est le *finetuning* [Yosinski 2014, Raffel 2020], qui consiste à reprendre un modèle déjà appris sur une autre tâche et à l'utiliser comme point de départ pour un nouveau modèle. On remplace généralement la couche finale de classification par une nouvelle couche, puis on lance un apprentissage sur notre nouvelle tâche, qui dispose de peu d'exemples. Les modèles pré-entraînés peuvent varier en fonction des modalités traitées ; par exemple, on peut utiliser des architectures ResNet [He 2015] ou ViT [Dosovitskiy 2021] pour les images, et BERT [Devlin 2019], RoBERTa [Zhuang 2021], EuroBERT [Boizard 2025], BART [Lewis 2020a], T5 [Ravaut 2024], GPT-4

[OpenAI 2024], Mistral-7B [Jiang 2023], LLama [Touvron 2023] pour le texte...

Cependant, les ressources de calcul nécessaires à l'apprentissage des modèles peuvent rester importantes, et de nouvelles approches ont été proposées pour remédier à ce problème. Parmi celles-ci, on trouve les adaptateurs [Houlsby 2019], le *Low-Rank Adaptation* (LoRA) [Hu 2021], le prompt tunning [Lester 2021] et le préfix-tuning [Li 2021a]. Les adaptateurs ajoutent des modules supplémentaires au modèle pré-entraîné, permettant ainsi de capturer des représentations spécifiques à la nouvelle tâche sans modifier les poids du modèle original. Le *Low-Rank Adaptation* (LoRA) consiste à ajuster uniquement une petite partie des poids du modèle, en utilisant des matrices de bas rang, ce qui permet de réduire considérablement la mémoire et les calculs nécessaires. Le prompt tuning consiste à ajuster des séquences de texte d'entrée (prompts) qui guident le modèle vers des réponses spécifiques sans modifier les poids du modèle. Enfin, le *préfix-tuning* consiste à ajouter un préfixe de paramètres supplémentaires à chaque couche du modèle, qui est ensuite ajusté pendant l'apprentissage, tout en gardant les poids du modèle pré-entraîné inchangés.

Dans la continuité de ces idées, des approches de construction de prompt efficace ont été proposés. [Lewis 2020b] introduit le *Retrieval-Augmented Generation* (RAG). Cette approche combine des modèles de génération de texte avec des systèmes de récupération d'informations, permettant de produire des réponses précises et contextuellement riches en sélectionnant des passages pertinents à partir d'une vaste base de connaissances, compensant ainsi la faible quantité de données d'entraînement disponibles. De plus l'intégration d'information externe par le prompt permet de surmonter les limitations des modèles appris sur les données potentiellement obsolètes.

Le choix de la meilleure initialisation peut également être fait en utilisant des techniques de meta-apprentissage [Finn 2017].

Certains articles s'intéressent à améliorer non pas l'initialisation de l'algorithme d'apprentissage mais directement les étapes de ce dernier. Au lieu d'utiliser une descente de gradient, ils proposent d'utiliser un algorithme appris par méta-learning pour prédire la prochaine étape de descente. On peut citer [Andrychowicz 2016] et [Ravi 2017].

2.1.2 Contributions sur l'apprentissage cross-domaine

Références des travaux associés : [Bhattarai 2016]

Mes travaux sur l'apprentissage cross-domaine [Bhattarai 2016]], présentés ici, s'inscrivent dans le cadre de la frugalité des données abordée précédemment. Plus spécifiquement, ils visent à pallier le manque de données annotées pour un domaine cible en exploitant les données d'un domaine source, ce qui relève principalement de l'introduction d'a priori au niveau des données et du modèle tels que catégorisés dans notre état de l'art (Sections 2.1.1.2 et 2.1.1.2). En effet, nous cherchons à la fois à transformer/aligner les données des deux domaines et à apprendre un modèle de régression adapté à cette représentation commune. On peut donc faire le lien avec les approches constrastive et l'utilisation

FIGURE 2.1 – L'apprentissage d'un estimateur d'âge se fait sur plusieurs domaines (un domaine majoritaire et un domaine cible peu représenté). Le test est effectué uniquement sur le domaine cible.

d'une tâche prétexe.

Un défi inhérent à la construction de bases de données est la présence de biais de représentation : certaines catégories ou caractéristiques sont souvent sur- ou sous-représentées, nuisant à la généralisation des modèles. Dans les travaux [Bhattarai 2016], nous nous sommes attaqués à ce problème dans le contexte de l'estimation de l'âge à partir de photographies de visages.

Ce type de problème était déjà bien étudié à l'époque, notamment par les travaux [Han 2013, Chen 2013, Song 2011, Thukral 2012]. Comme le souligne [Guo 2014], un problème majeur pour l'apprentissage dans ce contexte est le déséquilibre de représentation de certaines catégories d'individus dans les bases d'images. Il est impossible d'avoir toute la variété humaine équitablement répartie, car certains traits sont plus présents dans la population et peuvent varier d'une région à une autre. Cela entraîne des différences de performance importantes entre les individus en fonction des critères représentés dans la base d'apprentissage. Par exemple, un modèle entraîné uniquement sur des visages féminins aura de mauvaises performances sur des images masculines.

Bien que des études comme [Lou 2018] aient exploré la diversité dans l'apprentissage, elles se limitaient souvent à de petits jeux de données. Notre objectif était différent : comment transférer efficacement les performances d'un modèle entraîné sur un domaine source riche en données vers un domaine cible pour lequel nous ne disposons que de très peu d'exemples. (cf. Figure 2.1) ? Nous nous sommes alors demandé comment transférer les performances d'un modèle entraîné sur un domaine à un nouveau domaine, tout en limitant les exemples du nouveau domaine. Par exemple, cela revient à adapter notre estimateur d'âge, entraîné sur des visages féminins, aux visages masculins à partir de quelques photos d'hommes (cf. Fig. 2.1).

Pour répondre à cette question, nous avons proposé une approche optimisant *conjointement* le régresseur d'âge et une projection visant à aligner les deux domaines dans un espace latent commun. S'inspirant de travaux antérieurs du laboratoire sur l'apprentissage de métriques par projection [Mignon 2012], notre méthode repose sur une fonction de coût composite : une partie apprend une métrique de type Mahalano-

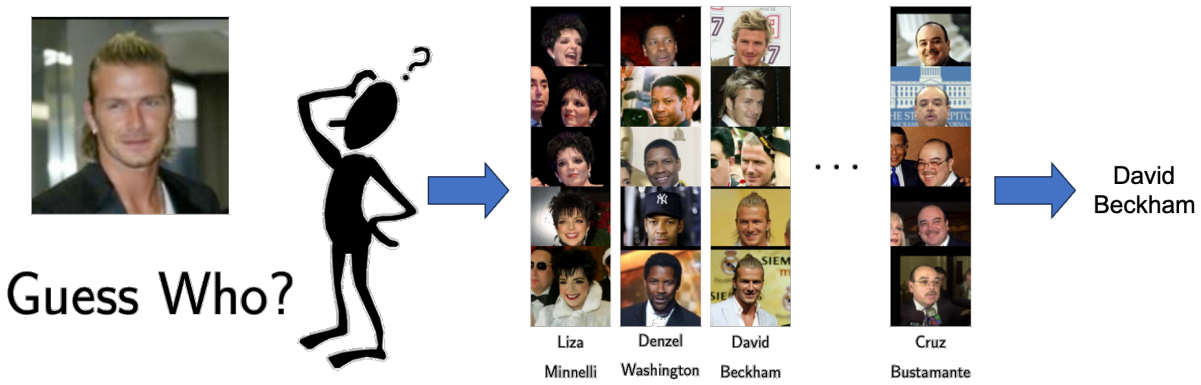


FIGURE 2.2 – Identification de personne à l'aide d'une métrique dans une base de visages.

bis pour rapprocher les distributions des deux domaines, tandis que l'autre optimise la régression dans cet espace aligné. L'optimisation est réalisée via une descente de gradient alternant entre ces deux objectifs.

Nous avons validé notre approche sur la base de données MORPH-II [Ricanek 2006], en suivant le protocole de [Guo 2014]. Cette base contient des images de visages diverses, tant par leur âge, leur genre que leur origine ethnique. Nous proposons un protocole spécifique pour montrer l'intérêt de notre approche dans le cas d'un apprentissage avec des domaines inégalement répartis.

2.1.3 Contributions sur l'apprentissage de métriques pour l'appariement de visage

Références des travaux associés : [Negrel 2015, Negrel 2016]

La constitution d'une base d'apprentissage exhaustive représentant toute la diversité des données potentiellement rencontrées en phase d'inférence est souvent illusoire. Prenons l'exemple d'un système de reconnaissance faciale pour le contrôle d'accès (Fig. 2.2) : il est impossible d'inclure a priori tous les futurs utilisateurs dans la base d'entraînement, sauf à se limiter à un nombre très restreint d'utilisateurs identifiés.

L'apprentissage "few-shot" ou "one-shot" offre une solution élégante en permettant d'apprendre une nouvelle identité à partir d'un nombre limité, voire d'un seul exemple d'apprentissage. Ces approches, dont nous avons présenté un état de l'art dans la section 2.1.1.2, reposent typiquement sur l'apprentissage d'une *métrique* permettant de mesurer une distance ou une similarité sémantique pertinente entre les exemples. Une fois cette métrique apprise, l'identification consiste simplement à comparer l'image d'entrée aux images de référence stockées pour chaque identité, sans nécessiter de ré-apprentissage global. L'ajout d'une nouvelle personne se réduit à l'ajout de ses images de référence. La clé réside donc dans la nécessité d'apprendre une métrique à la fois discriminante et rapide à calculer, même sur de très grandes bases.

Dans le cadre des travaux post-doctoraux de Romain Négrel [Negrel 2015, Negrel 2016], nous avons développé MLBoost [Negrel 2015], une méthode d'apprentissage de métrique par *boosting* spécifique-

ment conçue pour la recherche de visages à grande échelle. Cette méthode s'inscrit dans la continuité des travaux sur l'apprentissage de métriques de l'époque, tels que [Mignon 2012, Xiong 2014, Weinberger 2009, Guillaumin 2012, Davis 2007, Koestinger 2012]. Notre méthode apprend une métrique de type Mahalanobis entre les visages, pouvant être vue comme une projection dans un nouvel espace de représentation suivie d'une distance euclidienne. Elle serait aujourd'hui à rapprocher aux travaux réalisés sur les approches constratives présenté en section 2.1.1.2.

La solution que nous proposons repose sur une méthode de boosting dont le paradigme permet de ne pas avoir d'hyperparamètres à régler et offre une solution simple et robuste au sur-apprentissage. Inspiré par des travaux antérieurs sur le boosting de métriques [Shen 2012, Bi 2011], MLBoost construit l'espace de projection de manière itérative à partir de quadruplets d'images. Une contribution majeure a été de réduire la complexité de calcul (qui ne dépend plus que du nombre de paires d'images) grâce à une astuce inspirée de RankBoost [Freund 2003]. Pour l'inférence rapide, et suivant [Bhattarai 2014], la métrique apprise est intégrée à un processus de clustering hiérarchique semi-supervisé. L'approche a été validée sur LFW [Huang 2007], augmentée d'un million de distracteurs.

Nous avons ensuite développé une version améliorée de MLBoost dans [Negrel 2016], introduisant des solutions pour accélérer les temps d'inférence et réduire la consommation mémoire. MLBoost étant une méthode de boosting, elle consiste à combiner des métriques "faibles" afin d'en construire une plus "forte". Il est important pour une méthode de Boosting que les éléments faibles soient calculés rapidement. Par conséquent, nous avons cherché à réduire le coût de calcul de nos métriques faibles. Nous avons pour cela limité les métriques faibles à des métriques construites sur des matrices creuses de rang 1 (seule une partie aléatoire des vecteurs de représentation de données est utilisée par métrique faible). Nous proposons également de contrôler le rang de la métrique forte apprise à chaque itération en projetant la solution trouvée dans l'espace des matrices de rang R fixée. Ceci est fait en deux étapes. Tout d'abord, on approxime la métrique de Mahalanobis par une métrique de rang au plus égale à R. La métrique forte peut alors être vue comme une combinaison de R métriques faibles de rang 1. Nous calculons ensuite la meilleure pondération pour ces R métriques avant de passer aux itérations suivantes de l'algorithme de boosting. Cela peut être vu comme l'ajout d'une connaissance a priori dans l'algorithme d'apprentissage, qui à l'époque a été fixé à la main, mais qui pourrait être aujourd'hui trouvé par méta-learning (section 2.1.1.3).

2.1.4 Discussion

Rétrospectivement, et pour resituer nos contributions par rapport à la taxonomie de l'état de l'art de la frugalité en données (section 2.1.1), mes travaux sur **l'apprentissage cross-domaine** [Bhattarai 2016] agissaient principalement sur *le modèle* pour introduire un a priori visant à compenser le manque de données sur le domaine cible. Ils abordaient un problème précoce mais fondamental : l'adaptation de modèles face aux biais de distribution, ici dans le contexte de l'estimation d'âge. Son originalité résidait dans l'apprentissage *conjoint* d'une projection d'alignement des domaines et du régresseur spécifique à la

tâche, une approche moins courante à l'époque où les étapes étaient souvent découplées. Cela permettait d'optimiser directement la performance sur le domaine cible peu représenté. Cependant, cette approche reposait sur des descripteurs de type LBP, conçus avant l'avènement du *deep learning*. Sa capacité à gérer des écarts de domaine très importants ou des données complexes était limitée.

L'apprentissage de métriques par boosting (MLBoost) [Negrel 2015, Negrel 2016] se concentrait également sur l'*apprentissage d'un modèle* adapté au *few-shot*, mais avec une forte composante *algorithmique* (le boosting) pour l'efficacité, anticipant ainsi les préoccupations actuelles sur l'efficience (abordées plus en détail dans la section 2.2). L'objectif de ces travaux était double : obtenir une métrique performante pour l'appariement de visages *few-shot* et le faire de manière efficace pour des recherches à grande échelle. L'utilisation du boosting pour l'apprentissage de métriques était novatrice, offrant une alternative aux méthodes à base de SVM et méthodes à noyaux, sans hyperparamètres à régler et avec une certaine robustesse au sur-apprentissage. Les contributions sur la réduction de la complexité via l'astuce de type RankBoost, sur le contrôle explicite du rang [Negrel 2016] et l'utilisation de projecteurs faibles parcimonieux, abordaient directement l'enjeu de l'efficacité calculatoire. Là encore, l'utilisation de descripteurs non-profonds limitait la performance intrinsèque comparée aux approches actuelles basées sur des *embeddings* profonds. La scalabilité, bien qu'améliorée, restait un défi face aux volumes de données massives traitées aujourd'hui. Ces travaux ont démontré la pertinence du boosting pour l'apprentissage de métriques et ont mis l'accent sur l'efficacité algorithmique, un aspect qui demeure central avec la taille croissante des modèles. L'idée de contrôler le rang ou la sparsité des représentations apprises reste d'actualité dans les techniques de compression ou de quantification de modèles profonds. Ils illustrent la transition entre les approches "classiques" d'apprentissage de métriques et la nécessité d'intégrer des considérations d'efficacité dès la conception de l'algorithme.

Les contributions présentées dans cette section, bien que développées avant la prédominance actuelle du *deep learning*, abordaient déjà des questions centrales qui restent pertinentes aujourd'hui. Elles se concentraient sur l'apprentissage de métriques ou de projections pour créer des espaces de représentation adaptés, reposant sur des propriétés intéressantes, anticipant ainsi l'idée fondamentale qu'apprendre la représentation est aussi crucial qu'apprendre la tâche elle-même. Ces travaux peuvent être vus comme des précurseurs des méthodes actuelles visant à enrichir ou contraindre les représentations apprises par des connaissances a priori (décrites en section 2.1.1.2). À l'époque, nos approches reposaient sur des données supervisées. Depuis, le domaine a massivement évolué vers des méthodes d'apprentissage de représentations non-supervisées ou auto-supervisées, exploitant des tâches prétextes ingénieuses pour extraire des caractéristiques riches et généralisables sans supervision directe.

Après une période où les architectures *end-to-end* semblaient éclipser l'utilité de l'extraction de caractéristique, on observe un retour notable vers des approches en deux étapes : apprentissage d'une représentation (souvent auto-supervisée), puis apprentissage d'un modèle simple (souvent linéaire) sur cette représentation pour la tâche cible. L'enjeu n'est plus de concevoir manuellement des descripteurs comme ce fut le cas dans le passé, mais d'apprendre directement des représentations dotées de propriétés intéressantes. La question de l'apprentissage de représentations universelles se pose alors et fait l'objet

de recherche [Tamaazousti 2020].

Il faut cependant veiller à ne pas oublier les bonnes propriétés apprises lors de la conception des représentations. Certains scénarios, tels que l'apprentissage incrémental [Castro 2018, Petit 2024], où les nouvelles classes sont apprises les unes après les autres sans retour en arrière, illustrent bien les difficultés que peut engendrer un "oubli catastrophique" [French 1999] des premières classes rencontrées et les solutions à apporter pour s'en prémunir.

La question de l'apprentissage sur plusieurs domaines reste un enjeu majeur. La notion d'adaptation de domaine [Wilson 2020] consistant à adapter un modèle à de nouvelles données disponibles sur un nouveau domaine s'est enrichie du concept de généralisation de domaine [Zhou 2023, Ganin 2016, Li 2018a, Sun 2016] ne nécessitant pas de disposer de données du nouveau domaine. L'adaptation et la généralisation de domaine visent à rendre un algorithme robuste face à un changement de distribution entre les données d'entraînement et les données de test. Cette problématique est particulièrement pertinente lorsqu'il existe un écart significatif entre ces distributions, ce qui peut nuire aux performances des modèles. Cela peut par exemple arriver lorsque l'on utilise des données synthétiques pour apprendre ou lorsque les données disponibles sont d'un style, d'une modalité ou d'une perspective différentes des données de test.

L'avènement des modèles de fondation multimodaux, notamment ceux alignant les représentations textuelles et visuelles comme CLIP [Radford 2021] ou CoCa [Yu 2022], ainsi que leurs variantes [Li 2022a, Li 2023a, Xiao 2024] a marqué une avancée spectaculaire pour la généralisation de domaine grâce à l'utilisation de descriptions textuelles [Kwon 2022, Gal 2022, Zhou 2022b, Zhou 2022a, Vidit 2023, Fahes 2023]. Leur capacité remarquable à généraliser, due en grande partie aux vastes bases de données utilisées pour leur entraînement (comme DataComp [Gadre 2023] et LAION [Schuhmann 2022]), leur permet de produire des représentations de haute qualité qui capturent bien la distribution des "images du web". Cependant, leur efficacité doit encore être améliorée pour des applications spécifiques, notamment dans des contextes industriels où les données sont moins accessibles et mal représentées dans les bases de données généralistes académiques. De plus l'adaptation frugale de ces grands modèles reste un défi majeur.

2.2 Frugalité des ressources

La montée en puissance des réseaux de neurones profonds a révolutionné de nombreux domaines, mais elle a également mis en lumière des défis significatifs liés à l'efficacité des ressources. Les architectures de deep learning modernes, bien qu'extrêmement performantes, sont souvent gourmandes en calculs et en mémoire, limitant leur utilisation dans des environnements aux ressources contraintes (systèmes embarqués, objets connectés, smartphones...). Réduire l'empreinte calculatoire et mémoire des modèles est donc devenu une priorité de recherche, visant un équilibre optimal entre performance et efficacité.

Cette section explore les stratégies développées pour répondre à ce besoin de frugalité en ressources.

Après une revue de la littérature, je présenterai mes contributions récentes centrées sur l'apprentissage d'architectures multi-sorties, une approche permettant d'adapter dynamiquement le coût de calcul en fonction des contraintes ou de la difficulté de l'entrée.

2.2.1 État de l'art

Les réseaux de neurones efficaces en ressources ont émergé comme un domaine de recherche majeur ces dernières années, motivés par le besoin aussi bien académiques qu'industriels, de modèles capables de fonctionner efficacement sur des plateformes à ressources limitées. Ces réseaux sont conçus pour trouver un équilibre délicat entre la taille du modèle, les ressources utilisées, les exigences en mémoire et les performances. L'importance des réseaux de neurones efficaces en ressources réside dans leur capacité à offrir des solutions pratiques aux défis posés par les modèles traditionnels d'apprentissage profond, qui souffrent souvent de coûts de calculs toujours plus grands et nécessite de plus en plus de mémoire. En réponse à ces limitations, les chercheurs ont exploré diverses techniques et conceptions architecturales, la quantification [Jung 2019, Hubara 2016, Rastegari 2016, Jacob 2018], l'élagage de modèle existant [Han 2016, He 2019, Yang 2021, Liu 2017], et des optimisations spécialisées pour le matériel [Sarah 2022]. On peut également noter des architectures directement conçues pour être légères dont figurent MobileNet [Howard 2017, Howard 2019], ShuffleNet [Zhang 2018b], GhostNet [Han 2020] et SqueezeNet [Iandola 2016] ou des astuces de reparamétrisation [Ding 2021] permettant d'alléger les modèles lors des inférences en phase d'exploitation. De plus, les avancées récentes en recherche d'architecture neuronale (NAS) [Elsken 2019, Ren 2021] et en distillation de connaissances [Hinton 2015, lan 2018, Ba 2014, Gao 2023a] ont par ailleurs joué un rôle clé dans la production de modèles compacts et efficaces sans dégradation significative des performances. La distillation permet ainsi de transmettre les connaissances apprises sur de grandes modèles à des architectures plus légères via des mécanismes d'apprentissage enseignant/élève.

Une autre famille de méthode est important de citer : les techniques multi-sorties avec sorties anticipées [Huang 2018, Li 2019a, Yang 2020, Wang 2020b, Han 2022b, Han 2023, Phuong 2019]. Elles permettent aux modèles appris de sortir précocement pour certaines entrées, réduisant ainsi les calculs inutiles. La sortie anticipée dynamique utilise des critères adaptatifs, tels que des seuils de confiance [Teerapittayanon 2016] ou des mesures d'incertitude [Meronen 2024], pour décider s'il faut sortir précocement du processus d'inférence pour une entrée donnée. Ces approches présentent plusieurs avantages, notamment la réduction du temps d'inférence [Huang 2018], l'amélioration de l'évolutivité [Han 2022a] pour les environnements à ressources limitées et des économies d'énergie potentielles [Laskaridis 2021]. Bien que la sortie anticipée dynamique ait attiré l'attention pour ses avantages, des défis subsistent dans la recherche des bons critères pour équilibrer les gains de calculs et la préservation de la performance. Cependant, des recherches récentes ont montré des résultats prometteurs et ont été appliquées à plusieurs domaines, notamment la vision par ordinateur [Laskaridis 2021], le traitement du langage naturel [Hedderich 2021] et la reconnaissance vocale [Prabhavalkar 2023]. Certains auteurs ont tenté de s'écarter

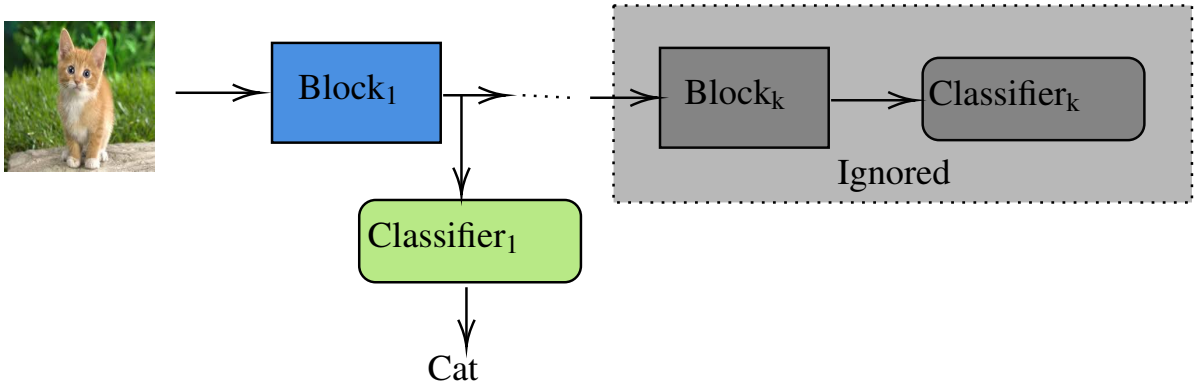


FIGURE 2.3 – Exemple d'architecture multi-sorties.

du cadre multi-sorties classiques, par exemple dans [Yu 2023], où les auteurs ont proposé une architecture de réseau de neurones s'appuyant sur des techniques de boosting.

Les approches les plus performantes dans les architectures multi-sorties sont basées sur BranchyNet [Teerapittayanon 2016], qui était l'un des premiers articles à proposer une telle architecture efficace, améliorée ensuite par MSDNet [Huang 2018]. Dans [Yang 2020], le *Resolution Adaptive Network* (RANet) a introduit l'idée de réaliser un apprentissage adaptatif de la résolution. En revanche, dans L2W [Han 2022b], les auteurs ont observé que MSDNet traite tous les échantillons pour toutes les sorties pendant l'entraînement, ignorant le comportement de sortie anticipée qui se produit pendant le test, et ont proposé de compenser cela en pondérant les échantillons d'entraînement en fonction de leur difficulté. Dans un article récent [Meronen 2024], les auteurs ont proposé une nouvelle méthode pour estimer l'incertitude dans les réseaux de neurones dynamiques, ce qui permet de mieux distinguer les exemples faciles et difficiles. Cette question a également été étudiée dans [Agarwal 2022] et nos travaux de [Addad 2024]. Il est également intéressant de mentionner l'approche présentée dans [Lee 2023], qui propose un mécanisme de distillation de connaissances pour les réseaux multi-sorties. Dynamic Vision Transformer (DVT) [Wang 2021b] et le *Coarse-to-Fine Vision Transformer* (CF-ViT) [Chen 2022a] propose des approches multi-sorties adapté aux *transformers*. Ces deux méthodes s'appuient sur le fait qu'il est sous-optimal de traiter tous les échantillons avec le même nombre de tokens. *Dynamic Perceiver* [Han 2023], préconise de dissocier les branches d'extraction de caractéristiques et de classification en raison du problème d'interférence du classifieur.

2.2.2 Contributions sur l'apprentissage multi-sorties

Références des travaux associés : [Addad 2023, Addad 2024, Addad 2025]

Comme souligné dans la section précédente, l'efficacité énergétique et calculatoire des modèles de réseaux de neurones est une question fondamentale pour leur déploiement à grande échelle. Cette question de la frugalité des ressources a été étudiée dans la littérature sous plusieurs angles, mais il apparaît

important de pouvoir construire des architectures dynamiquement adaptables aux ressources disponibles et/ou aux données traitées. En effet, si on ne tient pas compte de ces éléments, on est confronté à certaines limites qui peuvent être illustrées par les cas d'usages suivants :

— **Ressources de calcul variables :** Imaginons un modèle de deep learning fonctionnant sur un serveur de calcul partagé. Si la charge du serveur augmente ou si des calculs prioritaires arrivent, il serait souhaitable de réduire les ressources allouées au réseau de neurones tout en maintenant sa disponibilité, quitte à perdre légèrement en performance. Ce scénario est complexe à gérer avec des approches classiques, car il nécessite plusieurs architectures entraînées pour chaque niveau de ressource possible, ainsi que la capacité de les charger et décharger dynamiquement.

— **Difficulté variable des exemples :** On souhaite disposer d'un modèle capable d'adapter les ressources de calcul en fonction d'un critère de "difficulté" des exemples. Afin de diminuer les coûts de calcul, on cherche à utiliser la puissance de calcul minimale tout en maximisant les performances sur la tâche cible. Un modèle léger ne pourrait pas répondre à cette exigence, car il manquerait de performance, tandis qu'un modèle plus lourd consommerait plus de ressources que nécessaire pour la plupart des exemples.

Afin de répondre aux besoins évoqués précédemment, des architectures à plusieurs sorties ont été proposées [Rahmath P 2024, Bajpai 2025, Huang 2018, Li 2019a, Yang 2020, Wang 2020b, Han 2022b, Han 2023, Phuong 2019] (Fig. 2.3). Le principe est d'apprendre une unique architecture possédant plusieurs points de sortie (classifieurs) à différentes profondeurs. Si peu de ressources sont disponibles ou si l'exemple est jugé "simple" (basé sur un critère de confiance, par exemple), la décision peut être prise à une sortie précoce. Pour de meilleures performances ou pour des exemples "difficiles", les calculs se poursuivent jusqu'à des sorties plus profondes.

Dans [Addad 2023], nous proposons un modèle d'architecture multi-sortie inspiré par MSDNet [Huang 2018], qui est un modèle de classification d'images multi-échelles et multi-sorties conçu pour être efficace en termes du rapport coût de calcul/performances de classification. Nos contributions visent à améliorer les performances pour des budgets de calcul fixés, en recherchant le meilleur compromis entre performance et coût de calcul.

Pour cela, nous introduisons une couche d'amorçage initiale qui compresse l'image originale en une représentation plus compacte, mais tout aussi informative, réduisant ainsi le nombre total d'opérations en virgule flottante (FLOP). L'ajout de quelques FLOP pour cette couche est compensé par le gain généré sur le reste de l'architecture, qui traite des représentations plus légères.

À la suite de la couche d'amorçage, notre architecture reprend la succession de blocs basés sur la logique de MSDNet, traitant les images à plusieurs échelles de résolution. Cela nous permet de prendre en compte à la fois les informations globales (pour les petites échelles) et locales (pour les échelles les plus proches de la résolution initiale). Chaque échelle est traitée par une séquence de couches similaire à un réseau DenseNet [Huang 2017], organisées au sein de blocs successifs. Afin de transférer des informations entre les échelles, une fusion des représentations est effectuée des échelles les plus résolues vers les moins résolues. Notre deuxième contribution consiste à modifier la position de ce mécanisme de

fusion. Contrairement à MSDNet, qui fusionne les échelles à la même profondeur, nous proposons de décaler cette fusion d'une couche pour les échelles de meilleures résolutions. Cette opération augmente le coût de calcul au niveau de la fusion en général. Cependant, pour les couches au niveau des transitions entre blocs, ce report entraîne une fusion après une couche de compression des canaux qui réduit considérablement le nombre des calculs qui suivent, compensant ainsi les calculs supplémentaires introduits par les fusions des couches au sein des blocs. Nous démontrons la pertinence de notre approche par des expériences réalisées sur les bases *Cifar* et *ImageNet*.

Dans la continuité de ces travaux, nous avons amélioré l'architecture [Addad 2023] dans [Addad 2025]. Inspirés par le succès des *Transformers* introduit dans [Vaswani 2017] pour le texte et dans [Dosovitskiy 2021] pour les images, mais conscients de leur coût, notamment lié au grand nombre de *tokens* pour les images haute résolution, nous avons adopté une approche hybride similaire aux approches EfficientFormer [Li 2022b], LeViT [Graham 2021] et Mobile-Former [Chen 2022b]. Les premières sorties (moins coûteuses) utilisent une architecture convolutionnelle classique, tandis que la dernière sortie (la plus coûteuse, pour les cas difficiles) bénéficie de couches *Transformer* pour améliorer la performance. Le surcoût reste limité car ces couches n'interviennent qu'en fin de réseau. Nous avons également intégré des mécanismes de type *Squeeze-and-Excitation* [Hu 2019], qui peuvent s'apparenter à l'attention, au niveau des transitions entre blocs, permettant un recalibrage adaptatif des canaux pour focaliser le réseau sur les informations les plus pertinentes.

Enfin, dans [Addad 2024], nous avons proposé EEN (*Early Exit Neural Network*), une architecture où la politique de sortie (c'est-à-dire la décision de sortir à un niveau donné pour un budget cible) est *apprise conjointement* avec la fonction de classification durant l'entraînement. Contrairement à des approches comme Calibrated-DNN [Meronen 2024] (calibration post-entraînement) ou EENet [Ilhan 2024] (estimation d'incertitude pour guider la sortie), notre méthode intègre directement les contraintes budgétaires dans le processus d'optimisation. Nous proposons une fonction de coût qui optimise à la fois la classification (cross-entropie) et la probabilité de sortir précocement pour chaque exemple, trouvant ainsi un compromis explicite entre performance et coût de calcul moyen.

2.2.3 Discussion

Les travaux sur l'apprentissage multi-sorties présentés ici sont récents et s'inscrivent dans une thématique de recherche très active. Les résultats obtenus sont prometteurs et ouvrent la voie à des applications pratiques où l'adaptation dynamique des ressources est essentielle. Néanmoins, plusieurs questions et pistes d'évolution restent ouvertes.

Généralisation à d'autres problèmes que la classification

Actuellement, la recherche sur les architectures multi-sorties s'est majoritairement concentrée sur la classification. Leur potentiel pour d'autres tâches reste largement à explorer. Des travaux pionniers existent, comme [Bakhtiarnia 2021] qui adapte les mécanismes aux ViT pour la régression (comptage

de personnes dans des foules sur la base DISCO [Hu 2020]), ou [Xin 2021] qui apprend un score de confiance en fonction des zones des espaces de représentations de chaque sortie d'un réseau multi-sorties dérivé de BERT. [Kouris 2022] a également appliqué le concept à la segmentation sémantique.

L'application de l'apprentissage multi-sorties à d'autres tâches que la classification reste cependant embryonnaire et repose principalement sur des idées valables pour la classification. De nouvelles études sont à mener et ainsi d'autres applications que la classification pourraient tirer parti de ces techniques.

Application aux modèles de fondation multi-modaux

L'intégration d'architectures multi-sorties avec les grands modèles de fondation multimodaux (comme CLIP [Radford 2021], BLIP [Li 2022a], BLIP-2 [Li 2023a]¸, LLaVA [Liu 2023], SAM [Kirillov 2023], Florence-2 [Xiao 2024]...) représente une piste d'avenir particulièrement intéressante. Ces modèles, bien que très puissants, sont aussi très coûteux à l'inférence. Permettre une sortie précoce pour des requêtes simples ou en fonction des ressources disponibles pourrait considérablement améliorer leur efficacité et élargir leur champ d'application.

Propriétés pour l'apprentissage par transfert et l'adaptation de domaine

L'utilisation croissante de grands modèles pré-entraînés, ensuite spécialisés (*fine-tuning*) pour des tâches spécifiques, soulève la question des propriétés de transfert des architectures multi-sorties. Cela est d'autant plus indispensable dans le cas de scénario où les données d'entraînement sont peu abondantes et où un apprentissage directe ne donnerait pas de performance correcte. Il est alors important de se demander comment ces architectures se comportent-elles lorsqu'elles sont adaptées à de nouvelles tâches ou de nouveaux domaines ? Peu d'études ont abordé cette question. [Xin 2021] a identifié des difficultés pour le *fine-tuning* classique de modèles BERT multi-sorties et a proposé une stratégie d'apprentissage alternée. L'article propose une nouvelle méthode de *fine-tuning*. Classiquement, pour faire du *fine-tuning* sur du multi-sorties, on peut soit appliquer uniquement le *fine-tuning* sur la dernière sortie, puis *fine-tuner* chaque sortie avec un *backbone* gelé, soit faire un *fine-tuning* sur toutes les sorties en même temps en faisant une somme de toutes les fonctions de coûts. Ces deux solutions ne sont pas idéales, car dans un cas, seule la dernière sortie impacte l'ensemble du réseau, et dans l'autre cas, les différents classifieurs peuvent interférer négativement entre eux. Les auteurs proposent proposent une solution intermédiaire en alternant un *fine-tuning* uniquement de la dernière sortie et un *fine-tuning* sur toutes les sorties sans partie gelée. [Gao 2023b] suggère l'utilisation d'*adaptateurs* plutôt qu'un *fine-tuning* complet.

L'adaptation de domaine, qui consiste à adapter un modèle à une nouvelle distribution de données sur laquelle on dispose d'exemples non labellisés, a été très peu étudiée dans le contexte des architectures multi-sorties. [Jiang 2020] proposent une nouvelle méthode pour rendre l'adaptation de domaine opérationnelle sur des modèles tenant compte de la difficulté des exemples rencontrés dans l'usage de leurs ressources de calculs. Les auteurs se basent sur l'architecture multi-sorties, MSDNet [Huang 2018], et l'adaptent aux techniques de transfert de domaine. Les auteurs partent du constat que les méthodes

traditionnelles d'adaptation de domaine utilisant l'apprentissage d'un *discriminateur* de domaine se concentrent principalement sur une adaptation des représentations issues des dernières couches du réseau, négligeant l'adaptation des premières couches du modèle, ce qui est un souci pour une architecture multi-sorties. Ce choix est dû à des différences de transférabilité en fonction du niveau de profondeur des couches, comme le montre [Yosinski 2014], en partie dû à des problèmes de *vanishing gradient*. Pour surmonter cette limitation, les auteurs introduisent une technique de distillation du descripteur issu de la dernière couche, adapté aux nouveaux domaines, vers les autres sorties. Cette méthode permet d'améliorer la transférabilité des couches précoces, rendant ainsi l'adaptation de domaine possible pour les architectures multi-sorties. Ce type d'approche pourrait être utilisé sur les architectures que nous avons proposées qui sont des évolutions de MSDNet.

Mécanismes d'explicabilité

L'explicabilité des modèles d'apprentissage automatique est devenue un enjeu crucial, en particulier dans les applications critiques où les décisions des modèles ont un impact direct sur les individus. Les modèles dits «boîtes noires» souffrent d'un manque de transparence, ce qui pose des problèmes éthiques et pratiques. En effet, ces modèles peuvent parvenir à la bonne réponse pour de mauvaises raisons, en se basant sur des corrélations trompeuses ou en reproduisant des biais et des préjugés présents dans les bases de données utilisées pour leur apprentissage [Pedreschi 2019]. Cette problématique est d'autant plus préoccupante que les décisions automatisées influencent de plus en plus de domaines sensibles, tels que la santé, la justice et les finances.

Pourtant, cette question cruciale n'est actuellement pas abordée dans le contexte des architectures multi-sorties. Les méthodes traditionnelles d'explicabilité, conçues pour des modèles à sortie unique, ne garantissent pas nécessairement des solutions cohérentes entre les différentes sorties d'un modèle multi-sorties. Il est essentiel de développer des techniques spécifiques pour évaluer et expliquer les décisions prises à chaque sortie, afin de s'assurer que le modèle fonctionne de manière équitable et compréhensible à tous les niveaux.

Par ailleurs, il serait également intéressant de comprendre quelles parties d'une image ou d'un ensemble de données permettent de considérer qu'un exemple est "simple" ou "difficile". Cette compréhension pourrait conduire à des stratégies d'allocation des ressources plus efficaces. Par exemple, en prétraitant les données pour identifier les exemples simples, il serait possible de les faire sortir plus précocement des architectures multi-sorties, réduisant ainsi la charge computationnelle et améliorant l'efficacité globale du système.

En somme, l'explicabilité des modèles multi-sorties est un domaine de recherche inexploré qui pourrait néanmoins répondre à des questions nécessaires en cas d'usage réel de ces méthodes.

CHAPITRE 3

Apprentissage multimodal

Sommaire

Notre perception du monde réel est intrinsèquement multimodale. Nous ne l'appréhendons pas directement, mais nous en construisons une représentation mentale riche et nuancée en intégrant constamment les informations issues de nos différentes perceptions sensorielles : la vue, l'ouïe, le toucher, etc. Aucune de ces modalités, prise isolément, ne suffirait à nous offrir une compréhension complète de notre environnement. Pour illustrer cela, nous pouvons prendre l'exemple d'un cylindre vu en 2 dimensions. Ce dernier peut être perçu comme un cercle ou un rectangle selon l'angle de vue que l'on adopte. Avoir accès à la diversité des perspectives enrichit donc notre compréhension et la fusion des sources d'information peut permettre de mieux comprendre les propriétés de l'objet étudié. C'est précisément la *complémentarité* et la *redondance* entre ces flux d'informations qui nous permettent de lever les ambiguïtés, d'affiner nos jugements et, in fine, de mieux interagir avec le monde.

Paradoxalement, l'apprentissage automatique a longtemps reposé sur un traitement essentiellement unimodal de l'information, privant ainsi les machines d'une partie essentielle des indices disponibles pour l'interprétation du monde. Cependant, la nécessité de traiter des données de plus en plus complexes et hétérogènes a conduit à un intérêt croissant pour les méthodes d'apprentissage multimodal et de fusion de données. Ces approches visent précisément à exploiter la synergie entre différentes sources d'information pour améliorer la performance, la robustesse et la pertinence des modèles.

La nature des données à fusionner peut varier considérablement :

— **Données homogènes :** Issues de capteurs de même nature (par exemple, plusieurs caméras, plusieurs microphones, ou plusieurs électrodes EEG comme nous le verrons) ou de différentes parties d'une même donnée (par exemple, différentes régions d'une image). La fusion vise ici souvent à améliorer la couverture spatiale, à analyser plusieurs points de vue ou à réduire le bruit.

— **Données hétérogènes :** Provenant de sources de nature très différente, comme la fusion d'images et de texte, de vidéo et d'audio, ou de données tabulaires et de séries temporelles. La fusion cherche ici à exploiter la complémentarité fondamentale entre les modalités.

Au-delà de la nature des données, la question cruciale du *moment* et du *niveau* de la fusion se pose. Schématiquement, on distingue :

— **La fusion précoce (*early fusion*) :** Les données brutes ou des caractéristiques de bas niveau issues des différentes modalités sont combinées dès le début du processus, souvent par simple concaténation ou projection dans un espace commun.

— **La fusion tardive (*late fusion*) :** Chaque modalité est traitée indépendamment par un modèle spécifique jusqu'à l'obtention d'une décision ou d'un score. Ces décisions sont ensuite combinées (par vote, moyenne pondérée, etc.) pour obtenir la décision finale.

— **La fusion intermédiaire ou hybride :** Des stratégies cherchent à combiner les informations à différents niveaux d'abstraction au sein de l'architecture du modèle, tentant de bénéficier à la fois des interactions de bas niveau et de la robustesse des décisions unimodales.

Au cours de mes recherches, j'ai exploré ces différentes facettes de l'apprentissage multimodal. Ce chapitre présente plusieurs contributions illustrant divers aspects de la fusion d'informations. Après une revue de l'état de l'art (section 3.1), je décrirai nos travaux sur la fusion de signaux EEG issus de multiples électrodes pour l'analyse du sommeil (section 3.2.1). J'aborderai ensuite l'estimation de pose relative de caméras exploitant différentes modalités visuelles (section 3.2.2). Puis, je présenterai notre architecture *CentralNet*, conçue pour apprendre la position optimale de la fusion au sein d'un réseau de neurones (section 3.2.3). Enfin, je conclurai par nos travaux récents sur une méthode de fusion tardive robuste aux défaillances de capteurs, basée sur la théorie de Dempster-Shafer (section 3.2.4). Une discussion générale (section 3.3) synthétisera ces travaux et les replacera dans le contexte des évolutions actuelles.

3.1 État de l'art

La notion de multimodalité est un phénomène que nous expérimentons au quotidien. Le cerveau humain traite simultanément des informations provenant de multiples sources sensorielles, comme la vue, l'ouïe, le toucher, l'odorat et le goût... Cette interaction entre les sens permet une compréhension plus riche et nuancée de notre environnement. Par exemple, lors d'une interaction avec une personne, nous intégrons des éléments comme l'expression faciale, les gestes, l'intensité de la voix et le contenu des paroles, ce qui améliore notre interprétation des intentions et des émotions, réduisant ainsi les risques de malentendus. La gestion de la multimodalité est donc un mécanisme essentiel à notre perception du monde, dépassant les capacités de chaque sens pris isolément. Notre compréhension du monde se construit par l'agrégation d'informations provenant de diverses sources sensorielles. Chaque sens capte une facette unique de la réalité, mais c'est en combinant ces facettes que nous accédons à une représentation plus fidèle et complète de notre environnement. Ainsi, la multimodalité ne se contente pas de

compléter les informations ; elle enrichit notre perception en offrant une vue plus complète des concepts et des objets que nous rencontrons. Il est donc naturel d'étudier sa prise en compte dans le cadre de l'apprentissage machine.

[Baltrušaitis 2018] propose de définir 5 types de problèmes faisant appel à la multimodalité :

— **la représentation :** représenter l'information de la même manière quelque soit les données d'entrées.
— **la traduction :** la conversion d'une modalité en une autre. Cela englobe également les tâches de génération guidée par une modalité comme la synthèse de légende d'image ou la génération de contenu guidé par le texte.
— **l'alignement :** identification des parties communes entre modalités
— **la fusion :** joindre les informations des modalités afin d'améliorer les performances sur une tâche donnée.
— **le co-apprentissage :** transférer la connaissance d'une modalité à une autre.

Dans les travaux réalisés pour cette HDR, je me suis essentiellement consacré à la thématique de la *fusion*. Dans son livre «Fusion d'informations en traitement du signal et des images» de 2003, Isabelle Bloch énonce «La fusion d'informations consiste à combiner des informations issues de plusieurs sources afin d'améliorer la prise de décision.». Ces informations peuvent être issues de sources de données homogènes (parties différentes d'une même donnée, données issues d'un même capteur ou données de même nature...) ou de sources hétérogènes (ex : fusion image/texte, vidéo/son...). Même si l'objectif des méthodes de fusion est d'améliorer la prise de décision, cela peut se faire en exploitant la complémentarité des modalités ou en cherchant de l'information dans d'autres modalités pour combler de données manquantes, dégradés ou bruitées.

La littérature sur la fusion est composée de nombreux articles dont des articles de synthèses [Zhang 2021a, Atrey 2010, Cui 2023, Kalamkar 2023, Baltrušaitis 2018, Zong 2024, Zhao 2024, Zhang 2024]. Elle s'articule principalement autour des trois questions suivantes :

— Quand/Où faire la fusion ?
— Comment réaliser la fusion ?
— Comment effectuer l'apprentissage d'un modèle de fusion ?

Quand/Où faire la fusion ? La fusion de données multimodale peut se réaliser à différents niveaux. Une première approche consiste à assembler directement les données brutes, avant tout traitement, en superposant des canaux pour construire des images hyper-spectrales [Vivone 2023]. Bien que cette méthode soit simple à mettre en œuvre, elle peut devenir complexe à gérer si les données sont hétérogènes ou non alignées. Une autre technique de fusion consiste à se positionner au niveau des caractéristiques [Bhowmik 2014, Dechesne 2017, Ehatisham-Ul-Haq 2019]. Un premier traitement permet de construire des descripteurs pour chaque modalité, qui sont ensuite combinés en une seule représentation pouvant être traitée comme une seule modalité. Cette solution de fusion précoce offre plus de souplesse que la précédente, mais peut montrer ses limites lorsque certaines modalités sont trop prépondérantes ou, à

l'inverse, fortement dégradées [Zhang 2024].

La fusion peut également être effectuée de manière tardive [Zhang 2019, Cheng 2017, Guo 2024, Tong 2021a], au niveau de la prise de décision. Dans ce cas, on dispose d'une prédiction par modalité qui sont ensuite combinées. Ce type d'approche ne permet généralement pas d'avoir d'une interaction entre les modalités et de bénéficier de leurs complémentarités.

Enfin, certains auteurs [Zhang 2023a] proposent des solutions hybrides qui mélangent à la fois des fusions précoces et des fusions tardives. Certaines architectures, comme l'approche CentralNet que nous avons proposée, recherchent elles-mêmes les meilleures positions de fusion. Des approches similaires ont ainsi été proposées pour les approches de types *Transformers* par exemple dans [Nagrani 2021].

Il est cependant à noter que la frontière entre les trois paradigmes de fusion précoce, de fusion tardive et de fusion hybride tend à devenir de plus en plus floue avec les modèles de deep learning, comme le souligne [Zhao 2024].

Comment réaliser la fusion ? La fusion de données multimodale peut être réalisée de diverses manières, chacune ayant ses propres avantages et inconvénients. Parmi les méthodes les plus courantes, on trouve les opérations élémentaires telles que la concaténation, l'addition, la soustraction, la multiplication et la sélection du maximum. Ces opérateurs permettent de combiner directement les données brutes ou les caractéristiques extraites, offrant une solution simple, mais parfois limitée en termes de flexibilité et de robustesse.

Une autre approche consiste à utiliser des méthodes basées sur la classification, où chaque modalité est d'abord traitée individuellement pour produire des prédictions ou des scores, qui sont ensuite combinés pour une décision finale. Ce type de fusion tardive peut être réalisé à l'aide d'approche bayésienne, où les données de différentes modalités sont utilisées pour estimer une distribution de probabilité commune. Ces approches permettent de combiner les informations de manière probabiliste, en tenant compte des incertitudes et des corrélations entre les modalités. Il est également possible d'utiliser la théorie des fonctions de croyances pour réaliser ce type de fusion comme dans [Denoeux 2000, Tong 2021a].

Récemment, la fusion à l'aide de mécanismes d'attention a gagné en popularité [Wang 2020a, Huang 2020, Nagrani 2021, Prakash 2021, Rodríguez Bribiesca 2021]. Ces mécanismes permettent aux modèles de se concentrer sur les parties les plus pertinentes des données de chaque modalité, en attribuant des poids dynamiques en fonction de leur importance relative. Certain modèle tel que Florence 2 [Xiao 2024] représente chaque modalité au travers de tokens qui sont traités ensuite par des modèles de textes classiques.

Comment effectuer l'apprentissage d'un modèle de fusion ? L'optimisation d'un modèle de fusion multimodale peut être abordée de deux manières principales : l'apprentissage mono-tâche, qui se concentre sur une seule tâche, et l'apprentissage multi-tâche, qui vise à optimiser le modèle pour plusieurs tâches simultanément. Le choix de l'approche dépendra de la tâche à accomplir et de la nature des données. En apprentissage multi-tâche, la fonction de coût, qui mesure l'erreur du modèle, est souvent

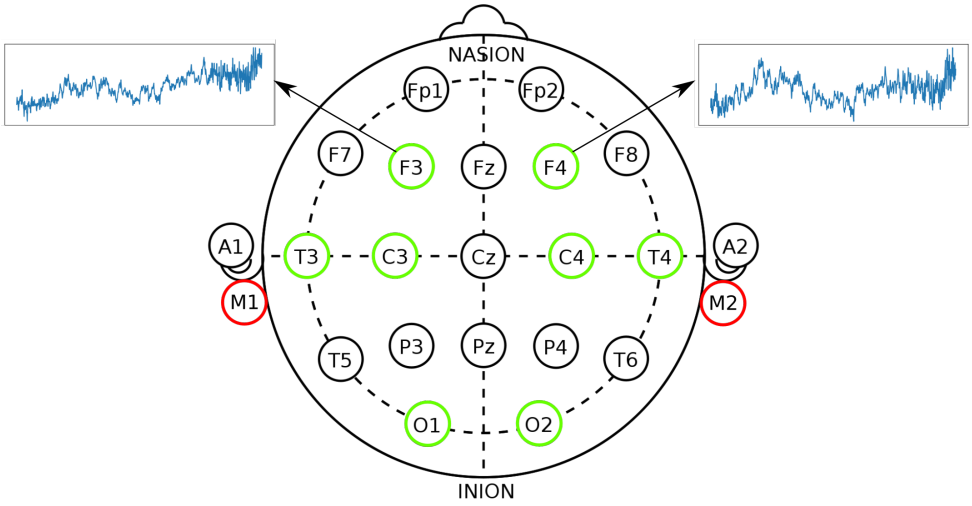


FIGURE 3.1 – Position en vert des 8 électrodes étudiées pour la détection du sommeil.

définie comme une combinaison pondérée des fonctions de coût de chaque tâche individuelle.

3.2 Contributions

3.2.1 Fusion de capteurs homogènes : l'exemple de l'estimation du sommeil

Références des travaux : [Dequidt 2023, Seraphim 2023b, Seraphim 2023a, Seraphim 2024b, Seraphim 2024a]

Dans le cadre de la thèse de Mathieu Séraphin et des travaux post-doctoraux de Paul Dequidt, nous nous sommes intéressés à la détection des phases de sommeil à partir de signaux électroencéphalogramme (EEG). L'étude du sommeil, processus physiologique essentiel, repose classiquement sur l'examen de polysomnographie (PSG), qui enregistre simultanément plusieurs signaux tels que l'électroencéphalogramme (EEG), l'électrooculogramme (EOG) et l'électromyogramme (EMG). Les études du sommeil sont généralement réalisées manuellement par des experts en annotant des périodes de 30 secondes selon cinq stades de sommeil : l'éveil (W), le sommeil paradoxal (REM) et le sommeil non-REM (N1, N2, N3) selon les normes AASM [Troester 2023]. Cette annotation manuelle est une tâche longue et fastidieuse, surtout pour des enregistrements de longue durée comme une nuit complète de sommeil. Pour automatiser ce processus, diverses approches ont été proposées, allant de l'authentification de motifs caractéristiques et l'apprentissage de modèles traditionnels [Güneş 2010, Herrera 2011, Liang 2012, Van Der Donckt 2023, Diykh 2016, Koley 2012] à des méthodes d'apprentissage profond [Supratak 2017, Seo 2020, Jia 2020, Phan 2022, Kontras 2024] qui tendent à se généraliser de plus en plus. La performance sur les classes minoritaires reste souvent un défi pour une application en clinique fiable.

Dans nos travaux, nous nous sommes principalement concentrés sur la modalité EEG, en fusionnant les informations issues de 8 électrodes spécifiques (Figure 3.1). Ce problème peut être vu comme une

fusion de données homogènes, où l'objectif est d'exploiter la redondance et la complémentarité des informations capturées par différents capteurs observant le même phénomène sous des angles légèrement différents.

Bien que les performances globales publiées dans la littérature soient aujourd'hui élevées, nous avons remarqué que certaines classes moins représentées pouvaient être mal classifiées remettant en cause l'usage en clinique de ces approches. En partant du constat du déséquilibre des classes des bases de données et ses conséquences sur les performances rapportées, nous avons dans un premier temps rigoureusement ré-évalué plusieurs méthodes de la littérature en utilisant des métriques adaptées (comme le *F1-score* macro-moyen et l'*accuracy* macro). Nous avons ensuite comparé ces approches à une méthode de *fusion précoce* simple présentée dans [Dequidt 2023] : les spectrogrammes temps-fréquence des 8 signaux EEG sont concaténés pour former une image multi-canaux, ensuite traitée par un VGG-16 modifié. Cette approche simple s'est avérée supérieure aux méthodes existantes sur les métriques équilibrées.

Explorant une voie différente, nous avons ensuite étudié la pertinence de la connectivité fonctionnelle pour cette tâche. Ce concept médical analyse les corrélations d'activité entre différentes zones cérébrales. Elle est particulièrement utilisée dans des études sur les états du sommeil [Wu 2012, Tagliazucchi 2012, Langheim 2011, Bouchard 2020]. Cependant, il semblerait que ces approches n'ont pas été explorées pour construire des méthodes automatiques d'annotation des états du sommeil. On peut néanmoins noter, l'utilisation de matrices de covariance des données EEG pour la réalisation d'interfaces cerveau-machine [Lotte 2007], et il nous est apparu intéressant de les étudier pour notre problématique. Ainsi, plutôt que de traiter directement les signaux bruts ou leurs spectrogrammes, nous avons proposé dans [Seraphim 2023b, Seraphim 2023a] d'utiliser les matrices de covariance 8×8 calculées entre les signaux EEG comme entrée d'un réseau de neurones. Ces matrices, qui sont semi-définies positives (SDP), capturent les relations de second ordre entre les électrodes. L'utilisation de matrices SDP nécessitant des architectures spécifiques, nous avons développé un modèle basé sur des *Transformers* [Vaswani 2017] adapté au traitement de séquences de telles matrices. Cette méthode peut être vue comme une fusion de niveau intermédiaire en deux temps. Par la suite, dans *SPDTransNet* [Seraphim 2024b, Seraphim 2024a], nous avons amélioré cette architecture en introduisant un mécanisme d'attention (*structure-preserving multi-head attention* SP-MHA), variante du mécanisme d'attention multi-tête, qui préserve explicitement la géométrie Riemannienne de l'espace des matrices SDP tout au long du réseau. Ces travaux ont montré l'intérêt en terme de performance de préserver la géométrie Riemannienne des données tous le long des traitements.

3.2.2 Estimation de pose relative et multi-modalité

Références des travaux associés : [En 2018a, En 2018b]

Après avoir présenté une méthode de fusion avec des données issues de capteurs homogènes, je vais maintenant présenter des travaux utilisant de données hétérogènes provenant de capteurs différents ou représentant des points de vue différents.

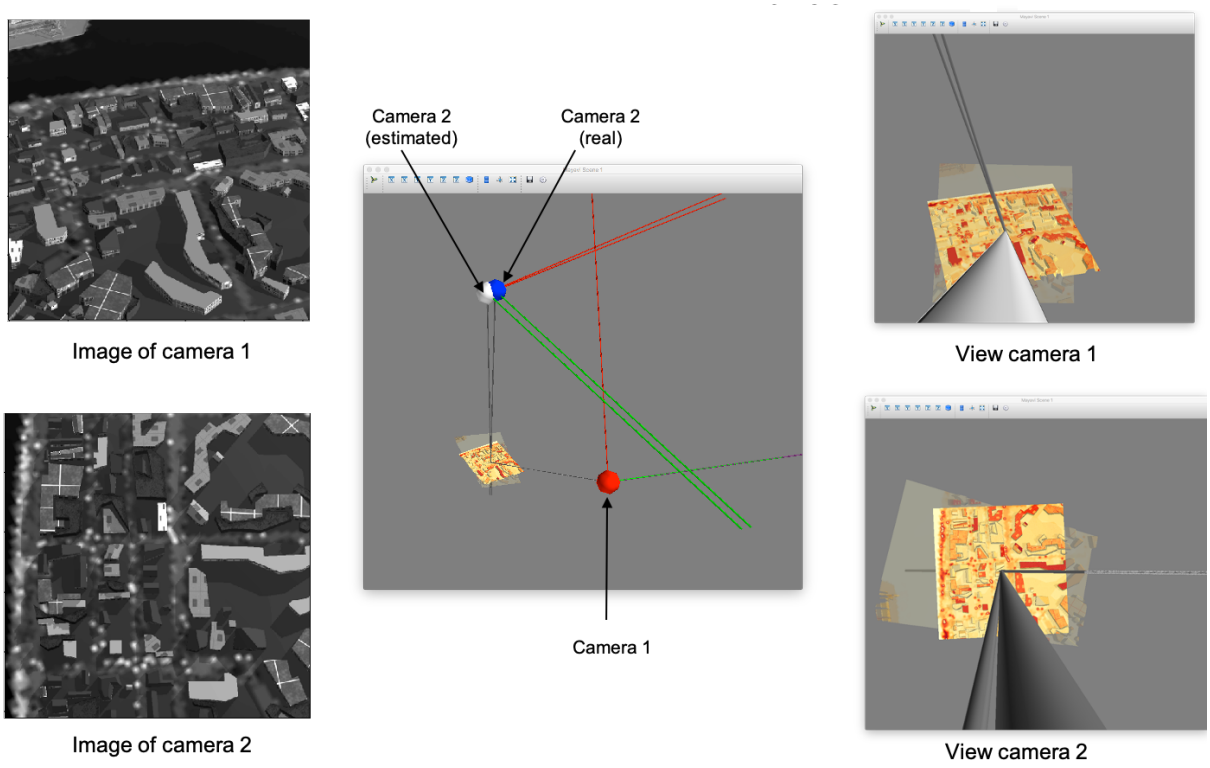


FIGURE 3.2 – Estimation de la pose relative entre deux caméras. Les images à gauche correspondent aux entrées du modèle, les images à droite correspondent aux vues reconstruites des deux caméras. Par exemple, l'image en haut à droite est la reconstruction de l'image en bas à gauche vue de l'angle de la caméra 1.

La localisation basée sur la vision (*Visual-Based localization*, VBL) est un domaine de recherche visant à estimer la pose d'un système (sa position et son orientation) à partir d'images prises de son environnement. Ce domaine joue un rôle important dans de nombreuses applications telles que l'initialisation des systèmes de réalités augmentée, l'asservissement visuel, l'odométrie visuelle, ou encore dans le cadre de la «*Structure from Motion (SfM)*», qui consiste à estimer la structure 3D d'un objet à partir d'une série d'images 2D et du « *Simultaneous Localisation And Mapping (SLAM)*», qui permet de cartographier et localiser un robot ou un agent autonome dans son environnement. Dans des environnements complexes où les systèmes de positionnement classiques comme le GPS échouent – notamment en milieu urbain dense ou en intérieur – la VBL s'impose comme une alternative robuste et précise. La localisation peut être faites à plusieurs niveaux d'échelles : à l'intérieur d'une pièce [Liang 2013], à l'échelle d'une rue [Kendall 2015], ou même sur la carte mondiale [Hays 2008, Vo 2017, Dufour 2025].

De nombreux travaux ont été proposés dans ce domaine et plusieurs synthèses de la littérature ont été réalisées. [Brejcha 2017] fournit une vaste revue des méthodes de géolocalisation visuelle existantes. Les auteurs soulignent que de nombreux travaux ont été réalisés dans les milieux urbains et constate que les environnements naturels ont été moins étudiés. Le livre [Zamir 2016] offre une perspective sur les techniques de géolocalisation visuelle à grande échelle, présentant différentes approches utilisées pour la géolocalisation, telles que les méthodes guidées par les données, les méthodes guidées par la sémantique et les approches basées géométrie. [Piasco 2018] met en lumière l'importance de l'utilisation de données hétérogènes pour améliorer la précision et la robustesse des systèmes de localisation.

La représentation des données est une étape cruciale dans les systèmes de localisation visuelle. Une première famille de méthode consiste à utiliser une représentation locale des données. Cela peut être fait par la détection de points d'intérêt et le calcul de vecteurs de description comme avec SIFT [Lowe 1999], SURF [Bay 2006], ORB [Rublee 2011], Daisy [Rublee 2011], ou LIFT [Yi 2016]. Cela peut également être réalisé par la représentation de patches comme dans [Dalal 2005, Gordo 2017]. D'autres approches cherche des caractéristiques géométriques tels que des lignes [Hays 2008, Arth 2015, Morago 2016, Ramalingam 2011] ou des contours [Russell 2011], qui peuvent être très fréquentes dans les environnements architecturaux urbains ou dans les scènes d'intérieur. Ces descripteurs géométriques peuvent être combinés avec des approches par points d'intérêt comme dans [Saurer 2016, Jegou 2008].

Une autre famille de méthode consiste à construire une représentation globale, soit à l'aide de descripteurs construits à la main comme GIST [Oliva 2001], soit en utilisant des méthodes d'apprentissage via des approches neuronales, telles que PoseNet [Kendall 2015], Bayesian PoseNet [Kendall 2016] et Posenet-LSTM [Walch 2017]...

[Piasco 2018] distingue deux catégories d'approches : les approches indirectes, qui cherchent dans une base d'images celles les plus similaires à une image requête, et les approches directes, qui régressent directement les six degrés de liberté correspondant à la pose de la caméra recherchée. Dans les travaux présentés dans cette section, nous nous intéresserons uniquement à la seconde catégorie. On peut principalement distinguer deux stratégies pour l'estimation de la pose. La première repose sur le calcul de la pose à partir de propriétés géométriques, tandis que la seconde utilise une régression directe des

paramètres définissant la pose.

La première stratégie suit classiquement le schéma suivant :

1. **Extraction de points d'intérêt :** Utilisation d'algorithmes tels que SIFT [Lowe 1999], SURF [Bay 2006], Daisy [Rublee 2011], LIFT [Yi 2016]...
2. **Mise en correspondance des points :** Les points d'intérêt détectés sont mis en correspondance avec des points dans des images de référence.
3. **Estimation de la pose :** Utilisation de l'algorithme des 5-points [Nistér 2004] ou l'algorithme des 8-points [Hartley 1997].

Généralement, un algorithme de filtrage par consensus de type RANSAC [Fischler 1981] est utilisé pour rejeter les valeurs aberrantes de manière robuste. Ces approches dépendent fortement de la qualité des mises en correspondance et présentent des performances limitées pour des objets faiblement texturés ou des images fortement bruitées. De plus, elles ne permettent pas d'estimer l'échelle de translation entre les deux caméras, se contentant uniquement de la direction.

La deuxième stratégie consiste à résoudre directement un problème de régression. Cela peut être réalisé à l'aide de méthodes apprises de bout en bout telles que PoseNet [Kendall 2015], Bayesian PoseNet [Kendall 2016] et PoseNet-LSTM [Walch 2017]. L'approche proposée par [Piasco 2019] pour la localisation visuelle explore l'idée de rendre un descripteur global d'image plus robuste aux conditions difficiles réalise également en y intégrant des informations issues d'une estimation de profondeur durant la phase d'entraînement. Les auteurs utilisent un réseau d'estimation de profondeur couplé à un encodeur pour construire une représentation de l'information de profondeur qui est ensuite concaténée aux vecteurs de représentation. L'ensemble est également appris de bout en bout. Cependant, toutes ces approches fournissent une estimation globale de la position, non relative à un autre point de vue.

Dans le cadre des activités post-doctorales de Sovann En, nous avons exploré l'estimation de la pose relative entre deux caméras utilisant différentes modalités (Fig. 3.2). Notre objectif est de déterminer la position d'une caméra par rapport à une autre en exploitant les images capturées par ces dernières.

Nous avons étudié deux stratégies possibles correspondant aux deux possibilités évoquées plus haut dans la littérature :

— **RPNet [En 2018a] :** un réseau de neurones de bout en bout pour l'estimation de pose,
— **TS-Net [En 2018b] :** une méthode d'appariement de patchs multimodale permettant ensuite l'utilisation de techniques de calcul de pose sans réseau de neurones.

Nous avons proposé une nouvelle approche d'estimation robuste directe de la pose relative de caméra (RPNet [En 2018a]), permettant une estimation plus stable et incluant l'échelle de translation, contrairement à [Yi 2018] qui estime uniquement la matrice de rotation entre les deux caméras. Cette solution repose sur un réseau de neurones siamois de bout en bout, estimant la position relative de deux caméras à partir d'une estimation absolue de leurs positions dans la scène. Contrairement aux approches classiques, il n'est pas nécessaire d'extraire des points d'intérêts et de les apparier. Les résultats de cette approche sont équivalents, voire meilleurs, que ceux des approches classiques, tout en ajoutant une estimation de

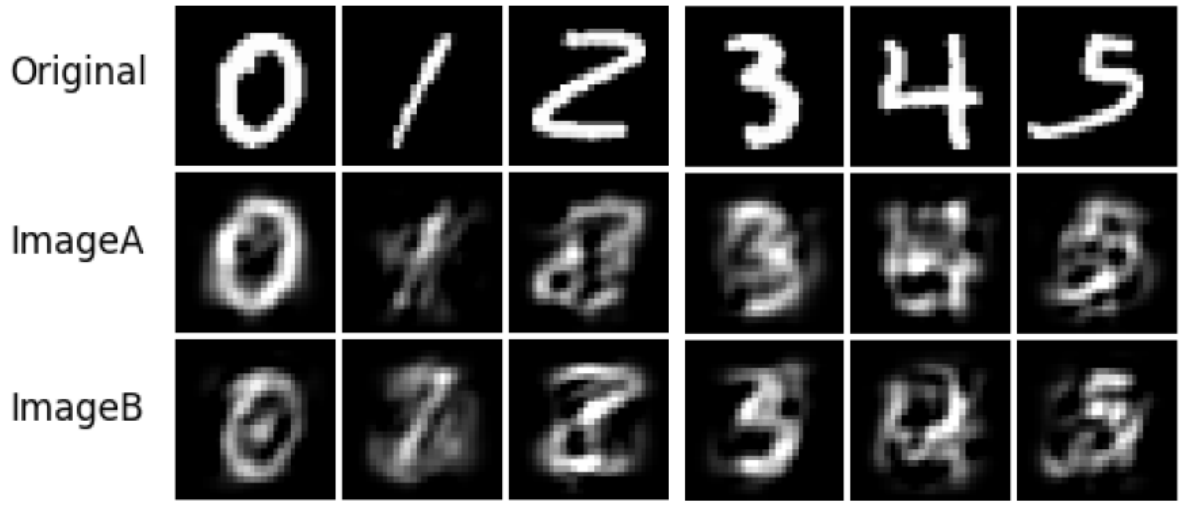


FIGURE 3.3 – Exemples de modalités générées pour la base MM-MINIST

l'échelle de translation.

Bien que la méthode RPNet présente des avantages, elle est difficile à appliquer dans un contexte multimodal où l'une des modalités est moins disponible et par conséquent dont les bases de données sont plus restreintes. Nous avons donc développé une nouvelle approche adaptée aux particularités du contexte multimodal du projet dans lequel dans lequel ces travaux s'inscrivaient. La solution que nous proposons, TS-NET [En 2018b], est un réseau de neurones permettant de comparer des patchs issus de différentes modalités. Notre approche vient en remplacement des méthodes d'appariement de points d'intérêt utilisées dans les approches traditionnelles d'estimation de pose et repose sur un réseau pseudo-siamois [Jahrer 2008, Zagoruyko 2015, Merkle 2017, Aguilera 2017] d'appariement de patchs. L'architecture proposée tire sa performance de l'apprentissage sur deux branches de traitement (basées sur un réseau siamois et pseudo-siamois) et de l'utilisation de plusieurs fonctions de coût pour contraindre le modèle à être correctement appris. L'utilisation de deux branches permet de séparer l'extraction de propriété commune à toutes les modalités (réseau siamois) des propriétés spécifiques à chacune (réseau pseudo-siamois). La fusion est double, une première est réalisée aux niveaux des réseaux siamois et pseudo-siamois, est peut-être qualifié de fusion de niveau intermédiaire. Une seconde fusion est effectuée à la fin de l'architecture entre les deux branches avant de réaliser la décision finale. Cette fusion est considérée comme tardive dans la taxonomie introduite dans le début de ce chapitre.

3.2.3 CentralNet : Apprendre automatiquement où faire la fusion

Références des travaux associés : [Vielzeuf 2018b]

Lorsqu'il s'agit de combiner des informations provenant de plusieurs sources, il est crucial de déterminer à quel moment dans le processus de traitement cette fusion doit se produire. Doit-on fusion-

ner les informations dès leur collecte, après des traitements spécifiques pour chaque source, ou après avoir résolu la tâche cible ? La question du moment de la fusion est un enjeu clé, sans réponse simple et définitive. Pour certains problèmes, des approches de *fusion précoce* [Arevalo 2017, Chen 2017] ont donné d'excellents résultats en permettant aux modalités d'interagir dès le début. Elles sont cependant sensibles au bruit ou à la dominance d'une modalité. Pour d'autres, des approches de *fusion tardive* [Kim 2017, Vielzeuf 2017] sont préférables. Généralement plus robuste, elles limitent les interactions profondes entre modalités. Les *fusions intermédiaires* [Yang 2016, Cangea 2017, Gu 2017, Kang 2017] semblent prometteuses mais la littérature manque de critères clairs pour déterminer a priori le niveau optimal pour un problème donné.

Dans le cadre de la thèse de Valentin Vielzeuf, nous avons proposé *CentralNet* [Vielzeuf 2018b], une architecture visant à répondre à cette question en apprenant la position idéale de la fusion au sein du réseau. L'idée maîtresse est d'introduire une "branche centrale" qui reçoit et fusionne les représentations intermédiaires issues des branches unimodales à *chaque niveau* de l'architecture. Cette branche centrale apprend ainsi une représentation multimodale progressivement enrichie. Pour garantir la convergence et forcer chaque branche unimodale à extraire des informations pertinentes, des fonctions de coût auxiliaires sont ajoutées sur les sorties des branches unimodales, les contraignant à résoudre (partiellement) la tâche cible de manière indépendante. La branche centrale effectue donc une fusion multi-niveaux, potentiellement capable de capturer des interactions complexes à différents degrés d'abstraction.

Pour évaluer CentralNet indépendamment des complexités liées aux données réelles, nous avons introduit deux bases de données synthétiques : *Multimodale-MNIST* (MM-MNIST) et *Audiovisual MNIST* (AV-MNIST). :

— *Multimodale-MNIST* simule deux modalités contenant une partie de l'information des images de la base MNIST [Deng 2012] et une certaine quantité de bruit (cf. Fig. 3.3). Cette nouvelle base nous permet de contrôler à la fois la quantité de bruit et la quantité d'information présente et partagée dans les modalités.

— *Audiovisual MNIST* est une base composée à la fois des chiffres de la base MNIST et de sons issus de la base *Free Spoken Digits Database* (`https://github.com/Jakobovski/free-spoken-digit-dataset`) augmentée de bruit venant de la base ESC-50 [Piczak 2015]. Afin de rendre la tâche de fusion plus complexe, nous conservons que 25% de l'information des images pour la modalité visuelle. Nous utilisons pour cela la même technique que pour la base Multimodale-MNIST.

CentralNet a ensuite été appliqué avec succès à plusieurs tâches diverses, détaillées dans les sous-sections suivantes.

3.2.3.1 Application à la détection d'émotions

Références des travaux associés : [Vielzeuf 2018a, Kervadec 2018, Vielzeuf 2019]

Durant la thèse de Valentin Vielzeuf, nous nous sommes intéressés à la détection et à la représen-

tation des émotions dans les vidéos. La reconnaissance automatique des émotions est un domaine qui a suscité un grand intérêt avec de nombreuses recherches et applications [Fan 2016, Gers 1999, Hu 2017, Knyazev 2018, Benitez-Quiroz 2016, Li 2017]. Afin de mesurer l'avancement des techniques, des challenges ont été mis en place, dont le challenge *Emotion in the Wild*[Dhall 2018], sur lequel nous avons été classés 3ème pour les éditions 2017 et 2018 (une présentation détaillée de notre contribution est faite dans [Vielzeuf 2018a]).

Nous avons débuté nos recherches en examinant les performances de la reconnaissance des émotions dans des séquences d'images. Dans l'article [Kervadec 2018], nous avons développé une méthode pour créer des représentations compactes et efficaces des émotions, en s'inspirant des modèles psychologiques décrivant les émotions selon les axes de *valence* (agréable/désagréable) et d'*arousal* (excitation physiologique allant de calme à excité). En entraînant un ResNet18 à régresser conjointement ces valeurs et à classifier les émotions discrètes, nous avons obtenu des représentations visuelles performantes.

Ces représentations visuelles ont ensuite servi de base pour notre participation au challenge EmotiW [Vielzeuf 2018a]. Nous avons implémenté une *fusion tardive* combinant ces caractéristiques visuelles avec des descripteurs audio classiques. L'utilisation d'ensembles de modèles (*bagging*) a permis en plus d'améliorer significativement les performances. Cette fusion, qui combine une modalité visuelle (49,4% d'accuracy) avec une modalité audio (35% d'accuracy), a permis d'améliorer les performances à 57,2%.

Dans l'article de journal [Vielzeuf 2019], nous présentons les résultats obtenus sur les données AFEW du challenge *Emotion in the Wild* en utilisant notre approche de fusion multi-niveau CentralNet [Vielzeuf 2018b]. La méthode de fusion CentralNet a permis d'augmenter l'accuracy de 55,3% en fusion tardive à 57,2%. Attention cependant, les résultats pour la base AFEW ne sont pas directement comparables entre nos articles [Vielzeuf 2018a] et [Vielzeuf 2019]. Dans [Vielzeuf 2018a], nous présentons les performances sur l'ensemble de test de la base AFEW, fourni par les organisateurs du challenge. En revanche, les résultats de [Vielzeuf 2019], réalisés après le challenge, ont été calculés sur un ensemble de validation, la vérité terrain de l'ensemble de test n'étant pas disponible.

3.2.3.2 Application au texte

Références des travaux associés : [Jha 2022]

Dans les travaux de Prince Jha [Jha 2022], nous avons exploré l'application de la fusion multimodale à une tâche d'évaluation de relation lexico-sémantique. Notre objectif est de déterminer si deux mots entretiennent une relation lexico-sémantique, telle que la synonymie, l'hyperonymie ou la co-hyponymie, ou s'ils n'ont aucune relation.

La synonymie désigne une relation entre deux mots ou expressions ayant des sens identiques ou très similaires. L'hyperonymie est une relation où un mot a un sens plus général qu'un autre, par exemple, "animal" est un hyperonyme de "chien". La co-hyponymie est une relation entre deux mots qui partagent un même hyperonyme, comme "chien" et "chat" qui sont tous deux des co-hyponymes sous l'hyperonyme "animal".

(a) image RGB (b) image polarimétrique

FIGURE 3.4 – La fiabilité des capteurs varie en fonction des conditions météorologiques.

Pour améliorer la résolution de cette tâche, nous proposons d'intégrer des informations visuelles en complément du texte. L'hypothèse sous-jacente, est que l'information visuelle associée aux mots peut compléter leur information sémantique textuelle. Dans notre étude [Jha 2022], nous présentons une stratégie visant à enrichir la représentation textuelle d'un mot en la fusionnant avec les représentations des mots de son voisinage ainsi que des représentations visuelles de ces mots. Ces représentations visuelles sont obtenues à partir des modèles VGG19 [Simonyan 2015] ou CLIP [Radford 2021], en utilisant les images issues du premier résultat de la recherche d'images Bing.

Nous avons comparé deux mécanismes de fusion pour combiner les représentations textuelles et visuelles ainsi obtenues : un mécanisme basé sur l'*attention* et une adaptation de notre architecture *CentralNet* [Vielzeuf 2018b]. Les résultats ont montré que l'ajout d'informations visuelles, même collectées automatiquement, améliorait les performances par rapport à l'utilisation des seules informations textuelles, validant notre hypothèse initiale. Cependant, la variabilité des résultats n'a pas permis de départager clairement les deux stratégies de fusion testées.

3.2.4 Fusion tardive robuste via la théorie de Dempster-Shafer

Références des travaux associés : [Deregnaucourt 2023, Deregnaucourta 2023, Deregnaucourt 2025]

Le développement de systèmes critiques comme les véhicules autonomes [Gupta 2021, Kaymak 2019] doivent reposer sur une perception fiable de l'environnement. Ces solutions s'appuient sur diverses technologies, notamment l'analyse automatique des scènes routières par segmentation sémantique.

Cependant, la fiabilité des capteurs (caméras RGB, lidars, radars, capteurs polarimétriques...) est fortement dépendante des conditions externes (météo [Zhang 2023b], luminosité, éblouissement, pannes [Fayyad 2020]) (cf. Fig. 3.4). Se fier à une seule modalité est donc risqué. La fusion d'informations issues de multiples capteurs, aux forces et faiblesses complémentaires, est essentielle pour garantir la robustesse. Par exemple, un capteur infrarouge sera plus performant la nuit, tandis qu'un capteur polari-

métrique le sera sous forte pluie [Deregnaucourt 2023].

Au-delà de la simple fusion, un défi majeur est d'assurer la robustesse face aux défaillances potentielles des capteurs et de tenir compte de l'incertitude associée à chaque mesure. Des approches comme [Besnier 2021] se sont intéressées à la mesure de l'incertitude des prédictions. Cela permet d'estimer si l'information prédite par le réseau de neurones peut être utilisée ou non, mais ne permet pas d'apporter de solution dans les zones incertaines.

Les approches classiques de fusion ne gèrent pas nativement cette incertitude. Une solution prometteuse à ce problème est la classification à valeurs d'ensemble. Ces méthodes permettent au modèle, en cas de forte incertitude, d'attribuer une donnée non pas à une classe unique mais à un ensemble de classes possibles.

La théorie de Dempster-Shafer (DST) [Dempster 2008, Shafer 1976] fournit un cadre mathématique rigoureux pour modéliser et raisonner avec l'incertitude et l'ignorance. Récemment, des efforts ont été faits pour intégrer la DST aux réseaux neuronaux profonds [Denoeux 2000, Sensoy 2018, Tong 2021b]. Un obstacle majeur est l'explosion combinatoire (passage à 2^K ensembles de classes pour K classes) inhérente à la théorie. Cette explosion combinatoire pose un véritable défi technique et ne permettait pas la généralisation de ce type d'approche pour un nombre important de classes. Dans [Deregnaucourta 2023], nous avons proposé une *solution algorithmique* pour contourner cette complexité exponentielle dans le cadre d'une *fusion tardive* basée sur la DST. En réarrangeant les équations et en introduisant des restrictions raisonnables, nous avons montré qu'il est possible de trouver le meilleur ensemble de classes (au sens de l'utilité espérée) en temps linéaire par rapport au nombre de classes, sans énumérer les 2^K possibilités.

Cela nous a permis de proposer le premier réseau de neurones réalisant une fusion tardive évidentialiste (basée sur la DST) capable de passer à l'échelle de grands jeux de données comme ImageNet [Deng 2009]. Ce réseau a été utilisé pour la détection d'exemples hors-distribution [Deregnaucourt 2023] et, dans [Deregnaucourt 2025], pour la construction de systèmes de segmentation sémantique robustes. Dans ce dernier travail, nous avons introduit un mécanisme adaptatif qui pondère la contribution de chaque modalité lors de la fusion en fonction d'une mesure de conflit (basée sur la distance de Jousselme [Jousselme 2001]) entre les fonctions de masse produites par les modèles unimodaux [Martin 2012]. Cela permet au système de réduire dynamiquement l'influence d'un capteur jugé peu fiable ou en conflit avec les autres.

3.3 Discussion

Les travaux présentés dans ce chapitre ont exploré diverses facettes et défis de la fusion multimodale, en lien direct avec les questions et catégories identifiées dans l'état de l'art (section 3.1). En portant un regard rétrospectif, nous pouvons analyser la portée et les limites de ces contributions et les resituer dans l'évolution rapide du domaine, notamment par rapport à la nature des données (homogènes/hétérogènes), au moment de la fusion (précoce/tardif/intermédiaire ou mixte) et aux mécanismes de fusion employés.

Analyse critique des contributions architecturales

Mes premières explorations architecturales, avec **TS-Net** [En 2018b] puis **CentralNet** [Vielzeuf 2018b], s'attaquaient à des questions fondamentales : comment combiner efficacement les informations *spécifiques* à chaque modalité et celles qui leur sont *communes* (TS-Net) ? Et surtout, comment déterminer le *niveau optimal* pour réaliser cette fusion (CentralNet) ? Ces questions correspondent respectivement à la recherche de mécanismes de fusion efficaces (**comment fusionner ?**) et à la détermination du niveau optimal (**où fusionner ?**).

Originalité et forces : L'originalité de TS-Net résidait dans sa structure à trois branches modélisant explicitement ces deux types d'informations (informations spécifiques ou communes). CentralNet poussait l'idée plus loin en proposant d'apprendre automatiquement une stratégie de fusion multi-niveaux via une branche centrale dédiée, se distinguant des approches antérieures qui imposaient un niveau de fusion fixe (précoce, tardif ou intermédiaire). L'introduction de bases de données synthétiques (MM-MNIST, AV-MNIST) a permis une analyse contrôlée de ces mécanismes.

Limites et positionnement : Développées avant l'avènement massif des *Transformers*, ces architectures nécessitent des adaptations pour passer à ce nouveau paradigme. Cependant, l'idée fondamentale d'une fusion multi-niveaux et adaptative reste pertinente. Des travaux plus récents comme ceux de [Nagrani 2021], bien que reposant sur les mécanismes d'attention *Transformers*, poursuivent cette même intuition qu'une fusion efficace doit potentiellement opérer à plusieurs niveaux d'abstraction. Les auteurs introduisent un goulot d'étranglement multi-niveaux où la fusion est réalisée tout au long de l'architecture. Contrairement à l'utilisation d'une branche centrale, cet article propose l'utilisation de tokens supplémentaires, appelés *fusion bottleneck tokens*, pour effectuer les fusions. Les attentions ne sont pas réalisées directement entre les tokens des différentes modalités (comme cela est le cas dans [Cai 2023, Rodríguez Bribiesca 2021]), mais entre les tokens de chaque modalité et les tokens réservés de fusion.

Mes travaux illustrent ainsi une transition conceptuelle, passant d'architectures de fusion rigide vers des approches cherchant plus à découvrir automatiquement la meilleur stratégie de fusion.

Analyse critique des contributions applicatives

L'application de ces concepts à des domaines spécifiques comme l'analyse des signaux **EEG pour le sommeil** [Dequidt 2023, Seraphim 2023b, Seraphim 2023a, Seraphim 2024b, Seraphim 2024a] ou l'**estimation de pose relative** [En 2018a, En 2018b] a permis de valider leur pertinence tout en soulevant des défis propres à chaque domaine.

Originalité et forces : Pour l'EEG, l'utilisation directe des matrices de covariance comme entrée et le développement de *Transformers* (SPDTransNet) préservant explicitement la géométrie Riemannienne de ces données constituent une approche originale et performante par rapport aux méthodes classiques basées sur les spectrogrammes. Pour la pose relative, RPNet [En 2018a] fut l'une des premières tentatives d'estimation *end-to-end* de la pose *complète* (rotation et translation), apportant une solution à l'ambiguïté

d'échelle des méthodes géométriques traditionnelles.

Limites et positionnement : Ces travaux montrent l'intérêt d'adapter les architectures à la nature spécifique des données (matrices SDP) ou aux contraintes de la tâche (géométrie de la pose).

Analyse critique des contributions sur la fusion robuste

Enfin, mes travaux les plus récents sur la **fusion tardive via la théorie de Dempster-Shafer (DST)** [Deregnaucourt 2023, Deregnaucourta 2023, Deregnaucourt 2025] s'attaquent à une dimension souvent critique mais négligée dans de nombreuses approches de fusion : la **fiabilité** et la gestion de l'**incertitude**.

Originalité et forces : La contribution majeure ici est la proposition d'une solution algorithmique [Deregnaucourta 2023] qui rend la prise de décision évidentialiste (classification à valeurs d'ensemble) traitable en termes de complexité calculatoire (complexité linéaire) même pour un grand nombre de classes, levant ainsi le verrou de l'explosion combinatoire propre à ce mécanisme de *fusion tardive* basé sur la DST. Cela a permis l'application de ce cadre théorique puissant à des problèmes à grande échelle (ImageNet, segmentation sémantique). L'ajout d'un mécanisme d'affaiblissement dynamique des sources basé sur le conflit [Deregnaucourt 2025] améliore encore la robustesse face aux défaillances potentielles de capteurs.

Limites et positionnement : Ces travaux ouvrent une voie prometteuse vers des systèmes multimodaux plus fiables et «conscients de leur incertitude», répondant à un besoin crucial pour les applications critiques. Ils démontrent l'intérêt de sortir du cadre probabiliste traditionnel et d'étudier les possibilités de la DST pour l'apprentissage profond. Les limites actuelles résident dans la nécessité de disposer de modèles unimodaux capables d'estimer des fonctions de masse et dans l'utilisation d'une version encore simplifiée de la DST (considérant principalement les singletons et l'ignorance totale Ω). Explorer des cadres DST plus riches reste une perspective d'avenir.

Connexion aux tendances actuelles et questions ouvertes

Ces différentes contributions s'inscrivent dans un paysage de la fusion multimodale en pleine effervescence. Si les architectures dédiées comme CentralNet ou celles basées sur les *Attention Bottlenecks* [Nagrani 2021] montrent l'intérêt des fusions multi-niveaux, elles souffrent souvent d'un manque de flexibilité : elles nécessitent généralement toutes les modalités et l'ajout dynamique d'une source est complexe. Cela explique en partie l'attrait pour des stratégies alternatives. Les approches par **alignement d'espaces latents** (typiquement via une fusion tardive), popularisées par CLIP [Radford 2021] pour le couple image/texte, offrent une grande flexibilité mais peuvent limiter la profondeur des interactions inter-modales. Des architectures comme Uniter [Chen 2020b] proposent un compromis avec une fusion intermédiaire via un *Transformer* traitant conjointement les représentations de chaque modalité.

L'émergence récente de modèles de fondation capables de traiter nativement plusieurs modalités et de générer du texte, comme Florence-2 [Xiao 2024], ouvrent de nouvelles pistes de recherche. Ces modèles tendent à convertir toutes les entrées des différentes modalités en une séquence de *tokens* uniformes,

traitée ensuite par un grand modèle de langage. La recherche sur la construction d'une "tokénisation universelle" pourrait potentiellement simplifier radicalement la fusion multimodale en la ramenant à un problème de traitement de séquences. Est-il encore pertinent de chercher des mécanismes de fusion complexes à tous les niveaux si un alignement initial avec des représentations textuelles suffit ? La création de tokens multimodaux sémantiquement riches, traitables par des modèles de langage standards, est une piste de recherche qui me semble très intéressante à continuer à explorer.

Cependant, ces approches basées sur les grands modèles de langage ou l'alignement d'espaces latents (comme CLIP) négligent souvent une dimension cruciale : l'**incertitude** et la **fiabilité variable** des différentes modalités. Nos travaux sur la fusion tardive via la théorie de Dempster-Shafer s'attaquent spécifiquement à ce problème, en proposant un cadre théorique fondé pour modéliser et exploiter l'incertitude lors de la prise de décision.

Dès lors, plusieurs questions stimulantes pour l'avenir de la fusion multimodale demeurent :

— La fusion doit-elle nécessairement se faire à tous les niveaux, ou des fusions tardives bien conçues peuvent-elles être aussi, voire plus, efficaces et robustes ?

— La "tokénisation universelle" est-elle une voie viable pour remplacer les architectures de fusion dédiées, ou restera-t-elle limitée à certains types d'interactions ou de tâches ?

— Comment intégrer *efficacement* et *théoriquement fondé* la gestion de l'incertitude et de la fiabilité au cœur des systèmes de fusion, qu'ils reposent sur des CNNs, des Transformers, des approches précoces, tardives ou hybrides ?

— Comment intégrer au coeur des architectures de fusion multimodales des contraintes d'efficacité énergétique en limitant la perte en performance ?

Explorer ces questions, en cherchant à concilier performance, efficacité et fiabilité, sera essentiel pour développer les systèmes multimodaux robustes et dignes de confiance dont nous aurons besoin demain.

CHAPITRE 4

Conclusion et Perspectives

Sommaire

4.1 Conclusion

Le présent manuscrit a offert une présentation synthétique et réflexive de mes activités de recherche depuis l'obtention de mon doctorat en 2012. Ces travaux se sont principalement articulés autour de deux domaines de recherche majeurs et complémentaires : **l'apprentissage frugal** et **l'apprentissage par fusion de modalités**.

L'exploration de l'**apprentissage frugal** répond à un double constat : d'une part, la limitation intrinsèque des données disponibles (en quantité ou en qualité d'annotation) dans de nombreux domaines applicatifs ; d'autre part, les contraintes croissantes en ressources de calcul et les préoccupations sociétales et environnementales associées au coût énergétique des modèles d'intelligence artificielle. Sur le sujet, le *Conseil Économique, Social et Environnemental* écrivait en février 2024 : « La consommation énergétique de l'entrainement des modèles d'IA, notamment l'apprentissage profond (deep learning), est appelée à s'accroitre considérablement dans les années à venir. Il conviendra de penser ensemble l'accroissement de l'empreinte carbone, des matières premières matérielles (silicium, terres rares), des besoins de ressources (consommation en eau et empreinte carbone des engins de chantiers) du numérique lié à l'IA et les gains des retombées de son usage.» (source : Saisine pour étude, Intelligence artificielle et environnement). Face à ces défis, j'ai contribué au développement de méthodes visant à :

— Pallier le manque de données via l'apprentissage *cross-domaine* ou l'apprentissage de métriques permettant une généralisation efficace à partir de peu d'exemples (*few-shot learning*).

— Réduire l'empreinte calculatoire via des architectures multi-sorties capables d'adapter dynamiquement leur complexité à la difficulté de la tâche ou aux contraintes de disponibilité des ressources de calculs.

Ces travaux s'inscrivent dans une démarche visant à développer une IA plus durable, plus accessible et mieux adaptée aux contraintes du monde réel.

Parallèlement, mes recherches sur la **fusion de données** et l'**apprentissage multimodal** s'attaquent à la nécessité d'intégrer des informations provenant de sources multiples et hétérogènes pour construire une compréhension plus complète et plus précise de notre environnement. La perception humaine est intrinsèquement multimodale, et doter les machines de capacités similaires est essentiel pour de nombreuses applications. Mes travaux dans ce domaine ont porté sur :

- La fusion de données homogènes, notamment pour l'analyse de signaux physiologiques (EEG) où l'exploitation des corrélations entre capteurs s'est avérée pertinente.
- L'étude des différents niveaux de fusion (précoce, tardive, intermédiaire) et le développement d'architectures (comme CentralNet) capables d'apprendre la stratégie de fusion optimale.
- La conception de méthodes de fusion tardive robustes, notamment via la théorie de Dempster-Shafer, pour garantir la fiabilité des systèmes même en présence de données incertaines ou de capteurs défaillants, un aspect crucial pour les applications critiques comme le véhicule autonome.

La fusion de données permet non seulement d'améliorer la performance et la robustesse, mais aussi, potentiellement, de travailler avec des données moins riches prises individuellement, rejoignant ainsi les préoccupations de la frugalité.

Ces deux axes, loin d'être disjoints, se révèlent profondément interconnectés. L'exploitation de sources multiples (multimodalité) enrichit l'information disponible, ouvrant potentiellement la voie à des apprentissages plus frugaux en données. Réciproquement, la complexité inhérente aux systèmes multimodaux rend la frugalité en ressources (via des architectures efficaces comme les multi-sorties) particulièrement pertinente. Mes travaux ont ainsi cherché à naviguer à l'intersection de ces deux enjeux majeurs de l'IA contemporaine.

4.2 Perspectives à court et moyen terme

Dans la continuité directe des travaux présentés, mes perspectives de recherche à court et moyen terme visent à approfondir et étendre nos contributions sur les modèles adaptables et fiables, en capitalisant sur les projets et encadrements en cours.

Concernant la **frugalité des ressources**, les travaux initiés dans la thèse de Youva Addad sur les architectures multi-sorties [Addad 2023, Addad 2024, Addad 2025] offrent une base solide. L'objectif immédiat est d'étendre ces architectures au-delà de la simple classification d'images et d'explorer leur potentiel pour des tâches multimodales. Une question de recherche clé sera : *Comment adapter efficacement les mécanismes de sortie précoce et les politiques de décision apprises à des architectures multimodales complexes, comme celles basées sur CLIP [Radford 2021], afin d'optimiser l'équilibre performance/coût pour la classification ou la génération multimodale ?*

Concernant la **fiabilisation des modèles via la fusion multimodale**, les travaux de Lucas Deregnaucourt sur la fusion tardive basée sur la théorie de Dempster-Shafer [Deregnaucourt 2023, Deregnaucourta 2023, Deregnaucourt 2025] ouvrent des perspectives prometteuses pour la gestion de

l'incertitude. Il serait particulièrement intéressant d'étudier une synergie avec les architectures multi-sorties : *Une version multi-sorties de notre approche de fusion évidentialiste permettrait-elle d'adapter dynamiquement non seulement le coût de calcul, mais aussi le niveau de confiance et la granularité de la prédiction (classe unique vs. ensemble de classes) en fonction de la fiabilité estimée des modalités disponibles à chaque sortie ?* Cela impliquerait d'apprendre conjointement la politique de sortie et la stratégie de fusion robuste.

Par ailleurs, les collaborations interdisciplinaires initiées (par ex. analyse du regard sur des œuvres d'art avec la thèse de Raphaelle Lemaire, la segmentation d'images médicales avec la thèse de Jérôme Cartier, l'analyse d'otolithes avec la thèse d'Éric Hu...) seront poursuivies, car elles ancrent mes recherches méthodologiques dans des problématiques applicatives concrètes et soulèvent de nouvelles questions à explorer. Une piste à explorer pourrait être de mieux intégrer les connaissances expertes ou les contraintes physiques dans les modèles que j'étudie.

4.3 Perspectives à long terme

Aborder les perspectives à long terme en apprentissage machine impose une certaine humilité tant le domaine évolue rapidement. Les progrès récents, notamment autour des modèles de fondation, redéfinissent constamment le paysage scientifique et applicatif. Plutôt que de prédire l'avenir, il s'agit d'identifier des directions de recherche qui me semblent porteuses de sens scientifique et sociétal.

Une direction majeure me semble être la convergence vers des **modèles de fondation multimodaux frugaux et fiables**. Si l'intégration de multiples modalités et la capacité à interagir en langage naturel sont des objectifs clairs, les défis de la frugalité (comment entraîner et adapter ces modèles géants ?) et de la fiabilité (comment leur faire confiance ?) restent immenses. Le développement de techniques de "tokénisation universelle" efficaces et sémantiquement riches pour chaque modalité, combiné à des stratégies d'apprentissage auto-supervisé adaptées et à des architectures nativement économes (potentiellement inspirées des multi-sorties ou d'autres paradigmes frugaux), constituera un axe de recherche majeur.

Parallèlement, la nécessité d'une **IA plus fiable, explicable et alignée sur des contraintes externes** va s'accroître, notamment pour favoriser son adoption dans des domaines scientifiques ou industriels critiques. Cela passe par une meilleure intégration des connaissances issues d'autres disciplines (physique, biologie, médecine, sciences humaines...) directement au sein des architectures neuronales, par exemple via l'intégration dans l'apprentissage de contraintes physiques (PINNs [Raissi 2019], FNOs [Li 2021b]) ou logiques ([Ledaguenel 2024]) ou par des mécanismes d'attention guidés par des connaissances *Retrieval-Augmented Generation* [Lewis 2020b]. Le développement de méthodes d'explicabilité spécifiques aux modèles multimodaux et frugaux sera également indispensable pour permettre aux experts d'autres domaines de comprendre, valider et s'approprier ces outils.

Enfin, ces perspectives scientifiques ne peuvent être dissociées du **rôle et de l'évolution de la recherche académique**. L'émergence d'une IA "plus mature" et l'investissement massif du secteur privé

créent une dynamique nouvelle. Si la collaboration est essentielle, la recherche publique a, à mon sens, un rôle crucial à jouer pour :

- Garantir le développement d'une **IA ouverte et accessible**, évitant les monopoles technologiques et favorisant l'innovation au bénéfice de tous (PME, services publics, autres disciplines scientifiques...).
- Aborder des **questions sociétales fondamentales** parfois négligées par une approche purement industrielle : protection de la vie privée, équité, éthique, impact environnemental, biais...
- Développer des **solutions pour des domaines "non rentables"** mais d'intérêt public majeur (santé, environnement, éducation, culture...).

Pour remplir ces missions, la recherche académique française et européenne aura besoin de se structurer davantage, en mutualisant les ressources (calcul, données, plateformes) et en favorisant les collaborations interdisciplinaires à grande échelle, peut-être à l'image des "grands équipements" existant dans d'autres disciplines comme la physique. Maintenir une recherche publique forte, indépendante et connectée aux besoins sociétaux me semble essentiel pour accompagner sereinement les prochaines révolutions de l'intelligence artificielle.

Deuxième partie

Sélection de publications

A JOINT LEARNING APPROACH FOR CROSS DOMAIN AGE ESTIMATION

Binod Bhattarai[1], *Gaurav Sharma*[2], *Alexis Lechervy*[1], *Frederic Jurie*[1]

[1]CNRS UMR 6072, University of Caen Normandy, ENSICAEN, France
[2]Max Planck Institute for Informatics, Saarbrücken, Germany

ABSTRACT

We propose a novel joint learning method for cross domain age estimation, a domain adaptation problem. The proposed method learns a low dimensional projection along with a regressor, in the projection space, in a joint framework. The projection aligns the features from two different domains, i.e. source and target, to the same space, while the regressor predicts the age from the domain aligned features. After this alignment, a regressor trained with only a few examples from the target domain, along with more examples from the source domain, can predict very well the ages of the target domain face images. We provide empirical validation on the largest publicly available dataset for age estimation i.e. MORPH-II. The proposed method improves performance over several strong baselines and the current state-of-the-art methods.

1. INTRODUCTION AND RELATED WORK

Automatic age estimation from face images has become a popular research problem [1, 2, 3, 4, 5]. It has various important applications such as age specific human-computer interaction [6], business intelligence [7] etc. Previous studies [8, 9, 10, 11] have shown that the rate of aging among different groups of people is different. This is because, aging patterns are directly affected by genes, dieting habits, culture, weather, race, gender etc. Thus, it has been more challenging to design an age prediction model which generalizes for people from such different categories. In addition, it has been shown that, training a single model on all different groups together, affect the performance that separate specialized models for different groups can give, due to the differences in aging patterns [9].

Training separate model for each and every group of people has its own limitations. It is difficult, expensive and time consuming to collect and annotate face images. Moreover, due to privacy related concerns, people may not be keen to share information about them such as ages, race etc. Thus, it would be ideal to utilise the training examples available for one group of people to improve performance in another group which has a very limited number of training examples. In this paper we are interested in such a setting as illustrated in Fig. 1.

This project is funded in part by the ANR (grant ANR-12-SECU-0005)

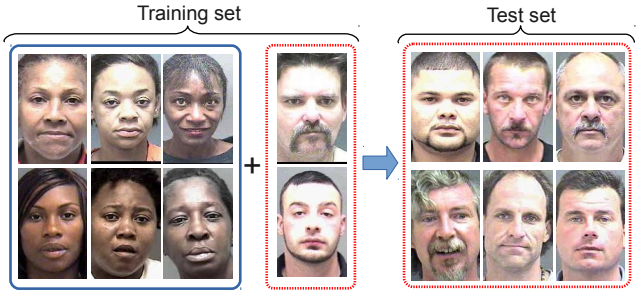


Fig. 1. Illustration of the proposed setting of cross domain age estimation. The algorithm learns a projection and a regressor jointly, to align source and target face domains and predict ages in the target domain. The training is mainly with source domain examples complemented very few target domain examples, while testing is done on target domain images only. The source and target domains may differ in age range, sex, race etc.

As explained above, we are interested in the problem of estimating age from face images, in a cross-population setting i.e. we have a large number of training examples available in one domain (the source domain) but only a very few ones in another domain (the target domain). We would like to utilise the training examples of the source domain to improve the performance of age estimation on the target domain. This problem was first posed and addressed by Guo et al. [12]. In their approach, they used a variant of LDA (Linear Discriminant Analysis) to learn common projection matrix which aligns aging patterns from source and target. However, they need a large number of target instances to learn target domain aging pattern, which are often not available in practice. Similarly, Alnajar et al. [13] proposed a method to do cross expression age estimation. But, the datasets they used for their experiments, FACES and LifeSpan are rather small and do not reflect the situation where abundant training data is available in the source domain.

We propose a joint learning method which (i) learns a subspace for aligning features from source and target domain and (ii) learns a regressor in this subspace for predicting ages. Our projection learning approach is similar to the metric learning method of Mignon and Jurie [14] – the projection matrix is

learnt to satisfy sparse pairwise (dis)similar constraints and age prediction based constraints simultaneously. We show empirically that the proposed method is consistently better than several strong baselines including those based on discriminative metric learning. We obtain state-of-the-art performance on the largest publicly available age estimation dataset. In the following, we discuss the proposed method in Sec. 2 then in Sec. 3 we provide the experimental results and, finally, in we conclude in Sec. 4.

2. PROPOSED METHODS

We now explain the proposed method in detail. We first introduce Metric Learning (ML) in general and then we explain how it can be used for learning a projection to align features from source and target domains. Finally, we explain the proposed Joint Learning (JL) algorithm.

2.1. Metric Learning and its application to cross-domain classification

Metric Learning (ML) has been quite successful in various facial analysis tasks such as face recognition [14, 15] and face retrieval [16]. Mahalanobis-like ML can be seen as learning a projection to map high dimensional features into a lower dimensional subspace where the pairwise constraints are better satisfied. For a pair of descriptors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{d'}$, ML involves the task of learning a Mahalanobis-like metric of the form $D^2_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top M(\mathbf{x}_i - \mathbf{x}_j)$, parameterized by positive semi-definite matrix M. As M is PSD, it can be decomposed as $M = L^\top L$. The problem can then be re-formulated as that of finding a linear subspace, into which features are first mapped and then compared with Euclidean distance i.e.,

$$D^2_L(\mathbf{x}_i, \mathbf{x}_j) = ||L\mathbf{x}_i - L\mathbf{x}_j||^2_2 \tag{1}$$

In the present case, we are given a training set of face images represented by their feature vectors and annotated with their ages i.e. $\mathcal{T} = \{(X, Y) : X \in \mathbb{R}^{d' \times N}, Y \in \mathbb{N}^N\}$. We construct two other sets from this information, set of *similar* vectors $\mathcal{S}$ annoated as $y_{ij} = 1$ and that of *dissimilar* ones $\mathcal{D}$, annotated as $y_{ij} = -1$, given by

$$\mathcal{S} = \{(i, j) : |y_i - y_j| \leq \delta\} \tag{2}$$

$$\mathcal{D} = \{(i, j) : |y_i - y_j| > \delta\} \tag{3}$$

with $\delta = 0$. We are interested in learning a mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$ to predict the age of new test faces where $d \ll d'$. We impose pairwise similarity and dissimilarity constraints, in the present case, and formulate the learning similar to the approach of Mignon and Jurie [14] i.e. optimize the objective function given as,

$$\min_L \mathcal{L}(\mathcal{T}, \mathcal{S}, \mathcal{D}; L) = \sum_{\mathcal{S} \cup \mathcal{D}} \ell_L(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) \tag{4}$$

$$\ell_L(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) = \max[0, m - y_{ij}(b - D^2_L(\mathbf{x}_i, \mathbf{x}_j))],$$

Algorithm 1 Joint learning of projection and regressor

Input: (i) Projection matrix, L ; Regressor $\mathbf{w}$, ii) Set of face features $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d' \times N}$, Set of age annotations $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^N$ (ii) Sparse pairwise age annotation $\mathcal{S}, \mathcal{D}$ (iii) maximum iterations `max-iters`, $\epsilon, m, \beta, \gamma$, `learn-rate:`$r$
Output: L, $\mathbf{w}$
while $t <$ `max-iters` **do**
  Choose $(\mathbf{x}_i, \mathbf{x}_j) \in X^2$ randomly
  $L_t \leftarrow L_{t-1}$, $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1}$
  $\Delta y_i \leftarrow |\mathbf{w}_{t-1} L_{t-1} \mathbf{x}_i - y_i|$
  if $\Delta y_i > \epsilon$ **then**
    $L_t \leftarrow L_t - r\beta \mathbf{w}_{t-1} \mathbf{x}_i^\top$
    $\mathbf{w}_t \leftarrow \mathbf{w}_t - r(\beta L_{t-1} \mathbf{x}_i + \mathbf{w}_{t-1})$
  end if
  $\Delta y_j \leftarrow |\mathbf{w}_{t-1} L_{t-1} \mathbf{x}_j - y_j|$
  if $\Delta y_j > \epsilon$ **then**
    $L_t \leftarrow L_t - r\beta \mathbf{w}_{t-1} \mathbf{x}_j^\top$
    $\mathbf{w}_t \leftarrow \mathbf{w}_t - r(\beta L_{t-1} \mathbf{x}_j + \mathbf{w}_{t-1})$
  end if
  $D^2_{L_{t-1}}(\mathbf{x}_i, \mathbf{x}_j) \leftarrow \|L_{t-1}\mathbf{x}_i - L_{t-1}\mathbf{x}_j\|^2$
  if $y_{ij}(1 - D^2_{L_{t-1}}(\mathbf{x}_i, \mathbf{x}_j)) < m$ **then**
    $L_t \leftarrow \left(1 - r\gamma y_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right) L_t$
  end if
end while

using stochastic gradient descent. In this equation, m and b are called margin and bias respectively and are free parameters. We generate the pairwise constraints from the large number of examples from source domain and a limited number of examples from the target domain. This is similar to the approach of Saenko et al. [17], who use ML for cross-domain image classification. It is important to note here that, the pairs they generated were from the examples belonging to two different domains. In [17], after learning projection matrix, training examples are projected into this subspace and classifier is trained in this subspace.

2.2. Proposed joint learning for cross-domain regression

An immediate extension of the approach of Saenko et al. [17] for regression could be similar ML projection followed by regressor learning. The problem with such approach is that it would not directly address the main goal of minimizing the absolute age difference between the ground truth age and predicted age. Moreover, pairwise constraints try to bring images belonging to same age categories together but push away the images belonging to different age categories. They push dissimilar pair away equally i.e. without taking into consideration the difference in their ages. For example, two pairs of images with the ages (25, 26) and (25, 55) are equally pushed apart. Unlike classification tasks, it is important to address this issue in regression tasks. Incorporating the re-

gressor while learning projection matrix address this problem by pushing the ages with lesser difference comparatively less farther.

We are thus interested in learning a projection L and a regressor $\mathbf{w}$, in the resulting space, jointly. We propose to minimize the following objective for learning $\mathbf{w}, L$,

$$\min_{L,\mathbf{w}} \mathcal{L}(\mathcal{T}, \mathcal{S}, \mathcal{D}; L, \mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \beta \sum_k \ell_{\mathbf{w}}(L\mathbf{x}_k, y_k) + \gamma \sum_{\mathcal{S}\cup\mathcal{D}} \ell_L(\mathbf{x}_i, \mathbf{x}_j, y_{ij}) \quad (5)$$

where, the first term is ℓ^2 regularization on $\mathbf{w}$, $\beta, \gamma \in \mathbb{R}$ are free parameters controlling the relative contributions of the different terms, $\ell_{\mathbf{w}}$ is the support vector regression loss which aims to bring the predicted age within $\pm\epsilon \in \mathbb{R}^+$ of the true age, given by:

$$\ell_{\mathbf{w}}(L\mathbf{x}, y) = \max(0, |\mathbf{w}^\top L\mathbf{x} - y| - \epsilon) \quad (6)$$

where $\ell_L(\mathbf{x}_i, \mathbf{x}_j, y_{ij})$ is the loss which aims at bringing similar age pairs together while pushing dissimilar age pairs away from each other. In practice, we optimize the objective using a stochastic gradient based solver, which is detailed in Alg. 1.

3. EXPERIMENTS

Dataset. We use the largest publicly available dataset for age estimation, the MORPH-II dataset, to evaluate the proposed method. We followed the experimental setup of Guo et al. [12] and compared the performance of our method with their method. We computed Local Binary Patterns (LBP) [18] of face images instead of Biologically Inspired Features (BIF) [10] which they used for their experiments. The database contains around 55k images from different races ('Black', 'White', 'Caucasian', etc.) and genders ('Male', 'Female'). Similar to [12], we took randomly sampled subsets of the database for the experiments. We took images from two races 'Black', and 'White', and two genders 'Male', and 'Female'. This subset contains 2,570 White Female (WF), 7,960 White Male (WM), 2,570 Black Female (BF), and 7,960 Black Male (BM) face images. Each of these categories is called a domain. From each of these domains, 50% of randomly sampled images are used for training and validation purposes and the rest 50% are used for testing. We used SVM regressor for predicting ages. The performance is calculated by Mean Absolute Error (MAE). MAE is the mean of absolute difference between the ground truth age and the predicted age.

Face Description. We used Viola and Jones face detector [19] to compute the bounding boxes of faces. These bounding boxes were resized to the size of 250×250. We computed facial landmarks using publicly available state-of-art facial landmark detector [20][1]. With the help of these facial landmarks we align the faces if required. The aligned faces are then centre cropped into the size of 160×100. We then compute local binary patterns (LBP) for each of these images using the publicly available `vlfeat` [21] library. We set cell size is equal to 10 as parameter and obtain a signature for each of the images which are of 9280 dimensions. Note however, the proposed method can work with other types of features e.g. LQP [22], LHS [23] or Fisher Vectors [24].

[1] https://github.com/soundsilence/FaceAlignment

3.1. Baselines

As a first reference we used the full features without any projection learning and hence without any compression. In addition, we compared with the following competitive baselines.

Unsupervised compression. We used Whitened Principal Components (WPCA) to compress high dimensional LBP to 64 dimensions. For training and testing, these representations are very efficient but suboptimal, as they may remove some discriminative information for age prediction.

Supervised Compression with ML. We used ML to learn compact representation of images which retains some discriminative information. We initialized with WPCA and learned the projection with stochastic gradient descent. This approach not only samples features that are useful for age estimation, but also aligns the features between the source and target domains.

After compressing, and potentially aligning the domains, for all these baselines, we use the publicly available SVR from `scikit-learn` [25] to learn the model on projected features to predict the ages. For all the experiments reported, we chose a linear kernel. We split train set into two halves for cross-validation. We set $\epsilon = 0.1$ and select the C parameter for SVR by cross-validation.

3.2. Proposed joint approach

Joint Learning (JL) learns the regressor and projection in with an integrated objective function. The advantage of JL in comparison to ML is that it takes care of dissimilarity constraint between the ages. As mentioned before in the Section 2, ML pushes the dissimilar images equally farther irrespective of difference between the ages. We trained JL identically cf. ML; we used the same training pairs that were used for ML and initialized the projection matrix with WPCA and regressor by mean of the principal components of WPCA. Since we learned a projection matrix of dimensions 64, our regressor has 64 dimensions. The initial values of regressor are mean values of 64 principal components. We set learning rate to 0.001, the margin m to 0.2 and the number of maximum iterations to 2×10^5. For the regressor, we set $\epsilon = 0.1$, similar to that of standard SVR we used for all the baselines.

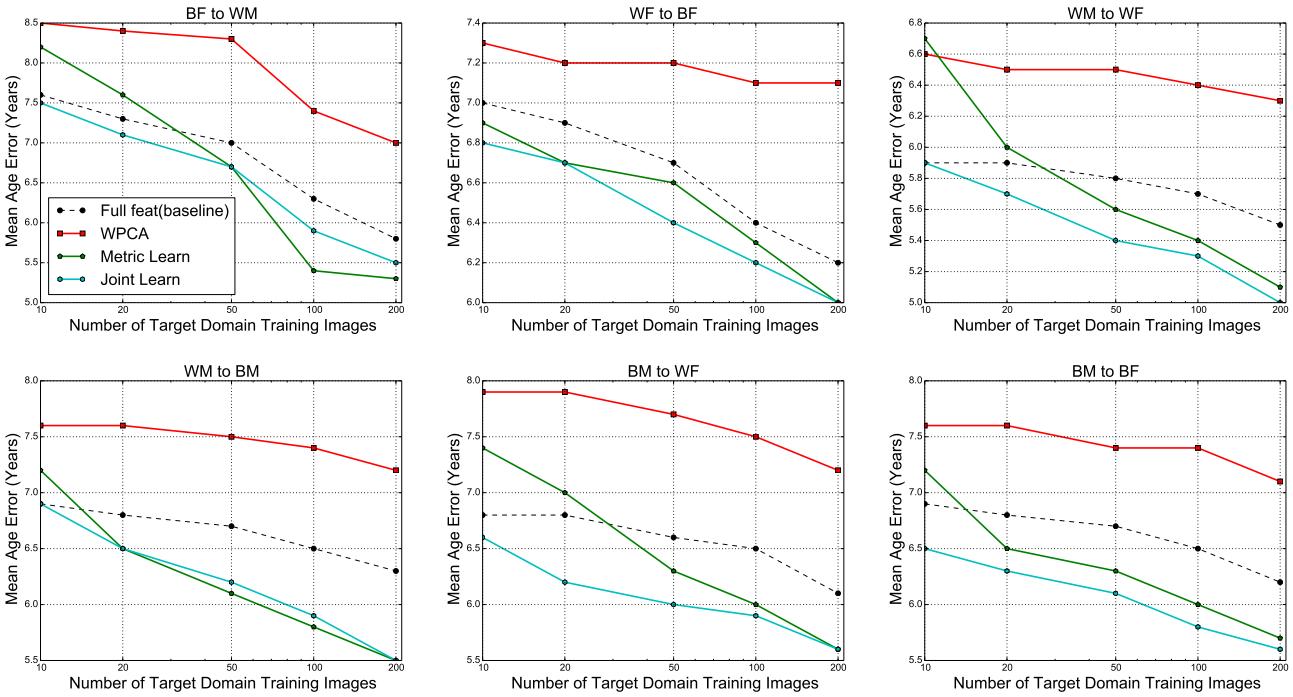


Fig. 2. Graphs showing performance of different approaches vs. the number of target training examples.

3.3. Experimental Results

Fig. 2 shows the performance of all the baselines and the one of our approach w.r.t. the size of the number of target training examples in 6 unique domain pairs. When we exchange the role of source and target of these 6 pairs, we get 12 domain pairs, which constitutes the total number domain pairs in our experiments. Tab. 1 shows the performances of our method along with those of the baselines and the current state-of-art method of Guo et al. [12]. The values in the table shows the Mean (over 12 domain pairs) of the MAE (mean average error over examples) in years in relation with the number of Target Training Examples (TTE) used. It usually requires a large number of labeled examples per class to compute scatter matrix using LDA, so we assume Guo et al. used more than 200 examples. In the domains, WF and BF, 200 examples counts around $(200/1285) \times 100 = 15.6\%$ and in WM, BM, it counts $(200/3980) \times 100 = 5\%$ of the training examples.

We note that, in comparison to the baselines i.e. LBP and WPCA, the proposed method consistently performs better. In comparison to ML, it performs better when the training examples from target domain is very small; whereas ML performs even worse than WPCA in such case (e.g. source target pair WM and WF). ML overfits when the positive training pairs are very small in number. This is an important practical use case, as often obtaining annotated examples of a new target domain is expensive. With the increasing size of target examples, the performance of ML ultimately converges to that of JL. Finally, the proposed approach clearly out-performs previous state-of-the-art method [12] by just taking 20 training examples from the target domain.

Method	LBP	WPCA	ML	JL
Dimensions	9280	64	64	64

TTE	Method	Mean of MAE (y)
>200	[12]	**6.6 ± 1.0**
0	LBP	6.8 ± 0.8
	WPCA	7.4 ± 0.7
10	LBP	6.8 ± 0.7
	WPCA	7.4 ± 0.7
	ML	7.2 ± 0.7
	JL	**6.7 ± 0.7**
20	LBP	6.7 ± 0.7
	WPCA	7.3 ± 0.7
	ML	6.7 ± 0.5
	JL	**6.5 ± 0.6**
50	LBP	6.5 ± 0.5
	WPCA	7.3 ± 0.7
	ML	6.2 ± 0.4
	JL	**6.1 ± 0.4**
100	LBP	6.2 ± 0.4
	WPCA	7.0 ± 0.6
	ML	**5.8 ± 0.4**
	JL	**5.8 ± 0.4**
200	LBP	5.9 ± 0.5
	WPCA	6.8 ± 0.6
	ML	**5.5 ± 0.4**
	JL	**5.5 ± 0.4**

Table 1. Performance comparison between different baselines, our approach and previous state-of-art method [12].

4. CONCLUSIONS

We propose a novel joint learning method for cross-domain age estimation. We have evaluated our method on the largest publicly available dataset. The proposed experimental validation shows that our method outperforms wide ranges of strong baselines, improves the performance over the previous state-of-art algorithm and attains a state-of-art performance.

5. REFERENCES

[1] Hu Han, Christina Otto, and Anubhav K Jain, "Age estimation from face images: Human vs. machine performance," in *ICB*, 2013, pp. 1–8.

[2] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy, "Cumulative attribute space for age and crowd density estimation," in *CVPR*, 2013, pp. 2467–2474.

[3] Zheng Song, Bingbing Ni, Dong Guo, Terence Sim, and Shuicheng Yan, "Learning universal multi-view age estimator using video context," in *ICCV*, 2011, pp. 241–248.

[4] Pavleen Thukral, Kaushik Mitra, and Rama Chellappa, "A hierarchical approach for human age estimation," in *ICASSP*, 2012, pp. 1529–1532.

[5] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung, "A ranking approach for human ages estimation based on face images," in *ICPR*, 2010, pp. 3396–3399.

[6] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles, "Automatic age estimation based on facial aging patterns," *PAMI*, vol. 29, no. 12, pp. 2234–2240, 2007.

[7] Caifeng Shan, Fatih Porikli, Tao Xiang, and Shaogang Gong, *Video Analytics for Business Intelligence*, vol. 409, Springer, 2012.

[8] Guodong Guo and Guowang Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *CVPR*, 2011, pp. 657–664.

[9] Guodong Guo, Guowang Mu, Yun Fu, Charles Dyer, and Thomas Huang, "A study on automatic age estimation using a large database," in *ICCV*, 2009, pp. 1986–1991.

[10] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S Huang, "Human age estimation using bio-inspired features," in *CVPR*, 2009, pp. 112–119.

[11] Guodong Guo and Guowang Mu, "Human age estimation: What is the influence across race and gender?," in *CVPR Workshops*, 2010, pp. 71–78.

[12] Guodong Guo and Chao Zhang, "A study on cross-population age estimation," in *CVPR*, 2014, pp. 4257–4263.

[13] Fares Alnajar, Zhongyu Lou, Jose Alvarez, and Theo Gevers, "Expression-invariant age estimation," in *BMVC*, 2014.

[14] Alexis Mignon and Frédéric Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.

[15] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? metric learning approaches for face identification," in *ICCV*, 2009, pp. 498–505.

[16] Binod Bhattarai, Gaurav Sharma, Frederic Jurie, and Patrick Pérez, "Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval," in *ECCV Workshops*, 2014, pp. 160–172.

[17] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *ECCV*. 2010, pp. 213–226, Springer.

[18] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen, *Computer vision using local binary patterns*, vol. 40, Springer, 2011.

[19] Paul Viola and Michael J Jones, "Robust real-time face detection," *IJCV*, vol. 57, no. 2, pp. 137–154, 2004.

[20] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.

[21] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[22] Sibt Ul Hussain, Thibault Napoléon, Fréderic Jurie, et al., "Face recognition using local quantized patterns," in *BMVC*, 2012.

[23] G. Sharma, S. ul Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *ECCV*, 2012.

[24] Karen Simonyan, Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Fisher vector faces in the wild," in *BMVC*, 2013.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-learn: Machine learning in python," *JMLR*, vol. 12, pp. 2825–2830, 2011.

MLBoost Revisited: A Faster Metric Learning Algorithm for Identity-Based Face Retrieval

Romain Negrel
romain.negrel@unicaen.fr
Alexis Lechervy
alexis.lechervy@unicaen.fr
Frederic Jurie
frederic.jurie@unicaen.fr

Normandie Univ, UNICAEN, ENSICAEN, CNRS
France

Abstract

This paper addresses the question of metric learning, *i.e*. the learning of a dissimilarity function from a set of similar/dissimilar example pairs. This domain plays an important role in many machine learning applications such as those related to face recognition or face retrieval. More specifically, this paper builds on the recent MLBoost method proposed by Negrel *et al*. [25]. MLBoost has been shown to perform very well for face retrieval tasks, but this algorithm relies on the computation of a weak metric which is very time consuming. This paper demonstrates how, by introducing sparsity into the weak projectors, the convergence time can be reduced up to a factor of 10× compared to MLBoost, without any performance loss. The paper also introduces an explicit way to control the rank of the so-obtained metrics, allowing to fix in advance the dimension of the (projected) feature space. The proposed ideas are experimentally validated on a face retrieval task with three different signatures.

1 Introduction

This paper focuses on the task of *identity-based face retrieval*. This has been a very dynamic research field over the past five years, raising many interesting challenges and producing a variety of interesting methods. Identity-based face retrieval heavily depends on the quality of the similarity function used to compare faces. Instead of using standard or handcrafted similarity functions, the most popular way to address this problem is to learn adapted metrics from sets of similar/dissimilar example pairs. It is usually equivalent to projecting face signatures into an adapted (possibly low-dimensional) space in which similarity can be measured with the Euclidean distance. For large scale applications, the dimensionality of this subspace should be as small as possible to limit the storage requirements, while the projection should also be fast to compute. Interestingly, the Euclidean metric fulfill the second requirement, which explains why producing face representations adapted to the Euclidean metric is interesting. However, such representations are usually of large

size. Several methods have been proposed to learn projection matrices reducing the size of the signatures while preserving the performance. This paper builds on such approaches.

More precisely, this paper proposes two improvements over MLBoost – MLBoost [25] is a supervised metric leaning method based on boosting – one of the state-of-the-art Mahalanobis metric leaning methods. These two contributions are:

- The introduction of a new way to compute the weak metrics at a lower computational cost;
- The introduction of a new approach to control the rank of the learned metrics, allowing to fix the dimensions of the low-dimensional space in which the images are represented.

The rest of the paper is as follows: after reviewing some metric learning techniques in Section 2 and giving more details on MLBoost [25] in Section 3, the proposed contributions are presented in Section 4. Section 5 compares the proposed method with state-of-the-art competitors and shows its benefits.

2 Related Works

During the last decades, many Metric Learning (ML) approaches have emerged and have been used in diverse applications such as tracking, image retrieval, face verification, person re-identification, *etc*. ML also plays an important role in many machine learning, pattern recognition or data mining techniques as learning metrics from data is usually better than designing hand crafted metrics. In practice, not only should the metric be good in terms of performance, but also it has to be fast, not memory demanding and computationally scalable.

The literature on ML is too vast to be fully covered here, and the interested reader is referred to the recent book of Bellet *et al*. [2]. We can, however, mention a few of the most notable approaches such as: DDML [15], RBML [20], Structural ML [41], PCCA [22], rPCCA [40], LMNN [37], LDML [14], ITML [11], KISSME [17], RS-KISSME [34], SML [7], MLBoost [25]. Most of these supervised approaches learn a distance or a similarity function based on the Mahalanobis distance. The Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j \in \mathbb{R}^D$ is defined as:

$$D_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{W} (\mathbf{x}_i - \mathbf{x}_j), \tag{1}$$

where $(\mathbf{x}_i, \mathbf{x}_j)$ denotes the pair of samples to compare and $\mathbf{W} \in \mathcal{M}_{D\times D}$ is a positive semi-definite matrix. The seminal work of [39] estimated $\mathbf{W}$ by solving a convex quadratic programming problem, by satisfying constraints defined by some given training pairs.

However guaranteeing the positive semi-definiteness of $\mathbf{W}$ is computationally expensive. To reduce this cost, several works suggested to factorize $\mathbf{W}$ as $\mathbf{W} = \mathbf{L}\mathbf{L}^\top$ with $\mathbf{L} \in \mathcal{M}_{D\times d}$. In this case, $\mathbf{W}$ is by construction a positive semi-definite matrix and $\mathbf{L}$ defines an implicit projection matrix ($\mathbf{y}_i = \mathbf{L}^\top \mathbf{x}_i$). thus, it is possible to impose rank constraints to regularize the model and learn a smaller feature space ($d << D$).

In the following, we denote by $(\mathbf{p}_{1i}, \mathbf{p}_{2i}) \in \mathcal{P}$ the set of positive pairs (two samples belonging to the same class) and by $(\mathbf{n}_{1j}, \mathbf{n}_{2j}) \in \mathcal{N}$ the set of negative pairs (two samples belonging to different classes). We also write $\mathrm{D}_{\mathbf{L}}$ instead of $D_{\mathbf{L}\mathbf{L}^\top}$, for simplicity.

In [3], Bellman highlighted the phenomenon called the *curse of dimensionality*: when the dimensionality of the feature space increases, the data representation becomes sparse. In general, this sparsity is problematic, in particular for any method that requires statistical significance. This is why a lot of ML techniques have proposed to reduce the dimension of the data space [10, 12, 33, 35]. For example, [23] and [24] proposed different (unsupervised and supervised) methods to reduce the dimension of large-size descriptors (from thousands to millions dimensions). PCCA [22, 40] proposed to learn a matrix $\mathbf{L}$ used to project the signatures into a low-dimensional space where the distance between similar pairs are smaller than those of dissimilar pairs. To do this, the authors suggested to solve the following optimization problem:

$$\arg\min_{\mathbf{L}} \sum_{\mathcal{P}} \ell_\beta \left(\mathrm{D}_{\mathbf{L}} \left(\mathbf{p}_{1i}, \mathbf{p}_{2i}\right) - 1\right) + \sum_{\mathcal{N}} \ell_\beta \left(1 - \mathrm{D}_{\mathbf{L}} \left(\mathbf{n}_{1i}, \mathbf{n}_{2i}\right)\right) + \lambda \left\|\mathbf{L}\right\|_F^2, \tag{2}$$

with $\ell_\beta(x) = \frac{1}{\beta} \log\left(1 + \exp\left(\beta x\right)\right)$, where β and λ are two hyper-parameters.

Tuning these hyper-parameters is not an easy task and is application dependent. Interestingly, several methods don't use any hyper-parameters. This is the case of the KissMe method, introduced in [17], formulating the learning problem as a likelihood-test between two Gaussian distributions (one for similar and one for dissimilar pairs). Consequently, it is easy to compute $\mathbf{W}$ such as:

$$\mathbf{W} = \Sigma_{\mathcal{P}}^{-1} - \Sigma_{\mathcal{N}}^{-1}, \tag{3}$$

with $\Sigma_{\mathcal{P}} = \sum_{\mathcal{P}} (\mathbf{p}_{1i} - \mathbf{p}_{2i})(\mathbf{p}_{1i} - \mathbf{p}_{2i})^\top$ and $\Sigma_{\mathcal{N}} = \sum_{\mathcal{N}} (\mathbf{n}_{1j} - \mathbf{n}_{2j})(\mathbf{n}_{1j} - \mathbf{n}_{2j})^\top$. Despite this method is very fast and is not requiring any hyper-parameters, it cannot guarantee that the metric is positive-definite (*i.e.*, distances are not necessarily positive). The authors proposed to project $\mathbf{W}$ on the cone of positive semi-definite matrices when $D_{\mathbf{W}}$ is not exactly a metric.

Recently, several researchers investigated the use of Boosting algorithms [30] for ML. Boosting algorithms are interesting as they do not have, in general, any hyper-parameters and are not prone to overfitting [29]. Strong metrics can be obtained by combining several weak metrics (generally rank-one metrics) to solve an optimization problem with triplet-wise constraints [6, 21, 31, 32]. Negrel *et al.* [25] introduced MLBoost, showing how to learn a boosted metric using pairwise constraints only, in a fast and scalable way.

Several Boosting methods have been developed with computational and storage efficiency in mind. A first strategy is to reduce the computational cost for learning weak learners. This is rather natural as, in boosting, it is better to have simple weak classifiers; a good example is the Haar basis functions introduced in [36]. A second strategy consists in evaluating less or using less weak learners. In [36], a cascade approach is introduced to reduce the average number of weak classifiers evaluated during the test stage. FloatBoost [19] uses a backtracking mechanism: in the training phase, after each iteration of AdaBoost, some weak classifiers are removed. As the number of weak classifiers selected does not change, the time required to compute the metric is controlled. Furthermore, removing some weak classifiers allows to remove the bad ones, improving both convergence and performance. Finally, using a fixed number of weak learners [1, 13] or updating the weak learners after their selection [27] have been studied a lot in the tracking literature.

In this paper, we propose two contributions for reducing the learning cost. First, we propose a novel fast weak ML algorithm; second, we add rank constraints on the strong metric, allowing us to fix the maximal dimension of the so-produced feature space, even when the number of boosting iteration increases.

Algorithm 1 Efficient MLBoost implementation

procedure MLBOOST($\mathbf{X}, \mathcal{P}, \mathcal{N}, itersMax$)
  $t \leftarrow 1$
  $\mathbf{L}^{(1)} \leftarrow \varnothing$
  Initialize weights: $\forall i,\ u_i^{(1)} = 1/|\mathcal{P}|$; $\forall j,\ v_j^{(1)} = 1/|\mathcal{N}|$.
  repeat
    Compute weak metric $\mathbf{z}^{(t)}$ with equation (4).
    Choose the best $\alpha^{(t)}$ with equation (6).
    if $\alpha^{(t)} \leq 0$ **then**
      break
    $\mathbf{L}^{(t+1)} \leftarrow \left[\mathbf{L}^{(t)}, \sqrt{\alpha^{(t)}}\mathbf{z}^{(t)}\right]$.
    Update weights $u_i^{(t+1)}$ and $v_j^{(t+1)}$ with equations (7).
  until $t < itersMax$
  return $\mathbf{L}$

3 Boosted Metric Learning (MLBoost)

This section briefly summarizes the recent MLBoost approach – an efficient technique allowing to learn metrics with Boosting – such as introduced in [25]. MLBoost learns a decomposition of a Mahalanobis-based metric $\mathbf{L}$. Like other boosting techniques, MLBoost combines the weak learners obtained at each iteration to form a strong classifier.

At the beginning, all the pairs are initialized with the identical weights ($u_i^{(1)} = 1/|\mathcal{P}|$ for positive pairs and $v_j^{(1)} = 1/|\mathcal{N}|$ for negative pairs). The weak metric $D_{\mathbf{z}^{(t)}}$ is then obtained by solving the following optimization problem:

$$\begin{aligned} \mathbf{z}^{(t)} = &\arg\max_{\mathbf{z}} \mathbf{z}^\top \mathbf{A}^{(t)} \mathbf{z}, \\ &\text{s.t. } \|\mathbf{z}\|_2 = 1, \text{with} \end{aligned} \tag{4}$$

$$\mathbf{A}^{(t)} = \sum_{\mathcal{N}} v_j^{(t)} \left((\mathbf{n}_{1j} - \mathbf{n}_{2j})(\mathbf{n}_{1j} - \mathbf{n}_{2j})^\top \right) - \sum_{\mathcal{P}} u_i^{(t)} \left((\mathbf{p}_{1i} - \mathbf{p}_{2i})(\mathbf{p}_{1i} - \mathbf{p}_{2i})^\top \right). \tag{5}$$

We note that solving problem (4) is equivalent to the computation of the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}^{(t)}$. Once the weak metric $D_{\mathbf{z}^{(t)}}$ is computed, the algorithm chooses the best weights $\alpha^{(t)}$ by solving the following problem:

$$\alpha^{(t)} = \arg\min_{\alpha} \left(\sum_{\mathcal{P}} u_i^{(t)} \mathrm{e}^{\alpha \left(\mathrm{D}_{\mathbf{z}^{(t)}}(\mathbf{p}_{1i}, \mathbf{p}_{2i}) \right)} \right) \left(\sum_{\mathcal{N}} v_j^{(t)} \mathrm{e}^{-\alpha \left(\mathrm{D}_{\mathbf{z}^{(t)}}(\mathbf{n}_{1j}, \mathbf{n}_{2j}) \right)} \right). \tag{6}$$

At the end of each iteration, the weights of the training pairs are updated by:

$$u_i^{(t+1)} = \frac{u_i^{(t)} \mathrm{e}^{\alpha^{(t)} \mathrm{D}_{\mathbf{z}^{(t)}}(\mathbf{p}_{1i}, \mathbf{p}_{2i})}}{w_{\mathcal{P}}^{(t)}}, \ \forall i \qquad v_j^{(t+1)} = \frac{v_j^{(t)} \mathrm{e}^{-\alpha^{(t)} \mathrm{D}_{\mathbf{z}^{(t)}}(\mathbf{n}_{1j}, \mathbf{n}_{2j})}}{w_{\mathcal{N}}^{(t)}}, \ \forall j. \tag{7}$$

Algorithm 2 Low cost weak metrics

procedure LCWEAKMETRIC($\mathbf{X}, \mathcal{P}, \mathcal{N}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, J$)
  Select randomly a subset I of $J < D$ indices $I = \{i_1, \cdots, i_J\}$.
  $\mathbf{X}' \leftarrow \mathbf{X}_I$.
  Compute $\mathbf{A}^{(t)} \in \mathcal{M}_{J\times J}$ with equation (5) and $\mathbf{X}'$.
  Solve equation (4), *i.e.* compute $\mathbf{z}'^{(t)}$, the first eigenvector of $\mathbf{A}^{(t)}$ matrix.
  Create a vector of zero entries: $\mathbf{z}^{(t)} \in \mathbb{R}^W$.
  Set the value of $\mathbf{z}^{(t)}$ in indices by I: $\mathbf{z}^{(t)}_I \leftarrow \mathbf{z}'^{(t)}$.
  return $\mathbf{z}^{(t)}$

The different steps of this algorithm are summarized in Algorithm 1.

MLBoost is robust to overfitting and is free of any hyper-parameters. However, one of its drawbacks is that the final size of the so-obtained feature space can be very large. Furthermore, computing the weak learners is very expensive.

4 Faster MLBoost

Our contribution for faster MLBoost is twofold: first, we introduce a new way of building weak learners; second we propose a better way to control the rank (and consequently the dimension of the signature) of the Mahanalobis matrix. The two contributions are presented in the two following subsections.

4.1 Producing weak metrics at lower cost

As explained previously, MLBoost relies on the computation of a weak metric, which is computationally expensive. This cost depends on two parameters: the dimensionality of the input features and the numbers of positive and negative pairs. More precisely, the weak metric is computed in two steps: first, matrix $\mathbf{A}^{(t)}$ is computed using equation (5); second, the Rayleigh quotient of equation (4) is obtained by computing the first eigenvector of $\mathbf{A}^{(t)}$. These two steps have (at least) a quadratic complexity with respect to size of the signature and hence become intractable for large signatures.

In order to reduce the computational cost of the weak metric, we propose to sparsify the weak metric projectors. We do it by arbitrarily setting some of the components of the projectors to zero, allowing to consider only the dimensions of the signatures corresponding to the non-null dimensions of the projectors. These non-null components of the weak metric projectors are randomly selected and uniformly distributed. Algorithm 2 summarizes this strategy. For clarity purposes, we introduce τ, *i.e.* the ratio of non-zero dimensions defined as: $\tau = J/D$, where $J < D$ is the number of non-null components and D is the size of the descriptors. The ratio τ can be seen as the proportion of the non-null components.

The so-computed sparse weak metrics are weaker than those of [25] and more boosting iterations are necessary to reach convergence. However, in the end, the speedup of each iteration is so important that the overall learning time is drastically reduced. We can explain the overall

gain by the fact that sampling only a few components reduces the time required to learn the weak classifiers in a quadratic way. On the other hand, we observe that the components are correlated, explaining why keeping only a fraction of them does not result in a strong degradation of the performance. In addition, the proposed random sampling ensures more diversity than optimally selecting the components (*e.g.* using PCA).

4.2 Explicitly Controlling the Rank of MLBoost

As discussed in the related work section (Section 2), controlling the dimensionality of the image signatures is very interesting for practical reasons. This can be done by controlling the rank of the Mahalanobis matrix. As MLBoost adds a new weak metric at each iteration, the rank of the Mahalanobis matrix is increased, iteration by iteration. The only way to control the rank is then to fix the number of boosting rounds, *e.g.* to $\mathbf{T}^{(t)} = \left[\mathbf{L}^{(t)}, \sqrt{\alpha^{(t)}}\mathbf{z}^{(t)}\right]$ which is inconsistent with the general principle of Boosting (*i.e.* the combination of lots of weak learners to obtain a strong learner).

We argue, in this paper, that a better way to control of the rank ($\text{rank}(\mathbf{W}) \leq R$) consists in adding an extra projection at each iteration. This projection is done in two steps: (i) we approximate the current Mahalanobis metric by a Mahalanobis metric with a rank lower or equal to R; (ii) we compute the best weighting of the new metric before using it in the boosting process.

The current Mahalanobis metric $\mathrm{D}_{\mathbf{T}^{(t)}}(\cdot,\cdot)$ can be approximated by solving:

$$\mathbf{P}^{(t)} = \underset{\mathbf{P}\in\mathcal{M}_{W\times R}}{\arg\min} \sum_{ij} \left(\mathrm{D}_{\mathbf{T}^{(t)}}(\mathbf{x}_i,\mathbf{x}_j) - \mathrm{D}_{\mathbf{P}}(\mathbf{x}_i,\mathbf{x}_j)\right)^2, \tag{8}$$

where $\mathbf{x}_i$ denotes the training samples.

This problem (Eq. (8)) is a standard Multi-Dimensional Scaling (MDS) problem [9]. Moreover as the Mahalanobis metric (1) can be seen as a Euclidean metric in the reduced subspace, then we solve this problem easily by using a Principal Component Analysis (PCA) in the reduced subspace:

$$\mathrm{Cov}(\mathbf{Y}) = \mathbf{V}\Lambda\mathbf{V}^\top, \tag{9}$$

with $\mathbf{Y} = \mathbf{T}^{(t)\top}\mathbf{X}$ and $\mathbf{X}$ the matrix containing the training pairs (vectors of differences), $\mathbf{V}$ the eigenvectors of the covariance, and Λ the diagonal matrix containing eigenvalues of the covariance matrix. The optimal matrix $\mathbf{P}^{(t)}$ is computed by combining the current Mahalanobis matrix $\mathbf{T}^{(t)}$ with the R eigenvectors corresponding to the largest eigenvalues $\mathbf{V}_{\{1,\cdots,R\}}$:

$$\mathbf{P}^{(t)} = \mathbf{T}^{(t)}\mathbf{V}_{\{1,\cdots,R\}}. \tag{10}$$

In this case, $\mathbf{P}^{(t)}$ is the best R-dimensional approximation of $\mathbf{T}^{(t)}$. However, it is not possible to directly replace $\mathbf{T}^{(t)}$ by $\mathbf{P}^{(t)}$ in the next steps of MLBoost. As in the first boosting step, we need to compute the weights of the metric:

$$\mathbf{L}^{(t+1)} = \sqrt{\alpha_2^{(t)}}\mathbf{P}^{(t)}, \tag{11}$$

Sign.	Method	Final Dim.	n= 1	n= 10	n= 20	n= 50	n= 100
LBP	-	9860	31.9	53.7	60.5	68.8	74.7
	PCA	16	10.2	24.8	34.5	44.7	55.3
		32	16.5	34.5	44.7	55.3	66.0
		128	28.4	46.6	54.6	65.7	72.1
		512	31.2	51.5	59.6	67.4	74.7
		585	36.4	57.7	64.3	74.2	79.7
	KissMe	-	24.3	53.6	59.5	69.7	78.3
	MLBoost	585	**40.2**	**60.8**	**66.7**	**74.9**	**81.1**
AlexNet	-	4096	78.3	92.2	94.8	97.2	97.9
	PCA	16	53.7	82.7	89.1	94.3	96.7
		32	70.7	90.5	92.4	96.2	97.6
		128	75.7	91.7	94.6	96.9	98.1
		383	78.7	92.7	94.8	97.4	**98.3**
		512	78.7	92.4	94.8	97.4	**98.3**
	KissMe	-	76.6	92.4	95.3	96.9	97.8
	MLBoost	383	**81.3**	**93.9**	**96.0**	**97.9**	**98.3**
VGG-Face	-	4096	89.6	96.9	97.4	**98.1**	98.3
	PCA	16	75.7	91.3	93.9	96.2	97.6
		32	87.7	95.0	95.7	97.2	97.9
		128	91.3	96.5	97.2	97.6	98.3
		191	**91.5**	96.5	97.2	97.6	98.6
		512	91.3	96.7	96.9	97.6	98.3
	KissMe	-	90.1	96.7	97.2	97.6	**98.8**
	MLBoost	191	**91.5**	**97.2**	**97.9**	**98.1**	98.3

Table 1: Baseline performance of 3 types of descriptors with (i) Euclidean metric (ii) Euclidean metric after PCA reduction (iii) KissMe [17] (iv) MLBoost [25].

Sign.	Final Dim.	n= 1	n= 10	n= 20	n= 50	n= 100
LBP	1226	41.8	61.4	68.6	75.4	80.9
AlexNet	1128	81.8	94.3	95.7	97.9	98.1
VGG-Face	538	**91.7**	**96.5**	**97.4**	**98.3**	**98.8**

Table 2: Performance of MLBoost with low-cost weak metrics ($\tau = 5\%$), for the three types of signatures.

Sign.	Final Dim.	n= 1	n= 10	n= 20	n= 50	n= 100
LBP	16	18.7	43.5	52.7	64.3	74.0
	32	31.4	57.0	63.1	72.1	77.3
	128	36.4	54.8	62.9	71.6	77.5
	512	**38.5**	**58.6**	**63.6**	**74.0**	**79.2**
AlexNet	16	60.0	97.9	91.3	93.4	94.8
	32	73.5	92.2	95.3	**97.6**	98.1
	128	78.0	93.9	**95.7**	**97.6**	97.9
	512	**79.0**	**94.1**	**95.7**	**97.6**	**98.3**
VGG-Face	16	82.0	94.1	96.7	97.6	98.6
	32	89.4	96.2	97.4	98.1	98.6
	128	90.8	95.7	97.2	98.1	**98.8**
	512	**92.4**	**96.7**	**97.6**	**98.3**	98.6

Table 3: Performance of MLBoost with low-cost weak metrics ($\tau = 5\%$) and rank constraints ($R \in \{16, 32, 128, 512\}$).

where $\alpha_2^{(t)}$ denotes the weights. Indeed, at the end of each boosting iteration, weighting the training pairs makes the previous weak metric performing as well as a random metric. To compute α_2, we solve (via line search) the following problem:

$$\alpha_2^{(t)} = \arg\min_{\alpha} \left(\sum_{\mathcal{P}} e^{\alpha \left(D_{\mathbf{P}^{(t)}}(\mathbf{p}_{1i}, \mathbf{p}_{2i}) \right)} \right) \left(\sum_{\mathcal{N}} e^{-\alpha \left(D_{\mathbf{P}^{(t)}}(\mathbf{n}_{1j}, \mathbf{n}_{2j}) \right)} \right). \tag{12}$$

Finally, we update the weights of the training pairs as follows:

$$u_i^{(t+1)} = \frac{e^{D_{\mathbf{L}^{(t+1)}}(\mathbf{p}_{1i}, \mathbf{p}_{2i})}}{w_{\mathcal{P}}^{(t+1)}}, \ \forall i \qquad v_j^{(t+1)} = \frac{e^{-D_{\mathbf{L}^{(t+1)}}(\mathbf{n}_{1j}, \mathbf{n}_{2j})}}{w_{\mathcal{N}}^{(t+1)}}, \ \forall j \tag{13}$$

with $w_{\mathcal{P}}^{(t+1)}$ and $w_{\mathcal{N}}^{(t+1)}$ the normalization factors chosen such that $\sum u_i^{(t+1)} = 1$ and $\sum v_i^{(t+1)} = 1$.

5 Experiments

The two contributions of this paper are experimentally evaluated on the identity-based face retrieval task, *i.e.* given a face query, the objective is to find a face of the same person in a set of known-identity face images and hence predict the identity of the query face. The criterion used to evaluate the performance is the one used in [4, 25], *i.e.*, the mean k-call@n (such as defined in [8]), with $k = 1$. for $n \in \{1, 10, 20, 50, 100\}$.

Datasets and learning pairs. We use the aligned version [38] of the Labeled Faces in the Wild (LFW) database by Huang *et al*. [16]. It contains more than 13,000 images of over 4,000 different persons. In our experiments, we use the same set of images/queries as [4, 25]. Only the identities having at least five examples are used; the others are not used during the learning of metrics nor during their evaluations. This results in a subset of 5,985 images of 423 different persons. The query set is composed of one image of each identity while the training set contains the remaining images. To learn the metrics, we build a set of similar pairs and a set of dissimilar pairs in such a way that all the identities are used equally.

Image descriptions. We evaluate the methods with three types of image signatures: (i) LBP [26]: we use the same signatures as in [4, 25] (signatures of 9860 dimensions). (ii) AlexNet descriptors [18]: we use the same descriptors as Bhattarai *et al*. [5](signatures of 4096 dimensions). (iii) VGG-Face CNN descriptors [28]: we use the publicly available source code[1] (signatures of 4096 dimensions).

MLBoost learning parameters. We learn the metrics using $2^{17} \approx 131,000$ positive and negative examples pairs. Boosting is stopped when the objective function is lower than 10^{-9} or the maximum number of iterations is reached, *i.e*., 2048 iterations. To evaluate the metric learned with MLBoost, we project the signatures on the projectors $\mathbf{y}_i = \mathbf{L}^{*\top}\mathbf{x}_i$ and we normalize (ℓ_2) the reduced signatures $\mathbf{y}'_i = \mathbf{y}_i/\|\mathbf{y}_i\|$. We then use the Euclidean metric to compare the queries with the images of the test set.

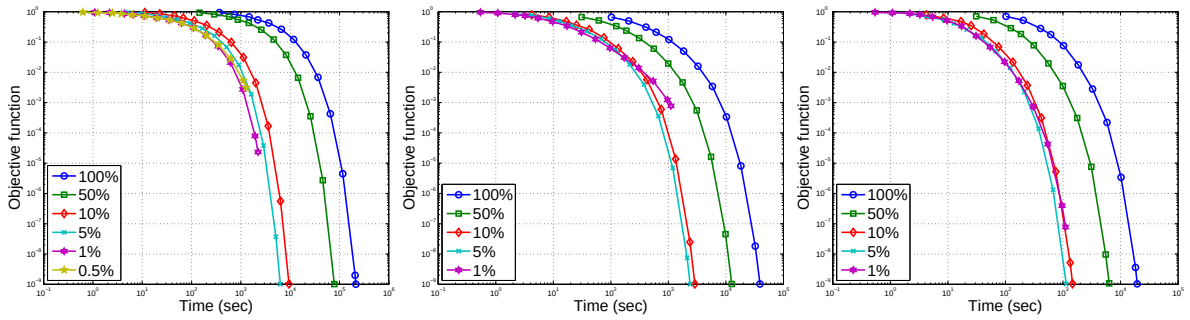
Baseline results. We use as a baseline the performance obtained with: (i) raw signatures (without metric learning) / Euclidean distance; (ii) signatures reduced by PCA; (iii) KissMe [17]; (iv) MLBoost [25]. The results are reported in Table 1, which compares the performance obtained with the three types of signatures (LBP, AlexNet and VGG-Face). The performances are given in terms of the percentage of the mean 1-call@n. To learn the metric with KissMe, we use the signatures reduced to 128 dimensions with PCA, and we use only $2^{14} \approx 16000$ positive and negative pairs (setting giving the best performance).

We can see that the recent CNN signatures provide much better performance than LBP. We also note that for AlexNet and VGG-Face, PCA can improve the performance (for 128-d or more projections). We can finally see that the metric learned with MLBoost constantly improves the performance, for all types of signatures.

5.1 Low cost weak metric performance

To analyze the effects of our low-cost weak metric on the convergence speed and metric performance, we learn the metrics for the different types of signatures and for various ratios of non-zeros dimensions $\tau \in \{100\%, 50\%, 10\%, 5\%, 1\%\}$. We note that $\tau = 100\%$ is equivalent to the original MLBoost of [25]. Figures 1(a), 1(b) and 1(c) illustrate the convergence of the algorithm for the different ratios of non-zeros components. The vertical axis corresponds to the objective function

[1] http://www.robots.ox.ac.uk/~vgg/software/vgg_face/



(a) LBP (b) AlexNet (c) VGG-Face

Figure 1: Objective as the function of the accumulated time spent on learning the weak metrics, for different values of τ. τ is the parameter fixing the ratio of non-zeros dimensions in the low-cost MLBoost weak metric.

while the horizontal axis corresponds to the accumulated time spent on computing the weak metrics during boosting. We see that for any type of signatures, the overall time spent in computing the weak metrics before the objective function reaches 10^{-9} is significantly reduced. For a ratio of 5% of non-zeros dimensions, the total time is at least divided by a factor of 10. Table 2 gives the performance of the metrics learned with our low-cost weak metric with 5% of non-zeros dimensions. The performance is very similar to those of the weak metric proposed in [25] (see Table 1). However, the dimension of the final signature is larger, due to the larger number of iterations needed reach convergence.

5.2 Adding Rank Constraints

In this section, we focus on the evaluation of our second contribution, *i.e.* the method proposed to limit the rank of the Mahalanobis matrix. We perform these experiments with our low-cost week metric with 5% of non-null components ($\tau = 0.05$), for the following rank constraint: $R \in \{16, 32, 64, 128, 256, 512\}$. Figure 2 illustrates the convergence of the algorithm (using LBP signatures) for the different rank constraints. The vertical axis corresponds to the objective function while the horizontal axis corresponds to the number of boosting iterations. The blue curve shows the convergence of MLBoost without rank constraints. We see that for strong rank constraints (*e.g.*, $R = 16$), the convergence speed is reduced. However, for $R = 64$, $R = 128$ and $R = 256$, we note that we need fewer iterations to converge than without the rank constraint. We report, in Table 3, the performance given by the metrics learned with MLBoost combined with our low-cost weak metric and the rank constraint. We see that the performance increases with R. In comparison to the original MLBoost (see Table 1), and for any type of signature, we always obtain better performance. The conclusion is that not only is the proposed method faster, but it is also better in terms of performance.

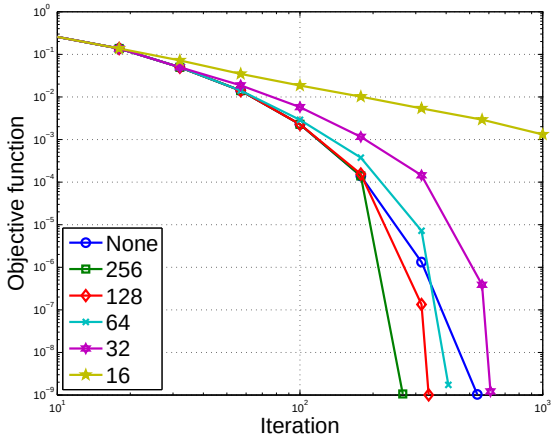


Figure 2: Effect of the rank constraints on MLBoost as a function of the number of iterations

6 Conclusions

This paper introduces two improvements to the state-of-the-art MLBoost method [25]. The first one addresses the prohibitive computational cost required to learn weak metrics in the presence of high-dimensional signatures. The second contribution allows us to limit the rank of the Mahalanobis matrix and, thus, to fix the dimension of the final signatures. The proposed experimental validation not only show a more than $10\times$ speedup but also a significant improvement of the performance. In addition, the paper shows that the size of the final signature can significantly be reduced with only a small loss in performance.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-12-SECU-0005 (project PHYSIONOMIE).

References

[1] Shai Avidan. Ensemble tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):261–271, February 2007. ISSN 0162-8828.

[2] Aurélien Bellet, Amaury Habrard, and Marc Sebban. *Metric Learning*. Morgan & Claypool Publishers, 2015.

[3] Richard E Bellman. *Adaptive control processes: a guided tour*. Princeton university press, 2015.

[4] Binod Bhattarai, Gaurav Sharma, Frederic Jurie, and Patrick Pérez. Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In *European Conference on Computer Vision (ECCV) Workshops*, pages 1–13, 2014.

[5] Binod Bhattarai, Gaurav Sharma, and Frédéric Jurie. Cp-mtml: Coupled projection multi-task metric learning for large scale face retrieval. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] Jinbo Bi, Dijia Wu, Le Lu, Meizhu Liu, Yimo Tao, and Matthias Wolf. AdaBoost on low-rank psd matrices for metric learning. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2617–2624. IEEE, 2011.

[7] Qiong Cao, Yiming Ying, and Peng Li. Similarity metric learning for face recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2408–2415. IEEE, 2013.

[8] Harr Chen and David R Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 429–436. ACM, 2006.

[9] Michael A.A. Cox and Trevor F. Cox. *Handbook of Data Visualization*, chapter Multidimensional Scaling, pages 315–347. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[10] John P. Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16:2859–2900, 2015.

[11] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 209–216, New York, NY, USA, 2007. ACM.

[12] Imola K. Fodor. A survey of dimension reduction techniques. Technical report, Technical report, Lawrence Livermore National Laboratory, 2002.

[13] Helmut Grabner and Horst Bischof. On-line boosting and vision. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 260–267, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2597-0.

[14] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Face recognition from caption-based supervision. *International Journal of Computer Vision*, 96(1): 64–82, 2012.

[15] Junlin Hu, Jiwen Lu, and Yap P. P. Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882, 2014.

[16] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[17] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Stan Z Li and ZhenQiu Zhang. Floatboost learning and statistical face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1112–1123, 2004.

[20] Venice E. Lion, Jiwen Lu, and Yongxin Ge. Regularized bayesian metric learning for person re-identification. In *ECCV Workshop on Visual Surveillance and Re-Identification*, 10 2014.

[21] Meizhu Liu and Baba C. Vemuri. A robust and efficient doubly regularized metric learning approach. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*, pages 646–659, 2012.

[22] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2666–2672. IEEE, 2012.

[23] Romain Negrel, David Picard, and Philippe-Henri Gosselin. Web scale image retrieval using compact tensor aggregation of visual descriptors. *IEEE MultiMedia*, 20(3):24–33, March 2013.

[24] Romain Negrel, David Picard, and Philippe-Henri Gosselin. Dimensionality reduction of visual features using sparse projectors for content-based image retrieval. In *IEEE International Conference on Image Processing*, pages 2192–2196, Paris, France, October 2014.

[25] Romain Negrel, Alexis Lechervy, and Frederic Jurie. Boosted metric learning for efficient identity-based face retrieval. In *British Machine Vision Conference*, volume 13, pages 1007–1036, 2015.

[26] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[27] Toufiq Parag, Fatih Porikli, and Ahmed Elgammal. Boosting adaptive linear weak classifiers for online learning and tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[28] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[29] Lev Reyzin and Robert E. Schapire. How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 753–760, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2.

[30] Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2014. ISBN 0262526034.

[31] Chunhua Shen, Alan Welsh, and Lei Wang. Psdboost: Matrix-generation linear programming for positive semidefinite matrices learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1473–1480. Curran Associates, Inc., 2009.

[32] Chunhua Shen, Junae Kim, Lei Wang, and Anton Van Den Hengel. Positive semidefinite metric learning using boosting-like algorithms. *The Journal of Machine Learning Research*, 13(1):1007–1036, 2012.

[33] C. O. S. Sorzano, J. Vargas, and A. Pascual Montano. A survey of dimensionality reduction techniques, 2014.

[34] Dapeng Tao, Lianwen Jin, Yongfei Wang, Yuan Yuan, and Xuelong Li. Person re-identification by regularized smoothing kiss metric learning. *Circuits and Systems for Video Technology, IEEE Transactions on*, 23(10):1675–1685, 2013.

[35] Laurens.J.P. van der Maaten, Eric. O. Postma, and H. Jaap van den Herik. Dimensionality reduction: A comparative review. Technical report, Technical report, Tilburg University, 2009.

[36] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518, 2001.

[37] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[38] Lior Wolf, Tal Hassner, and Yaniv Taigman. Similarity scores based on background samples. In *Computer Vision–ACCV 2009*, pages 88–97. Springer, 2009.

[39] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance Metric Learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, Vancouver, Bristish Columbia, December 2002.

[40] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.

[41] Gang Yuan, Zhaoxiang Zhang, and Yunhong Wang. Enhancing person re-identification by robust structural metric learning. In *Image and Graphics (ICIG), 2013 Seventh International Conference on*, pages 453–458. IEEE, 2013.

RPNet: an End-to-End Network for Relative Camera Pose Estimation

Sovann En, Alexis Lechervy, and Frédéric Jurie

Normandie Univ, UNICAEN, ENSICAEN, CNRS — UMR GREYC
firstname.lastname@unicaen.fr

Abstract. This paper addresses the task of relative camera pose estimation from raw image pixels, by means of deep neural networks. The proposed RPNet network takes pairs of images as input and directly infers the relative poses, without the need of camera intrinsic/extrinsic. While state-of-the-art systems based on SIFT + RANSAC, are able to recover the translation vector only up to scale, RPNet is trained to produce the full translation vector, in an end-to-end way. Experimental results on the Cambridge Landmark data set show very promising results regarding the recovery of the full translation vector. They also show that RPNet produces more accurate and more stable results than traditional approaches, especially for hard images (repetitive textures, textureless images, *etc.*). To the best of our knowledge, RPNet is the first attempt to recover full translation vectors in relative pose estimation.

Keywords: relative pose estimation · pose estimation · posenet

1 Introduction

In this paper, we are interested in *relative camera pose estimation* — a task consisting in accurately estimating the location and orientation of the camera with respect to another camera's reference system. Relative pose estimation is an essential task for many computer vision problems, such as Structure from Motion (SfM), Simultaneous Localisation And Mapping (SLAM), *etc.* Traditionally, this task can be accomplished by i) extracting sparse keypoints (ex. SIFT, SURF), ii) establishing 2D correspondences between keypoints and iii) estimating the essential matrix using 5-points or 8-point algorithms [13]. RANSAC is very often used to reject outliers in a robust manner.

This technique, although it has been considered as the de facto standard for many years, presents two main drawbacks. First, the quality of the estimation depends heavily on the correspondence assignment. This is to say, too few correspondences (textureless objects) or too many noisy correspondences (repetitive texture or too much viewpoint change) can lead to surprisingly bad results. Second, the traditional method is able to estimate the translation vector only up to scale (directional vector).

In this paper, our objective is three folds: i) we propose a system producing more stable results ii) recovering the full translation vector iii) and we provide insights regarding relative pose inference (*i.e.* from absolute pose, regressor *etc.*).

As pointed out in [20], CNN based methods are able to produce pretty good results in some cases where SIFT-based methods fail (*i.e.* texture less images). This is the reason why we opted for a global method based on CNN. Inspired by the success of PoseNet [9], we propose a modified Siamese PoseNet for relative camera pose estimation, dubbed as RPNet, with different ways to infer the relative pose. To the best of our knowledge, [12] is the only end-to-end system aiming at solving relative camera pose using deep learning approach. However, their system estimate the translation vector up to scale, while ours produces full translation vectors.

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 introduces the network architecture and the training methodology. Section 4 discusses the datasets and presents the experimental validation of the approach. Finally, Section 5 concludes the paper.

2 State of the Art

Local keypoint-based approaches. They address relative camera pose estimation using the epipolar geometry between 2D-2D correspondences of keypoints. Early attempts aimed at better engineering interest point detectors to focus on interesting image properties such as corners [6], blobs in scale-space [10], regions [11], or speed [2, 18, 16] *etc.* More recently, there is a growing interest to train interest point detectors together with the matching function [5, 23, 17, 4, 19]. LIFT [21] adopted the traditional pipeline combining a detector, an orientation estimator, and a descriptor, tied together with differentiable operations and learned end-to-end. [1] proposed a multitask network with different sub-branches to operate on varying input sizes. [4] proposed a bootstrapping strategy by first learning on simple synthetic data and increasing the training set with real images in a second time.

End-to-End pose estimation. The first end-to-end neural network for camera pose estimation from single RGB images is PoseNet [9]. It is based on GoogLeNet with two output branches to regress translations and rotations. PoseNet follow-up includes: Baysian PoseNet [7], Posenet-LSTM [20] where LSTM is used to model the context of the images, Geometric-PoseNet where the loss is calculated using the re-projection error of the coordinates using the predicted pose and the ground truth [8]. Since all the 3D models used for comparisons are created using SIFT-based techniques, traditional approach seems more accurate. [20] showed that the classical approaches completely fail with less textured datasets such as the proposed TMU-LSI dataset. [14] is an end-to-end system for pose regression taking sparse keypoint as inputs. Regarding relative pose estimation, [12] is the only system we are aware of. Their network is based on ResNet35 with FCs layers acting as pose regressor. Similar to the previous networks, the authors formulate the loss function as minimising the L2-distances between the ground truth and the estimated pose. Unfortunately, several aspects of their results (including their label generation, experimental methodology and the baseline system) make comparisons difficult. Along side with pose regression problems, another promis-

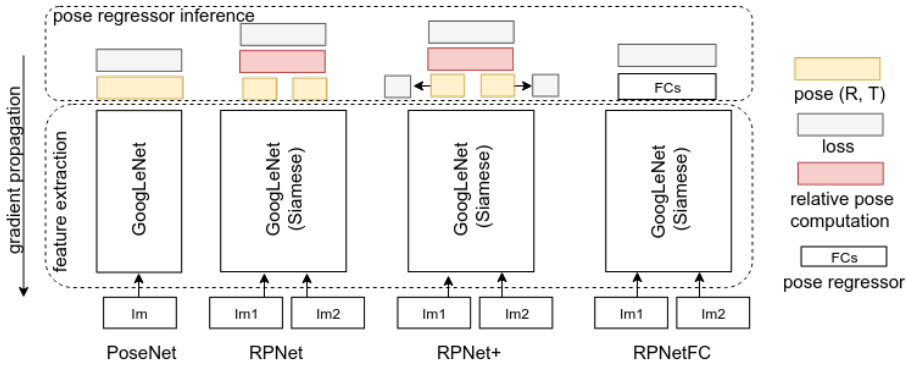


Fig. 1. Illustration of the proposed system

ing works from [15] showed that an end-to-end neural network can effectively be trained to regress to infer the homography between two images. Finally, two recent papers [22, 3] made useful contributions to the training of end-to-end systems for pose estimation. [22] proposed a regressor network to produce essential matrix which can be then used to find the relative pose. However, their system is able to find the translation up to scale which is completely different from our objective. In [3], a differentiable RANSAC is proposed for outlier rejection and can be a plug-and-play component into an end-to-end system.

3 Relative pose inference with RPNet

Architecture. The architecture of the proposed RPNet, illustrated Fig. 1, is made of two building blocks: i) a Siamese Network with two branches regressing one pose per image, ii) a pose inference module for computing the relative pose between the cameras. We provide three variants of the pose inference module: (1) a parameter-free module, (2) a parameter-free module with additional losses (same as PoseNet loss [9]) aiming at regressing the two camera poses as well as the relative pose, and (3) a relative pose regressor based on FC layers. The whole network is trained end-to-end for relative pose estimation. Inspired by PoseNet [9], the feature extraction network is based on the GoogLeNet architecture with 22 CNN layers and 6 inception modules. We only normalize the quaternion during test time. It outputs one pose per image.

For RPNet and RPNet$^+$, the module for computing the relative pose between the cameras is straightforward and relies on simple geometry. Following the convention of OpenCV, the relative pose is calculated in the reference system of the 2nd camera. Let (R_1, t_1, R_2, t_2) be the rotation matrices and translation vectors used to project a point X from world coordinate system to a fixed camera system (camera 1 & 2). Let (q_1, q_2) be the corresponding quaternions of (R_1, R_2). The relative pose is calculated as followed:

$$R_{1,2} = q_2 \times q_1^* \quad \text{and} \quad T_{1,2} = R_2(-R_1^T t_1) + t_2 \tag{1}$$

Table 1. Number of training and testing pairs for Cambridge Landmark dataset. SE stands for spatial extent, measured in meter.

Scene	Train	Test	SE	Scence	Train	Test	SE
Kings College	9.1k	2.4k	140x40	Shop Facade	1.6k	0.6k	35x25
Old Hospital	6.5k	1.2k	50x40	St Marys Church	11k	4.1k	80x60

where q_1^* is the conjugate of q_1, and $\times$ denotes the multiplication in the quaternion domain. Both equations are differentiable. For RPNetFC, the pose inference module is a simple stacked fully connected layers with $relu$ activation. To limit over-fitting, we modified the output of the Siamese network by reducing its output dimension from 2048 to 256. This results in almost 50% reduction of the number of parameters compared to PoseNet, RPNet and RPNet$^+$ network. The pose regressor network contains two FC layers (both with 128 dimensions).

Losses. The loss function uses the Euclidean distance to compare predicted relative rotation $\hat{q_{1,2}}$ and translation $\hat{T}_{1,2}$ with ground truth $\hat{q}_{1,2}$ and $q_{1,2} : loss = \sum_i (||\hat{T}^i_{1,2} - T^i_{1,2}||_2 + \beta * ||\hat{q}^i_{1,2} - q^i_{1,2}||_2)$. Quaternions are unit quaternions. The original PoseNet has a β term in front of quaternions to balance the loss values between the translation and rotation. To find the most suitable value of β, we cross-validated on our validation set. Please refer to our codes for different hyper-parameter values on different subsets.

4 Experimentations

4.1 Experimental Setup

Dataset. Experimental validation is done on the Cambridge Landmark dataset[1]. Each image is associated with a ground-truth pose. We provide results on 4 of the 5 subsets (scenes). As discussed by several people, the 'street' scene raises several issues[2].

Pair generation. For each sequence of each scene, we randomly pair each image with eight different images of the same sequence. For a fair comparison with SURF, the pair generation is done by making sure that they overlap enough. We followed the train-test splits defined with the data set. Images are scaled so that the smallest dimension is 256 pixels, keeping its original aspect ratio. During training, we use 224*224 random crops and feed them into the network. During test time, we center crop the image.

Baseline. The baseline is a traditional keypoint-based method (SURF). The focal length and the principle point are provided by the dataset. Other parameters are cross-validated on the validation set. For a fair comparison, we provide two scenarios for baselines: (1) the image are scaled to be 256*455 pixels, followed by a center-crop (224*224 pixels) to produce the same image pairs as tested with our networks and (2) the original images without down-sampling. We named

[1] `http://mi.eng.cam.ac.uk/projects/relocalisation`

[2] `https://github.com/alexgkendall/caffe-posenet/issues/2`

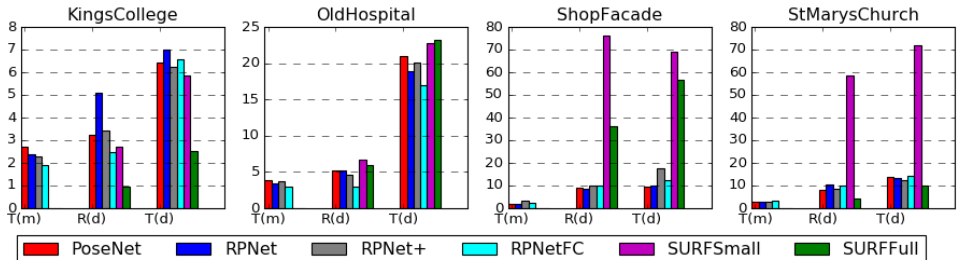


Fig. 2. Translation and Rotation errors (median) of the different approaches

these two scenarios as 'SURFSmall' and 'SURFFull'. All the camera parameters are adapted to the scaling and cropping we applied.

Evaluation metric. We measured 3 different errors: i) translation errors, in meters ii) rotation errors, in degrees and iii) translation errors in degrees. We report the median for all the measurements.

4.2 Experimental results

Relative pose inference module. Fig. 2 compares the performance of the different systems and test scenarios. Based on these experimental results, RPNetFC and RPNet$^+$ are the most efficient ways to recover the relative pose. On easy dataset (*i.e.*. KingsCollege and OldHospital), where there is no ambiguity textures, using pose regressor (RPNetFC) produces slightly better results than inferring the relative pose from the two images (PoseNet/RPNet/RPNet$^+$). On the contrary, on hard datasets (*i.e.*. ShopFacade and StMarysChurch), RPNet-family outperforms RPNetFC. This behavior is also true for relative rotation and relative translation measured in degree. Globally, RPNetFC produces the best results followed by RPNet$^+$, PoseNet and finally, RPNet. The differences of their results are between 0 and 8 degrees. Regarding technical aspect, RPNetFC is a lot easier to train than RPNet$^+$/RPNet since it does not involve multiple hyper-parameters to balance the different losses. It also converges faster.

Comparison with traditional approaches. We will start by discussing the SURFSmall scenario first. In general, the error on both translation and rotation can be reduced between 5 to 70% using RPNet family, except on KingsCollege where the traditional approach slightly outperforms RPNet-based methods. We observed that the performance of the traditional approaches varies largely from one subset to another, while RPNet$^+$/RPNetFC are more stable. In addition, the traditional approach requires camera information for each image in order to correctly estimate the pose. In contrast, RPNet-based does not require any specific information at all. Using the original image size (SURFFull) significantly boost the performance of the traditional approach. However, RPNetFC still enjoy a significant gain in performance on OldHospital and ShopFacade, while performing slightly worse than SURFFull on KingsCollege and StMarysChurch. The difference in performance between SURFFull and RPNetFC is even more significant when the images contain large view point changes (see Fig. 3).

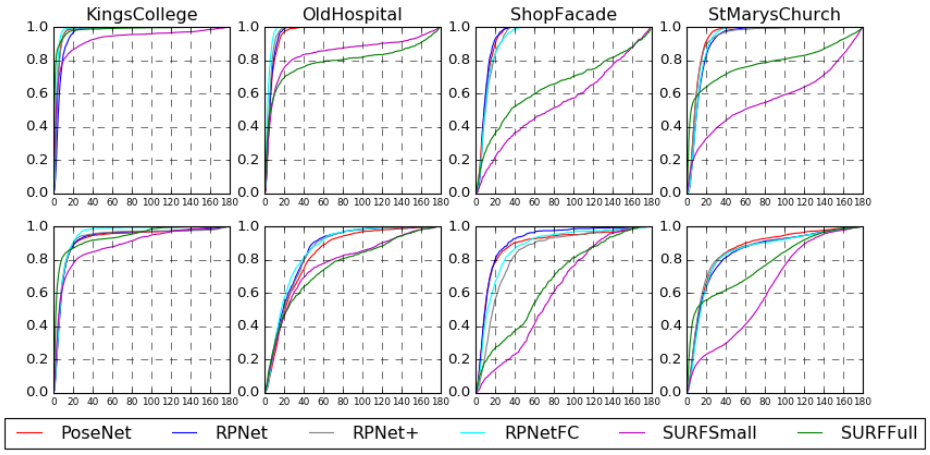


Fig. 3. Accumulative hist. of errors in rotation (1st row, d), translation (2nd row, m).

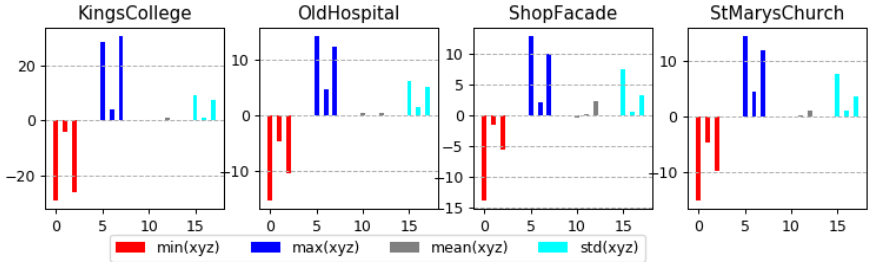


Fig. 4. Min/Max/Mean/STD relative translations (ground truth), w.r.t. XYZ axis (m).

Full translation vector. One of our objectives is to provide a system able to estimate the full translation vector. On average, we observed that the median error ranges between 2 to 4 meters, using RPNetFC. Fig. 4 gives an idea of ground truth translations w.r.t. reference axes (xyz). For instance, on KingsCollege, the values of X-axis can range from -29m to 30m with an STD of 7 meters. Interestingly, our network has a translation error of only 2.88 meters.

5 Conclusions

This paper proposed a novel architecture for estimating full relative poses using an end-to-end neural network. The network is based on a Siamese architecture, which was experimented with different ways to infer the relative poses. In addition, to produce competitive or better results over the traditional SURF-based approaches, our system is able to produce an accurate full translation vector. We hope this paper will provide more insight and motivate other researchers to focus on global end-to-end system for relative pose regression problems.

Acknowledgements. This work was partly funded by the French–UK MCM ITP program and by the ANR-16-CE23-0006 program.

References

1. Altwaijry, H., Veit, A., Belongie, S.J.: Learning to detect and match keypoints with deep architectures. In: Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016 (2016)
2. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European conference on computer vision. pp. 404–417 (2006)
3. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: Dsac-differentiable ransac for camera localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 3 (2017)
4. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. arXiv preprint arXiv:1712.07629 (2017)
5. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 3279–3286 (2015)
6. Harris, C.G., Stephens, M.: A combined corner and edge detector. In: Proceedings of the Alvey Vision Conference, AVC 1988, Manchester, UK, September, 1988. pp. 1–6 (1988)
7. Kendall, A., Cipolla, R.: Modelling uncertainty in deep learning for camera relocalization. In: Robotics and Automation (ICRA), 2016 IEEE International Conference on. pp. 4762–4769 (2016)
8. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6555–6564. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.694
9. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 2938–2946 (2015)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
11. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and vision computing **22**(10), 761–767 (2004)
12. Melekhov, I., Ylioinas, J., Kannala, J., Rahtu, E.: Relative camera pose estimation using convolutional neural networks. In: International Conference on Advanced Concepts for Intelligent Vision Systems. pp. 675–687 (2017)
13. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE transactions on pattern analysis and machine intelligence **26**(6), 756–770 (2004)
14. Purkait, P., Zhao, C., Zach, C.: Spp-net: Deep absolute pose regression with synthetic views. arXiv preprint arXiv:1712.03452 (2017)
15. Rocco, I., Arandjelovic, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 39–48. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.12
16. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: Computer Vision (ICCV), 2011 IEEE international conference on. pp. 2564–2571 (2011)
17. Tian, Y., Fan, B., Wu, F., et al.: L2-net: Deep learning of discriminative patch descriptor in euclidean space. In: Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2 (2017)

18. Tola, E., Lepetit, V., Fua, P.: Daisy: An efficient dense descriptor applied to wide-baseline stereo. IEEE transactions on pattern analysis and machine intelligence **32**(5), 815–830 (2010)
19. Trujillo, L., Olague, G.: Using evolution to learn how to perform interest point detection. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. vol. 1, pp. 211–214 (2006)
20. Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using lstms for structured feature correlation. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 627–637. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.75
21. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European Conference on Computer Vision. pp. 467–483 (2016)
22. Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). vol. 3 (2018)
23. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 4353–4361 (2015)

TS-NET: COMBINING MODALITY SPECIFIC AND COMMON FEATURES FOR MULTIMODAL PATCH MATCHING

Sovann En, Alexis Lechervy, Frédéric Jurie

Normandie Univ, UNICAEN, ENSICAEN, CNRS – UMR GREYC

ABSTRACT

Multimodal patch matching addresses the problem of finding the correspondences between image patches from two different modalities, *e.g.* RGB vs sketch or RGB vs near-infrared. The comparison of patches of different modalities can be done by discovering the information common to both modalities (Siamese like approaches) or the modality-specific information (Pseudo-Siamese like approaches). We observed that none of these two scenarios is optimal. This motivates us to propose a three-stream architecture, dubbed as TS-Net, combining the benefits of the two. In addition, we show that adding extra constraints in the intermediate layers of such networks further boosts the performance. Experimentations on three multimodal datasets show significant performance gains in comparison with Siamese and Pseudo-Siamese networks[†].

***Index Terms*—** Multimodal Patch Matching, Siamese network, Deep Metric Learning

1. INTRODUCTION AND RELATED WORK

Patch matching, the task consisting in determining the correspondences between image patches, is essential for many computer vision problems, *i.e.*, multi-view reconstruction, structure from motion, object-instance recognition, *etc.* In this work, we aim to study the problem of matching patches in a multimodal setting where input patches come from different sources, *i.e.* RGB images vs hand-drawn sketches or RGB vs near-infrared images.

Broadly speaking, there are two main ways to design local patch matching systems, either by employing handcrafted features or through machine learning techniques. Pioneer works in patch matching [1] are based on handcrafted features such as the SIFT descriptor/detector or some variants, *e.g.* [2], DAISY, [3], *etc.* Such approaches usually use conventional distance to measure patch similarity, *e.g.* the Euclidean distance, which usually does not provide an optimal solution for matching purposes. This family of approaches relies heavily on human expertise.

In contrast with feature engineering, another approach to patch matching consists in using supervised algorithms to find adapted features or adapted similarity functions, for given datasets. Machine learning allows to find optimal projections minimizing (or maximizing) the distances between positive patches (negative patches respectively) [4, 5, 6].

Recent breakthroughs in deep learning have strongly contributed to this field. One of the first works in deep metric learning is the one of Jahrer *et al.* [7] introducing a Siamese networks inspired by the LeNet5 networks, and comparing the so-obtained features with the Euclidean distance. Since then, Siamese networks have been very popular in the literature. Several variants have been proposed, differing by their weight-sharing strategy [8] (Siamese vs Pseudo Siamese), combinations of the inputs [8, 9] (two channels input images vs multi-scale images), similarity functions (conventional distance [8, 10, 9, 11] or using metric layers [12]).

Another important aspect when training deep networks for patch matching is the objective function. It can be (a) the cross entropy (binary classification loss) [12] (b) the hinge loss [8] (c) the triplet loss [13, 14] to incorporate the notion of relative distance, relative distance [9] (d) the global loss which models the loss as two distributions (positive and negative) to be pushed away from each other [14].

More specifically, the question of multimodal patch matching has been investigated recently by several authors. [15, 16] suggested to concatenate the different modalities as different channels of the input data. [17] experimented the use of Siamese networks for the matching of visible/SAR patches. [18] studied the quadratic network, a variant of the Siamese network that takes 4 patches as input. In the context of cross-spectral face recognition [19] proposed two components (one before and another one after the feature extraction network) to allow the system to transform the NIR images into the VIS spectrum. As we write, Siamese networks are still seen as a reference for multimodal patch matching.

One important aspect of Siamese architectures is that the weights of feature extraction towers are shared between the inputs. This is to say that the network is trained to extract characteristics present in both modalities. In case of the Pseudo-Siamese architectures, the feature towers are not shared: contrarily to the Siamese networks, the motivation is to extract modality specific information in order to better discriminate the pair of inputs. Our motivation in this paper is to take advantage of these two complementary aspects and

[†]Codes and resources available at `http://github.com/ensv/TS-Net`

propose a novel architecture, dubbed as TS-Net. It consists of two sub-networks, one Siamese and one Pseudo-Siamese networks. Their outputs are combined with a fully connected layer, acting as a weighting scheme between the modality specific information and the common information present in the input patches. The overall architecture is given Figure 1.

Our second contribution is to show that adding a constraint on the feature embedding, by means of a contrastive loss in the feature extraction tower, helps to boost the performance further. The idea is to encourage the network to bring projections of positive pairs closer in the Euclidean space. In the extreme case, this is equivalent to make two clusters of input pairs at the metric layers, allowing to easily separate them with an hyperplane instead of having to learn an arbitrarily complex boundary.

The rest of the paper is organized as follows: Section 2 introduces the network architecture and the training methodology. Section 3 discusses the datasets and presents the experimental validation of the approach. Finally, Section 4 concludes the paper.

2. THE PROPOSED THREE-STREAM NETWORK

As explained before, the proposed architecture for multimodal patch matching, denoted as the TS-Net architecture, is intended to combine the advantages of both Siamese and Pseudo-Siamese networks. The overall architecture of TS-Net is given Fig. 1(c). Each sub-network has 2 main parts: two feature extraction towers and a metric learning module. In the case of the Siamese network, the parameters of the feature extraction towers are shared, while for Pseudo-Siamese networks they are distinct. TS-Net takes a pair of patches as input, one from each modality, and predicts independently in each sub-network whether they are similar or not. Finally, the outputs of each sub-network are combined by an additional fully connected (FC) layer to produce the final prediction. In the next paragraphs, the different components of TS-Net are described and commented.

Feature extraction network. Each tower is based on convolutional and pooling layers to hierarchically extract high-level information from the input patches. We use max-pooling layers to reduce the dimensions of the feature maps by a factor of 2. At the end of the tower, we use a bottleneck (fully connected) layer to produce a compact output vector with 128 dimensions. Inspired from [12], we use $Relu$ activation as a non-linear activation function.

Tower Fusion. We observed in our experiments that subtracting the layers produced better performance than concatenating them, as in the original MatchNet. So the output of the feature extraction tower are element-wise subtracted before they are fed to the metric network.

Metric network. The metric learning part of the network consists of three fully connected layers. It takes a vector of 128 dimensions and produces a vector of dimension two, suitable for binary classification.

Losses. We treat patch matching as a binary classification problem, as we observed it performs better (also observed by [9]) than learning a similarity function. Therefore, Siamese and Pseudo-Siamese parts of TS-Net are trained with binary cross-entropy loss functions.

One contribution of this paper is to introduce additional constraints, at the feature level, by means of a contrastive loss [20] enforcing the features coming from the two feature towers to be close to each other if the pair is positive. This will enable the features of positive pairs to be in the hypersphere and the features of the negative pairs to be outside the hypersphere.

The fusion of Siamese and Pseudo-Siamese networks is done by introducing an additional cross-entropy loss on the top of the two.

More formally, let (x_1, x_2) be the input pair of patches and y the class label. $y = 1$ means the pair is positive (similar patches), $y = 0$ means the pair is negative (different patches). We denote by L_{en} and L_{con} the cross-entropy and the contrastive loss, with:
(a) $L_{en} = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})$ where $\hat{y}$ is output of the *Softmax* layer, and
(b) $L_{con} = y\frac{2}{Q}D^2 + (1 - y)2Qe^{\frac{-2.77}{Q}D}$ where D is the Euclidean distance between features. Q is the margin to be optimized. The overall loss function is then given by:
$L = L_{tsnet_{en}} + L_{siam_{en}} + L_{pseudo_{en}} + \lambda L_{siam_{con}} + \beta L_{pseudo_{con}}$, with λ and β two cross-validated parameters in $[0, 1]$.

In multimodal settings, it is not always guaranteed that the two modalities can be projected into the same subspace. In practice, we observed that optimal performance is obtained for λ and β set to 10^{-2} (values obtained by cross validating the parameters on the validation set).

Implementation details. We initialize the weights of each convolutional layer using the Xavier initialization and all the FC layers with a truncated normal distribution ($stddev = 0.005$ and $mean = 0, bias = 0.1$). While the original MatchNet is trained with plain stochastic gradient descent, we found that training with 0.95 momentum produce equal or better performance. We train the network with $lr = 10^{-3}$ with L2 regularization of 10^{-3} with neither dropout nor $batchnorm$. Q is optimized experimentally on VeDAI validation set and set to be 50 for the other two datasets. During training, we observe that the λ and β parameters should be carefully set and the best performance we obtain is for $\lambda = 10^{-2}, \beta = 10^{-4}$ or $\beta = 10^{-2}$ on CUHK and NIR Scene (cross validation experiments). We use batch size of 32 and train with at least 150 epochs. All the experimentations are done using Tensorflow 1.4 with NVIDIA P100 or K80 GPU. Patches are normalized

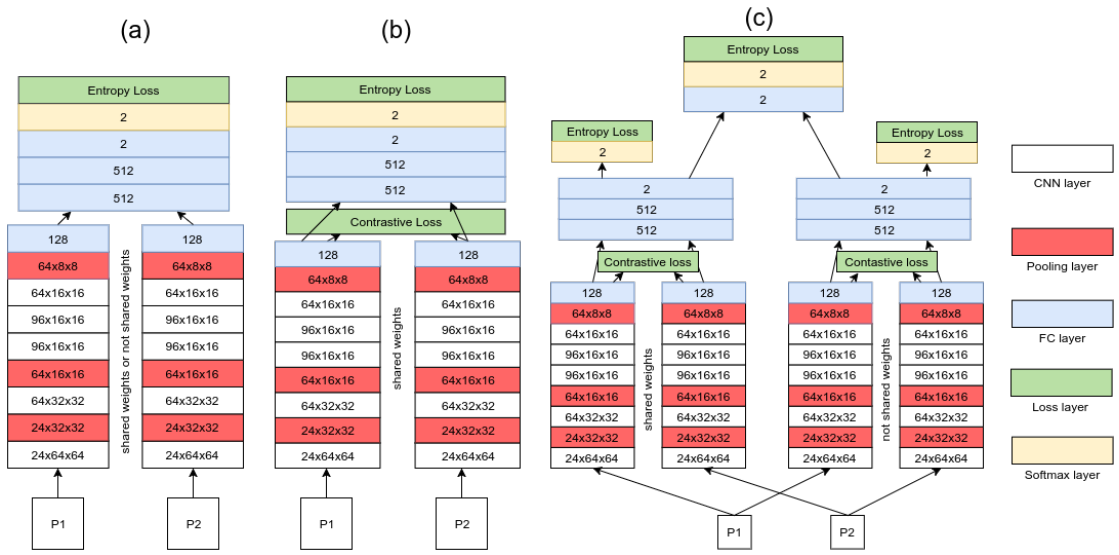


Fig. 1. The detailed architectures of (a) standard Siamese networks (b) Siamese networks with the proposed additional loss on the feature towers (c) the proposed TS-Net network with additional losses on the feature extraction tower and on the metric network. The numbers on each rectangle indicate the output size of this layer.

to have zero mean and unit standard deviation for each modality.

3. EXPERIMENTATIONS

Our aim in this section is to provide insights about TS-Net, its behavior and, more importantly, to draw a comparison with Siamese and Pseudo-Siamese networks, which are considered as a reference to this task. First, we run a series of experiments on the VeDAI dataset to validate TS-Net. It consists in evaluating different ways to fuse information either in the metric or after the feature extraction network. Next, we show that the gain in performance is not due to an increase of the number of parameters. Finally, we run experiments on three public datasets to experimentally validate our network and compare it to Siamese and Pseudo-Siamese networks. To report the performance, we employ the standard evaluation protocol defined in [6], namely the *95% error rate* criteria, abbreviated 95%ErrRate, which is the percentage of false matches present when 95% of all correct matches are detected. For each experimentation, we report the average performance with its standard deviation on at least 3 runs (Table 2) and 8 runs (Table 3).

Datasets The proposed approach is experimentally validated on three different datasets: VeDAI [*], RGB-NIR Scene [†] and CUHK [‡]. These 3 datasets contain images from two different modalities. It is worth mentioning that these 3 datasets have been created for different tasks. Therefore, it will provide an opportunity to test and compare performance on a variety of fields. For instance, VeDAI is generally used for Vehicle Detection in Aerial Imagery while CUHK for face sketch synthesis/recognition. VeDAI, CUHK and RGB-NIR Scene contain respectively a total of 1246, 188 and 477 pairs of images.

Pairs of Patch Generation. For each dataset, the images are given as sets of aligned pairs (one image from each modality). To extract patches and form pairs, we uniformly sample each image using grid-like layout where each cell has a width and height of 64 $\times$ 64 pixels. This gives us a collection of corresponding positive patches. We randomly choose patches coming from different images to form negative pairs.

To make our patch matching experiments more realistic and challenging, we artificially augment our datasets by introducing some random affine transformations between the images of the same pair. For each pair, we generate three additional pairs using a random combination of: (i) Rotation (-12 to 12 degrees), (ii) Translation (-5 to 5 pixels on both axes) and (iii) Scale (0.8 to 0.99). For the validation and test set, we keep only one pair among the four, chosen randomly. Table 1 summarizes the number of train, test and validation pairs of patches. Half are positive, half are negative.

Combining Siamese and Pseudo-Siamese networks. Our motivation is to find an efficient way to combine the information coming from the two sub-networks. We consider four options depending on whether this fusion (element-wise subtraction) is done (a) after the feature extraction tower (b) after the first (c) second or (d) third layer of the metric network.

[*] https://downloads.greyc.fr/vedai/

[†] https://ivrl.epfl.ch/supplementary_material/cvpr11/

[‡] http://mmlab.ie.cuhk.edu.hk/archive/facesketch.html

Table 1. Number of pairs of patches in the train, test and validation set, for each dataset. Each set contains 50% of positive pairs and 50% of negative ones.

Dataset	Train (70%)	Test (20%)	Validation (10%)
VeDAI	448k	128k	64k
CUHK	113k	32k	16k
NIR Scene	427k	122k	61k

Table 2. 95%ErrRate on VeDAI validation set using TS-Net. Rows: tower fusion after the feature extraction network (bottleneck layer), FC1, FC2 or FC3 of the metric layer. '1 Entropy' means there is only one classification loss at the top of the network. '3 Entropy': each sub-network also has his own classification loss. S*: Matchnet Network with the same number of parameters as TS-Net.

	3 Entropy losses	1 Entropy loss
FC3 (TS-Net)	$\mathbf{0.52 \pm 0.07}$	0.93 ± 0.05
FC2	0.62 ± 0.13	0.92 ± 0.05
FC1	0.74 ± 0.07	1.03 ± 0.06
Feature tower	n/a	1.05 ± 0.07
S*	n/a	1.01 ± 0.11

In the case of early fusion, all the following layers are kept as in MatchNet. Table 2 shows the performance given by each alternative. It also compares the performance obtained when 1 unique entropy loss ($tsnet_{en}$) is used, on the top of the network, with the performance obtained when each sub-network has, in addition, its own loss ($L_{tsnet_{en}} + L_{siam_{en}} + L_{pseudo_{en}}$). Based on these results, it is clear that the additional losses are important. The two additional losses help to guarantee the Siamese and the Pseudo-Siamese network learn complementary representation of the modalities. Consequently, this is the reason why having a late fusion (after FC3) is more beneficial. In addition, to guarantee that the gain in performance of TS-Net is not due to a larger number of parameters, we also provide the performance of MatchNet (noted as S* in Table 2) when we increase the number of parameters in the feature tower by a factor of 1.45 and the bottleneck by 2 to have exactly the same number of parameters as in TS-Net. Experimental results suggest that this is roughly equivalent to the performance of TS-Net without additional losses with fusion at the FC1 layer.

Table 3. 95%ErrRate on the 3 datasets, for Siamese network alone (S), Pseudo-Siamese network alone (PS), TS-NET, without/with the additional contrastive loss (C).

Dataset	Vedai	CUHK	NIR Scene
S	1.16 ± 0.07	5.07 ± 0.46	14.35 ± 0.20
PS	1.50 ± 0.08	5.56 ± 0.36	16.05 ± 0.30
TS-Net	0.52 ± 0.07	3.58 ± 0.14	12.40 ± 0.34
S+C	0.84 ± 0.05	3.38 ± 0.20	13.17 ± 0.86
PS+C	1.37 ± 0.08	3.70 ± 0.14	15.60 ± 0.28
TS-Net+C	$\mathbf{0.45 \pm 0.05}$	$\mathbf{2.77 \pm 0.07}$	$\mathbf{11.86 \pm 0.27}$

Influence of the contrastive loss. Table 3 presents the experimental results given by the three architectures: Siamese, Pseudo-Siamese and TS-Net network with/without the additional contrastive loss. In general, we observed that the error can be reduced by up to 30 % by adding this loss, for any architecture and dataset. More importantly, this gain can be obtained with negligible computing costs and with little effort. During training, we found that the margin Q and the weighting value λ and β are crucial and need to be carefully cross-validated. We also consider replacing it by the classical contrastive loss. However it turned out to be very sensitive to gradient explosion. In addition, to make these parameters less sensitive during training, we tried to normalize the features before feeding into the loss function in order to maintain a fixed range of distances. Unfortunately, we observed some (marginal) drop in performance.

Comparison to Siamese and Pseudo-Siamese network. Intuitively, the Pseudo-Siamese network has more parameters and degree of freedom to project the two modalities onto the new subspace. Hence, it should produce better results compared to the Siamese network (See Table 3). However, in practice, we observed the opposite. We perform a grid search on the different parameters, regularization techniques (dropout, L2/L1), different losses (entropy/contrastive loss) with different strategy of combining the two towers (concatenation/subtraction). In all the experimentations, the Siamese network always outperform the Pseudo-Siamese network. This behavior has also been observed by [8, 15, 17]. When combining the Siamese and Pseudo-Siamese network, we notice significant improvement over the 3 datasets. On VeDAI and CUHK, the error is reduced by almost 50% not counting the additional loss at the feature level. On the three datasets, our approach outperforms the Siamese and Pseudo-Siamese networks. This fully justifies the competitiveness of our approach.

4. CONCLUSIONS

We proposed a novel architecture, called TS-Net, for multimodal patch matching. TS-Net consists of two sub-networks: a Siamese and Pseudo-Siamese network. Each of them is responsible for learning different types of complementary characteristics from both modalities. In addition, we showed that an additional loss, at the intermediate feature level, is beneficial at the price of only a small additional computational costs. Experimental results demonstrate the superiority of our approach over Siamese and Pseudo-Siamese networks.

Acknowledgements. This work has been performed by GREYC in partnership with MBDA, DSTL and the DGA, and funded under the MCM ITP and by the ANR-16-CE23-0006 programme. The authors thank Shivang Agarwal for proofreading the manuscript.

5. REFERENCES

[1] David G Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *ECCV*, 2006.

[3] Engin Tola, Vincent Lepetit, and Pascal Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE PAMI*, vol. 32, no. 5, pp. 815–830, 2010.

[4] Jared Heinly, Enrique Dunn, and Jan-Michael Frahm, "Comparative evaluation of binary features," in *ECCV*. 2012.

[5] Prateek Jain, Brian Kulis, Jason V Davis, and Inderjit S Dhillon, "Metric and kernel learning using a linear transformation," *JMLR*, vol. 13, no. Mar, pp. 519–547, 2012.

[6] Matthew Brown, Gang Hua, and Simon Winder, "Discriminative learning of local image descriptors," *IEEE PAMI*, vol. 33, no. 1, pp. 43–57, 2011.

[7] Michael Jahrer, Michael Grabner, and Horst Bischof, "Learned local descriptors for recognition and matching," in *Computer Vision Winter Workshop*, 2008, vol. 2.

[8] Sergey Zagoruyko and Nikos Komodakis, "Learning to compare image patches via convolutional neural networks," in *CVPR*, 2015.

[9] Bin Fan Yurun Tian and Fuchao Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," in *CVPR*, 2017.

[10] Jure Zbontar and Yann LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *JMLR*, vol. 17, no. 1-32, pp. 2, 2016.

[11] Hani Altwaijry, Eduard Trulls, James Hays, Pascal Fua, and Serge Belongie, "Learning to match aerial images with deep attentive architectures," in *CVPR*, 2016.

[12] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *CVPR*, 2015.

[13] Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk, "Pn-net: conjoined triple deep network for learning local image descriptors," *arXiv preprint arXiv:1601.05030*, 2016.

[14] BG Kumar, Gustavo Carneiro, Ian Reid, et al., "Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions," in *CVPR*, 2016.

[15] Cristhian A Aguilera, Francisco J Aguilera, Angel D Sappa, Cristhian Aguilera, and Ricardo Toledo, "Learning cross-spectral similarity measures with deep convolutional neural networks," in *CVPR Workshops*, 2016.

[16] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla, "Cross-spectral image patch similarity using convolutional neural network," in *ECMSM*. IEEE, 2017, pp. 1–5.

[17] Nina Merkle, Wenjie Luo, Stefan Auer, Rupert Müller, and Raquel Urtasun, "Exploiting deep matching and sar data for the geo-localization accuracy improvement of optical satellite images," *Remote Sensing*, vol. 9, no. 6, pp. 586, 2017.

[18] Cristhian A Aguilera, Angel D Sappa, Cristhian Aguilera, and Ricardo Toledo, "Cross-spectral local descriptors via quadruplet network," *Sensors*, vol. 17, no. 4, pp. 873, 2017.

[19] José Lezama, Qiang Qiu, and Guillermo Sapiro, "Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding," in *CVPR*, 2017, pp. 6807–6816.

[20] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.

Combining Vision and Language Representations for Patch-based Identification of Lexico-Semantic Relations

Prince Jha*
princekumar_1901cs42@iitp.ac.in
Indian Intitute of Technology Patna
India

Gaël Dias
gael.dias@unicaen.fr
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
France

Alexis Lechervy
alexis.lechervy@unicaen.fr
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
France

Jose G. Moreno
jose.moreno@irit.fr
Université de Toulouse, IRIT UMR 5505 CNRS
France

Anubhav Jangra†*
anubhav0603@gmail.com
Indian Institute of Technology Patna
India

Sebastião Pais
sebastiao@di.ubi.pt
University of Beira Interior
Portugal

Sriparna Saha
sriparna@iitp.ac.in
Indian Institute of Technology Patna
India

ABSTRACT

Although a wide range of applications have been proposed in the field of multimodal natural language processing, very few works have been tackling multimodal relational lexical semantics. In this paper, we propose the first attempt to identify lexico-semantic relations with visual clues, which embody linguistic phenomena such as synonymy, co-hyponymy or hypernymy. While traditional methods take advantage of the paradigmatic approach or/and the distributional hypothesis, we hypothesize that visual information can supplement the textual information, relying on the apperceptum subcomponent of the semiotic textology linguistic theory. For that purpose, we automatically extend two gold-standard datasets with visual information, and develop different fusion techniques to combine textual and visual modalities following the patch-based strategy. Experimental results over the multimodal datasets show that the visual information can supplement the missing semantics of textual encodings with reliable performance improvements[1].

CCS CONCEPTS

• **Computing methodologies** → **Lexical semantics**; **Image representations**; **Supervised learning by classification**.

*Work done during internship at Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC.

†Now at Google Research.

[1]Code and datasets are available at https://github.com/Jhaprince/Combining-Vision-and-Language-Representations-for-Patch-based-Identification-of-Lexico-Semantic-Rela

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548299

KEYWORDS

Lexico-semantic relations, multimodal representations, early and hybrid fusion techniques, multimodal patch-based classification.

ACM Reference Format:
Prince Jha, Gaël Dias, Alexis Lechervy, Jose G. Moreno, Anubhav Jangra, Sebastião Pais, and Sriparna Saha. 2022. Combining Vision and Language Representations for Patch-based Identification of Lexico-Semantic Relations. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548299

1 INTRODUCTION

The ability to automatically identify lexico-semantic relations is an important issue for information retrieval and natural language processing applications such as question answering [14], query expansion [25], or text summarization [16]. Lexico-semantic relations embody linguistic phenomena such as synonymy (e.g. phone ↔ telephone), co-hyponymy (e.g. phone ↔ monitor), hypernymy (e.g. phone → speakerphone), but more can be enumerated [63]. To tackle this task, different strategies have been proposed that either define new specific features [1, 50, 59], build specific latent semantic spaces [37, 46, 64], conceptualize multitask architectures [3, 4], or augment input data with textual information [6, 23].

Although many different ideas have been proposed to classify whether two words are in lexico-semantic relation or not, two different input text representations have mostly been used. On the one hand, the paradigmatic approach represents the input data as the lexico-syntactic patterns that connect the two words in a pair [19, 27, 38, 48, 52, 55]. On the other hand, the distributional approach consists in characterizing the semantic relation that exists between two words based on their n-dimensional individual representations [7, 15, 18, 47, 52, 62, 63, 65].

Interestingly, some recent studies have emerged that tackle vision-grounded natural language representations [9, 28, 33, 34, 45] and applications [2, 21, 31, 35, 51, 56]. This idea is founded on the

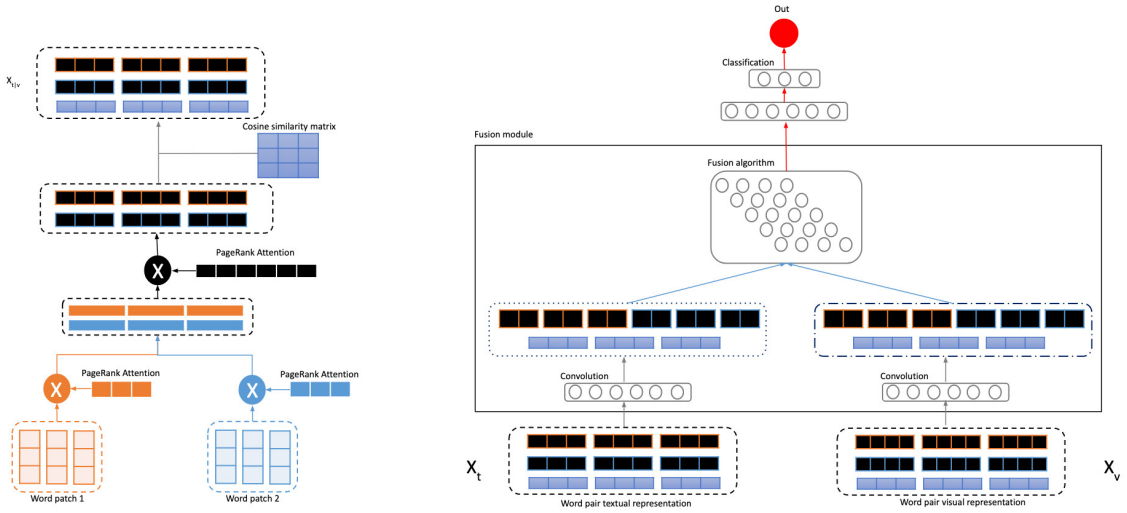


Figure 1: On the left, the patch-based architecture for individual modalities. On the right, the overall multimodal framework.

semiotic textology linguistic theory [17], which lists three subcomponents in order to consider how each textual media produces meaning and the relation between them: dictum (aka. denotation), evocatum (aka. as connotation), and apperceptum (mental images), the latter one embodying the vision-grounded analysis of textual content. However, little research has been endeavoured that combines textual and visual information for relational textual data, to the exception of recent studies on prepositional phrase attachment [11] and relation extraction for knowledge graphs [68].

In this paper, we propose the first attempt to use visual information to identify lexico-semantic relations between word pairs. In particular, we first augment two gold-standard datasets (RUMEN [4] and ROOT9 [49]) with visual information automatically gathered from a search engine. Then, two different fusion techniques, one based on attention fusion [22] and another one based on CentralNet [57], are experimented to combine the textual and visual modalities, where the textual distributional representations are encoded with GloVE [40], and the visual representations are encoded with VGG19 [54]. In order to take advantage of recent multimodal representations, we also propose to encode both modalities with CLIP [45] encodings. Finally, we test our hypothesis following the augmentation data paradigm proposed by [6], by increasing the initial words by their K most similar neighbors within some text representation space, here GloVE, which are then further combined with their visual information. Experiments over the extended multimodal datasets demonstrate that introducing visual information can supplement the missing semantics of textual information with reliable performance improvements.

2 RELATED WORK

Lexico-semantic relation identification. Four major research directions have been proposed for the identification of lexico-semantic relations: (1) feature engineering, (2) fine-tuned semantic spaces, (3) multitask architectures and (4) data augmentation. Within the first topic, [29, 63] propose similar evaluations to combine word input vectors. In particular, word pairs are encoded as the concatenation of the constituent word representations, their vector difference or their sum. [38, 52] propose to overcome domain dependency by representing contextual patterns as continuous vectors, thus successfully combining the paradigmatic approach with the distributional hypothesis. [1, 59] compute specific features over the distributional space (e.g. cosine similarity) in addition to the vector representations themselves, leading to significant improvements. The second research direction aims to build fine-tuned neural latent semantic spaces that embody relational information. [37, 60] learn new embeddings from a background knowlegde of word pairs. To generalize this idea, [8, 24, 64] learn explicit specialization functions that are further injected in the embedding learning process. The third approach tackles this task from the architecture point of view. As semantic relations are known to be closely semantically related, it is likely that multitask learning may improve the decision process. For that purpose, [3] propose a coarse-grained model through a multitask convolutional neural network, while [4] propose a fine-grained methodology, which aims to determine whether the learning process of a given semantic relation can be improved by the concurrent learning of another relation. The fourth strategy aims to augment the initial word pair input with semantically close terms. Within this context, [23] propose a set cardinality-based method, which exploits the WordNet [36] graph, while [6] define a patch-based approach, which augments each constituent word from a latent semantic space.

Vision-grounded language applications and representations. The combination of new multimodal datasets [42] with the definition of new multimodal machine learning models [5] has fostered research in the broad field of multimodal natural language processing [20] and multimodal computer vision [66]. In particular, multimodal machine learning has enabled a wide range of applications, such as multimedia content indexing and retrieval [10], video summarization [51], multimodal sentiment [56] and emotion

[35] analysis, visual question answering [2], image captioning [21], and multimodal dialogue systems [31], to name but a few. Another research direction aims to learn how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. In the specifc field of vision-grounded language representations, different models have been proposed [9, 28, 33, 34, 45]. [28] extend the skip-gram model by taking visual information into account. As such, for a restricted set of words, the model is exposed to the visual representations of the objects they denote, and must predict linguistic and visual features jointly. [34] extend the BERT architecture [13] to a multimodal two-stream model by processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers. [9] introduce UNITER, which includes four pretraining tasks over transformers: masked language modeling conditioned on image, masked region modeling conditioned on text, image-text matching, and word-region alignment. [33] present DiMBERT, which takes both visual features from images and textual features from sentences as input, and then apply a single cross-modal transformer to learn vision-language grounded representations. [45] study the behaviors of image classifiers trained with natural language supervision at large scale. Enabled by the large amounts of publicly available data of image-text form on the internet, they create a new dataset of 400 million pairs and demonstrate that a simplified version of ConVIRT [67] trained from scratch, which they call CLIP, is an efficient method of learning from natural language supervision.

Multimodal lexical semantics. Although a wide range of applications have been proposed in the wide field of multimodal natural language processing, very few works have been tackling multimodal relational lexical semantics, with the rare exceptions of [11, 68]. [11] propose to score alternative prepositional phrase attachments from the caption of an image, previously syntactically-parsed, based on how much the attachments are coherent with the corresponding image. The set of attachments that yields the best score is identified and the corresponding tree is output. [68] present the multimodal relation extraction task that consists in identifying the semantic relations that link two entities in a sentence with visual clues. For that purpose, they propose a multimodal neural network with a graph alignment method that incorporates structural similarity and semantic agreement between visual objects in an image and textual entities in a sentence. Experiments show that improved results can be obtained compared to the concatenation of visual and textual representations. In this paper, we present the first study that tackles multimodal lexico-semantic relation identification.

3 MULTIMODAL METHODOLOGY

The main task at hand consists in deciding whether a given lexico-semantic relation (i.e. synonymy, hypernymy, co-hyponymy) holds between a pair of words (w_0, w_1) or not (i.e. random). For that purpose, we present our methodology, illustrated in Figure 1, which consists in adapting the patch-based approach proposed by [6] in a multimodal environment, thus relying on fusion techniques.

Patch-based Representation. The idea of patch-based classification has been introduced by [6, 23], and consists in augmenting each word in a pair with its K most semantically-related words in some semantic space. While [23] use WordNet for the augmentation, [6] rely on GloVE embeddings. Based on our experiments, we follow the strategy of [6] as it outperforms the one of [23].

Formally, a patch consists of the K most similar words w_j to a source word w_0 in terms of cosine similarity in some latent semantic space, and it is defined in Equation 1. Thus, each input pair (w_0, w_1) is transformed into its patch-based representation $(P^K_{w_0}, P^K_{w_1})$.

$$P^K_{w_0} = \{w_0\} \cup \left\{ w_j | \underset{K}{\text{argmax}} \cos(w_0, w_j) \right\} \tag{1}$$

All words within a patch are then subject to a fixed attention mechanism, which integrates the notion of centrality. This ensures that the most central words within a patch receive higher attention. This process is performed through the PageRank algorithm [39] over the undirected weighted[2] patch graph, which results in a vector of $(K+1)$ dimensions, where each word within the patch receives a centrality score in $\mathbb{R}$, and it is noted $\langle \alpha_{w_0^0}, \alpha_{w_0^1}, \alpha_{w_0^2}, \ldots, \alpha_{w_0^K} \rangle$.

A second attention mechanism spotlights on word centrality between patches to acknowledge, which words are central to both concepts. The same process is applied with the PageRank algorithm based on the graph that comprises of all $2 \times (K+1)$ words as vertices and links all vertices belonging to different patches. This process results in a vector of $2 \times (K+1)$ dimensions, where each word of both patches receives a centrality score in $\mathbb{R}$, and it is noted $\langle \beta_{w_0^1}, \beta_{w_0^2}, \ldots, \beta_{w_0^K}, \beta_{w_1^0}, \beta_{w_1^1}, \ldots, \beta_{w_1^K} \rangle$.

Both attention mechanisms are then combined into a unique learning representation, which is defined in Expression 2, where w^i_x represents a word embedding of patch $P^K_{w_x}$, and $\oplus$ is the concatenation operator. Note that the embeddings are in descending order of cosine similarity with their source word.

$$A_1 = \left(\bigoplus_{i=0}^{K} \alpha_{w_0^i} . \beta_{w_0^i} . w_0^i \right) \oplus \left(\bigoplus_{i=0}^{K} \alpha_{w_1^i} . \beta_{w_1^i} . w_1^i \right) \tag{2}$$

In order to account for domain independence [61], the cosine similarity is measured between all components of both patches, which concatenation is defined in Expression 3.

$$A_2 = \bigoplus_{i=0}^{K} \bigoplus_{j=0}^{K} cos(w_0^i, w_1^j) \tag{3}$$

Finally, each input pair (w_0, w_1) receives two different learning representations, namely X_t for the textual modality and X_v for the visual modality, generically defined in Equation 4. Such representations are then fed to the multimodal fusion module.

$$X_{t|v} = A_1 \oplus A_2 \tag{4}$$

While Bannour et al. [6] exclusively focus on textual data augmentation, we need to deal with visual augmentation. For that purpose, we propose that textual data drives the augmentation process[3]. As such, each word pair (w_0, w_1) is transformed into its patch-based representation $(P^K_{w_0}, P^K_{w_1})$, based on finding the K most similar words within some textual semantic space, here GloVE [40], in terms of cosine similarity. Then, each word present in a patch

[2]The weight corresponds to the cosine similarity value.
[3]Other strategies are possible but they remain for future work.

is sent to a search engine, here the Bing Image Search API[4], and the highest ranked image returned by the search engine is taken as the augmented visual information (cf. §4 for more details). Once visual augmentation is performed, the process of [6] is replicated in the exact same way for the visual information but relying on visual n-dimensional representations, VGG19 [53] or CLIP [45].

Multimodal fusion networks. In order to reduce each modality representation X_t and X_v to the same dimension, a reduction process is first performed. Then, two fusion techniques are implemented to combine modalities: early [22] and hybrid fusions [58].

Attention fusion network. Both visual and textual modalities may not equally be relevant for the identification of lexico-semantic relations. This motivates the introduction of an attention fusion network, which weights each modality independently, in the same line of [41, 43]. The attention fusion network is shown in Figure 2.

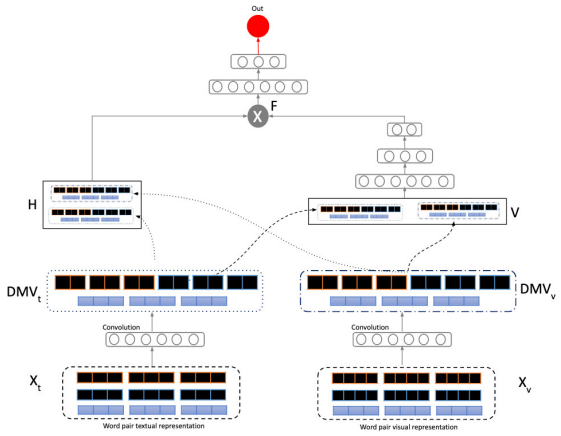


Figure 2: Attention fusion network.

Formally, the input to the attention fusion network is noted MV, the set of modality representations, where the dimension of a modality vector $MV_k \in MV$ is d_k. The first step consists in giving the same dimension d to all the elements of MV. This process is referred to a reduction process, and it is done using a stack of dense layers. The resultant vectors are denoted DMV, such that the reduced modality representations $DMV_k \in DMV$. All DMV_k are then concatenated into a vector V, which is passed through a set of dense layers followed by sigmoid activation layer to calculate attention scores. These attention values weight each modality, and the resulting modality representations are concatenated to build the early fusion vector F. This process is recaped in Equation 5.

$$\begin{aligned} MV &= [MV_t = X_t, MV_v = X_v] \\ DMV_i &= ReLU(W^T_{MV_i} MV_i + b) \\ Att_i &= \sigma(W^T_{DMV_i} V + b) \\ F &= Att_t.V_t \bigoplus Att_v.V_v \end{aligned} \tag{5}$$

[4]https://www.microsoft.com/en-us/bing/apis/bing-image-search-api

F is then passed through further dense layers for the decision process, and the categorical cross-entropy loss function $L_{CE}(\hat{y}, y)$ is used to train the network parameters, where $\hat{y}_i^j$ is the predicted label and y_i^j is the true label.

$$L_{CE}(\hat{y}, y) = -\frac{1}{N} \sum_{j=1}^{C} \sum_{i=1}^{N} y_i^j log(\hat{y}_i^j) \tag{6}$$

CentralNet fusion network. CentralNet [58] is a hybrid fusion network, which mixes early and late fusions into a single architecture as illustrated in Figure 3.

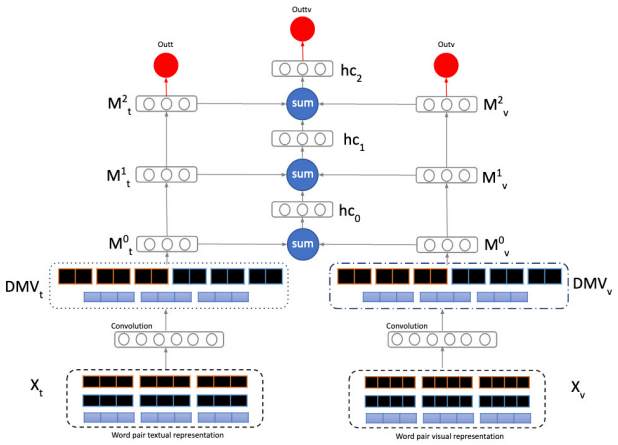


Figure 3: CentralNet fusion network.

The architecture consists of k independent networks corresponding to each modality, and one central network. In particular, the central network combines the features generated from the different modalities by considering the weighted sum of unimodal hidden representations, and its own previous layer. Such fusion layers are defined in Equation 7, where α_p are scalar trainable weights, M_k^i is the hidden representation of k^{th} modality at the i^{th} layer, and hc_i is the central hidden representation at the i^{th} layer. Note that fusing at a low-level layer stands for early fusion, while fusing at a last layer means late fusion.

$$hc_{i+1} = ReLU\left(\alpha_c hc_i \ + \sum_{k=t}^{v} \alpha_{M_i^k} M_i^k\right) \tag{7}$$

Each layer hc_{i+1} is fed to an operating layer composed of a dense layer followed by a $ReLU$ activation function. Note that the input to the first layer of the central network is only the weighted sum of the modalities hidden representations $M_0^t = DMV_t$ and $M_0^v = DMV_v$ as there is no previous central hidden representation. The final output representation of the central network represents the fusion vector F, which is used for the final prediction. In particular, we employ the categorical cross-entropy loss function (Equation 6) to train the network parameters, and the final $Loss$ function is defined in Equation 8, where L_{CE}^C is the loss computed from the output of central network, and $L_{CE}^{M^k}$ is the loss of modality k.

$$Loss(\hat{y}, y) = L_{CE}^{C}(\hat{y}, y) \; + \; \sum_{k=t}^{v} L_{CE}^{M^k}(\hat{y}, y) \tag{8}$$

Note that our model differs from the one presented in Vielzeuf et al. [58] in the sense that each unimodal network is first pre-trained independently, and then frozen to learn the central network. As such, only the central network is trainable, and the remaining parts of the architecture are kept non-trainable, i.e. frozen. Indeed, the frozen architecture showed stronger performances compared to the all-trainable model for the sake of our experiments.

Implementation details and experimental setups of all the modules of the methodology are given in Apppendix A.

4 NEW DATASETS: IXRUMEN AND IXROOT9

Since the community lacks a multimodal dataset for the task of lexico-semantic relation identification, we propose the extension of two gold-standard datasets, namely RUMEN [4] and ROOT9 [49]. RUMEN is a dataset comprising of 3213 instances for synonymy detection and 3375 instances for hypernymy detection, whereas ROOT9 comprises of 1636 instances for co-hyponymy detection and 1256 instances for hypernymy detection. As we follow the patch-based data augmentation strategy due to its empirical effectiveness, where each word instance is augmented by its K-nearest neighbors in the GloVE [40] embedding space, the visual augmentation must deal with the original words within the pair plus the K augmented words that form the respective patches.

To extend the two datasets in a multimodal setting, we propose to scrap the web for exemplar images by using the Bing Image Search API, such that for each of the $K+1$ words within a patch, we download exactly 3 images ordered by their retrieval rank[5]. This multimodal augmentation strategy is performed for a patch size up to $K = 5$, and we adopt lexical split [4, 6, 30], which avoids vocabulary intersection between the train and test splits, thus bypassing the lexical memorization issue [30]. As a consequence, it is clear that some of the initial word pairs contained in RUMEN and ROOT9 must be withdrawn from their original datasets, if they cannot provide up to 3 visual clues. The statistics for the image-extended RUMEN dataset (IxRUMEN) and the image-extended ROOT9 dataset (IxROOT9) can be found in Table 1.

Dataset	Train	Test	Total
RUMEN (Synonym)	2256	957	3213
RUMEN (Hypernym)	2638	737	3375
RUMEN (Random)	2227	969	3196
IxRUMEN (Synonym)	2031	860	2891
IxRUMEN (Hypernym)	2393	648	3041
IxRUMEN (Random)	2006	830	2836
ROOT9 (Co-hyponym)	1070	566	1636
ROOT9 (Hypernym)	826	430	1256
ROOT9 (Random)	381	129	510
IxROOT9 (Co-hyponym)	975	531	1506
IxROOT9 (Hypernym)	717	392	1109
IxROOT9 (Random)	335	103	438

Table 1: Statistics for RUMEN, ROOT9, IxRUMEN and IxROOT9 datasets.

[5]Note that we explored existing large-scale corpora like the MSCOCO dataset [32] for better reproducibility, but due to its limited lexical coverage, an open-domain retrieval strategy was opted for.

To overcome privacy concerns, we decided to release the image encodings for each image in the dataset over the actual image. For that purpose, we use VGG19 [53] and CLIP [44] embeddings. VGG19 [53] shows state-of-the-art performances in image classification tasks. It is 19 layers deep convolutional network, which is pre-trained on ImageNet [12] to predict 1000 object classes. Thus, VGG19 embeddings have the ability to represent robust visual concepts. Here, each image is encoded as a 4096-dimensional vector. CLIP (Contrastive Language-Image Pre-training) [44] is a pre-trained visual-linguistic model that can encode image-text pairs. CLIP was pre-trained on 400 million image-text pairs, where for a given batch of N (image,text) pairs, the model had to predict N correct matches out of $N \times N$ possible pairings. In particular, CLIP maximizes the cosine similarity of N real pairs by training image and text encoders together to create an efficient multimodal embedding space. Here, each image is encoded as a 512-dimensional vector and note that the image information is combined with the textual pattern "*a photo of <source word>*" as suggested in [44] to get full advantage of the contextualized multimodal model.

5 RESULTS AND DISCUSSION

In this section, we first present the results of the unimodal models, where each textual and visual modalities are taken individually for the decision process. Then, we present the results obtained for the early and hybrid fusions. Finally, we present a qualitative analysis that shows the benefits and drawbacks of the multimodal fusion.

5.1 Unimodal Models

Results for unimodal models are given in Table 2 for the textual modality and in Table 3 for the visual modality. For the textual modality, results confirm the findings of [6] and show that the patch-based approach outperforms the baseline strategy, where no word augmentation is performed, i.e. $K = 0$. In particular, larger values of K steadily improve results for the identification of symmetric relations (synonymy and co-hyponymy), while such is not true for asymmetric relations such as hypernymy. This can be explained by the fact that larger values of K might lead to concept shift for the hypernym relation, thus noising the input data. This is particularly true for GloVE embeddings, although such does not stand for CLIP embeddings. Results also show that CLIP multimodal embeddings do not provide a sustainable alternative for the sake of the identification of lexico-semantic relations, as results drastically drop, when compared to text-based embeddings, especially for the case of the IxRUMEN dataset. This can be explained by the fact that CLIP embeddings have been tuned to better represent visual information at the expense of textual information [45]. Moreover, as the IxRUMEN dataset contains a wide spectrum of abstract words [4] (e.g. destiny $\leftrightarrow$ fate), this might lead to difficulties in visually representing such information. As a consequence, multimodal embeddings might not correctly encode this information.

For the visual modality, different situations occur. While the use of VGG19 encodings clearly evidences the positive impact of using the patch-based strategy with steady improvements for high values of K independently of the lexico-semantic relation and the dataset at hand, similar results are not exactly observable for multimodal representations. Indeed, while the use of CLIP multimodal

			Synonym v/s Random				Hypernym v/s Random			
		Patch Size	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)
RUMEN	GLOVE	K=0	80.98/0.35/81.52	80.98/0.35/81.52	80.98/0.35/81.52	80.978/0.35/81.52	79.52/0.15/79.66	79.51/0.15/79.66	79.52/0.15/79.66	79.50/0.16/79.66
		K=1	84.17/0.23/84.53	84.17/0.23/84.52	84.18/0.23/84.54	84.19/0.22/84.53	82.24/0.27/82.47	82.20/0.27/82.44	82.24/0.27/82.47	82.19/0.27/82.43
		K=2	84.81/0.12/84.94	84.80/0.12/84.94	84.81/0.12/84.94	84.84/0.12/84.97	82.13/0.17/84.42	82.05/0.16/82.33	82.13/0.17/82.42	82.09/0.17/82.38
		K=3	83.97/0.07/84.06	83.96/0.07/84.05	83.97/0.07/84.06	84.02/0.07/84.12	82.33/0.19/82.59	82.27/0.19/82.53	82.33/0.19/82.59	82.29/0.19/82.55
		K=4	84.67/0.10/84.79	84.67/0.10/84.78	84.67/0.10/84.79	84.70/0.10/84.82	**82.55**/0.13/82.71	**82.49**/0.13/82.65	**82.55**/0.13/82.71	**82.50**/0.13/82.67
		K=5	**84.90**/0.04/84.94	**84.90**/0.04/84.94	**84.90**/0.04/84.94	**84.91**/0.04/84.95	81.82/0.08/81.95	81.74/0.07/81.86	81.82/0.08/81.95	81.78/0.08/81.91
	CLIP-Text	K=0	56.09/0.13/56.33	56.09/0.13/56.33	56.09/0.13/56.33	56.11/0.13/56.34	54.24/0.28/54.57	51.76/1.69/54.76	54.24/0.28/54.57	52.52/1.50/55.15
		K=1	59.54/0.09/59.61	59.54/0.09/59.60	59.55/0.09/59.61	59.54/0.09/59.60	59.60/0.30/59.85	57.48/0.31/57.75	59.60/0.30/59.85	58.62/0.35/58.91
		K=2	63.66/0.04/63.71	63.66/0.04/63.71	63.66/0.04/63.71	63.66/0.05/63.71	62.56/0.24/62.95	61.37/0.23/61.74	62.56/0.23/62.95	61.98/0.26/62.41
		K=3	64.99/0.13/65.21	64.99/0.13/65.21	64.99/0.13/65.21	64.99/0.13/65.21	**64.37**/0.18/64.54	**62.94**/0.20/63.19	**64.37**/0.18/64.54	64.07/0.21/64.23
		K=4	66.51/0.31/66.87	66.50/0.31/66.87	66.51/0.31/66.87	66.50/0.31/66.87	62.09/0.21/62.31	60.25/0.24/60.51	62.09/0.21/62.31	61.54/0.24/61.79
		K=5	**66.64**/0.10/66.72	**66.63**/0.10/66.71	**66.64**/0.10/66.72	**66.65**/0.10/66.73	64.35/0.19/64.48	62.59/0.20/62.75	64.35/0.19/64.48	**64.19**/0.23/64.39
			Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	GLOVE	K=0	92.17/0.25/92.46	92.24/0.24/92.56	92.16/0.25/92.46	92.47/0.22/92.74	85.06/0.07/85.11	85.15/0.06/85.2	85.06/0.07/85.11	85.32/0.30/85.54
		K=1	93.72/0.15/93.81	93.72/0.15/93.82	93.72/0.15/93.81	93.73/0.16/93.83	91.10/0.11/91.27	91.04/0.11/91.2	91.10/0.11/91.27	91.02/0.12/91.18
		K=2	93.86/0.07/93.94	93.87/0.08/93.95	93.86/0.07/93.94	93.88/0.08/93.96	91.86/0.15/92.09	91.82/0.15/92.05	91.86/0.15/92.09	91.80/0.15/92.03
		K=3	93.41/0.10/93.54	93.47/0.09/93.60	93.41/0.10/93.54	93.59/0.08/93.70	**92.69**/0.09/92.75	**92.64**/0.09/92.70	**92.69**/0.09/92.75	**92.63**/0.09/92.69
		K=4	93.46/0.12/93.67	93.52/0.12/93.74	93.46/0.12/93.67	93.62/0.13/93.85	92.42/0.12/92.59	92.43/0.11/92.59	92.42/0.12/92.59	92.45/0.11/92.60
		K=5	**94.00**/0.12/94.08	**94.05**/0.16/94.14	**94.00**/0.12/94.08	**94.16**/0.11/94.27	91.96/0.14/92.09	91.98/0.14/92.11	91.96/0.14/92.09	92.00/0.12/92.12
	CLIP-Text	K=0	74.32/0.62/75.24	75.90/0.54/76.68	74.32/0.63/75.24	80.05/0.29/80.34	67.41/0.39/67.87	67.65/0.45/68.2	67.41/0.39/67.87	67.93/0.53/68.58
		K=1	82.26/0.22/82.50	82.97/0.21/83.18	82.26/0.22/82.50	84.51/0.22/84.68	73.01/0.18/73.31	73.20/0.16/73.44	73.01/0.18/73.31	73.43/0.14/73.58
		K=2	86.00/0.17/86.14	86.40/0.15/86.52	86.00/0.17/86.14	87.21/0.11/87.28	75.98/0.14/76.11	76.09/0.13/76.21	75.98/0.14/76.11	76.22/0.11/76.32
		K=3	85.82/0.23/86.14	86.24/0.22/86.56	85.82/0.23/86.14	87.10/0.23/87.46	77.66/0.15/77.76	77.59/0.15/77.74	77.66/0.15/77.76	77.53/0.15/77.72
		K=4	87.92/0.31/88.29	88.26/0.29/88.61	87.92/0.31/88.29	89.02/0.29/89.33	77.92/0.12/78.09	77.89/0.13/78.07	77.92/0.12/78.09	77.85/0.14/78.05
		K=5	**89.10**/0.00/89.10	**89.43**/0.01/89.43	**89.10**/0.00/89.10	**90.24**/0.04/90.26	**79.61**/0.14/79.74	**79.55**/0.11/79.69	**79.61**/0.14/79.74	**79.51**/0.10/79.64

Table 2: Results for IxRUMEN and IxROOT9 based on textual unimodality represented either by GloVE or CLIP embeddings.

			Synonym v/s Random				Hypernym v/s Random			
		Patch Size	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)
RUMEN	VGG19	K=0	52.65 / 0.41 / 53.37	50.56 / 1.04 / 52.35	52.65 / 0.41 / 53.37	53.34 / 0.35 / 53.80	53.52 / 0.28 / 53.75	45.53 / 0.75 / 46.80	53.52 / 0.28 / 53.75	47.53 / 0.78 / 48.76
		K=1	56.49 / 1.51 / 57.84	56.15 / 1.91 / 57.80	56.49 / 1.51 / 57.84	56.59 / 1.37 / 57.90	53.89 / 0.34 / 54.34	50.10 / 0.08 / 50.16	53.89 / 0.34 / 54.34	51.09 / 0.28 / 51.45
		K=2	57.25 / 2.28 / 59.40	57.18 / 2.34 / 59.38	57.25 / 2.34 / 59.38	57.27 / 2.27 / 59.43	57.09 / 0.53 / 57.50	53.43 / 0.59 / 54.04	57.09 / 0.53 / 57.50	55.34 / 0.71 / 55.88
		K=3	59.21 / 0.30 / 59.50	59.17 / 0.31 / 59.46	59.21 / 0.30 / 59.50	59.28 / 0.30 / 59.56	56.62 / 0.62 / 57.56	56.57 / 0.68 / 57.64	56.62 / 0.62 / 57.56	57.21 / 0.78 / 57.91
		K=4	**60.60** / 0.40 / 61.01	**60.56** / 0.39 / 60.95	**60.60** / 0.40 / 61.01	**60.68** / 0.41 / 61.10	**58.49** / 0.16 / 58.68	57.41 / 0.76 / 57.82	**58.49** / 0.16 / 58.68	57.67 / 0.16 / 57.81
		K=5	60.59 / 0.44 / 60.90	**60.56** / 0.46 / 60.88	60.59 / 0.44 / 60.90	60.66 / 0.42 / 60.96	58.38 / 0.38 / 58.73	**57.92** / 0.45 / 58.36	58.38 / 0.38 / 58.73	**57.96** / 0.27 / 58.25
	CLIP-Image	K=0	**77.11** / 0.29 / 77.52	**77.11** / 0.29 / 77.52	**77.11** / 0.29 / 77.52	**77.12** / 0.29 / 77.53	**75.91** / 0.08 / 75.97	**75.94** / 0.08 / 75.99	**75.91** / 0.08 / 75.97	75.97 / 0.07 / 76.02
		K=1	75.41 / 0.20 / 75.65	75.46 / 0.30 / 75.90	75.41 / 0.20 / 75.64	75.46 / 0.21 / 75.72	73.77 / 0.42 / 74.09	73.56 / 0.44 / 73.85	73.77 / 0.42 / 74.09	73.61 / 0.42 / 73.97
		K=2	74.76 / 0.17 / 74.97	74.76 / 0.17 / 74.97	74.76 / 0.17 / 74.97	74.77 / 0.17 / 74.98	74.57 / 0.11 / 74.74	74.22 / 0.12 / 74.41	74.57 / 0.11 / 74.74	74.54 / 0.10 / 74.69
		K=3	74.15 / 0.15 / 74.25	74.15 / 0.15 / 74.25	74.15 / 0.15 / 74.25	74.15 / 0.15 / 74.25	74.86 / 0.07 / 74.91	74.38 / 0.07 / 74.44	74.86 / 0.07 / 74.91	75.01 / 0.06 / 75.07
		K=4	74.91 / 0.16 / 75.08	74.91 / 0.16 / 75.08	74.91 / 0.16 / 75.08	74.91 / 0.16 / 75.08	75.19 / 0.21 / 75.38	75.24 / 0.67 / 75.82	75.19 / 0.21 / 75.38	75.46 / 0.23 / 75.68
		K=5	76.52 / 0.10 / 76.69	76.52 / 0.10 / 76.69	76.52 / 0.10 / 76.69	76.52 / 0.10 / 76.69	75.66 / 0.05 / 75.73	75.40 / 0.50 / 75.94	75.66 / 0.05 / 75.73	**76.18** / 0.04 / 76.22
			Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	VGG19	K=0	62.19 / 3.13 / 65.95	65.00 / 2.83 / 68.42	62.26 / 3.00 / 65.95	75.46 / 0.74 / 75.97	62.41 / 1.39 / 64.42	60.30 / 0.48 / 60.73	62.41 / 1.39 / 64.42	58.96 / 0.27 / 59.32
		K=1	70.53 / 0.19 / 70.66	72.57 / 0.15 / 72.67	70.53 / 0.19 / 70.66	78.61 / 0.14 / 78.71	61.08 / 0.44 / 61.45	61.66 / 0.46 / 62.02	61.08 / 0.44 / 61.45	62.37 / 0.52 / 62.74
		K=2	77.93 / 0.25 / 78.20	79.08 / 0.20 / 79.30	77.93 / 0.25 / 78.20	81.86 / 0.07 / 81.98	66.39 / 0.20 / 66.56	66.60 / 0.21 / 66.74	66.39 / 0.20 / 66.56	66.82 / 0.24 / 66.95
		K=3	79.41 / 0.25 / 79.81	80.55 / 0.22 / 80.90	79.41 / 0.25 / 79.81	83.67 / 0.11 / 83.85	69.36 / 0.20 / 69.69	69.20 / 0.19 / 69.45	69.36 / 0.20 / 69.69	69.06 / 0.20 / 69.29
		K=4	79.65 / 0.07 / 79.68	80.75 / 0.07 / 80.79	79.65 / 0.07 / 79.68	83.68 / 0.13 / 83.79	70.25 / 0.38 / 70.51	69.84 / 0.39 / 70.17	70.25 / 0.38 / 70.51	69.53 / 0.41 / 69.89
		K=5	**82.00** / 0.06 / 82.10	**82.89** / 0.06 / 83.00	**82.00** / 0.06 / 82.10	**85.29** / 0.12 / 85.43	**73.38** / 0.15 / 73.48	**72.72** / 0.16 / 72.84	**73.38** / 0.15 / 73.48	**72.35** / 0.18 / 72.48
	CLIP-Image	K=0	90.79 / 0.31 / 91.12	90.97 / 0.29 / 91.27	90.79 / 0.31 / 91.12	91.34 / 0.26 / 91.59	84.48 / 0.39 / 85.17	84.55 / 0.37 / 85.20	84.48 / 0.39 / 85.17	84.63 / 0.34 / 85.22
		K=1	91.04 / 0.20 / 91.25	91.23 / 0.19 / 91.43	91.04 / 0.20 / 91.25	91.67 / 0.15 / 91.81	86.79 / 0.32 / 87.15	86.82 / 0.31 / 87.17	86.79 / 0.32 / 87.15	86.85 / 0.30 / 87.19
		K=2	92.38 / 0.20 / 92.73	92.46 / 0.20 / 92.80	92.38 / 0.20 / 92.73	92.59 / 0.19 / 92.91	**88.37** / 0.09 / 88.47	**88.37** / 0.09 / 88.47	88.34 / 0.09 / 88.47	**88.37** / 0.09 / 88.47
		K=3	91.79 / 0.13 / 91.92	91.92 / 0.13 / 92.05	91.79 / 0.13 / 91.92	92.19 / 0.11 / 92.32	88.04 / 0.09 / 88.14	88.14 / 0.08 / 88.23	88.04 / 0.09 / 88.14	88.30 / 0.08 / 88.38
		K=4	**92.47** / 0.18 / 92.76	**92.52** / 0.07 / 92.58	**92.41** / 0.07 / 92.46	**92.80** / 0.05 / 92.84	87.77 / 0.22 / 88.14	87.81 / 0.21 / 88.16	87.77 / 0.22 / 88.14	87.85 / 0.21 / 88.18
		K=5	91.63 / 0.11 / 91.79	91.80 / 0.11 / 91.96	91.63 / 0.11 / 91.79	92.16 / 0.11 / 92.33	**88.37** / 0.19 / 88.63	88.32 / 0.17 / 88.56	**88.37** / 0.19 / 88.63	88.29 / 0.17 / 88.52

Table 3: Results for IxRUMEN and IxROOT9 based on visual unimodality represented either by VGG19 or CLIP embeddings.

representations outperforms overall results compared to VGG19 encodings for a great margin, the impact of the patch-based strategy is more mitigated. For the IxRUMEN dataset, it is clear that the use of patches seems to be counter-productive, while the contrary is true for the IxROOT9 dataset, with best values obtained for higher values of K. Nevertheless, by looking closely at the results for IxRUMEN, we can observe that values for $K = 5$ are close to the ones of $K = 0$, although intermediate values of K show lower performances, but with a small increasing tendency along with higher values of K. We can then hypothesize that if higher values of K had been tested, CLIP representations would benefit more from the patch-based strategy.

It is clear that relying exclusively on visual information does not compete with text-based strategies, especially when the datasets include abstract words, i.e. IxRUMEN. Indeed, the results differences between the textual modality and the visual modality range from 7.79 points in F1 score for synonymy in IxRUMEN in favor of the textual modality to 1.53 points improvements of the textual modality in terms of F1 score for co-hyponymy in IxROOT9. This situation was expected due to the uncontrolled process of gathering extra visual information with search engine queries. Nevertheless,

			Synonym v/s Random				Hypernym v/s Random			
		Patch Size	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)
RUMEN	GLOVE+CLIP-Img	K=0	82.97 / 0.09 / 83.07	82.96 / 0.09 / 83.06	82.97 / 0.09 / 83.07	83.00 / 0.11 / 83.14	81.32 / 0.60 / 81.95	81.25 / 0.60 / 81.87	81.32 / 0.60 / 81.95	81.28 / 0.61 / 81.90
		K=1	**86.62** / 0.19 / 86.86	**86.61** / 0.19 / 86.85	**86.62** / 0.19 / 86.86	**86.72** / 0.18 / 86.96	82.77 / 0.21 / 83.06	82.62 / 0.21 / 82.89	82.77 / 0.21 / 83.06	82.83 / 0.23 / 83.18
		K=2	86.44 / 0.07 / 86.50	86.42 / 0.07 / 86.49	86.44 / 0.07 / 86.50	86.57 / 0.08 / 86.64	83.54 / 0.20 / 83.76	83.37 / 0.19 / 83.60	83.54 / 0.20 / 83.76	83.68 / 0.23 / 83.92
		K=3	85.09 / 0.21 / 85.25	85.06 / 0.21 / 85.23	85.09 / 0.21 / 85.25	85.26 / 0.22 / 85.44	83.48 / 0.16 / 83.76	83.30 / 0.16 / 83.57	83.48 / 0.16 / 83.76	83.62 / 0.23 / 83.99
		K=4	85.51 / 0.14 / 85.67	85.49 / 0.15 / 85.65	85.51 / 0.14 / 85.67	85.68 / 0.13 / 85.81	**84.07** / 0.08 / 84.17	**83.93** / 0.08 / 84.03	**84.07** / 0.08 / 84.17	**84.17** / 0.10 / 84.28
		K=5	85.50 / 0.14 / 85.72	85.48 / 0.14 / 85.70	85.50 / 0.14 / 85.72	85.68 / 0.14 / 85.90	82.89 / 0.12 / 83.06	82.73 / 0.11 / 82.90	82.90 / 0.11 / 83.06	82.97 / 0.13 / 83.16
			Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	GLOVE+CLIP-Img	K=0	94.51 / 0.31 / 97.89	94.54 / 0.30 / 94.91	94.51 / 0.31 / 94.89	94.57 / 0.28 / 94.93	87.64 / 0.26 / 87.97	87.62 / 0.30 / 88.00	87.64 / 0.26 / 87.97	87.61 / 0.33 / 88.04
		K=1	94.80 / 0.20 / 95.15	94.85 / 0.21 / 95.20	94.80 / 0.20 / 95.15	94.91 / 0.22 / 95.29	92.26 / 0.20 / 92.42	92.25 / 0.20 / 92.41	92.26 / 0.20 / 92.42	92.24 / 0.20 / 92.40
		K=2	**95.07** / 0.15 / 95.29	**95.11** / 0.15 / 95.33	**95.07** / 0.15 / 95.29	**95.19** / 0.14 / 95.40	**92.95** / 0.22 / 93.25	**92.92** / 0.22 / 93.22	**92.95** / 0.22 / 93.25	**92.91** / 0.22 / 93.20
		K=3	94.86 / 0.11 / 95.02	94.91 / 0.11 / 95.06	94.86 / 0.11 / 95.02	95.01 / 0.10 / 95.14	92.75 / 0.17 / 92.92	92.73 / 0.17 / 92.91	92.75 / 0.17 / 92.92	92.72 / 0.17 / 92.90
		K=4	94.89 / 0.10 / 95.02	94.94 / 0.09 / 95.07	94.89 / 0.10 / 95.02	95.07 / 0.06 / 95.17	92.62 / 0.07 / 92.75	92.63 / 0.08 / 92.77	92.62 / 0.07 / 92.75	92.66 / 0.08 / 92.81
		K=5	94.99 / 0.11 / 95.15	95.04 / 0.12 / 95.21	94.99 / 0.11 / 94.89	95.14 / 0.13 / 95.32	92.69 / 0.15 / 92.92	92.70 / 0.15 / 92.93	92.64 / 0.16 / 92.92	92.73 / 0.16 / 92.96

Table 4: Results for IxRUMEN and IxROOT9 for early fusion with the attention fusion network.

			Synonym v/s Random				Hypernym v/s Random			
		Patch Size	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)	Accuracy (avg/stddev/max)	F1 Score (avg/stddev/max)	Precision (avg/stddev/max)	Recall (avg/stddev/max)
RUMEN	GLOVE+CLIP-Img	K=0	81.67 / 0.19 / 81.98	81.67 / 0.19 / 81.98	81.67 / 0.19 / 81.98	81.68/ 0.20/81.99	78.76 / 0.21 / 78.96	78.71 / 0.23 / 78.95	78.76 / 0.21 / 78.96	78.71 / 0.23 / 78.95
		K=1	85.15 / 0.20 / 85.46	85.15 / 0.20 / 85.46	85.15 / 0.20 / 85.46	85.16/ 0.20/85.48	81.83 / 0.56 / 82.36	81.74 / 0.57 / 82.27	81.83 / 0.56 / 82.36	81.79 / 0.57 / 82.32
		K=2	85.25 / 0.41 / 85.88	85.25 / 0.41 / 85.87	85.25 / 0.41 / 85.88	85.27 / 0.42 / 85.91	82.55 / 0.27 / 83.00	82.42 / 0.28 / 82.88	82.55 / 0.27 / 83.00	82.56 / 0.28 / 83.02
		K=3	84.76 / 0.19 / 84.94	84.76 / 0.19 / 84.94	84.76 / 0.19 / 84.94	84.76/ 0.19/84.94	82.64 / 0.26 / 82.94	82.46 / 0.26 / 82.77	82.64 / 0.26 / 82.94	82.75 / 0.27 / 83.07
		K=4	85.27 / 0.28 / 85.62	85.27 / 0.28 / 85.62	85.27 / 0.28 / 85.62	85.27/ 0.28/85.63	**82.86** / 0.30 / 83.18	**82.77** / 0.32 / 83.16	**82.86** / 0.30 / 83.18	**82.85** / 0.27 / 83.14
		K=5	**85.49** / 0.34 / 85.83	**85.49** / 0.34 / 85.82	**85.49** / 0.34 / 85.83	**85.50**/ 0.34/85.83	82.17 / 0.66 / 82.65	82.01 / 0.70 / 82.53	82.17 / 0.66 / 82.65	82.23 / 0.64 / 82.68
			Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	GLOVE+CLIP-Img	K=0	93.73 / 0.20 / 93.94	93.78 / 0.18 / 93.97	93.73 / 0.20 / 93.94	93.87/ 0.14/94.01	87.38 / 0.83 / 88.14	87.42 / 0.81 / 88.16	87.38 / 0.83 / 88.14	87.48 / 0.80 / 88.18
		K=1	94.51 / 0.36 / 94.89	94.54 / 0.36 / 94.91	94.51 / 0.36 / 94.89	94.59/ 0.34/94.93	91.23 / 0.18 / 91.49	91.20 / 0.17 / 91.40	91.23 / 0.18 / 91.43	91.18 / 0.17 / 91.38
		K=2	**95.18** / 0.15 / 95.42	**95.20** / 0.13 / 95.41	**95.20** / 0.15 / 95.42	**95.24** / 0.11 / 95.41	92.82 / 0.22 / 93.08	92.76 / 0.22 / 93.03	92.82 / 0.22 / 93.08	92.75 / 0.22 / 93.02
		K=3	94.89 / 0.33 / 95.15	94.90 / 0.33 / 95.17	94.89 / 0.33 / 95.15	94.92/ 0.34/95.20	**93.31** / 0.19 / 93.57	**93.24** / 0.18 / 93.50	**93.31** / 0.19 / 93.57	**93.26** / 0.19 / 93.53
		K=4	94.24 / 0.36 / 94.75	94.29 / 0.38 / 94.80	94.24 / 0.36 / 94.75	94.35/ 0.39/94.87	92.89 / 0.30 / 93.25	92.88 / 0.29 / 93.24	92.89 / 0.30 / 93.25	92.88 / 0.29 / 93.23
		K=5	94.51 / 0.15 / 94.75	94.54 / 0.16 / 94.80	94.51 / 0.15 / 94.75	94.59/ 0.18/94.87	92.72 / 0.74 / 93.41	92.70 / 0.70 / 93.36	92.72 / 0.74 / 93.41	92.70 / 0.70 / 93.36

Table 5: Results for IxRUMEN and IxROOT9 for hybrid fusion based on the CentralNet fusion network.

unimodal visual results tend to show that visual information provides some useful information that can be used in a multimodal decision process. This is confirmed by the values of the error analysis proposed in Table 6, that evidence the complementarity between both modalities. Indeed, although there is a great deal of learning instances that are correctly classified by the textual modality and misclassified by the visual modality, the opposite is also true for a non negligible set of learning instances. It is also interesting to note that the patch-based strategy for the visual modality with CLIP embeddings shows the worst complementary performance with the textual modality, as the highest percentage of correct guesses made by the visual modality, that are misclassified by the textual modality is evidenced for $K = 0$. This is in line with the results discussed in the above paragraph about the impact of the patch-based strategy on the visual modality, when encoded with multimodal representations.

As a consequence of these preliminary results, we propose to combine the textual unimodal models based on GloVE with the visual unimodal models based on CLIP to apply early fusion and hybrid fusion techniques.

		Synonym v/s Random				Hypernym v/s Random			
	Patch Size	VC,TC	VC,TI	VI,TC	VI,TI	VC,TC	VC,TI	VI,TC	VI,TI
RUMEN	K=0	1259	221	289	157	1110	185	245	166
	K=1	1295	157	326	148	1149	109	254	192
	K=2	1297	156	338	135	1151	134	248	163
	K=3	1302	142	329	153	1153	136	262	155
	K=4	1289	158	344	135	1153	138	250	165
	K=5	1320	155	314	137	1159	139	239	169
		Co-hyponym v/s Random				Hypernym v/s Random			
ROOT9	K=0	639	33	47	24	462	53	57	35
	K=1	667	11	30	35	509	18	44	36
	K=2	668	19	30	26	511	25	47	24
	K=3	663	22	32	26	504	22	59	22
	K=4	639	13	58	33	510	20	52	25
	K=5	663	17	36	27	513	23	45	26

Table 6: Error Analysis on IxRUMEN and IxROOT9 for the GloVE textual model and the CLIP visual model. VC (resp. TC) stands for visual (resp. textual) correct predictions, and VI (resp. TI) stands for visual (resp. textual) incorrect guesses.

5.2 Multimodal Models

Results for multimodal models are given in Table 4 for the early fusion with the attention fusion network (AFN), and in Table 5 for the hybrid fusion with the CentralNet fusion network (CFN). The first conclusion to be drawn is that both fusion techniques allow to achieve higher results in terms of F1 score for all experimental configurations when compared to the best unimodal model (here textual modality encoded with GloVE). In particular, 1.71 point improvement is obtained for synonymy on IxRUMEN with AFN, 1.44 point for hypernymy on IxRUMEN with AFN, 1.16 point for co-hyponymy on IxROOT9 with CFN, and 0.60 point for hypernymy on IxROOT9 with CFN. Interestingly, the early fusion provides better results for IxRUMEN, while the hybrid fusion presents stronger results for IxROOT9. For instance, for IxRUMEN, the difference between AFN and CFN ranges from 1.16 points for hypernymy to 1.12 points for synonymy, while for IxROOT9, the difference is smaller to the advantage of CFN, with values ranging from 0.09

point for co-hyponymy to 0.32 point for hypernymy. Note that these results are obtained for similar values of K, to the exception of synonymy for IxRUMEN, where the AFN provides best results for $K = 1$, while the CFN achieves highest performance for $K = 5$. Nevertheless, when closely looking at the results, it is clear that the difference between the AFN and the CFN is marginal for IxROOT9, while it is clearly in favour of the AFN for IxRUMEN. Note that the best configuration of CFN is presented here, which freezes the unimodal results. Indeed, lower experimental results were obtained for the all-trainable architecture proposed in [58].

The patch-based strategy is also beneficial for the multimodal models. Indeed, all result values for $K > 0$ steadily exceed the figures obtained for $K = 0$, in all experimental setups, i.e. for all datasets, lexico-semantic relations and fusion techniques. Note that within this paper, we propose to use the same number of K for both modalities. This can be an obstacle for further improvements as it has been shown in section 5.1 that the textual modality and the visual modality behave differently with respect to patch size[6]. The other particularity of the multimodal models is that they tend to produce higher results for less number of patches for the symmetric relations (i.e. synonymy and co-hyponymy). As such, they rely on less information for each modality, but take advantage of the diversity of the representations. In particular, for synonymy in IxRUMEN, best results are obtained for $K = 1$, while the best unimodal model provides highest results for $K = 5$. For co-hyponymy in IxROOT9, highest results are evidenced for $K = 2$, while the best unimodal model relies on $K = 5$ to achieve the maximum performance. Note that this situation does not hold for asymmetric relations (i.e. hypernymy), as similar values of K are needed to reach highest results.

5.3 Qualitative Analysis

In order to better understand the quantitative results, we provide a qualitative analysis between unimodal models and multimodal models, by looking at specific successful and unsuccessful cases. In Table 7, we first show learning examples that have been correctly identified by the multimodal fusion model and misclassified by both the unimodal models, and where a specific lexico-semantic relation holds (i.e. synonymy, co-hyponymy, hypernymy). These examples show that when the set of images of both words are closely related in terms of visual content and the respective words non polysemous, positive decisions can be made by the multimodal architecture. Note that in this study, we refer to the multimodal model with the attention fusion network, and both GloVE and CLIP-Image unimodal models.

Word pair	Dataset	Relation	Unimodal
(labour, toil)	RUMEN	synonym	random
(rub, snag)	RUMEN	synonym	random
(walk, paseo)	RUMEN	hypernym	random
(rebate, discount)	RUMEN	hypernym	random
(bowl, tumbler)	ROOT9	co-hyponym	random
(falcon, crow)	ROOT9	hypernym	random

Table 7: Pairs identified by multimodal fusion but misclassified by unimodal models, where a semantic relation holds.

[6]This line of work remains for future work.

However, this situation is relatively rare in the IxRUMEN dataset, while more frequent in the IxROOT9 dataset. But, visual information can also help in disambiguating wrong guesses from the unimodal models for the random relation. Indeed, unimodal models show a high rate of false positives that the multimodal model is capable of handling, as shown in Table 8. Note that most of the result improvements by the multimodal architecture come from this situation. In this case, while the unimodal models infer a lexico-semantic relation, the multimodal model correctly classifies the learning input as random. This situation stands if visual contents are clearly unrelated and word pairs non polysemous.

Word pair	Dataset	Relation	Unimodal
(trafficker, trading)	RUMEN	random	synonym
(esr, keyboard)	RUMEN	random	synonym
(jog, trot)	RUMEN	random	hypernym
(spouse, mate)	RUMEN	random	hypernym
(bettle, ant)	ROOT9	random	hypernym
(flute, saxophone)	ROOT9	random	hypernym

Table 8: Pairs identified by multimodal fusion but misclassified by unimodal models, where a random relation holds.

Finally, some good predictions made by the multimodal model are difficult to interpret based on the associated multimodal information, as illustrated in Table 9. This clearly shows that the proposed model is still subject to deep improvements, especially when the word pair is polysemous and when the quality of the visual information is not controlled, or difficult to retrieve in the case of abstract words.

Word pair	Dataset	Relation	Unimodal
(chalk, trash)	RUMEN	synonym	random
(chest, bureau)	RUMEN	synonym	random
(slob, pig)	RUMEN	synonym	random
(cardholder, clef)	RUMEN	hypernym	random
(bite, snack)	RUMEN	hypernym	random
(bang, fringe)	RUMEN	hypernym	random

Table 9: Pairs identified by multimodal fusion but misclassified by unimodal models, where a semantic relation holds, but interpretation is hard.

6 CONCLUSION

In this paper, we propose the first attempt to deal with the identification of lexico-semantic relations based on multimodal information, thus following the semiotic textology linguistic theory. For that purpose, we build the IxROOT9 and IxRUMEN datasets, the multimodal versions of the gold standards RUMEN and ROOT9, as well as we gather the necessary visual information to apply the augmentation data paradigm. To take advantage of the multimodal information, we implement two fusion techniques (early and hybrid), and extend the patch-based strategy to visual information. Experimental results demonstrate that introducing visual information can reliably supplement the missing semantics of textual information. In particular, improvements are observed that range from 1.71 point to 0.60 point in terms of F1 score depending on the dataset and the lexico-semantic relation. Nevertheless, improvements are still limited, essentially due to the automatic selection process of images, which cannot guarantee the quality of the visual information, as well as the inability to connect abstract words to reliable visual information.

REFERENCES

[1] Houssam Akhmouch, Gaël Dias, and Jose G. Moreno. 2021. Understanding Feature Focus in Multitask Settings for Lexico-semantic Relation Identification. In *Findings of the Association for Computational Linguistics (ACL/IJCNLP)*. ACL, Thailand, 2762–2772.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*. 2425–2433.

[3] Mohammed Attia, Suraj Maharjan, Younes Samih, Laura Kallmeyer, and Thamar Solorio. 2016. CogALex-V Shared Task: GHHH - Detecting Semantic Relations via Word Embeddings. In *Workshop on Cognitive Aspects of the Lexicon*. 86–91.

[4] Georgios Balikas, Gaël Dias, Rumen Moraliyski, Houssam Akhmouch, and Massih-Reza Amini. 2019. Learning Lexical-Semantic Relations Using Intuitive Cognitive Links. In *41st European Conference on Information Retrieval (ECIR)*. 3–18.

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[6] Nesrine Bannour, Gaël Dias, Youssef Chahir, and Houssam Akhmouch. 2020. Patch-Based Identification of Lexical Semantic Relations. In *42nd European Conference on Information Retrieval (ECIR)*. 126–140.

[7] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung chieh Shan. 2012. Entailment Above the Word Level in Distributional Semantics. In *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 23–32.

[8] Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *34th AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 7456–7463.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision (ECCV)*. 104–120.

[10] Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E Losada, G Heinatz Bürki, Linda Cappellato, and Nicola Ferro. 2019. Experimental IR Meets Multilinguality, Multimodality, and Interaction. In *10th International Conference of the CLEF Association (CLEF)*, Vol. 11696. Springer.

[11] Sebastien Delecraz, Leonor Becerra-Bonache, Benoît Favre, Alexis Nasr, and Frédéric Béchet. 2021. Multimodal Machine Learning for Natural Language Processing: Disambiguating Prepositional Phrase Attachments with Images. *Neural Processing Letters* 53, 5 (2021), 3095–3121.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186.

[14] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to Paraphrase for Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 886–897.

[15] R. Fu, J. Guo, B. Qin, W. Che, H. Wang, and T. Liu. 2015. Learning Semantic Hierarchies: A Continuous Vector Space Approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 3 (2015), 461–471.

[16] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.

[17] Benito García-Valero. 2020. The Legacy of János S. Petőfi. Text Linguistics, Literary Theory and Semiotics. *Journal of Literary Semantics* 49, 1 (2020), 61–64.

[18] Goran Glavas and Ivan Vulic. 2019. Generalized Tuning of Distributional Word Vectors for Monolingual and Cross-Lingual Lexical Entailment. In *57th Conference of the Association for Computational Linguistics (ACL)*. 4824–4830.

[19] Marti Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th Conference on Computational Linguistics (COLING)*. 539–545.

[20] Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences* 23, 8 (2019), 639–652.

[21] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *Comput. Surveys* 51, 6 (2019), 1–36.

[22] Glyn W Humphreys and Jie Sui. 2016. Attentional control and the self: the Self-Attention Network (SAN). *Cognitive neuroscience* 7, 1-4 (2016), 5–17.

[23] Sergio Jimenez, Fabio A. Gonzalez, Alexander Gelbukh, and George Duenas. 2019. Word2set: WordNet-Based Word Representation Rivaling Neural Word Embedding for Lexical Similarity and Sentiment Analysis. *IEEE Computational Intelligence Magazine* 14, 2 (2019), 41–53.

[24] Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing Distributional Vectors of All Words for Lexical Entailment. In *4th Workshop on Representation Learning for NLP (RepL4NLP)*. 72–83.

[25] Neha Kathuria, Kanika Mittal, and Anusha Chhabra. 2017. A Comprehensive Survey on Query Expansion Techniques, their Issues and Challenges. *International Journal of Computer Applications* 168, 12 (2017).

[26] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations (ICLR)*, Yoshua Bengio and Yann LeCun (Eds.).

[27] Zornitsa Kozareva and Eduard Hovy. 2010. A Semi-supervised Method to Learn and Construct Taxonomies Using the Web. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1110–1118.

[28] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HTL)*. 153–163.

[29] Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations?. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 970–976.

[30] Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do Supervised Distributional Methods Really Learn Lexical Inference Relations?. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 970–976.

[31] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *26th ACM International Conference on Multimedia (MM)*. 801–809.

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014).

[33] Fenglin Liu, Xian Wu, Shen Ge, Xuancheng Ren, Wei Fan, Xu Sun, and Yuexian Zou. 2021. DiMBERT: Learning Vision-Language Grounded Representations with Disentangled Multimodal-Attention. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 1 (2021), 1–19.

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems (NeurIPS)* 32 (2019).

[35] Catherine Marechal, Dariusz Mikolajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, and Katarzyna Wegrzyn-Wolska. 2019. Survey on AI-Based Multimodal Methods for Emotion Detection. *High-performance modelling and simulation for big data applications* 11400 (2019), 307–324.

[36] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3, 4 (1 January 1990), 235–244.

[37] Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 233–243.

[38] Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 76–85.

[39] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab.

[40] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*. 1532–1543.

[41] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *IEEE International Conference on Data Mining (ICDM)*. 1033–1038.

[42] James Pustejovsky, Eben Holderness, Jingxuan Tu, Parker Glenn, Kyeongmin Rim, Kelley Lynch, and Richard Brutti. 2021. Designing Multimodal Datasets for NLP Challenges. *CoRR* abs/2105.05999 (2021). arXiv:2105.05999

[43] Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask Representation Learning for Multimodal Estimation of Depression Level. *IEEE Intelligent Systems* 34, 5 (2019), 45–52.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021).

[46] Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring Lexical Entailment with a Supervised Directional Similarity Network. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 638–643.

[47] Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet Selective: Supervised Distributional Hypernymy Detection. In *25th International Conference on Computational Linguistics (COLING)*. 1025–1036.

[48] Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora. In *56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 358–363.
[49] Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. Nine Features in a Random Forest to Learn Taxonomical Semantic Relations. In *10th International Conference on Language Resources and Evaluation (LREC)*. 4557–4564.
[50] Enrico Santus, Vered Shwartz, and Dominik Schlechtweg. 2017. Hypernyms under Siege: Linguistically-motivated Artillery for Hypernymy Detection. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 65–75.
[51] Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal Video Summarization via Time-Aware Transformers. In *29th ACM International Conference on Multimedia (MM)*. 1756–1765.
[52] Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. In *54th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2389–2398.
[53] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[54] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations (ICLR)*.
[55] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning Syntactic Patterns for Automatic Hypernym Discovery. In *17th International Conference on Neural Information Processing Systems (NeurIPS)*. 1297–1304.
[56] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.
[57] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. CentralNet: A Multilayer Approach for Multimodal Fusion. In *European Conference on Computer Vision (ECCV)*. 575–589.
[58] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.
[59] Tu Vu and Vered Shwartz. 2018. Integrating Multiplicative Features into Supervised Distributional Methods for Lexical Entailment. In *7th Joint Conference on Lexical and Computational Semantics (*SEM)*. 160–166.
[60] Ivan Vulić and Nikola Mrkšić. 2018. Specialising Word Vectors for Lexical Entailment. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 1134–1145.
[61] Ivan Vulic and Nikola Mrksic. 2018. Specialising Word Vectors for Lexical Entailment. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 1134–1145.
[62] Ivan Vulic, Nikola Mrksic, Roi Reichart, Diarmuid Ó Séaghdha, Steve J. Young, and Anna Korhonen. 2017. Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules. In *55th Annual Meeting of the Association for Computational Linguistics (ACL)*. 56–68.
[63] Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning. In *54th Annual Meeting of the Association for Computational Linguistics (ACL)*. 1671–1682.
[64] Chengyu Wang and Xiaofeng He. 2020. BiRRE: Learning Bidirectional Residual Relation Embeddings for Supervised Hypernymy Detection. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 3630–3640.
[65] Julie Weeds, Daoud Clarke, Jeremy Reffin, David J. Weir, and Bill Keller. 2014. Learning to Distinguish Hypernyms and Co-Hyponyms. In *5th International Conference on Computational Linguistics (COLING)*. 2249–2259.
[66] Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. 2021. Towers of babel: Combining images, language, and 3D geometry for learning multimodal vision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 428–437.
[67] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *CoRR* abs/2010.00747 (2020). arXiv:2010.00747
[68] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal Relation Extraction with Efficient Graph Alignment. In *29th ACM International Conference on Multimedia (MM)*. 5298–5306.

A EXPERIMENTAL SETUPS

Within this first attempt to combine visual and textual information for the identification of lexico-semantic relations, only the highest ranked image for each word has been taken into account, the two less ranked images being withdrawn from the process[7]. In order to train each model, a random split of 90% training and 10% validation instances is built from the original training set. Note that at validation, lexical split is not performed. All models are run 5 times for patch size ranging from 0 (no augmentation) to 5 (5 extra words form the patch), to produce average performance results with corresponding standard deviation values and maximum performance scores. All models are trained with a batch size of 32 for up to 200 epochs with early stopping (patience = 10). Adam optimizer [26] is used with a learning rate = 10^{-5}, β_1 = 0.9 and β_2 = 0.999. With respect to encodings, GloVE embeddings are of size 300, CLIP embeddings are 512-dimensional vectors and VGG19 encodings are of size 4096.

[7]The use of this extra information remains for future work.

Multi-Exit Resource-Efficient Neural Architecture for Image Classification with Optimized Fusion Block

Youva Addad Alexis Lechervy Frédéric Jurie

Normandy University, ENSICAEN, UNICAEN, CNRS, GREYC, France
{youva.addad, alexis.lechervy,frederic.jurie}@unicaen.fr

Abstract

In this paper, we propose a test-time resource-efficient neural architecture for image classification. Building on MSDNet [12], our multi-exit architecture excels in both anytime classification, which allows progressive updates of predictions for test examples and facilitates early output, and budgeted batch classification, which allows flexible allocation of computational resources across inputs to classify a set of examples within a fixed budget. Our proposed multi-exit architecture achieves state-of-the-art performance on CIFAR10 and CIFAR100 in these two critical scenarios, thanks to a novel feature fusion building block combined with an efficient stem block.

1. Introduction and related works

State-of-the-art convolutional neural networks (CNNs) such as EfficientNet [25, 26], ResNet [7], and DenseNet [13] possess remarkable network depth, enabling them to achieve exceptional accuracy. However, these deep models often incur significant computational costs, making real-time inference unattainable on resource-constrained platforms such as smartphones, wearable health monitoring devices, or robotic platforms.

In recent years, considerable efforts have been made to improve the inference efficiency of deep CNNs. Various approaches have been explored, including efficient architecture design [4, 34, 11, 10], network pruning [5, 8, 31, 20], weight quantization [17, 14, 23, 16], knowledge distillation [9, 18, 1], and adaptive inference [28, 12, 6, 30, 19]. Notable contributions in this area include MobileNet [11], ShuffleNet [34], and SqueezeNet [15], which introduced innovative strategies such as depth-wise convolutions, channel shuffling, and fire modules to minimize computational effort while maintaining satisfactory accuracy. In addition, recent advances in neural architecture search (NAS) [29, 2] and knowledge distillation [3, 9] have also played a key role in producing compact and efficient models without significant performance degradation. As the importance of efficient neural networks continues to grow, this paper aims to enrich ongoing efforts by presenting a novel approach that further enhances model efficiency in specific real-world applications.

Adaptive inference aims to reduce computational redundancy on ”easy” examples. Specifically, this method involves designing a model with the ability to intelligently select specific segments of the network to execute during test time, depending on the input it receives. ”Easy” samples require less computation than ”hard” ones. An example of an adaptive inference technique is early exit [12, 30, 22, 6].

Dynamic Early Exit Networks create multiple classifiers within the depth of a network, enabling rapid prediction of high-confidence samples at early stages (easy samples) without activating deeper layers. This recognizes that not all samples require the same model complexity for accurate prediction. Unlike traditional architectures, dynamic early-exiting networks introduce branching points [27] at different depths to assess prediction confidence early. Shallow classifiers provide fast predictions for simple samples, saving computation, while complex samples progress through deeper layers for accurate predictions at a slightly higher cost, balancing efficiency and accuracy. In addition, these networks adapt inference based on resource and latency requirements, prioritizing early exit for speed or exploring deeper branches for improved accuracy.

MSDNet, proposed by Huang et. al. [12], is a leading approach for dynamic early exit that effectively addresses two main challenges to achieve resource-efficient image classification. The first challenge, where classifiers change the internal representation, is solved by using dense connectivity [13], which prevents the dominance of a single early exit and balances the tradeoff between early and later classification through the loss function. The second challenge, the lack of coarse-scale features in early layers, is addressed by employing a multiscale network structure. At each layer,

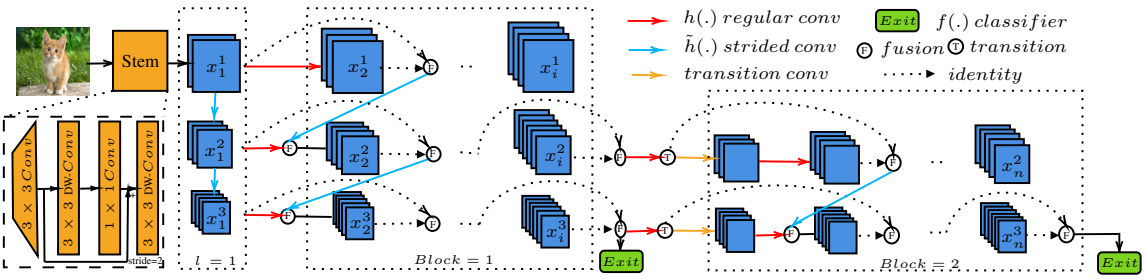


Figure 1: The proposed architecture operates with $S = 3$ scales. It consists of two blocks, with a reduction in the number of scales within each successive block. F represents the concatenation operator, while T denotes the transition operator.

the network produces features of all scales, from fine to coarse. MSDNet has served as the basis for a number of other works, such as those by Meronen et. al. [21].

One limitation of MSDNet lies in the way it reuses previously computed features, as the last features of each scale are not shared with subsequent scales. To overcome this limitation, we present a novel fusion layer that improves the reuse of previous features. In addition, we introduce a novel stem that effectively limits the overall computation of the network, further improving its efficiency.

2. Method

The proposed method extends the foundational principles of MSDNet [12]. The architecture of our model is given in Fig. 1. As mentioned in the introduction, it includes a novel operator for effectively fusing layers across scales and depth, and a novel stem block that further enhances the model's capabilities.

An improved fusion layer. As shown in Fig. 1, the key issue with multi-exit architectures is knowing which features (across scales and depths) to combine at each level, and how to combine them. Our model relies on a feature fusion technique that incorporates both local and global context, enabling the model to make well-informed predictions and decisions. The integration of features through concatenation and strided convolutions plays a central role in enhancing the model's ability to learn complex patterns, leading to significant improvements in its overall performance.

In our model, the output x_l^s of layer l at the s^{th} scale is obtained using the concatenation operator denoted by $[...]$. The transformation $h_l^s(.)$ represents a regular convolution operation, while $\tilde{h}_l^s(.)$ corresponds to a strided convolutional operation. $\tilde{h}_1^1(.)$ correspond to the stem layer which will be described in the next paragraph.

More specifically, the fusion is done according to the following equations:

x_l^s	$l=1$	$l=2$	$l=3$	$l=4$
$s=1$	$\tilde{h}_1^1(x_0^1)$	$h_2^1(x_1^1)$	$h_3^1(x_1^1, x_2^1)$	$h_4^1(x_1^1, x_2^1, x_3^1)$
$s=2$	$\tilde{h}_1^2(x_1^1)$	$\begin{bmatrix}\tilde{h}_2^2([x_1^1, x_2^1]) \\ h_2^2([x_1^2])\end{bmatrix}$	$\begin{bmatrix}\tilde{h}_3^2([x_1^1, x_2^1, x_3^1]) \\ h_3^2([x_1^2, x_2^2])\end{bmatrix}$	$\begin{bmatrix}\tilde{h}_4^2([x_1^1, x_2^1, x_3^1, x_4^1]) \\ h_4^2([x_1^2, x_2^2, x_3^2])\end{bmatrix}$
$s=3$	$\tilde{h}_1^3(x_1^2)$	$\begin{bmatrix}\tilde{h}_2^3([x_1^2, x_2^2]) \\ h_2^3([x_1^3])\end{bmatrix}$	$\begin{bmatrix}\tilde{h}_3^3([x_1^2, x_2^2, x_3^2]) \\ h_3^3([x_1^3, x_2^3])\end{bmatrix}$	$\begin{bmatrix}\tilde{h}_4^3([x_1^2, x_2^2, x_3^2, x_4^2]) \\ h_4^3([x_1^3, x_2^3, x_3^3])\end{bmatrix}$

Although this mechanism may seem similar to MSDNet's, it has one important difference: it makes greater use of features computed at the previous scale level. While this results in better performance, it has the potential disadvantage of increasing the number of computations required to process the features. For this reason, we have also introduced a transition layer (which is equivalent to the original version, except it disregards the previous scale in the convolution process) that is designed to effectively optimize the reduction of spatial dimensions and channel numbers. Instead of merging concatenated features and halving the number of channels, we directly halve the number of channels within the current scale. This reduction is achieved using a 1x1 convolution in our transition layer. The key advantage of this approach is that we strategically downsample features within the fine-scale branch. In this way, we optimize feature processing and facilitate a seamless flow of information within the current scale. The transition layer ensures that the fine-scale features are properly prepared before being fed into the current scale.

An efficient stem layer. The stem layer plays a crucial role in improving the efficiency of the architecture, as it is the key element responsible for effectively extracting essential features from the input data. Fig. 1 illustrates our proposed stem layer, which consists of four successive convolutional layers. It starts with a 3x3 standard convolutional layer to capture initial patterns, followed by a 3x3 Depthwise Separable Convolution layer to extract spatial information. The 3rd layer uses a 1x1 convolution to compress and refine the features. To improve the model's ability to capture important patterns and structures, the fourth layer uses a 3x3 Depthwise Separable Convolution (DW-Conv) with a step size of 2 to downsample the feature maps, making it highly suitable for our specific dataset. In addition, the stem incorporates a residual connection between the output of the

first convolutional layer and the output of the third convolutional layer. This connection allows the model to retain and propagate essential information, promoting effective feature extraction and overall performance improvement.

In addition, the architectural design includes a dedicated pair of convolutional layers, designed exclusively for the classifier function, each with 128 output channels. The first layer in this pair performs a 3×3 convolutional process, smoothly incorporating a downsampling step of 2. This is followed by a 1x1 convolution with a step value of 1. These layers are then followed by the introduction of adaptive average pooling, a technique skilfully used to reshape the spatial attributes into a streamlined 1x1 framework.

3. Experiments

We experimentally validate the effectiveness of our method on two widely recognized image classification datasets: CIFAR-10 and CIFAR-100, and compare our performance with state-of-the-art architectures, namely MSDNet [12] and RANet [30]. To ensure a fair comparison, the experimental settings adopted are consistent with those described in the original MSDNet and RANet papers. We also include two other competing models, namely $ResNet^{MC}$ and $DenseNet^{MC}$ [13]. While we do not give ablative comparisons due to space limitations, each of the two contributions (the fusion layer and the stem layer) result in performance gains of the same order of magnitude. The stem serves the dual purpose of extracting initial valuable features while efficiently reducing floating point operations (FLOPs). In addition, the merging process plays a critical role in significantly improving the overall classification performance, making it a key component in this context. We stress that in order to compare the performance of our model with existing approaches, we used the implementation provided by their authors.

Datasets. The CIFAR-10 and CIFAR-100 datasets consist of 32 × 32 natural RGB images and include 10 and 100 classes, respectively, with each dataset containing 50,000 training images and 10,000 test images. To ensure consistency with previous studies [12], we specifically selected 5,000 images from the training set to form a validation set. This validation set played a crucial role in our research, as it allowed us to fine-tune and optimize the parameters of our method, ultimately identifying the optimal confidence threshold required for adaptive inference. By carefully validating our approach, we were able to improve its performance and generalization capabilities, thereby producing more accurate and reliable results in real-world image classification scenarios.

Training and Data Augmentations. We train the proposed models for 300 epochs using stochastic gradient descent (SGD) with an initial learning rate of 0.1. After 20 epochs of linear warm-up, the schedule transitions to cosine decay. We use a batch size of 512, a momentum of 0.9, and a weight decay of $1e^4$. For data augmentation, we follow the approach described in [7], which involves randomly cropping images to 32×32 pixels after zero padding (4 pixels on each side). In addition, images are flipped horizontally with a 50% probability, and RGB channels are normalized by subtracting their respective channel mean and dividing by their standard deviation. To further improve performance, we integrate popular schemes such as Mixup [33], Cutmix [32], and network regularization with label smoothing [24].

Experiments on Anytime prediction These experiments highlight a model's ability to make predictions with varying degrees of accuracy and computational cost. In traditional image classification tasks, a model processes an input image and generates a single class prediction. In contrast, in an anytime prediction scenario, the model predicts at different computational stages, with each prediction becoming more accurate as additional computational resources are allocated. The classification accuracies are shown in Figure 2. The evaluation includes three classifiers: MSDNet, represented by a black line, RANet, represented by a yellow-green line, and our method, shown in yellow. While MSDNet and RANet show similar performance, RANet performs better when computational resources are limited. However, our model performs significantly better on both the CIFAR-10 and CIFAR-100 datasets. For CIFAR-10, our model achieves an impressive 94.1% accuracy for the last classifier, requiring 32% fewer FLOPs than RANet's 93.8% last classifier, and 15% fewer FLOPs than MSDNet's 93.6% last classifier. Moreover, when the computational budget ranges from 0.25×10^8 FLOPs to 0.64×10^8 FLOPs, the accuracies of the various classifiers in our method consistently exceed those of MSDNet or RANet by 0.5% to 1%. In the case of CIFAR-100, our model achieves a remarkable 76.3% accuracy for the last classifier, using 32% fewer FLOPs than RANet's 74.28% last classifier, and 15% fewer FLOPs than MSDNet's 74.3% last classifier. Again, the accuracies of the various classifiers in our method outperform those of MSDNet or RANet by 1% to 2% when the computational budget is in the range of 0.25×10^8 FLOPs to 0.64×10^8 FLOPs.

Budgeted Batch Classification Experiments. This approach efficiently processes image batches within a predefined computational, memory, or time budget. It divides the batch into smaller subsets and processes them sequentially with different computational resources. Partial predictions for each subset are aggregated to produce the final batch predictions. This strategy enables reasonably accurate classifications with limited resources, making it valuable for applications on resource-constrained devices or systems. Figure 3 shows the results obtained with the budgeted batch classification setting. To identify the optimal model for each budget, we evaluate its accuracy on the test set and

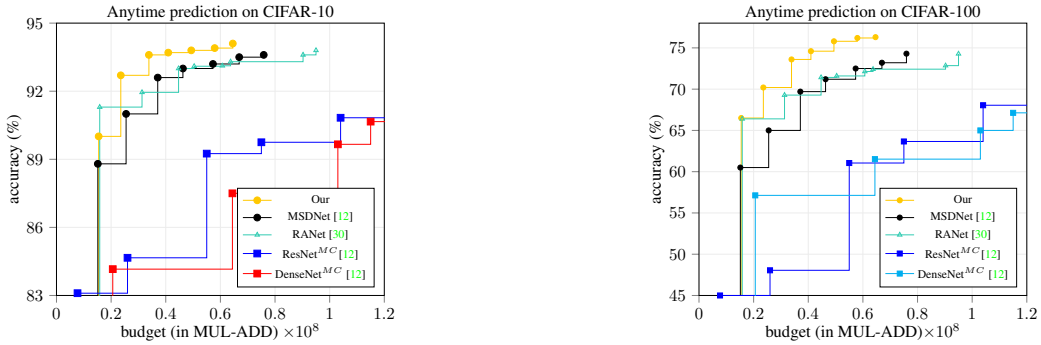


Figure 2: Accuracy (top-1) of anytime prediction models as a function of computational budget on CIFAR-10 (left) and CIFAR-100 (right). Higher is better.

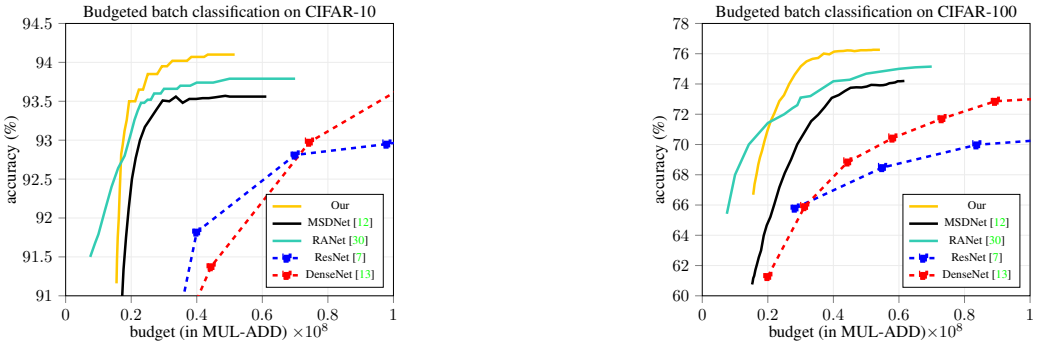


Figure 3: Accuracy (top-1) of budgeted batch classification models as a function of average computational budget per image on CIFAR-10 (left) and CIFAR-100 (right). Higher is better.

plot the corresponding curves for MSDNet, RANet, and our model. For both CIFAR datasets, our model consistently outperforms MSDNets, RANets, and other baseline models for all budgets, except for RANet's low flops performance, which slightly outperforms ours. In particular, networks with a multiscale dense connection architecture consistently achieve significantly higher accuracy than other baseline models with equivalent computational cost, underscoring the strengths of our approach in the budgeted batch classification setting. For computational budgets above 0.2×10^8 FLOPs on CIFAR-10, our proposed model requires 32% fewer FLOPs to achieve a classification accuracy of 93.5% compared to MSDNet and 15% fewer FLOPs than RANet. Similarly, on CIFAR-100, our model achieves a classification accuracy of 74.01% with only about 53% fewer FLOPs than MSDNet and 34% fewer FLOPs than RANet. While RANet and MSDNet perform similarly on CIFAR-10 within the computational budget range of 0.15×10^8 to 0.5×10^8, our model outperforms them, requiring only 0.4×10^8 to reach 94%. On CIFAR-100, the classification accuracies of our model consistently exceed those of MSDNet and RANet by 1% to 2% in the median and high budget intervals (over 0.2×10^8 FLOPs). Furthermore, our model achieves an accuracy of 94.1% when the budget exceeds 0.4×10^8 FLOPs, outperforming MSDNet and RANet by 0.5% and 0.3%, respectively, under the same computational budget conditions. In addition, the experiments show that our model is up to 5 times more efficient than ResNets on both CIFAR-10 and CIFAR-100 datasets. This efficiency further underscores the superiority of our proposed model in the context of budgeted batch classification settings.

4. Conclusions

We have proposed a resource-efficient neural architecture for image classification based on MSDNet. Preliminary results show that this multi-exit design excels in anytime and budget batch classification, achieving state-of-the-art performance on CIFAR10 and CIFAR100. Key contributions include a novel feature fusion block and an efficient stem block. Our approach, which seems promising for real-world scenarios with limited resources, still needs to be validated on more challenging and diverse tasks.

Acknowledgments. Research reported in this paper was supported by the ANR under award number ANR-19-CHIA-0017 and was performed using computing resources of CRIANN.

References

[1] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 1

[2] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. 2018. 1

[3] Minghong Gao. A survey on recent teacher-student learning studies, 2023. 1

[4] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[5] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2015. 1

[6] Yizeng Han, Yifan Pu, Zihang Lai, Chaofei Wang, Shiji Song, Junfeng Cao, Wenhui Huang, Chao Deng, and Gao Huang. *Learning to Weight Samples for Dynamic Early-Exiting Networks*, pages 362–378. 11 2022. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 3, 4

[8] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[9] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 1

[10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[11] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 1

[12] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 4

[13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3, 4

[14] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 1

[15] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size, 2016. 1

[16] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[17] Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang, and Changkyu Choi. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[18] xu lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1

[19] Hao Li, Hong Zhang, Xiaojuan Qi, Ruigang Yang, and Gao Huang. Improved techniques for training adaptive deep networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[20] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[21] Lassi Meronen, Martin Trapp, Andrea Pilzer, Le Yang, and Arno Solin. Fixing overconfidence in dynamic neural networks. *CoRR*, abs/2302.06359, 2023. 2

[22] Mary Phuong and Christoph H. Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[23] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 525–542, Cham, 2016. Springer International Publishing. 1

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[25] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. 1

[26] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. 2021. 1

[27] Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. Branchynet: Fast inference via early exiting from deep neural networks, 2017. 1

[28] Andreas Veit and Serge Belongie. Convolutional networks with adaptive inference graphs. 2018. 1

[29] Colin White, Mahmoud Safari, Rhea Sukthanker, Binxin Ru, Thomas Elsken, Arber Zela, Debadeepta Dey, and Frank Hutter. Neural architecture search: Insights from 1000 papers, 2023. 1

[30] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3, 4

[31] Le Yang, Haojun Jiang, Ruojin Cai, Yulin Wang, Shiji Song, Gao Huang, and Qi Tian. Condensenet v2: Sparse feature reactivation for deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3569–3578, June 2021. 1

[32] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[33] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 3

[34] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

Temporal Sequences of EEG Covariance Matrices for Automated Sleep Stage Scoring with Attention Mechanisms *

Mathieu Seraphim[1 [0000−0002−9367−1190]], Paul Dequidt[1 [0000−0002−8362−7735]], Alexis Lechervy[1 [0000−0002−9441−0187]], Florian Yger[2,1 [0000−0002−7182−8062]], Luc Brun[1 [0000−0002−1658−0527]], and Olivier Etard[3 [0000−0003−3661−0233]]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
[2] LAMSADE, UMR CNRS 7243, Université Paris-Dauphine, PSL, France
[3] Université de Caen Normandie, INSERM, COMETE U1075, CYCERON, CHU de Caen, Normandie Univ, F-14000 Caen, France

Abstract. Electroencephalographic (EEG) data is commonly used in sleep medicine. It consists of a number of cerebral electrical signals measured from various brain locations, subdivided into segments that must be manually scored to reflect their sleep stage. These past few years, multiple implementations aimed at an automation of this scoring process have been attempted, with promising results, although they are not yet accurate enough with respect to each sleep stage to see clinical use. Our approach relies on the information contained within the covariations between multiple EEG signals. This is done through temporal sequences of covariance matrices, analyzed through attention mechanisms at both the intra- and inter-epoch levels. Evaluation performed on a standard dataset using an improved methodological framework show that our approach obtains balanced results over all classes, this balancing being characterized by a better MF1 score than the State of the Art.

Keywords: Sleep analysis · EEG · Deep Learning · Attention · Symmetric Positive Definite matrices.

1 Introduction

To study sleep patterns in the field of sleep medicine, the gold standard is the polysomnography (PSG) study, which usually includes electroencephalography (EEG), electrooculography (EOG), electromyography (EMG) and electrocardiography (ECG) recordings, corresponding to brain, eye, muscle and heart electrical activity, respectively. These signals are derived from the voltage existing between electrodes over time, often with one being set as a reference. In this paper, the term "signal" shall refer exclusively to such a voltage.

* This work has been co-funded by the Normandy Region and the French National Research Agency (ANR) through a HAISCoDe Ph.D. grant. It was granted access to the HPC resources of IDRIS under the allocation 2022-AD010613618 made by GENCI, and to the computing resources of CRIANN (Normandy, France).

Table 1. Frequency bands that we use for EEG data analysis

	Delta	Theta	Alpha	Beta_{low}	Beta_{high}	Gamma
Hz	[0.5, 4[	[4, 8[	[8, 12[	[12, 22[	[22, 30[	[30, 45[

The set of norms most often used to analyze PSG signals is the one defined by the American Academy of Sleep Medicine (AASM) [4]. This analysis is done by subdividing the signals into 30 second epochs, sometimes called "sleep epochs" in this paper. These may be manually scored (labeled) as being in one of five stages: wakefulness, rapid eye movement (REM) sleep, and three stages of non-REM sleep (N1, N2 and N3).

In this paper, we study the relevance of cerebral functional connectivity as a tool for the automated classification of sleep stages, through a study of co-variations between EEG signals. In particular, we aim to obtain a high level of class-wise performance. For that purpose, we analyze timeseries of covariance matrices, computed for various frequency bands (Table 1). We base our analysis on an existing model architecture [14], itself based on successive Transformer encoders. After an overview of the existing State of the Art (SOA) in Section 2, we shall explain our method in Section 3. Finally, in Section 4, we present our results on a commonly used dataset, including a comparison with SOA methods.

2 State of the Art

Some approaches consider that a single signal contains enough information to classify sleep epochs [12, 14, 21]. A common strategy is to combine an EEG and an EOG signal with the same reference electrode by subtracting them [15–17]. Other approaches use a multitude of input signals, often including EOG or EMG signals to said input, in addition to EEG. Phan et al. [11] use one signal of each type (EEG, EOG and EMG) as input, whereas Jia et al. [7, 8] use multiple of each, and additionally include one ECG signal. Given the same dataset, the latter approaches seem to yield better results.

A common approach in EEG preprocessing pipelines is the extraction of relevant frequency components, since sleep stages are characterized by events with specific frequential components [4]. As such, Phan et al. [11,12,14] compute time-frequency images to use as input of their model.

Manual scoring of a sleep epoch takes into consideration said epoch's context - i.e. information contained in neighboring sleep epochs. Similarly, the architectures of models used for this task often include contextual information in the classification process. Such sequence-based models can be divided into two sections: intra-epoch (extracting features from each epoch in the input sequence) and inter-epoch (combining said features). Convolutional neural networks (CNNs) can be used at the intra-epoch level, usually followed at the inter-epoch level by recurrent neural networks (RNNs) [12, 15–17]. Phan et al. expand on both the RNN and attention mechanism approaches. In [11,12], they utilize bi-directional

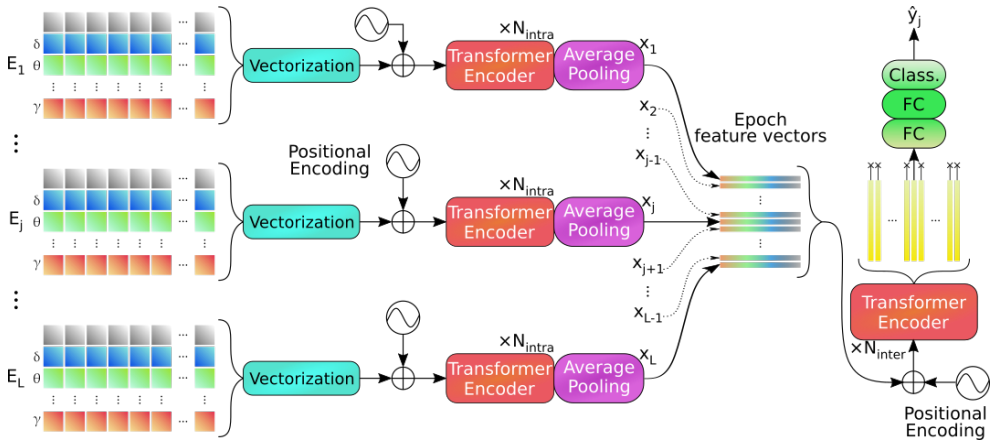


Fig. 1. Our model. $(E_1, ..., E_L)$ is the input sequence, with E_j referring to the central epoch. $\hat{y}_j$ is the output classification of the model. N_{intra} and N_{inter} refer to the number of sublayers in our intra- and inter-epoch Transformer encoders.

RNNs at both the intra-epoch and inter-epoch levels, whereas they use Transformer encoder-based attention mechanisms [18] in [14]. Similarly, Zhu et al. [21] use attention blocs inspired by said encoders at both levels, together with convolutions and other more classic layers. It has been stated that the performance of sequence-based State of the Art automatic sleep scoring models is currently near perfect, with little room for improvement [13]. While we do not dispute that claim in absolute terms, we have noticed a discrepancy in class-wise performance, particularly regarding the N1 stage (see Section 4.4). Therefore, our main focus is to correct for this discrepancy.

Our chosen axis of analysis concerns functional connectivity. In other words, one may study the connectivity between different brain regions through correlations detected between them, often independently of the structural (i.e. physical) connectivity between said regions [6]. In the context of sleep studies, it has been proven that sleep induces a characteristic cerebral response, describable in terms of functional connectivity [5]. Jia et al. [7,8] explicit these inter-region relationships through graph timeseries. Their intra-epoch section is a graph learning model, with each node corresponding to an electrode. These graphs are then convolved both spatially and temporally in the inter-epoch section. Note that most graph convolution methods do not assign a specific weight to each node, nor do they use the relative positioning of said nodes. For the proposed graphs, however, each node corresponds to an electrode, so ignoring node specificity in such a way might actually be a drawback.

In this paper, we perform an analysis of functional connectivity, estimated through the covariations of brain signals. For this purpose, we analyze covariance matrices computed from multiple simultaneous EEG signals, excluding other signal types (EOG, EMG...) in order to focus exclusively on brain activity. Covariance matrices are guaranteed to be symmetric positive semi-definite, but tend

to be fully symmetric positive definite (SPD) when computed from real-world data. The set of all SPD matrices in $\mathbb{R}^{n\times n}$ is a Riemannian manifold (metered curved space), and we postulate that preserving this geometry in our model would be advantageous to our classification, as similar approaches using SPD matrices have already been implemented in the field of EEG signal analysis, most notably in brain-computer interfaces (BCI) [19].

3 Method

3.1 From EEG signals to covariance-derived SPD matrices

As do Zhu et al. [21], we apply a z-score normalization to our EEG signals, in order to harmonize their means and standard deviations. Moreover, according to the AASM [4], the signal components indicative of the current sleep stage have specific frequential properties. In order to allow the network to more effectively analyze them, we filter our EEG signals along the six frequency bands presented in Table 1. This is done through a fourth-order Butterworth bandpass filter.

The discrete events indicative of a sleep epoch's proper classification are around one second in length. To capture them, we elected to subdivide our recordings into one second segments. Each sleep epoch is therefore subdivided into 30 non-overlapping segments. On each segment, we compute a covariance matrix between the n electrodes. We verify that the resulting matrices are properly SPD, and add the matrix $\mathbb{I}_n \times 10^{-5}$ to those who aren't. This is done on the unfiltered and filtered signals, resulting in a total of 7 data channels.

Two main families of metrics have been defined on the set of SPD matrices. The so-called affine invariant metrics [10] are invariant to affine transformations, but have some drawbacks - for instance, it is impossible to compute an algebraic mean using such a metric, though algorithmic approximations do exist. LogEuclidean metrics [2] do not showcase the same invariance properties, but are significantly easier to work with. The LogEuclidean distance between two SPD matrices A and B is defined as:

$$\delta_{LE}^{P}(A,B) = \|log(P^{-1/2}AP^{-1/2}) - log(P^{-1/2}BP^{-1/2})\|_F \tag{1}$$

This metric relies on the bijection existing between the manifold and its tangent space, the space of symmetric matrices, by way of the matrix logarithm and exponential functions. The parameter P may be interpreted as a center of projection onto said space.

Given a covariance matrix, the only mono-signal information stored is the variance of the signal along the segment. Additional signal-specific features may be added using equation 2, which "augments" a covariance matrix C, preserving its SPD property while adding a feature vector V (referred to as a "side vector"), weighted by a factor α (with $V_\alpha = \alpha V$):

$$M = \left(\begin{array}{c|c} C + V_\alpha V_\alpha^T & V_\alpha \\ \hline V_\alpha^T & 1 \end{array}\right) \tag{2}$$

Each epoch entering the model is thus represented by 7 channels of 30 SPD covariance matrices, and their associated side vectors. Multiple side vectors may be computed per matrix, such as its mean, maximum value, or average power spectral density (PSD) over the corresponding one second segment.

Being biological, our EEG data is marked by the specificities inherent to each recording, that are then transferred to our covariance matrices. In order to reduce said specificities, we compute every recording-wise covariance matrix G, and use them to apply a whitening operation [20] onto the relevant matrices:

$$M' = G^{-1/2} M G^{-1/2} \tag{3}$$

The idea is to operate a "transport" of the data M centered around G to be centered around $\mathbb{I}_n$ instead. We perform this shift for each recording and compute distances between centered SPD matrices using equation 1, with $P = \mathbb{I}_n$. If need be, both M and G are augmented with the relevant side vectors.

3.2 The model

Our model architecture uses Transformer encoders at the intra- and inter-epoch levels, as does [14]. It takes as input a timeseries of sleep epochs, composed of a central epoch and l epochs on either side, for a total of $L = 2l+1$. These sequences are constructed with maximum overlap, with classification on the central epoch. Thus, the first and last l epochs of each recording are not classified.

Our model starts with a vectorization layer. It performs the augmentation of matrices by their weighted side vectors (equation 2), followed by the whitening operation. The nature of the side vectors V, and the value of their weight α, are model hyperparameters. Using n electrodes, we project our SPD matrices of $\mathbb{R}^{(n+1)\times(n+1)}$ onto their tangent set (Section 3.1), and vectorize the upper triangular of the resulting symmetric matrix onto $\mathbb{R}^{\frac{(n+1)(n+2)}{2}}$ [2]. These operations being bijective, all Euclidean operations on these vectors are interpretable as LogEuclidean operations on the augmented matrices.

These vectors undergo a positional encoding [18]. The channels are then concatenated and fed to a first, intra-epoch Transformer encoder, composed of a number of sequential sublayers. The fully connected layers present in each encoder sublayer allow for a mixing of the elements of each input vector, and therefore a mixing of the original channels. In order to obtain a single feature vector per sleep epoch, the output of the intra-epoch encoder layer passes through an average pooling layer. The resulting L epoch feature vectors are then fed through another positional encoding layer, followed by an inter-epoch encoder. Only the output vector corresponding to the central sleep epoch is preserved, passing through two fully connected layers, each followed by a ReLU activation and a dropout layer. A final fully connected "classification" layer reduces the output to the desired 5 data points (one per class), and this classification is then fed to a softmax-including cross-entropy loss function.

We optimize this model using the Adam algorithm, with the function parameters β_1, β_2 and ϵ set to 0.9, 0.999 and 10^{-7} respectively. The weight decay is a

hyperparameter, and so are the model's learning rate λ and the corresponding exponential decay parameter γ_λ.

Our architecture can be seen in Fig. 1. The number of sublayers and attention heads of each encoder, the size of parameter tensors for the fully connected layers and the various dropout probabilities are all hyperparameters. Our hyperparameter-obtaining strategy is described in Section 4.2 , and the obtained values are presented in the annex.

4 Experiments

4.1 Dataset used

We chose to validate our model on the SS3 subset of the Montreal Archive of Sleep Studies (MASS) dataset [9], as it is heavily utilized within the SOA and contains a large number of electrodes to choose from for our analysis. Said subset is made up of 62 subjects, with a single full-night recording per subject and 20 EEG channels in common. Each EEG signal went through a notch filter at 60 Hz as well as a lowpass and highpass filter with cutoff frequencies of 0.30 Hz and 100 Hz respectively. This dataset is unbalanced, with the largest and smallest classes (N2 and N1) respectively containing 50.24% and 8.16% of its sleep epochs.

In order to capture a significant range of signals, and to limit redundancy between neighboring electrodes, we chose electrodes F3, F4, C3, C4, T3, T4, O1 and O2. This selection has a relatively homogeneous distribution with regards to the cranium, with inter-hemispheric symmetry to capture relevant variations along that axis. All of these signals are captured with a common reference electrode, located behind the left ear.

4.2 Model validation

As is best practice, we subdivide our database into three subsets: training, validation and test. We utilize a k-fold cross-validation scheme, using the same fold-wise subset separation as Seo et al. [15] in order to facilitate comparisons. Each of the $k = 31$ folds are divided into 50, 10 and 2 recordings for each training, validation and testing set respectively. The 31 folds' testing sets add up to

Table 2. Ablation study and comparison to the SOA.

		Balanced statistics		Unbalanced statistics	
	Method	MF1	Macro accuracy	General accuracy	Kappa
0	SleepTrans. [14]	73.97 ± 3.50	76.37 ± 4.35	81.25 ± 3.54	0.722 ± 0.046
1	IITNet [15]	78.48 ± 3.15	**81.88** ± 2.89	83.90 ± 3.03	0.763 ± 0.043
2	DeepSleepNet [16]	78.14 ± 4.12	80.05 ± 3.47	84.81 ± 3.70	0.773 ± 0.052
3	GraphSleepNet [8]	75.58 ± 3.75	79.75 ± 3.41	80.97 ± 4.35	0.724 ± 0.057
4	Our method	**79.78** ± 4.56	81.76 ± 4.61	**85.05** ± 4.97	**0.776** ± 0.069
5	No covariance	77.39 ± 5.82	79.76 ± 4.95	82.61 ± 6.01	0.741 ± 0.081
6	No side vectors	78.14 ± 4.10	80.56 ± 3.95	83.38 ± 4.16	0.753 ± 0.060

Table 3. F1 scores per class.

	Method	N3 F1	N2 F1	N1 F1	REM F1	Wake F1
0	[14]	74.26 ± 12.36	86.72 ± 3.28	47.60 ± 6.37	83.84 ± 6.99	77.40 ± 8.63
1	[15]	**81.97** ± 8.91	88.15 ± 2.84	56.01 ± 6.54	85.14 ± 5.64	81.11 ± 8.49
2	[16]	80.38 ± 9.35	**89.25** ± 3.12	53.52 ± 8.24	86.67 ± 5.34	80.86 ± 9.04
3	[8]	74.77 ± 12.12	84.84 ± 4.22	50.80 ± 8.06	85.09 ± 7.38	82.42 ± 7.43
4	Ours	78.17 ± 11.49	88.66 ± 4.59	**58.43** ± 6.41	**86.91** ± 7.79	**86.73** ± 6.42

the 62 recordings in SS3, with no overlap. We set the parameter l of our network to 10, as is done in [14]. We rebalance each fold's training set through oversampling, with each class having as many elements as N2 has. The validation and test sets aren't rebalanced, though test sets are further restricted (Section 4.3).

Every hyperparameter research is ran using the Tree-structured Parzen Estimator algorithm [3], as implemented by Optuna [1]. This research is done on the same randomly selected fold. The best hyperparameters are then utilized to train the model on all folds. We use the macro-averaged F1 score (MF1) as our main performance statistic, as it reflects imbalances in class-wise classification performance, and is widely used throughout the SOA. All statistics are summarized over the 31 folds by computing their mean and standard deviation.

4.3 Reproducing the State of the Art

In order to compare our results to the State of the Art, we selected four approaches. Three of those are DeepSleepNet [16], often used as a benchmark, IITNet [15], whose cross-validation folds we are using, and GraphSleepNet [8], which also analyses functional connectivity. The fourth, SleepTransformer [14], shall be discussed subsequently.

All three have their code available on GitHub, and were trained on MASS SS3 in their respective papers. IITNet, GraphSleepNet and DeepSleepNet use sequences of epochs as inputs, of size equal to 10, 5 and 25 respectively. Like us (Section 3.2), IITNet and GraphSleepNet use each sequence to classify a single epoch, respectively the last and central epoch of the sequence. In contrast, DeepSleepNet outputs one classification per epoch in their sequences, which are constructed without overlap. Because of this, for each recording, IITNet won't classify the first 9 epochs, GraphSleepNet will ignore the first and last 2, and DeepSleepNet might ignore up to 24 epochs at the end.

All three models use a similar results aggregation strategy. For each fold, the best trained parameters are used to compute predictions on the test set. Despite originating from different models, these predictions are concatenated, and statistics are computed over this unified predictions tensor. As the number of sleep epochs per recording is not homogeneous, neither are the test sets. This strategy therefore results in an implicit weighting effect, giving more importance to sets of parameters computed on folds with larger test sets.

In order to better compare these methods to our model, we retrained these models with our metrics, folds, and results summarizing methods (Section 4.2).

All methods were adapted to select their best fold learned parameters through their validation MF1 score. In the spirit of fairness, we rebalanced GraphSleepNet and IITNet's training sets through oversampling. DeepSleepNet already does this when pretraining its intra-epoch submodel, and its multi-label sequences can't be rebalanced in that way. We did not change any of their model architectures, and used their published hyperparameters.

The fourth SOA method presented is our reimplementation of the original SleepTransformer model. Compared to our model, this method uses a custom attention softmax layer instead of our average pooling. We also replicated their preprocessing using a recombined Fz-Cz signal from MASS SS3. It was trained with our methodology, including a hyperparameter research.

The obtained results (Tables 2 and 3) differ from those originally published, which may stem for the aforementioned methodological differences. To harmonize all test sets, we have elected to exclude the classification of the first and last 24 epochs of each recording. The training or validation sets remain, however, unchanged. This has been applied to all results presented in this paper.

4.4 Analysis of results

Aside from lines 1, 2 and 3 of Tables 2 and 3, all presented results are preceded by a hyperparameter research.

Line 0 of Tables 2 and 3 show us the results obtained through our reimplementation of SleepTransformer. As we can see, they are the lowest of all presented methods. Due to the similarities between our approaches, one might view these as the baseline for our architecture's performance.

As stated in Section 3.2, we tested multiple side vector types in our hyperparameter research. The one that consistently performed the best was the vector of mean PSDs. The other chosen hyperparameters are described in the annex.

The last 3 lines of Table 2 give an overview of the obtained results. Line 4 corresponds to our results, trained on the best hyperparameters mentioned above. A surprising hyperparameter is the value of α (Section 3.1) of 99.53. This implies that the side vectors have a large impact on the final classification, and thus that our network favors a signal-specific input (one not obtained through covariance). To assess the relevance of covariances altogether, we removed all covariance information from our data (i.e. the non-diagonal elements of the covariance matrices), and reran our model. As seen in line 5 of Table 2, all statistics but Kappa are lower than the ones of line 4 by about 2%. This is coherent with the literature, as decent performances have been obtained on MASS without relying on covariations. We also trained our model on the original covariance matrices themselves, with no side vector augmentation (as seen in line 6). We obtain similar results to line 4 and superior results to line 5, thus implying that considering covariations adds a net benefit.

When it comes to the rest of the reran State of the Art, lines 1 through 4 of Table 2 shows that our model performs better in all measured metrics except for macro-averaged accuracy, where we are a close second. In addition, Table 3 shows that our method outperforms the others in REM, Wake and N1

sleep classification. As seen by the scores and standard deviations, though, the quality of predictions varies much per class, for both the State of the Art and us. In particular, N1 sleep epochs seem particularly hard to classify, but our method shows a two points lead over the next best one in that regard. This lead would explain our ranking in terms of MF1 score (Table 2).

All-in-all, Tables 2 and 3 show that a method based in part on covariance information provides results either equivalent or superior to the State of the Art on this problem (relative to the chosen statistics), with notable improvements to performance on the N1 stage, though it also benefits from signal-specific inputs.

5 Conclusion

We have presented our novel approach for automatic scoring of sleep stages through an analysis of the covariations between EEG signals. Motivated by the high imbalance between the classification of said stages, we established a fairer methodology for training and validating models on this problem. The results validate our hypothesis on the relevance of such covariations in this context, and by extension, that of functional connectivity.

Appendix

The hyperparameters corresponding to the best version of our model are:
Side vectors: PSD; α: 99.53; intra-epoch encoder: 5 sublayers, 15 attention heads, fully connected components of size 1024, dropout of 6.2×10^{-5}; intra-epoch encoder: 6 sublayers, 5 attention heads, fully connected components of size 256, dropout of 8.1×10^{-3}; final fully connected layers: of size 2048, dropout of 1.4×10^{-3}; learning rate (λ): 4.9×10^{-5}, γ_λ at 0.94; weight decay at 1.76×10^{-6}.

Many thanks to Huy Phan [11–14] for answering all our questions.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
2. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic Resonance in Medicine **56**(2), 411–421 (2006)
3. Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. Advances in neural information processing systems **24** (2011)
4. Berry, R.B., Brooks, R., Gamaldo, C., Harding, S.M., Lloyd, R.M., Quan, S.F., Troester, M.T., Vaughn, B.V.: Aasm scoring manual updates for 2017 (version 2.4) (2017)
5. Bouchard, M., Lina, J.M., Gaudreault, P.O., Dubé, J., Gosselin, N., Carrier, J.: EEG connectivity across sleep cycles and age. Sleep **43**(3) (11 2019)

6. Eickhoff, S., Müller, V.: Functional connectivity. In: Toga, A.W. (ed.) Brain Mapping, pp. 187–201. Academic Press, Waltham (2015)
7. Jia, Z., Lin, Y., Wang, J., Ning, X., He, Y., Zhou, R., Zhou, Y., Lehman, L.w.H.: Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering **29**, 1977–1986 (2021)
8. Jia, Z., Lin, Y., Wang, J., Zhou, R., Ning, X., He, Y., Zhao, Y.: Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In: IJCAI. pp. 1324–1330 (2020)
9. O'reilly, C., Gosselin, N., Carrier, J., Nielsen, T.: Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. Journal of sleep research **23**(6), 628–635 (2014)
10. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. International Journal of Computer Vision **66**(1), 41–66 (Jan 2006)
11. Phan, H., Chén, O.Y., Tran, M.C., Koch, P., Mertins, A., De Vos, M.: Xsleepnet: Multi-view sequential model for automatic sleep staging. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5903–5915 (2022)
12. Phan, H., Lorenzen, K.P., Heremans, E., Chén, O.Y., Tran, M.C., Koch, P., Mertins, A., Baumert, M., Mikkelsen, K., Vos, M.D.: L-seqsleepnet: Whole-cycle long sequence modelling for automatic sleep staging (2023)
13. Phan, H., Mikkelsen, K.: Automatic sleep staging of eeg signals: recent development, challenges, and future directions. Physiological Measurement **43**(4), 04TR01 (apr 2022). https://doi.org/10.1088/1361-6579/ac6049, https://dx.doi.org/10.1088/1361-6579/ac6049
14. Phan, H., Mikkelsen, K., Chén, O.Y., Koch, P., Mertins, A., De Vos, M.: Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. IEEE Transactions on Biomedical Engineering **69**(8), 2456–2467 (2022)
15. Seo, H., Back, S., Lee, S., Park, D., Kim, T., Lee, K.: Intra- and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg. Biomedical Signal Processing and Control **61**, 102037 (2020)
16. Supratak, A., Dong, H., Wu, C., Guo, Y.: Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(11), 1998–2008 (Nov 2017)
17. Supratak, A., Guo, Y.: Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). pp. 641–644 (2020)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
19. Yger, F., Berar, M., Lotte, F.: Riemannian approaches in brain-computer interfaces: A review. IEEE Transactions on Neural Systems and Rehabilitation Engineering **25**(10), 1753–1762 (2017)
20. Yger, F., Sugiyama, M.: Supervised logeuclidean metric learning for symmetric positive definite matrices (2015)
21. Zhu, T., Luo, W., Yu, F.: Convolution-and Attention-Based Neural Network for Automated Sleep Stage Classification. Int J Environ Res Public Health **17**(11) (Jun 2020)

An Evidential Deep Network Based on Dempster-Shafer Theory for Large Dataset

Lucas Deregnaucourt
LITIS, Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE
Rouen, France
lucas.deregnaucourt@insa-rouen.fr

Alexis Lechervy
GREYC, Normandie Univ, UMR CNRS 6072, UNICAEN, ENSICAEN
Caen, France
alexis.lechervy@unicaen.fr

Hind Laghmara
LITIS, Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE
Rouen, France
hind.laghmara@insa-rouen.fr

Samia Ainouz
LITIS, Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE
Rouen, France
samia.ainouz@insa-rouen.fr

***Abstract*—We introduce a novel deep neural network architecture based on Dempster-Shafer theory capable of handling large image datasets with numerous classes, such as ImageNet. Our approach involves analyzing images through multiple experts, composed of convolutional deep neural networks that predict mass functions. These experts are then merged using the Dempster's rule, thereby returning a set of potential classes by selecting the best expected utility based on the previously computed mass functions. Our innovative algorithm can identify the best set of classes among the 2^K possible sets for K classes while maintaining a complexity of $O(K\log(K))$. To illustrate our approach, we apply it to an out-of-distribution example search problem, demonstrating its efficiency.**

***Index Terms*—Dempster-Shafer Theory, Evidence theory, belief function, Deep learning, Out-of-distribution**

I. Introduction

In recent years, image classification has made remarkable strides with the advent of deep neural networks (DNNs). However, high ambiguity in the feature vector may lead to missclassification due to the fact that multiple classes share similar expected probabilities. Moreover, a model only trained for precise classifiation may struggle to detect out-of-distribution (OOD) data.

One promising solution to this problem is set-valued classification [1], [2]. This method allows the model to assign a new data to a non-empty set of classes, particularly when uncertainty is high and precise classification is challenging.

In the context of Out-of-Distribution (OOD) detection, a prevalent approach is the utilization of a classification method with a reject option [3], [4], which can be seen as a special case of set-valued classification. Rejection is defined by assigning a data to the set of all possible classes, indicating a state of high uncertainty.

Recently, several works have sought to integrate the Dempster-Shafer theory (DST) into deep neural networks, aiming to leverage the power of evidential reasoning [5]–[7]. However, these attempts have been confined to relatively small and well-structured datasets such as MNIST [8] or CIFAR-10 [9]. The primary impediment has been the algorithmic complexity of DST, which scales exponentially with the size of the frame of discernment Ω, containing 2^K subsets where $K = |\Omega|$.

Based on [10], [11] proposed an end-to-end deep evidential neural network that allocates mass values only to singletons and Ω. This method addresses this computational bottleneck, effectively reducing the spatial complexity from $O(2^K)$ to $O(K+1)$ for the training phase. Nevertheless, the decision-making process for set-valued classification during the evaluation phase remains a computationally expensive task, requiring an exhaustive selection from all possible subsets of Ω, still operating at $O(2^K)$ complexity. Thus, they selected the possible subsets of Ω based on the distance between the classes derived from the confusion matrix.

We propose in this work an algorithmic solution to mitigate the $O(2^K)$ complexity, making set-valued decisions derived from a mass function output by a Convolutional Neural Network (CNN) feasible with linear complexity without intermediate steps to restrict the number of subsets. Additionally, we introduce mathematical optimizations to enhance numerical computations, enabling scalable implementation of set-valued classification evidential models. These contributions pave the way for the application of the DST theoretical framework to high-dimensional real-world datasets with many classes. They offer significant potential for improving the reliability of deep learning models in various applications such as OOD detection.

The remaining parts of this work are organized as follows. In section II we recall basics of Dempster-Shafer theory. In section III, we present the evidential neural network architecture we use and the algorithmic solution we propose to make set-valued decision in linear complexity. The experiments and preliminary results on large datasets are presented in section

IV. Finally, we conclude in section V.

II. Belief Theory

A. Background on belief functions

Belief function theory, called also Evidence theory or Dempster-Shafer theory [12], [13], is able to model and reason about imprecise and uncertain problems, and has more obvious advantages in the representation and combination of uncertain information.

To represent partial knowledge in the belief function theory, let consider the *frame of discernment* Ω as a finite set of variables ω which refers to K elementary events to a given problem ($\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$).

The power set of Ω is the set of all the 2^K possible subsets. It is presented as follows:

$$2^\Omega = \{\emptyset, \{\omega_1\}, ..., \{\omega_k\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, ..., \Omega\}, \quad (1)$$

where the $\{w_i\}$ elements are titled as singletons and $\emptyset$ denotes the empty set.

The key point of Dempster-Shafer theory is the basic belief assignment (*bba*) which represents the partial knowledge about the value of w. A *bba* is a function from 2^Ω to $[0, 1]$ defined as follows:

$$\begin{aligned} m : 2^\Omega &\to [0, 1] \\ A &\mapsto m(A) \end{aligned} \quad (2)$$

where m satisfies the following constraint:

$$\sum_{A \subseteq \Omega} m(A) = 1. \quad (3)$$

An element A of Ω is called a *focal element* when $m(A) > 0$, and the set containing all these elements is called a *body of evidence* (BOE). When each element in BOE is a singleton, m is named a *Bayesian bba*. On the other hand, when BOE contains only Ω as a focal element, we are in the *complete ignorance* situation and m is called vacuous belief function. However, when it contains only one singleton of Ω as a focal element, m is presented as a *Certain mass function*.

A *bba* function is normalized when the mass given to the empty set is constrained to be zero ($m(\emptyset) = 0$). In that case, it corresponds to the *closed-world assumption* [13]. A contrary explanation is that the frame of discernment Ω can be incomplete and the value of w can be taken outer Ω. Accordingly, the mass of belief that is not linked to Ω can allowed to be strictly positive ($m(\emptyset) > 0$). That case corresponds to the *open world assumption* [14].

B. Information fusion

The most common way to combine two *bba* m_1 and m_2 defined on the same frame of discernment Ω is the Dempster's rule [13], denoted as $\oplus$. It is defined by $m_{DS}(\emptyset) = 0$ and $\forall A \in 2^\Omega \backslash \{\emptyset\}$ by

$$m_{DS}(A) = (m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{\substack{B \cap C = A \\ B, C \in 2^\Omega}} m_1(B) m_2(C) \quad (4)$$

where κ is the degree of conflict between the two sources of evidence defined by:

$$\kappa = \sum_{\substack{B \cap C = \emptyset \\ B, C \in 2^\Omega}} m_1(B) m_2(C).$$

This fusion can be seen as the normalized version of the conjunctive rule which is defined by:

$$m_\cap(A) = \sum_{\substack{B \cap C = A \\ B, C \in 2^\Omega}} m_1(B) m_2(C). \quad (5)$$

C. Decision-making

The most common way of making decisions with belief functions is to apply the pignistic transformation [15] to obtain a probability vector of size K, then the predicted class corresponds to the argmax of this vector. However, such a strategy doesn't allow the model to predict a set of classes. To this end, [16] defines the lower and upper expected utilities of selecting $A \subseteq \Omega$ as follows:

$$\overline{\mathbb{E}}(f_A) = \sum_{B \subseteq \Omega} m(B) \max_{\omega_j \in B} u_{A,j} \quad (6)$$

$$\underline{\mathbb{E}}(f_A) = \sum_{B \subseteq \Omega} m(B) \min_{\omega_j \in B} u_{A,j} \quad (7)$$

where $u_{Aj} \in [0, 1]$ designates the utility of the act of selecting $A \subseteq \Omega$ denoted as f_A when the ground truth is ω_j. The utility matrix $U_{2^{|\Omega|} \times K}$ is computed following [17], [18] with a parameter $\gamma \in [0.5, 1]$ that represents the imprecision tolerance degree. If the true class is ω_j, the utility of assigning a sample to set A is calculated as an Ordered Weighted Average (OWA) aggregation [18] of the individual utilities associated with each precise assignment within A as follows:

$$u_{A,j} = g_{|A|} 1_{\{\omega_j \in A\}} \quad (8)$$

where $g \in \mathbb{R}^{|A|}$ is a weight vector whose elements represent the decision making strategy's tolerance to imprecision. For example if $g = (1, 0, \ldots, 0)$, then the decision making's strategy will be totally intolerant to imprecision, thus forcing the model to output only one class.

Following [17] and [19], this weight vector is obtained by maximizing the following entropy:

$$Ent(g) = \sum_{k=1}^{|A|} \log g_k \quad (9)$$

subject to constraints $\sum_{k=1}^{|A|} g_k = 1$, $\sum_{k=1}^{|A|} \frac{|A| - k}{|A| - 1} g_k = \gamma$ and $g_k \geq 0$ where γ is a parameter representing the tolerance to imprecision. An example of a utility matrix with $\gamma = 0.9$ and $\Omega = \{\omega_1, \omega_2, \omega_3\}$ is shown in Table I. As we can see, the values in the utility matrix are the same according to the cardinality of the selected set. This means that instead of computing every values of the utility matrix, we only need to compute a value U_k for each possible cardinality of the

subsets of Ω. In this example, we have $U_1 = 1$, $U_2 = 0.9$ and $U_3 = 0.8263$.

Since we have:

$$\min_{\omega_j \in A} u_{A,j} = \begin{cases} U_k & if\ A = \Omega \\ 0 & else \end{cases} \tag{10}$$

and

$$\max_{\omega_j \in A} u_{A,j} = U_{|A|} \tag{11}$$

the equations (6) and (7) can be simplified as illustrated in section III-C.

	ω_1	ω_2	ω_3
$f_{\{\omega_1\}}$	1	0	0
$f_{\{\omega_2\}}$	0	1	0
$f_{\{\omega_3\}}$	0	0	1
$f_{\{\omega_1,\omega_2\}}$	0.9	0.9	0
$f_{\{\omega_1,\omega_3\}}$	0.9	0	0.9
$f_{\{\omega_2,\omega_3\}}$	0	0.9	0.9
$f_{\{\Omega\}}$	0.8263	0.8263	0.8263

TABLE I
UTILITY MATRIX WITH $\gamma = 0.9$ AND $K = 3$.

The expected utility is then obtained using the generalized Hurwicz decision criterion [20], [21] as follows:

$$\mathbb{E}(f_A) = \nu \underline{\mathbb{E}}(f_A) + (1 - \nu)\overline{\mathbb{E}}(f_A). \tag{12}$$

Where $\nu \in [0, 1]$ is the pessimism index.

When $\gamma = 0.5$, the decision-making strategy is totally intolerant to imprecision so that $u_{ij} = 1$ if $\omega_i = \omega_j$, else $u_{Aj} = 0$. In this sense, we can see the expected utility as a generalized accuracy. The other extreme strategy is the totally tolerant which is achieved when $\gamma = 1$ where $u_{Aj} = 1$ if $\omega_j \in A$, else $u_{Aj} = 0$ so that a model that always outputs Ω will get an expected utility of 1.

We chose this decision-making strategy among all those proposed in [16] since it is the most general form of decision criterion resulting from Jaffray's axioms [21]. Moreover, the expression of the expected utility leads to interesting simplifications in the restricted framework where we only consider the singletons and Ω.

III. SCALABLE EVIDENTIAL NEURAL NETWORK

In this section, we present how the DST framework can be incorporated into a deep neural network architecture. Considering some asumptions on the *bbas* the model will output, we propose an algorithmic solution to make set-valued decision in linear complexity along with mathematical optimizations for a more scalable implementation.

A. Evidential deep neural network

As depicted in Figure 1, the proposed evidential neural network architecture is very similar to a probabilistic one. Our architecture is based on the evidential deep neural network architecture introduced in [11]. The main difference resides in the construction of the mass function. The given image of size $(C \times H \times W)$ first passes through the backbone of a convolutional neural network, resulting in a feature map of size $(C' \times 1 \times 1)$. This feature map captures the data's latent representation.

In the work presented in [11], the construction of mass functions involves the use of a distance-based layer. The classifier is composed of p prototypes t_i in $\mathbb{R}^P$, where P is the dimension of the feature map. In their method, the first step is to compute the distance-based support between the feature map x of a data and each prototype t_i. For the second step, the mass function m_i associated to t_i is computed by multiplying the distance-based support s_i by a weight h_{ij} which characterizes the degree of membership of prototype t_i to the class ω_i.

Our method for constructing the mass functions is more computer vision oriented and is inspired by mixture of experts approaches [22]. Instead of considering prototypes, we consider p experts that see the feature map of a data from different points of view. For this purpose, the classical fully connected layer is replaced by a depthwise convolution [23] with a kernel of size (1×1) and p groups. For a given feature map and a given number of experts p, the depthwise convolution will output a matrix of size $(p \times (K+1))$, namely one mass function per expert. Each mass function holds $|\Omega| + 1$ values, with one value dedicated to each singleton and an another one for the entire set Ω. This vector is then reshaped into a matrix of experts of size $p \times (|\Omega| + 1)$. We apply a softmax activation to satisfy the equation (3). In this matrix, the i-th row represents the mass function associated with expert p_i. The *bbas* of this matrix are then fused with Dempster's rule to obtain a final *bba* of size $|\Omega| + 1$ which we will present in the next section.

B. Computational optimization of Dempster's rule

As seen in the previous section, since our network is only considering the masses assigned to singletons and Ω, the expression of the conjunctive rule simplifies as shown in equation (13).

$$\begin{aligned} m_\cap(A) &= \sum_{\substack{B \cap C = A \\ B,C \in 2^\Omega}} m_1(B) m_2(C) \\ &= m_1(A)m_2(A) + m_1(A)m_2(\Omega) + m_1(\Omega)m_2(A) \end{aligned} \tag{13}$$

$\forall A \in \Omega$.

This brings us to an iterative algorithm for performing Dempster's rule as shown by the Algorithm 1. We define $\mu_1 = m_1$ and $\mu_{i+1} = m_\cap(\mu_i, m_i)$ where μ_i represents the mass function obtained by the fusion of the i first expert's mass functions by the conjunctive rule.

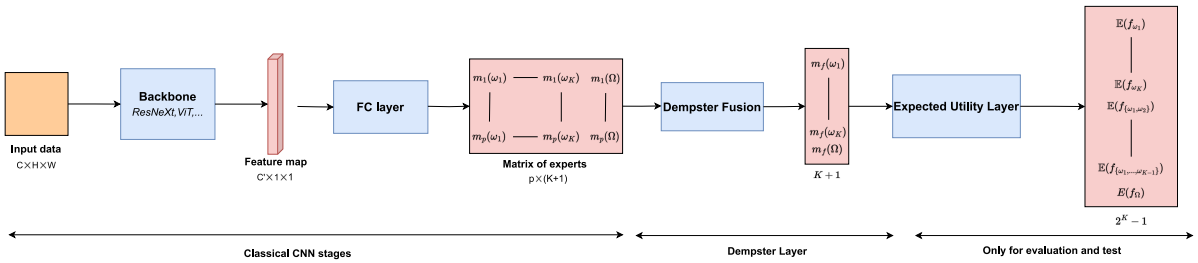


Fig. 1. Architecture of an evidential deep neural network.

Algorithm 1 Iterative Dempster's rule

Require: p mass functions $m_1, \ldots, m_p$

$\mu_1 \leftarrow m_1$

for $i = 2, \ldots, p$ **do**

for $j = 1, \ldots, K$ **do**

$\mu_i(\{\omega_j\}) = \mu_{i-1}(\{\omega_j\})m_i(\{\omega_j\}) + \mu_{i-1}(\{\omega_j\})m_i(\Omega) + \mu_{i-1}(\Omega)m_i(\{\omega_j\})$

end for

$\mu_i(\Omega) = \mu_{i-1}(\Omega)m_i(\Omega)$

end for

return μ_p/Z

where Z is a normalization term.

The expression of $\mu_i(\{\omega_j\})$ can be rewriten as follows:

$$\begin{aligned}\mu_i(\{\omega_j\}) &= \mu_{i-1}(\{\omega_j\})m_i(\{\omega_j\}) + \mu_{i-1}(\{\omega_j\})m_i(\Omega) \\ &\quad + \mu_{i-1}(\Omega)m_i(\{\omega_j\}) \\ &= (\mu_{i-1}(\{\omega_j\}) + \mu_{i-1}(\Omega)) \times (m_i(\{\omega_j\}) + m_i(\Omega)) \\ &\quad - \mu_{i-1}(\Omega)m_i(\Omega)\end{aligned} \tag{14}$$

which leads to an improved algorithm that only iterates on the number of classes K as presented in the Algorithm 2.

Algorithm 2 Scalable Dempster's rule

Require: p mass functions $m_1, \ldots, m_p$

$\mu_p(\Omega) = \prod_{i=1}^{p} m_i(\Omega)$

for $j = 1, \ldots, K$ **do**

$\mu_p(\{\omega_j\}) = \prod_{i=1}^{p} (m_i(\{\omega_j\}) + m_i(\Omega)) - \mu_p(\Omega)$

end for

return μ_p/Z where Z is a normalization term.

The algorithm 2 is highly parallelizable and each element of the loop can be calculated independently of the others, unlike the algorithm 1 where each element depends on the previous iteration. In practice, this second algorithm provides a very fast implementation of Dempster's rule in the restricted framework chosen where we only consider singletons and Ω.

C. Scalable decision making

Since we only consider the singletons and Ω for the construction of the mass function, we can simplify the equations (6) and (7) as follows:

$$\overline{\mathbb{E}}(f_A) = \sum_{\omega_i \in \Omega} (m(\{\omega_i\})u_{A,i}) + m(\Omega) \max_{\omega_k \in \Omega} u_{A,k}, \tag{15}$$

$$\underline{\mathbb{E}}(f_A) = \sum_{\omega_i \in \Omega} (m(\{\omega_i\})u_{A,i}) + m(\Omega) \min_{\omega_k \in \Omega} u_{A,k}. \tag{16}$$

During the training phase, we want f_A to be a singleton. That's to say $u_{ii} = 1$ and $u_{ij} = 0 \ \forall i \neq j$ which can be seen as the classical accuracy metric. Under those hypotheses, we can simplify the equations (15) and (16) as follows:

$$\overline{\mathbb{E}}(f_{\omega_i}) = m(\{\omega_i\}) + m(\Omega) \tag{17}$$

$$\underline{\mathbb{E}}(f_{\omega_i}) = m(\{\omega_i\}) \tag{18}$$

leading to this simplified expression of the expected utility:

$$\begin{aligned}\mathbb{E}(f_{\omega_i}) &= \nu m(\{\omega_i\}) + (1-\nu)\left(m(\{\omega_i\}) + m(\Omega)\right) \\ &= m(\{\omega_i\}) + (1-\nu)m(\Omega).\end{aligned} \tag{19}$$

This expression can be considered as a rewriting of the pignistic transformation in our restricted framework. Indeed, taking $\nu = 1 - \frac{1}{|\Omega|}$ in equation (19) leads to the pignistic probability expression when $m(A) = 0 \ \forall A \subset \Omega$ such that $|A| \geq 2$.

We propose to use the cross-entropy loss on the expected utilities vector for training our network:

$$-\sum_{i=1}^{n}\sum_{k=1}^{K} y_{i,k} \log\left(\mathbb{E}(f_{\omega_k}(x_i))\right) \tag{20}$$

with n is size of training dataset, $y_{i,k}$ is 1 if the label of example x_i is ω_k and 0 otherwise.

For decision-making during the evaluation and test phase, we want our network to be able to output a subset of Ω. The main obstacle is the algorithmic complexity since it would require to compute $2^{|\Omega|}$ expected utilities to choose the subset that maximizes it. To solve this issue, [11] proposes to compute the confusion matrix from the training set generated by an evidential deep neural network as explained above. Based on the distance between the classes, they only keep the classes and groups of classes that are similar enough by thresholding. Although in practice this strategy reduces the

number of expected utilities to be computed, it remains in $2^{|\Omega|}$ in the worst case (when the result is to be attributed to the Ω set). Furthermore, we are not convinced that this strategy is sufficient to scale to databases with a large number of classes such as ImageNet [24] where $|\Omega|$=1000. Moreover, it requires a costly intermediate step between the training phase and the evaluation and test phases.

To this end, we propose a very simple and computationally efficient iterative algorithm (3) to determine the argmax between all subsets of Ω without any *a priori* about the correlation between the classes nor intermediate step to restrict the number of subsets of Ω. The first step is to compute the expected utilities of singletons using the equation (19) and to sort them in a decreasing order. We then compare the higher singleton expected utility with the expected utility of the subset composed of the two best singletons using the equations (12),(15),(16) and so on until adding a new singleton to the subset decreases the expected utility. Let's consider $\Omega = \{\omega_1, \omega_2, \omega_3\, \omega_4\}$ with $\mathbb{E}(\omega_1) > \mathbb{E}(\omega_2) > \mathbb{E}(\omega_3) > \mathbb{E}(\omega_4)$. We then compute $\mathbb{E}(\{\omega_1, \omega_2\})$ and compare it with $\mathbb{E}(\omega_1)$. Let's suppose that $\mathbb{E}(\{\omega_1, \omega_2\})$ is effectively higher than $\mathbb{E}(\omega_1)$, we now have to compute $\mathbb{E}(\{\omega_1, \omega_2, \omega_3\})$. By considering that $\mathbb{E}(\{\omega_1, \omega_2\}) > \mathbb{E}(\{\omega_1, \omega_2, \omega_3\})$, we obtain $A^\star = \{\omega_1, \omega_2\}$. If $\mathbb{E}(A^\star) > \mathbb{E}(\Omega)$ then the model outputs $A^\star$, else it ouputs Ω.

Algorithm 3 Argmax of the Expected Utility

Require: sorted singletons expected utilities $\mathbb{E}(\{\omega_{\alpha_1}\}) \geq \mathbb{E}(\{\omega_{\alpha_2}\}) \geq \ldots \geq \mathbb{E}(\{\omega_{\alpha_K}\})$.
$A^\star \leftarrow \omega_{\alpha_1}$
for $i = 2, \ldots, K$ **do**
  $A^\star_{temp} \leftarrow \{A^\star, \omega_{\alpha_i}\}$
  if $\mathbb{E}(A^\star_{temp}) > \mathbb{E}(A^\star)$ **then** $A^\star \leftarrow A^\star_{temp}$
  end if
end for
return $A^\star$

This strategy allows the model to output a set of classes among all the possible subsets of Ω while maintaining a complexity of $O(K \log(K))$ without requiring any limitations on the number of subsets of Ω to compare their expected utilities.

IV. Experiments

To demonstrate the relevance of our model, we conducted several experiments. Firstly, we carry out a study on the impact of the various parameters on our model. Secondly, we sought to demonstrate the ability of our model to process large databases containing a large number of classes and compare our model with a standard problistic model for classification problem. Finally, we demonstrated the superiority of our approach over the standard probabilistic model for an OOD detection task.

In all our experiments, we assume that the backbone used is of type ResNext50 [25]. This applies both to our model and to the probabilistic models to which the comparison is conducted.

A. Datasets

We conducted our experiments using the following 3 databases: CIFAR-100, ImageNet and SVHN dataset.

CIFAR-100 [26] is a database of low-resolution 28×28 images. It contains $60,000$ images divided into 100 classes with 600 images per class.

ImageNet [24] contains 1.5 million images of 224×224 resolution, manually annotated in $1,000$ categories. The annotation is based on the WordNet hierarchical object categorisation structure (augmented by 120 dog categories).

The SVHN (Street View House Numbers) database [27] is a collection of 32×32 digital images that includes handwritten digits from photos of house numbers taken in street scenes. The database contains 10 classes, corresponding to digits from 0 to 9.

B. Ablation study

In this section, we present some experiments designed to measure the impact of the various parameters of our approach on its performances. We measure two metrics: *expected utility* and *average cardinality*.

Given that the accuracy is obtained by fixing the imprecision tolerance degree γ to 0.5 while computing the expected utility, we propose to evaluate the *expected utilities* across a range of γ values from 0.5 to 0.95.

We compute the *average cardinality* of the predictions according to γ as follows:

$$AC(T) = \frac{1}{|T|} \sum_{i=1}^{|T|} |A(i)| \tag{21}$$

where $T = \{x_1, \ldots, x_{|T|}\}$ is the test set and $A(i)$ is the set-valued output for the data $x_i \in T$. It is clear that for $\gamma = 1$, the model will always output f_Ω since $\mathbb{E}(\Omega) = 1$ and the *average cardinality* will be equal to the number of elements in Ω.

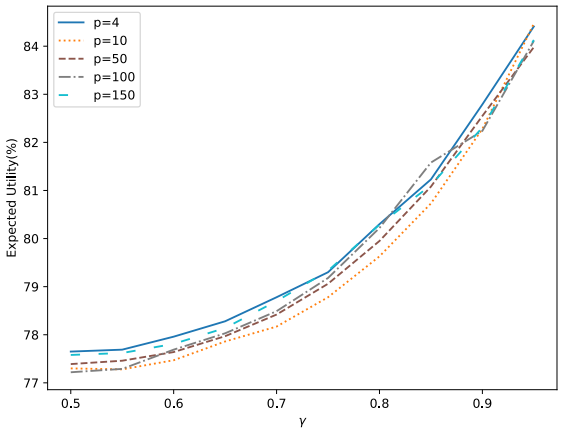


Fig. 2. Expected Utility according to the number of experts on CIFAR-100.

Firstly, we need to determine the hyperparameters of our model, namely the number of experts p and the degree of pessimism ν. Since this search process is quite time-intensive,

we restrict it to the CIFAR-100 dataset. To identify the optimal number of experts, we fix ν to 0.99 so that the equation(19) corresponds to the pignistic probability. As shown on Figure 2, the impact of the number of experts does not appear to be significant. This is mainly because there is no guarantee that the experts simulated by the fully connected layer will be independent. So we choose $p = 4$ as there is no need for a lot of experts. Then we search for the optimal ν by setting the number of experts $p = 4$. As depicted in Figure 3, the model learns in a similar way, independently of ν. Indeed, the model always outputs a value very close to zero for $m(\Omega)$ for precise classification task, so the impact of ν is not significant during the training phase. Consequently, we have selected $\nu = \frac{1}{|\Omega|}$, namely $\nu = 0.99$ for CIFAR-100 and $\nu = 0.999$ for ImageNet.

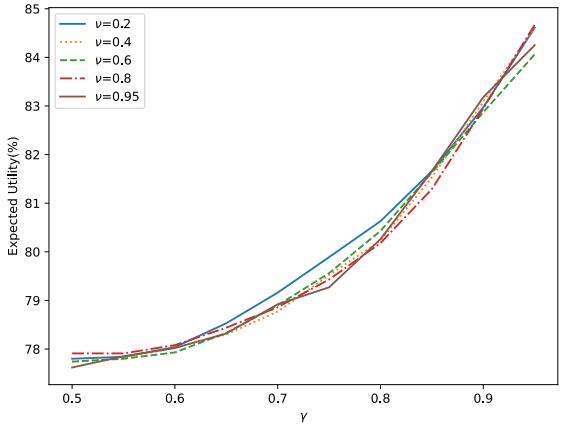


Fig. 3. Expected Utility according to ν on CIFAR-100.

C. Comparison with probabilistic approaches for image classification

Now that we have fixed the model hyperparameters, we can compare the evidential neural network with the probabilistic one on precision classification. As mentioned previously, the probabilistic model used corresponds to a ResNext50 type backbone. This is followed by a fully connected layer and a softmax.

For fair comparison between our method and the probabilistic approach, we have to allow the probabilistic network to output set-valued predictions in order to compute the expected utility. To do so, we considere the probability vector output by the model as a mass function with $m(\Omega) = 0$ and $m(\{\omega_j\}) = p(\omega_j) \ \ \forall j = 1, \ldots, K$.

The Expected Utility and Cardinality curves over 10 runs on CIFAR-100 are respectively presented in Figure 4 and Figure 5. The Expected Utility and Cardinality curves on ImageNet are respectively presented in Figure 6 and Figure 7. Due to the size of the database, we limited the ImageNet experiments to a single run and were therefore unable to calculate standard deviations. For both experiments, we can see that there is almost no difference between the two models from $\gamma = 0.5$ to $\gamma = 0.7$ where the decision-making strategy is quite intolerant

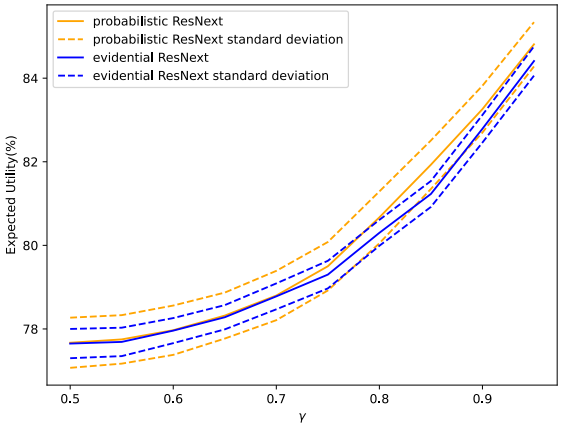


Fig. 4. Expected Utility on CIFAR-100.

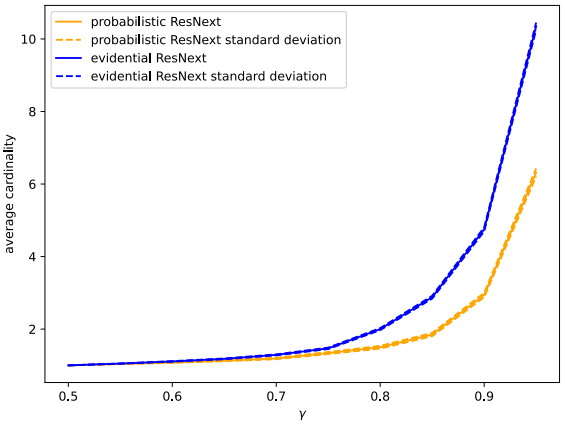


Fig. 5. Average Cardinality on CIFAR-100.

to uncertainty, forcing the model to output one or two classes. For $\gamma = 0.75$ to $\gamma = 0.95$ the evidential model is less confident than the probabilistic one and outputs sets with a higher cardinality, which decreases the Expected Utility. On Imagenet the performance of the probabilistic model is 77.77% in accuracy against 77.65%. The difference in performance is relatively small.

D. OOD detection

For OOD detection task, we want to evaluate the capability of the network to output Ω if, and only if, the data does not belong to the classes from the training set. For this purpose, we evaluate the rate of f_Ω by varying γ from 0.5 to 0.95. A good model has to get a high rate of f_Ω on out-of-distribution data and a low rate of f_Ω on in-distribution data. For $\gamma = 1$, the model will always predict Ω since all the non-zero values in the utility matrix will be equal to 1. So the f_Ω rate will always be equal to 100%.

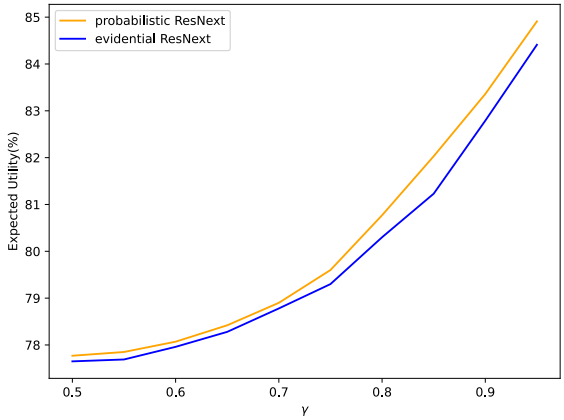


Fig. 6. Expected Utility on ImageNet.

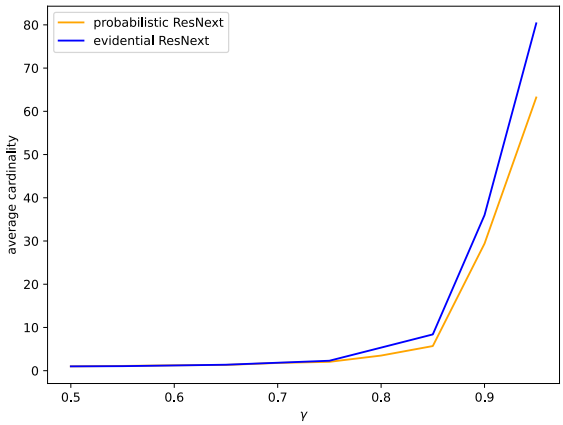


Fig. 7. Average Cardinality on ImageNet.

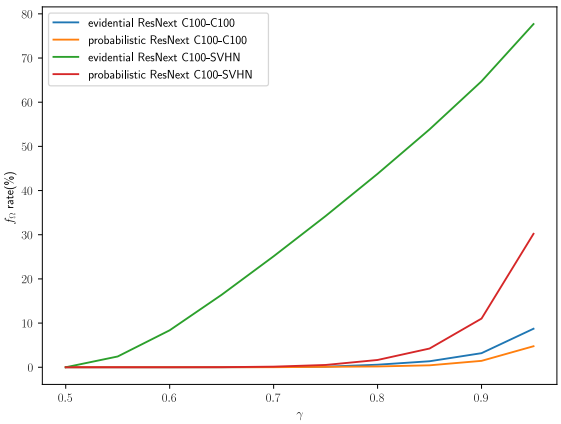


Fig. 8. f_Ω rate for OOD detection, CIFAR-100.

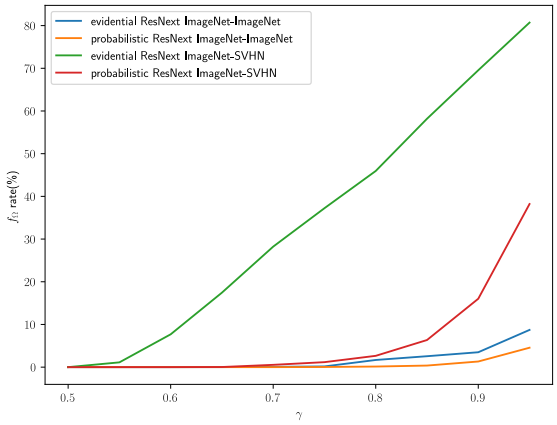


Fig. 9. f_Ω rate for OOD detection, ImageNet.

The results on the OOD detection task for the models trained on CIFAR-100 and ImageNet are respectively presented in Figure 8 and Figure 9. As expected, the f_Ω rate is very low for the evidential and the probabilistic models on in-distribution test set. However, it is clear that the evidential network outperforms the probabilistic network for OOD detection task when we evaluate them on the SVHN dataset.

V. Discussions and Conclusions

In this work, we have presented a novel deep neural network based on Dempster-Shafer theory capable of handling large datasets for image classification. Furthermore, we have introduced mathematical optimizations to improve numerical computations, facilitating a scalable implementation of evidential models for set-valued classification. This approach makes it possible to obtain results on databases with a large number of classes, while avoiding the problem of traversing the 2^K subset of possible classes.

The proposed evidential neural network shows similar results to the probabilistic one for precise classification task. One way to improve it can be to ensure the independence of the experts with a Deep Ensemble approach [28], [29].

However, our network clearly outperforms the probabilistic one for OOD detection task regarding the f_Ω rate. This illustrates that the proposed method overcomes one of the main problems of neural networks, namely the overconfidence even if the data is out-of-distribution. Of course, the scope of our method does not limit itself to image classification. We can adapt it to other computer vision tasks such as semantic segmentation and instance segmentation.

Another way of improving our method would be to also take into account the partial ignorance of the experts when fusing the mass functions and making a decision. This would require to overcome computational bottlenecks but would open the doors for other decision-making strategies and more optimal fusion rules.

ACKNOWLEDGEMENTS

The first author is supported by the French National Research Agency (ANR) and Region Normandie under grant HAISCoDe. We also thank our colleagues from the Criann who provided us some computation resources with Myria and Austral and by so enabled us to get our results efficiently and fast.

REFERENCES

[1] E. Chzhen, C. Denis, M. Hebiri, and T. Lorieul, "Set-valued classification – overview with a unified framework," *arXiv preprint arXiv:2102.12318*, 2021.

[2] E. Grycko, "Classification with set-valued decision functions," *Information and Classification (O. OPITZ, B. LAUSEN and R. KLAR, eds.)*, pp. 218–224, 1993.

[3] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.

[4] M. Pimentel, D. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[5] S. Xu, Y. Chen, C. Ma, and X. Yue, "Deep evidential fusion network for medical image classification," *International Journal of Approximate Reasoning*, vol. 150, pp. 188–198, 2022.

[6] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in Neural Information Processing Systems*, vol. 31, pp. 3179–3189, 2018.

[7] Z. Guo, Z. Wan, Q. Zhang, X. Zhao, Q. Zhang, L. Kaplan, A. Jøsang, D. Jeong, F. Chen, and J.-H. Cho, "A survey on uncertainty reasoning and quantification in belief theory and its application to deep learning," *Information Fusion*, vol. 101, 2024.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[9] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Ph.D. dissertation*, 2009.

[10] T. Denoeux, "A neural network classifier based on dempster-shafer theory," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, no. 2, pp. 131–150, 2000.

[11] Z. Tong, P. Xu, and T. Denœux, "An evidential classifier based on dempster-shafer theory and deep learning," *Neurocomputing*, vol. 450, pp. 275–293, 2021.

[12] A. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[13] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[14] P. Smets, "The combination of evidence in the transferable belief model," *Transactions on pattern analysis and machine intelligence*, vol. 12, no. 2, pp. 447–458, 1990.

[15] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, pp. 191–234, 1994.

[16] T. Denœux, "Decision-making with belief functions: a review," *International Journal of Approximate Reasoning*, vol. 109, pp. 87–110, 2019.

[17] L. Ma and T. Denœux, "Partial classification in the belief function framework," *Knowledge-Based Systems*, vol. 214, p. 106742, 2021.

[18] R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision-making," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, pp. 183–190, 1988.

[19] M. O'Hagan, "Aggregating template or rule antecedents in real-time expert systems with fuzzy set logic," in *Twenty-Second Asilomar Conference on Signals, Systems and Computers*, vol. 2, 1988, pp. 681–689.

[20] L. Hurwicz, "The generalized bayes minimax principle: a criterion for decision making under uncertainty," *cowles Commission Discussion Paper Statistics*, vol. 355, 1951.

[21] J.-Y. Jaffray, "Linear utility theory for belief functions," *Operations Research Letters*, vol. 8, no. 2, pp. 107–112, 1989.

[22] T. Baldacchino, E. J. Cross, K. Worden, and J. Rowson, "Variational bayesian mixture of experts models and sensitivity analysis for nonlinear dynamical systems," *Mechanical Systems and Signal Processing*, vol. 66-67, pp. 178–200, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327015002307

[23] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[25] S. Xie, R. Girshick, P. Doll´ar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5987–5995, 2017.

[26] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Technical report, University of Toronto, Toronto, Ontario, Tech. Rep. 0, 2009, backup Publisher: University of Toronto. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[27] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. [Online]. Available: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf

[28] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ense," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, and R. Ferg, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2

[29] O. Laurent, A. Lafage, E. Tartaglione, G. Daniel, J.-M. Martinez, A. Bursuc, and G. Franchi, "Packed-ensembles for efficient uncertainty estimation," in *ICLR*, 2023.

Publications de l'auteur

Les publications fournies dans ce document sont indiquées par une étoile.

Journaux internationaux

[Moreau 2025] N. Moreau, S. Valable, C. Jaudet, L. Dessoude, T. Leleu, R. Herault, R. Modzelewski, D. Stefan, J. Thariat, **A. Lechervy**, A. Corroyer-Dulmont, *Early characterization and prediction of glioblastoma and brain metastases treatment efficacy using medical imaging-based radiomics and artificial intelligence algorithms*, Frontiers in Oncology, 2025.

[Dessoude 2025] L. Dessoude, R. Lemaire, R. Andres, T. Leleu, A. G. Leclercq, A. Desmonts, T. Corroller, A. Fara Orou-Guidou, L. Laduree, L. Le Henaff, J. Lacroix, **A. Lechervy**, D. Stefan, A. Corroyer-Dulmont. *Development and routine implementation of deep learning algorithm for automatic brain metastases segmentation on MRI for RANO-BM criteria follow-up*, NeuroImage, 2025.

[Seraphim 2024a] M. Seraphim, **A. Lechervy**, F. Yger, L. Brun and O. Etard. *Automatic Classification of Sleep Stages from EEG Signals Using Riemannian Metrics and Transformer Networks*, SN Computer Science, 2024.

[Lemaire 2024a] R. Lemaire, C. Raboutet, T. Leleu, C. Jaudet, L. Dessoude, F. Missohou, Y. Poirier, P.-Y. Deslandes, **A. Lechervy**, J. Lacroix, I. Moummad, S. Bardet, J. Thariat, D. Stefan, A. Corroyer-Dulmont. *Artificial intelligence solution to accelerate the acquisition of MRI images : Impact on the therapeutic care in oncology in radiology and radiotherapy departments*, Cancer/Radiothérapie, 2024.

[Moummad 2022] I. Moummad, C. Jaudet, **A. Lechervy**, S. Valable, C. Raboutet, Z. Soilihi, J. Thariat, N. Falzone, J. Lacroix, A. Batalla, A. Corroyer-Dulmont, *The Impact of Resampling, Denoising Deep Learning Algorithms on Radiomics in Brain Metastases MRI*, Cancers, Juin 2022.

[Jaudet 2021] C. Jaudet, K. Weyts, **A. Lechervy**, A. Batalla, S. Bardet, A. Corroyer-Dulmont, The Impact of Artificial Intelligence CNN *Based Denoising on FDG PET Radiomics*, Frontiers in Oncology, 2021.

[Vielzeuf 2019] V. Vielzeuf, **A. Lechervy**, S. Pateux, F. Jurie, *Multi-Level Sensor Fusion with Deep Learning*, IEEE Sensors Letters, October 2018.

[Lechervy 2014] **A.Lechervy**, P-H Gosselin, F. Precioso. *Boosted Kernel for Image Categorization*. Multimedia Tools and Applications (MTAP), 2012.

Conférences internationales

[Addad 2025] Y. Addad, **A. Lechervy** and F. Jurie, *CHASE : Channel-Wise and Spatial Attention for Early Exiting in Image Classification*, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2025.

[Deregnaucourt 2025] L. Deregnaucourt, H. Laghmara, **A. Lechervy**, S. Ainouz, *A Conflict-Guided Evidential Fusion for Multimodal Semantic Segmentation*, Winter Conference on Applications of Computer Vision (WACV), 2025.

[Seraphim 2024b] M. Seraphim, **A. Lechervy**, F. Yger, L. Brun, O. Etard, *Structure-Preserving Transformers for Sequences of SPD Matrices*, European Signal Processing Conference (EUSIPCO 2024), 2024.

[Addad 2024] Y. Addad, **A. Lechervy**, F. Jurie, *Balancing Accuracy and Efficiency in Budget-Aware Early-Exiting Neural Networks*, International Conference on Pattern Recognition (ICPR), 2024.

* **[Addad 2023]** Y. Addad, **A. Lechervy**, F. Jurie, *Multi-Exit Resource-Efficient Neural Architecture for Image Classification with Optimized Fusion Block*, International Conference on Computer Vision (ICCV) Workshops, 2023.

* **[Seraphim 2023b]** M. Seraphim, P. Dequidt, **A. Lechervy**, F. Yger, L. Brun, O. Etard, *Temporal Sequences of EEG Covariance Matrices for Automated Sleep Stage Scoring with Attention Mechanisms*, International Conference on Computer Analysis of Images and Patterns (CAIP 2023), 2023.

[Dequidt 2023] P. Dequidt, M. Seraphim, **A. Lechervy**, I.I. Gaez, L. Brun, O. Etard, *Automatic Sleep Stage Classification on EEG Signals Using Time-Frequency Representation*,International Conference on Artificial Intelligence in Medicine (AIME 2023), 2023.

* **[Jha 2022]** P. Jha, G. Dias, **A. Lechervy**, A. Jangra, J. Moreno, S. Pais, S. Sriparna, *Combining Vision and Language Representations for Patch-based*, Identification of Lexico-Semantic Relations. 30th ACM International Conference on Multimedia (ACM MM)., 2022.

[Rane 2021] C. Rane, G. Dias, **A. Lechervy**, A. Ekbal, *Improving Neural Text Style Transfer by Introducing Loss Function Sequentiality*, SIGIR Conference on Research and Development in Information Retrieval, 2021.

[Quéau 2019] Y. Quéau, F. Leporcq, **A. Lechervy**, A. Alfalou, *Learning to classify materials using Mueller imaging polarimetry*, Fourteenth International Conference on Quality Control by Artificial Vision, May 2019.

* **[En 2018b]** S. En, **A. Lechervy**, F. Jurie, *TS-Net : Combining Modality Specific and Common Features for Multimodal Patch Matching*, IEEE International Conference on Image Processing (ICIP), Athen, Greece, October 2018.

[Vielzeuf 2018a] V. Vielzeuf, C. Kervadec, S. Pateux, **A. Lechervy**, F. Jurie, *An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Training Sets*, ICMI (EmotiW), Boulder, Colorado, United States, October 2018.

[Kervadec 2018] C. Kervadec, V. Vielzeuf, S. Pateux, **A. Lechervy**, F. Jurie, *CAKE : Compact and Accurate K-dimensional representation of Emotion*, Image Analysis for Human Facial and Activity Recognition (BMVC Workshop), September 2018.

* **[En 2018a]** S. En, **A. Lechervy**, F. Jurie, *RPNet : an End-to-End Network for Relative Camera Pose Estimation*, 4th International Workshop on Recovering 6D Object Pose (ECCV Workshop), Munich, Germany, September 2018.

[Vielzeuf 2018b] V. Vielzeuf, **A. Lechervy**, S. Pateux, F. Jurie, *CentralNet : a Multilayer Approach for Multimodal Fusion*, Multimodal Learning and Applications (ECCV Workshop), Munich, Germany, September 2018.

* **[Bhattarai 2016]** B. Bhattarai, G. Sharma, **A. Lechervy**, F. Jurie, *A Joint Learning Approach for Cross Domain Age Estimation*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

* **[Negrel 2016]** R. Negrel, **A. Lechervy**, F. Jurie, *MLBoost Revisited : A Faster Metric Learning Algorithm for Identity-Based Face Retrieval*, Proceedings of the British Machine Vision Conference (BMVC), September 2016.

[Negrel 2015] R. Negrel, **A. Lechervy**, F. Jurie, *Boosted Metric Learning for Efficient Identity-Based Face Retrieval*, Proceedings of the British Machine Vision Conference (BMVC), September 2015.

[Lechervy 2012a] **A.Lechervy**, P-H Gosselin, F. Precioso. *Boosting kernel combination for multi-class image categorization*. IEEE International Conference on Image Processing (ICIP), Orlando, Florida, U.S.A, September 2012.

[Lechervy 2012b] **A.Lechervy**, P-H Gosselin, F. Precioso. *Linear kernel combination using boosting*. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belguim, April 2012.

[Lechervy 2010a] **A.Lechervy**, P-H Gosselin, F. Precioso. *Active Boosting for interactive object retrieval*. International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, August 2010.

Conférences nationales

[Lemaire 2024b] R. Lemaire, **A. Lechervy** and Y. Chahir, *Appariement d'images d'oeuvres d'Art avec descriptions textuelles variées*, Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP), Lille, France, Juillet 2024.

[Seraphim 2023a] M. Seraphim, P. Dequidt, **A. Lechervy**, F. Yger , L. Brun and O. Etard, *Appariement d'images d'oeuvres d'Art avec descriptions textuelles variées*, Journée des jeunes chercheurs en vision par ordinateur, Caen, France, May 2023.

[Deregnaucourt 2023] L. Deregnaucourt, H. Laghmara, **A. Lechervy**, S. Ainouz, *Fusion multimodale centrale RGB-polarimétrique pour l'analyse de scènes routières dans des conditions météorologiques dégradées*, Journée des jeunes chercheurs en vision par ordinateur, Caen, France, May 2023.

[Moummad 2021] I. Moummad, C. Jaudet, **A. Lechervy**, S. Valable, C. Raboutet, J. Lacroix, A. Batalla, A. Corroyer-Dulmont, *Impact d'un algorithme de machine learning de resampling sur les radiomiques en IRM, Journée thématique Santé et Sciences du numérique*, Caen, France, Jun 2021.

[Lechervy 2010b] **A.Lechervy**, P-H Gosselin, F. Precioso. *Boosting actif pour la recherche interactive d'images*. Reconnaissance des Formes et Intelligence Artificielle (RFIA2010), Caen, France, Jan. 2010.

Chapitre de livre

* **[Deregnaucourta 2023]** L. Deregnaucourt, **A. Lechervy**, H. Laghmara, S. Ainouz, *An Evidential Deep Network Based on Dempster-Shafer Theory for Large Dataset*, Advances and Applications of DSmT for Information Fusion. Collected Works, Volume 5, October 2023.

Thèse

— **A. Lechervy**, *Apprentissage interactif et multi-classes pour la détection de concepts sémantiques dans des données multimédia*, In Université de Cergy-Pontoise. Cergy, France, Dec. 2012.

Bibliographie

[Addad 2023] Youva Addad, Alexis Lechervy et Frédéric Jurie. *Multi-Exit Resource-Efficient Neural Architecture for Image Classification with Optimized Fusion Block*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pages 1486–1491, Paris, France, Octobre 2023. (Cité en pages 24, 25, 26, 48 et 140.)

[Addad 2024] Youva Addad, Alexis Lechervy et Frédéric Jurie. *Balancing Accuracy and Efficiency in Budget-Aware Early-Exiting Neural Networks*. In International Conference on Pattern Recognition (ICPR), 2024. (Cité en pages 24, 26, 48 et 140.)

[Addad 2025] Y. Addad, A. Lechervy et F. Jurie. *CHASE : Channel-wise and Spatial Attention forEarly Exiting in Image Classification*. In International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE/CVF, 2025. (Cité en pages 24, 26, 48 et 140.)

[Agarwal 2022] Chirag Agarwal, Daniel D'souza et Sara Hooker. *Estimating Example Difficulty Using Variance of Gradients*. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10358–10368, New Orleans, LA, USA, Juin 2022. IEEE. (Cité en page 24.)

[Aguilera 2017] Cristhian A Aguilera, Angel D Sappa, Cristhian Aguilera et Ricardo Toledo. *Cross-Spectral Local Descriptors via Quadruplet Network*. Sensors, vol. 17, no. 4, page 873, 2017. (Cité en page 38.)

[Andrychowicz 2016] Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford et Nando de Freitas. *Learning to Learn by Gradient Descent by Gradient Descent*. In Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc., 2016. (Cité en page 17.)

[Antonelli 2022] Simone Antonelli, Danilo Avola, Luigi Cinque, Donato Crisostomi, Gian Luca Foresti, Fabio Galasso, Marco Raoul Marini, Alessio Mecca et Daniele Pannone. *Few-Shot Object Detection : A Survey*. ACM Comput. Surv., vol. 54, no. 11s, pages 242 :1–242 :37, Septembre 2022. (Cité en page 11.)

[Arevalo 2017] John Arevalo, Thamar Solorio, Manuel Montes-y-Gómez et Fabio A González. *Gated Multimodal Units for Information Fusion*. In ICLR Worshop, 2017. (Cité en page 39.)

[Arth 2015] Clemens Arth, Christian Pirchheim, Jonathan Ventura, Dieter Schmalstieg et Vincent Lepetit. *Instant Outdoor Localization and SLAM Initialization from 2.5D Maps*. IEEE Transactions on Visualization and Computer Graphics, vol. 21, no. 11, pages 1309–1318, Novembre 2015. (Cité en page 36.)

[Ashraf 2021] Hina Ashraf, Yoonsang Jeong et Chong Hyun Lee. *Underwater Ambient-Noise Removing GAN Based on Magnitude and Phase Spectra*. IEEE Access, vol. 9, pages 24513–24530, 2021. (Cité en page 12.)

[Assran 2023] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun et Nicolas Ballas. *Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15619–15629, 2023. (Cité en pages 14 et 15.)

[Atrey 2010] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik et Mohan S. Kankanhalli. *Multimodal Fusion for Multimedia Analysis : A Survey*. Multimedia Systems, vol. 16, no. 6, pages 345–379, Novembre 2010. (Cité en page 31.)

[Ba 2014] Jimmy Ba et Rich Caruana. *Do Deep Nets Really Need to Be Deep ?* In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence et K. Q. Weinberger, editeurs, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014. (Cité en page 23.)

[Bachman 2019] Philip Bachman, R Devon Hjelm et William Buchwalter. *Learning Representations by Maximizing Mutual Information Across Views*. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. (Cité en page 14.)

[Bajpai 2025] Divya Jyoti Bajpai et Manjesh Kumar Hanawal. *A Survey of Early Exit Deep Neural Networks in NLP*, Janvier 2025. (Cité en page 25.)

[Bakhtiarnia 2021] Arian Bakhtiarnia. *Multi-Exit Vision Transformer for Dynamic Inference*. In BMVC, Online, 2021. (Cité en page 26.)

[Baltrušaitis 2018] Tadas Baltrušaitis, Chaitanya Ahuja et Louis-Philippe Morency. *Multimodal Machine Learning : A Survey and Taxonomy*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pages 423–443, 2018. (Cité en page 31.)

[Bardes 2022] Adrien Bardes, Jean Ponce et Yann Lecun. *VICReg : Variance-Invariance-Covariance Regularization For Self-Supervised Learning*. In ICLR 2022 - International Conference on Learning Representations, Avril 2022. (Cité en page 15.)

[Bardes 2024a] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran et Nicolas Ballas. *Revisiting Feature Prediction for Learning Visual Representations from Video*, Fèvrier 2024. (Cité en page 15.)

[Bardes 2024b] Adrien Bardes, Jean Ponce et Yann LeCun. *VICRegL : Self-Supervised Learning of Local Visual Features*. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, pages 8799–8810, Red Hook, NY, USA, Avril 2024. Curran Associates Inc. (Cité en page 15.)

[Bay 2006] Herbert Bay, Tinne Tuytelaars et Luc Van Gool. *Surf : Speeded up Robust Features*. In European Conference on Computer Vision, pages 404–417, 2006. (Cité en pages 36 et 37.)

[Benaim 2018] Sagie Benaim et Lior Wolf. *One-Shot Unsupervised Cross Domain Translation*. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. (Cité en page 15.)

[Bendou 2022] Yassir Bendou, Yuqing Hu, Raphael Lafargue, Giulia Lioi, Bastien Pasdeloup, Stéphane Pateux et Vincent Gripon. *Easy-Ensemble Augmented-Shot-Y-Shaped Learning : State-of-The-Art Few-Shot Classification with Simple Components*. Journal of Imaging, vol. 8, no. 7, page 179, 2022. (Cité en page 15.)

[Bengio 2013] Yoshua Bengio, Aaron Courville et Pascal Vincent. *Representation Learning : A Review and New Perspectives*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pages 1798–1828, Août 2013. (Cité en page 13.)

[Benitez-Quiroz 2016] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan et Aleix M. Martinez. *EmotioNet : An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5562–5570, Juin 2016. (Cité en page 40.)

[Besnier 2021] Victor Besnier, David Picard et Alexandre Briot. *Learning Uncertainty for Safety-Oriented Semantic Segmentation in Autonomous Driving*. In 2021 IEEE International Conference on Image Processing (ICIP), pages 3353–3357, Septembre 2021. (Cité en page 42.)

[Bhattarai 2014] Binod Bhattarai, Gaurav Sharma, Frederic Jurie et Patrick Pérez. *Some Faces Are More Equal than Others : Hierarchical Organization for Accurate and Efficient Large-Scale Identity-Based Face Retrieval*. In European Conference on Computer Vision (ECCV) Workshops, pages 1–13, 2014. (Cité en page 20.)

[Bhattarai 2016] Binod Bhattarai, Gaurav Sharma, Alexis Lechervy et Frédéric Jurie. *A Joint Learning Approach for Cross Domain Age Estimation*. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, Mars 2016. (Cité en pages 17, 18, 20 et 141.)

[Bhowmik 2014] Neelanjan Bhowmik, Ricardo González V., Valérie Gouet-Brunet, Hélio Pedrini et Gabriel Bloch. *Efficient Fusion of Multidimensional Descriptors for Image Retrieval*. In 2014 IEEE International Conference on Image Processing (ICIP), pages 5766–5770, Octobre 2014. (Cité en page 31.)

[Bi 2011] Jinbo Bi, Dijia Wu, Le Lu, Meizhu Liu, Yimo Tao et Matthias Wolf. *AdaBoost on Low-Rank PSD Matrices for Metric Learning*. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On, pages 2617–2624. IEEE, 2011. (Cité en page 20.)

[Boizard 2025] Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei et Pierre Colombo. *EuroBERT : Scaling Multilingual Encoders for European Languages*, Mars 2025. (Cité en page 16.)

[Bouchard 2020] Maude Bouchard, Jean-Marc Lina, Pierre-Olivier Gaudreault, Jonathan Dubé, Nadia Gosselin et Julie Carrier. *EEG Connectivity across Sleep Cycles and Age*. Sleep, vol. 43, no. 3, page zsz236, Mars 2020. (Cité en page 34.)

[Bowles 2018] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw et Daniel Rueckert. *GAN Augmentation : Augmenting Training Data Using Generative Adversarial Networks*, Octobre 2018. (Cité en page 12.)

[Brejcha 2017] Jan Brejcha et Martin Čadík. *State-of-the-Art in Visual Geo-Localization*. Pattern Anal. Appl., vol. 20, no. 3, pages 613–637, Août 2017. (Cité en page 36.)

[Cai 2023] Gan Cai, Yu Zhu, Yue Wu, Xiaoben Jiang, Jiongyao Ye et Dawei Yang. *A Multimodal Transformer to Fuse Images and Metadata for Skin Disease Classification*. The Visual Computer, vol. 39, no. 7, pages 2781–2793, Juillet 2023. (Cité en page 43.)

[Cangea 2017] Catalina Cangea, Petar Velickovic et Pietro Liò. *XFlow : 1D-2D Cross-modal Deep Neural Networks for Audiovisual Classification*. CoRR, vol. abs/1709.00572, 2017. (Cité en page 39.)

[Cao 2023] Weipeng Cao, Yuhao Wu, Yixuan Sun, Haigang Zhang, Jin Ren, Dujuan Gu et Xingkai Wang. *A Review on Multimodal Zero-shot Learning*. WIREs Data Mining and Knowledge Discovery, vol. 13, no. 2, page e1488, Mars 2023. (Cité en page 16.)

[Caron 2020] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski et Armand Joulin. *Unsupervised Learning of Visual Features by Contrasting Cluster Assignments*. In Advances in Neural Information Processing Systems, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. (Cité en pages 14 et 15.)

[Caron 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski et Armand Joulin. *Emerging Properties in Self-Supervised Vision Transformers*. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9650–9660, 2021. (Cité en pages 14 et 15.)

[Castro 2018] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid et Karteek Alahari. *End-to-End Incremental Learning*. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu et Yair Weiss, editeurs, Computer Vision – ECCV 2018, pages 241–257, Cham, 2018. Springer International Publishing. (Cité en page 22.)

[Chen 2013] Ke Chen, Shaogang Gong, Tao Xiang et Chen Change Loy. *Cumulative Attribute Space for Age and Crowd Density Estimation*. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, pages 2467–2474, Juin 2013. (Cité en page 18.)

[Chen 2017] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh et Louis-Philippe Morency. *Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning*. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pages 163–171. ACM, 2017. (Cité en page 39.)

[Chen 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi et Geoffrey Hinton. *A Simple Framework for Contrastive Learning of Visual Representations*. In Proceedings of the 37th International

Conference on Machine Learning, pages 1597–1607. PMLR, Novembre 2020. (Cité en pages 14 et 15.)

[Chen 2020b] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng et Jingjing Liu. *UNITER : UNiversal Image-TExt Representation Learning*. In Andrea Vedaldi, Horst Bischof, Thomas Brox et Jan-Michael Frahm, editeurs, European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, pages 104–120, Cham, 2020. Springer International Publishing. (Cité en page 44.)

[Chen 2021] Xinlei Chen et Kaiming He. *Exploring Simple Siamese Representation Learning*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15750–15758, 2021. (Cité en pages 14 et 15.)

[Chen 2022a] Mengzhao Chen, Mingbao Lin, Ke Li, Yunhang Shen, Yongjian Wu, Fei Chao et Rongrong Ji. *CF-ViT : A General Coarse-to-Fine Method for Vision Transformer*. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37. arXiv, Novembre 2022. (Cité en page 24.)

[Chen 2022b] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan et Zicheng Liu. *Mobile-Former : Bridging MobileNet and Transformer*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5270–5279, Juin 2022. (Cité en page 26.)

[Chen 2023] Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Jeff Z. Pan, Yuan He, Wen Zhang, Ian Horrocks et Huajun Chen. *Zero-Shot and Few-Shot Learning With Knowledge Graphs : A Comprehensive Survey*. Proceedings of the IEEE, vol. 111, no. 6, pages 653–685, Juin 2023. (Cité en pages 11 et 16.)

[Cheng 2017] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao et Kaiqi Huang. *Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation*. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1475–1483, Juillet 2017. (Cité en page 32.)

[Chou 2020] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei et Da-Cheng Juan. *Remix : Rebalanced Mixup*. In Adrien Bartoli et Andrea Fusiello, editeurs, Computer Vision – ECCV 2020 Workshops, pages 95–110, Cham, 2020. Springer International Publishing. (Cité en page 12.)

[Cui 2023] Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A. Coburn, Keith T. Wilson, Bennett A. Landman et Yuankai Huo. *Deep Multimodal Fusion of Image and Non-Image Data in Disease Diagnosis and Prognosis : A Review*. Progress in Biomedical Engineering, vol. 5, no. 2, page 022001, Avril 2023. (Cité en page 31.)

[Dalal 2005] N. Dalal et B. Triggs. *Histograms of Oriented Gradients for Human Detection*. In Cordelia Schmid, Stefano Soatto et Carlo Tomasi, editeurs, IEEE International Conference on Computer Vision and Pattern Recognition, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, Juin 2005. (Cité en page 36.)

[Davis 2007] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra et Inderjit S. Dhillon. *Information-Theoretic Metric Learning*. In Proceedings of the 24th International Conference on Machine Learning, ICML '07, pages 209–216, New York, NY, USA, 2007. ACM. (Cité en page 20.)

[Dechesne 2017] Clément Dechesne, Clément Mallet, Arnaud Le Bris et Valérie Gouet-Brunet. *Semantic Segmentation of Forest Stands of Pure Species Combining Airborne Lidar Data and Very High Resolution Multispectral Imagery*. ISPRS Journal of Photogrammetry and Remote Sensing, vol. 126, pages 129–145, Avril 2017. (Cité en page 31.)

[Dempster 2008] Arthur P. Dempster. *Upper and Lower Probabilities Induced by a Multivalued Mapping*. In Roland R. Yager et Liping Liu, editeurs, Classic Works of the Dempster-Shafer Theory of Belief Functions, pages 57–72. Springer, Berlin, Heidelberg, 2008. (Cité en page 42.)

[Deng 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei. *ImageNet : A Large-Scale Hierarchical Image Database*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 248–255, Juin 2009. (Cité en page 42.)

[Deng 2012] Li Deng. *The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]*. IEEE Signal Processing Magazine, vol. 29, no. 6, pages 141–142, Novembre 2012. (Cité en page 39.)

[Denoeux 2000] T. Denoeux. *A Neural Network Classifier Based on Dempster-Shafer Theory*. IEEE Transactions on Systems, Man, and Cybernetics - Part A : Systems and Humans, vol. 30, no. 2, pages 131–150, 2000. (Cité en pages 32 et 42.)

[Dequidt 2023] Paul Dequidt, Mathieu Seraphim, Alexis Lechervy, Ivan Igor Gaez, Luc Brun et Olivier Etard. *Automatic Sleep Stage Classification on EEG Signals Using Time-Frequency Representation*. In Jose M. Juarez, Mar Marcos, Gregor Stiglic et Allan Tucker, editeurs, Artificial Intelligence in Medicine, pages 250–259, Cham, 2023. Springer Nature Switzerland. (Cité en pages 33, 34, 43 et 140.)

[Deregnaucourt 2023] L Deregnaucourt, H Laghmara, A Lechervy et S Ainouz. *Fusion Multimodale Centrale RGB-polarimétrique Pour l'analyse de Scènes Routières Dans Des Conditions Météorologiques Dégradées*. In 19è Journées Francophones Des Jeunes Chercheurs En Vision Par Ordinateur (ORASIS 2023), Carqueiranne, France, Mai 2023. (Cité en pages 41, 42, 44, 48 et 142.)

[Deregnaucourt 2025] L. Deregnaucourt, H. Laghmara, A. Lechervy et S. Ainouz. *A Conflict-Guided Evidential Fusion for Multimodal Semantic Segmentation*. In Winter Conference on Applications of Computer Vision (WACV). IEEE/CVF, 2025. (Cité en pages 41, 42, 44, 48 et 140.)

[Deregnaucourta 2023] Lucas Deregnaucourta, Alexis Lechervyb, Hind Laghmaraa et Samia Ainouza. *An Evidential Deep Network Based on Dempster-Shafer Theory for Large Dataset*. Advances and Applications of DSmT for Information Fusion, vol. 5, page 907, 2023. (Cité en pages 41, 42, 44, 48 et 142.)

[Dessoude 2025] Loïse Dessoude, Raphaëlle Lemaire, Romain Andres, Thomas Leleu, Alexandre G. Leclercq, Alexis Desmonts, Typhaine Corroller, Amirath Fara Orou-Guidou, Luca Laduree, Loic Le Henaff, Joëlle Lacroix, Alexis Lechervy, Dinu Stefan et Aurélien Corroyer-Dulmont. *Development and Routine Implementation of Deep Learning Algorithm for Automatic Brain Metastases Segmentation on MRI for RANO-BM Criteria Follow-Up*. NeuroImage, vol. 306, page 121002, Fèvrier 2025. (Cité en page 139.)

[Devlin 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova. *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT), pages 4171–4186, Minneapolis, Minnesota, Juin 2019. Association for Computational Linguistics. (Cité en pages 14 et 16.)

[DeVries 2017] Terrance DeVries et Graham W. Taylor. *Improved Regularization of Convolutional Neural Networks with Cutout*, Novembre 2017. (Cité en page 12.)

[Dhall 2018] Abhinav Dhall, Amanjot Kaur, Roland Goecke et Tom Gedeon. *EmotiW 2018 : Audio-Video, Student Engagement and Group-Level Affect Prediction*. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18, pages 653–656, New York, NY, USA, Octobre 2018. Association for Computing Machinery. (Cité en page 40.)

[Ding 2021] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding et Jian Sun. *Repvgg : Making Vgg-Style Convnets Great Again*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13733–13742, 2021. (Cité en page 23.)

[Diykh 2016] Mohammed Diykh, Yan Li et Peng Wen. *EEG Sleep Stages Classification Based on Time Domain Features and Structural Graph Similarity*. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 24, no. 11, pages 1159–1168, Novembre 2016. (Cité en page 33.)

[Dosovitskiy 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit et Neil Houlsby. *An Image Is Worth 16x16 Words : Transformers for Image Recognition at Scale*, Juin 2021. (Cité en pages 16 et 26.)

[Douze 2018] Matthijs Douze, Arthur Szlam, Bharath Hariharan et Hervé Jégou. *Low-Shot Learning with Large-Scale Diffusion*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3349–3358, 2018. (Cité en page 12.)

[Dufour 2025] Nicolas Dufour, David Picard, Vicky Kalogeiton et Loic Landrieu. *Around the World in 80 Timesteps : A Generative Approach to Global Visual Geolocation*. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. (Cité en page 36.)

[Ehatisham-Ul-Haq 2019] Muhammad Ehatisham-Ul-Haq, Ali Javed, Muhammad Awais Azam, Hafiz M. A. Malik, Aun Irtaza, Ik Hyun Lee et Muhammad Tariq Mahmood. *Robust Human Activity*

Recognition Using Multimodal Feature-Level Fusion. IEEE Access, vol. 7, pages 60736–60751, 2019. (Cité en page 31.)

[Elsken 2019] Thomas Elsken, Jan Hendrik Metzen et Frank Hutter. *Neural Architecture Search : A Survey*. arXiv :1808.05377 [cs, stat], vol. 20, no. 1, pages 1997–2017, Avril 2019. (Cité en page 23.)

[En 2018a] Sovann En, Alexis Lechervy et Frédéric Jurie. *RPNet : An End-to-End Network for Relative Camera Pose Estimation*. In European Conference on Computer Vision Workshops, Munich, Germany, Septembre 2018. (Cité en pages 34, 37, 43 et 141.)

[En 2018b] Sovann En, Alexis Lechervy et Frédéric Jurie. *TS-Net : Combining Modality Specific and Common Features for Multimodal Patch Matching*. In ICIP, Athens, Greece, Octobre 2018. IEEE International Conference on Image Processing. (Cité en pages 34, 37, 38, 43 et 140.)

[Fahes 2023] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez et Raoul de Charette. *PØDA : Prompt-driven Zero-shot Domain Adaptation*. In International Conference on Computer Vision (ICCV), Octobre 2023. (Cité en page 22.)

[Fan 2016] Yin Fan, Xiangju Lu, Dian Li et Yuanliu Liu. *Video-Based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks*. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI '16, pages 445–450, New York, NY, USA, Octobre 2016. Association for Computing Machinery. (Cité en page 40.)

[Fayyad 2020] Jamil Fayyad, Mohammad A. Jaradat, Dominique Gruyer et Homayoun Najjaran. *Deep Learning Sensor Fusion for Autonomous Vehicle Perception and Localization : A Review*. Sensors, vol. 20, no. 15, page 4220, Janvier 2020. (Cité en page 41.)

[Feng 2021] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura et Eduard Hovy. *A Survey of Data Augmentation Approaches for NLP*. In Chengqing Zong, Fei Xia, Wenjie Li et Roberto Navigli, editeurs, Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021, pages 968–988, Online, Août 2021. Association for Computational Linguistics. (Cité en page 12.)

[Finn 2017] Chelsea Finn, Pieter Abbeel et Sergey Levine. *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks*. In Proceedings of the 34th International Conference on Machine Learning, pages 1126–1135. PMLR, Juillet 2017. (Cité en page 17.)

[Fischler 1981] Martin A. Fischler et Robert C. Bolles. *Random Sample Consensus : A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Commun. ACM, vol. 24, no. 6, pages 381–395, Juin 1981. (Cité en page 37.)

[French 1999] Robert M. French. *Catastrophic Forgetting in Connectionist Networks*. Trends in Cognitive Sciences, vol. 3, no. 4, pages 128–135, Avril 1999. (Cité en page 22.)

[French 2020] Geoff French, Avital Oliver et Tim Salimans. *Milking CowMask for Semi-Supervised Image Classification*, Juin 2020. (Cité en page 12.)

[Freund 2003] Y. Freund, R. Yyer, R. E. Schapire et Y. Singer. *An Efficient Boosting Algorithm for Combining Preferences*. Journal on Machine Learning Research, vol. 4, pages 933–969, Novembre 2003. (Cité en pages 6 et 20.)

[Frid-Adar 2018] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger et Hayit Greenspan. *GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification*. Neurocomputing, vol. 321, pages 321–331, Décembre 2018. (Cité en page 12.)

[Gadre 2023] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar et Ludwig Schmidt. *DataComp : In Search of the next Generation of Multimodal Datasets*. In Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, Novembre 2023. (Cité en page 22.)

[Gal 2022] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik et Daniel Cohen-Or. *StyleGAN-NADA : CLIP-guided Domain Adaptation of Image Generators*. ACM Trans. Graph., vol. 41, no. 4, pages 141 :1–141 :13, Juillet 2022. (Cité en page 22.)

[Ganin 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March et Victor Lempitsky. *Domain-Adversarial Training of Neural Networks*. Journal of machine learning research, vol. 17, no. 59, pages 1–35, 2016. (Cité en page 22.)

[Gao 2023a] Minghong Gao. *A Survey on Recent Teacher-student Learning Studies*, Avril 2023. (Cité en page 23.)

[Gao 2023b] Xiangxiang Gao, Yue Liu, Tao Huang et Zhongyu Hou. *PF-BERxiT : Early Exiting for BERT with Parameter-Efficient Fine-Tuning and Flexible Early Exiting Strategy*. Neurocomputing, vol. 558, page 126690, Novembre 2023. (Cité en page 27.)

[Gatys 2016] Leon A. Gatys, Alexander S. Ecker et Matthias Bethge. *Image Style Transfer Using Convolutional Neural Networks*. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2414–2423, Juin 2016. (Cité en page 12.)

[Ge 2021] Songwei Ge, Shlok Mishra, Haohan Wang, Chun-Liang Li et David Jacobs. *Robust Contrastive Learning Using Negative Samples with Diminished Semantics*. In Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21, pages 27356–27368, Red Hook, NY, USA, 2021. Curran Associates Inc. (Cité en page 15.)

[Gers 1999] F.A. Gers, J. Schmidhuber et F. Cummins. *Learning to Forget : Continual Prediction with LSTM*. In 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), volume 2, pages 850–855 vol.2, Septembre 1999. (Cité en page 40.)

[Gharoun 2024] Hassan Gharoun, Fereshteh Momenifar, Fang Chen et Amir Gandomi. *Meta-Learning Approaches for Few-Shot Learning : A Survey of Recent Advances*. ACM Comput. Surv., vol. 56, no. 12, pages 294 :1–294 :41, Juillet 2024. (Cité en page 11.)

[Goodfellow 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio. *Generative Adversarial Nets*. Advances in neural information processing systems, vol. 27, 2014. (Cité en page 12.)

[Gordo 2017] Albert Gordo, Jon Almazán, Jerome Revaud et Diane Larlus. *End-to-End Learning of Deep Visual Representations for Image Retrieval*. International Journal of Computer Vision, vol. 124, no. 2, pages 237–254, Septembre 2017. (Cité en page 36.)

[Graham 2021] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou et Matthijs Douze. *LeViT : A Vision Transformer in ConvNet's Clothing for Faster Inference*. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 12239–12249, Octobre 2021. (Cité en page 26.)

[Grill 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos et Michal Valko. *Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning*. In Advances in Neural Information Processing Systems, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. (Cité en pages 14 et 15.)

[Gu 2017] Zepeng Gu, Bo Lang, Tongyu Yue et Lei Huang. *Learning Joint Multimodal Representation Based on Multi-fusion Deep Neural Networks*. In International Conference on Neural Information Processing, pages 276–285. Springer, 2017. (Cité en page 39.)

[Gu 2023] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp et Philip Torr. *A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models*, Juillet 2023. (Cité en page 16.)

[Guillaumin 2012] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek et Cordelia Schmid. *Face Recognition from Caption-Based Supervision*. International Journal of Computer Vision, vol. 96, no. 1, pages 64–82, 2012. (Cité en page 20.)

[Güneş 2010] Salih Güneş, Kemal Polat et Şebnem Yosunkaya. *Efficient Sleep Stage Recognition System Based on EEG Signal Using* K*-Means Clustering Based Feature Weighting*. Expert Systems with Applications, vol. 37, no. 12, pages 7922–7928, Décembre 2010. (Cité en page 33.)

[Guo 2014] Guodong Guo et Chao Zhang. *A Study on Cross-Population Age Estimation*. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 4257–4263, Juin 2014. (Cité en pages 18 et 19.)

[Guo 2024] Z. Guo, Z. Wan, Q. Zhang, X. Zhao, Q. Zhang, L.M. Kaplan, A. Jøsang, D.H. Jeong, F. Chen et J.-H. Cho. *A Survey on Uncertainty Reasoning and Quantification in Belief Theory and Its Application to Deep Learning*. Information Fusion, vol. 101, 2024. (Cité en page 32.)

[Gupta 2021] Abhishek Gupta, Alagan Anpalagan, Ling Guan et Ahmed Shaharyar Khwaja. *Deep Learning for Object Detection and Scene Perception in Self-Driving Cars : Survey, Challenges, and Open Issues*. Array, vol. 10, page 100057, Juillet 2021. (Cité en page 41.)

[Gutmann 2010] Michael Gutmann et Aapo Hyvärinen. *Noise-Contrastive Estimation : A New Estimation Principle for Unnormalized Statistical Models*. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 297–304. JMLR Workshop and Conference Proceedings, Mars 2010. (Cité en page 15.)

[Han 2013] Hu Han, Charles Otto et Anil K. Jain. *Age Estimation from Face Images : Human vs. Machine Performance*. In 2013 International Conference on Biometrics (ICB), pages 1–8, Juin 2013. (Cité en page 18.)

[Han 2016] Song Han, Huizi Mao et William J. Dally. *Deep Compression : Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding*. In Yoshua Bengio et Yann LeCun, editeurs, 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. (Cité en page 23.)

[Han 2020] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu et Chang Xu. *GhostNet : More Features From Cheap Operations*. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Juin 2020. (Cité en page 23.)

[Han 2022a] Y. Han, G. Huang, S. Song, L. Yang, H. Wang et Y. Wang. *Dynamic Neural Networks : A Survey*. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 44, no. 11, pages 7436–7456, Novembre 2022. (Cité en page 23.)

[Han 2022b] Yizeng Han, Yifan Pu, Zihang Lai, Chaofei Wang, Shiji Song, Junfeng Cao, Wenhui Huang, Chao Deng et Gao Huang. *Learning to Weight Samples for Dynamic Early-Exiting Networks*. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella et Tal Hassner, editeurs, Computer Vision – ECCV 2022, volume 13671, pages 362–378, Cham, Novembre 2022. Springer Nature Switzerland. (Cité en pages 23, 24 et 25.)

[Han 2023] Yizeng Han, Dongchen Han, Zeyu Liu, Yulin Wang, Xuran Pan, Yifan Pu, Chao Deng, Junlan Feng, Shiji Song et Gao Huang. *Dynamic Perceiver for Efficient Visual Recognition*. In ICCV 2023. arXiv, Août 2023. (Cité en pages 23, 24 et 25.)

[Hartley 1997] R.I. Hartley. *In Defense of the Eight-Point Algorithm*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 6, pages 580–593, Juin 1997. (Cité en page 37.)

[Hays 2008] James Hays et Alexei A. Efros. *IM2GPS : Estimating Geographic Information from a Single Image*. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, Juin 2008. (Cité en page 36.)

[He 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun. *Deep Residual Learning for Image Recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Décembre 2015. (Cité en page 16.)

[He 2019] Yang He, Ping Liu, Ziwei Wang, Zhilan Hu et Yi Yang. *Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Acceleration*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Juin 2019. (Cité en page 23.)

[He 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie et Ross Girshick. *Momentum Contrast for Unsupervised Visual Representation Learning*. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9726–9735, Juin 2020. (Cité en page 15.)

[Hedderich 2021] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen et Dietrich Klakow. *A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios*. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, pages 2545–2568, Online, Juin 2021. Association for Computational Linguistics. (Cité en page 23.)

[Hendrycks 2020] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer et Balaji Lakshminarayanan. *AugMix : A Simple Data Processing Method to Improve Robustness and Uncertainty*, Fèvrier 2020. (Cité en page 12.)

[Herrera 2011] L. J. Herrera, A. M. Mora, C. Fernandes, D. Migotina, A. Guillén et A. C. Rosa. *Symbolic Representation of the EEG for Sleep Stage Classification*. In 2011 11th International Conference on Intelligent Systems Design and Applications, pages 253–258, Novembre 2011. (Cité en page 33.)

[Hinton 2015] Geoffrey Hinton, Oriol Vinyals et Jeffrey Dean. *Distilling the Knowledge in a Neural Network*. In NIPS Deep Learning and Representation Learning Workshop, 2015. (Cité en page 23.)

[Houlsby 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan et Sylvain Gelly. *Parameter-Efficient Transfer Learning for NLP*. In Proceedings of the 36th International Conference on Machine Learning, pages 2790–2799. PMLR, Mai 2019. (Cité en page 17.)

[Howard 2017] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto et Hartwig Adam. *MobileNets : Efficient Convolutional Neural Networks for Mobile Vision Applications*, Avril 2017. (Cité en page 23.)

[Howard 2019] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le et Hartwig Adam. *Searching for MobileNetV3*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Octobre 2019. (Cité en page 23.)

[Hu 2017] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao et Yurong Chen. *Learning Supervised Scoring Ensemble for Emotion Recognition in the Wild*. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI '17, pages 553–560, New York, NY, USA, 2017. ACM. (Cité en page 40.)

[Hu 2018] Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu et Maosong Sun. *Few-Shot Charge Prediction with Discriminative Legal Attributes*. In Emily M. Bender, Leon Derczynski et Pierre Isabelle, editeurs, Proceedings of the 27th International Conference on Computational Linguistics, pages 487–498, Santa Fe, New Mexico, USA, Août 2018. Association for Computational Linguistics. (Cité en pages 15 et 16.)

[Hu 2019] Jie Hu, Li Shen, Samuel Albanie, Gang Sun et Enhua Wu. *Squeeze-and-Excitation Networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), page 10, Mai 2019. (Cité en page 26.)

[Hu 2020] Di Hu, Lichao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou et Xiao Xiang Zhu. *Ambient Sound Helps : Audiovisual Crowd Counting in Extreme Conditions*, Mai 2020. (Cité en page 27.)

[Hu 2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang et Weizhu Chen. *LoRA : Low-Rank Adaptation of Large Language Models*, Octobre 2021. (Cité en page 17.)

[Hu 2024] Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen et Qiu Guan. *A Comprehensive Survey on Contrastive Learning*. Neurocomputing, vol. 610, page 128645, Décembre 2024. (Cité en page 14.)

[Huang 2007] Gary B. Huang, Manu Ramesh, Tamara Berg et Erik Learned-Miller. *Labeled Faces in the Wild : A Database for Studying Face Recognition in Unconstrained Environments*. Rapport technique 07-49, University of Massachusetts, Amherst, Octobre 2007. (Cité en page 20.)

[Huang 2017] Gao Huang, Zhuang Liu, Laurens van der Maaten et Kilian Q. Weinberger. *Densely Connected Convolutional Networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Juillet 2017. (Cité en page 25.)

[Huang 2018] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten et Kilian Weinberger. *Multi-Scale Dense Networks for Resource Efficient Image Classification*. In International Conference on Learning Representations, Fèvrier 2018. (Cité en pages 23, 24, 25 et 27.)

[Huang 2020] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian et Mingyue Niu. *Multimodal Transformer Fusion for Continuous Emotion Recognition*. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3507–3511, Mai 2020. (Cité en page 32.)

[Hubara 2016] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv et Yoshua Bengio. *Binarized Neural Networks*. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon et R. Garnett,

editeurs, Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc., 2016. (Cité en page 23.)

[Iandola 2016] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally et Kurt Keutzer. *SqueezeNet : AlexNet-level Accuracy with 50x Fewer Parameters and <1MB Model Size*. ArXiv, vol. abs/1602.07360, 2016. (Cité en page 23.)

[Ilhan 2024] Fatih Ilhan, Ka-Ho Chow, Sihao Hu, Tiansheng Huang, Selim Tekin, Wenqi Wei, Yanzhao Wu, Myungjin Lee, Ramana Kompella, Hugo Latapie, Gaowen Liu et Ling Liu. *Adaptive Deep Neural Network Inference Optimization with EENet*. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1362–1371, Janvier 2024. (Cité en page 26.)

[Inoue 2018] Hiroshi Inoue. *Data Augmentation by Pairing Samples for Images Classification*, Avril 2018. (Cité en page 12.)

[Jacob 2018] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam et Dmitry Kalenichenko. *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Juin 2018. (Cité en page 23.)

[Jacob 2020] Pierre Jacob, David Picard, Aymeric Histace et Edouard Klein. *DIABLO : Dictionary-based Attention Block for Deep Metric Learning*. Pattern Recognition Letters, vol. 135, pages 99–105, Juillet 2020. (Cité en page 13.)

[Jadon 2023] Shruti Jadon et Aryan Jadon. *An Overview of Deep Learning Architectures in Few-Shot Learning Domain*, Avril 2023. (Cité en page 11.)

[Jahrer 2008] Michael Jahrer, Michael Grabner et Horst Bischof. *Learned Local Descriptors for Recognition and Matching*. In Computer Vision Winter Workshop, volume 2, 2008. (Cité en page 38.)

[Jaiswal 2021] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee et Fillia Makedon. *A Survey on Contrastive Self-Supervised Learning*. Technologies, vol. 9, no. 1, page 2, Mars 2021. (Cité en page 14.)

[Jang 2023] Taeuk Jang et Xiaoqian Wang. *Difficulty-Based Sampling for Debiased Contrastive Representation Learning*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24039–24048, 2023. (Cité en page 15.)

[Jaudet 2021] Cyril Jaudet, Kathleen Weyts, Alexis Lechervy, Alain Batalla, Stéphane Bardet et Aurélien Corroyer-Dulmont. *The Impact of Artificial Intelligence CNN Based Denoising on FDG PET Radiomics*. Frontiers in Oncology, vol. 11, page 692973, 2021. (Cité en page 139.)

[Jegou 2008] Herve Jegou, Matthijs Douze et Cordelia Schmid. *Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search*. In David Forsyth, Philip Torr et Andrew Zisserman, editeurs, Computer Vision – ECCV 2008, pages 304–317, Berlin, Heidelberg, 2008. Springer. (Cité en page 36.)

[Jha 2022] Prince Jha, Gaël Dias, Alexis Lechervy, José G. Moreno, Anubhav Jangra, Sebastião Pais et Sriparna Saha. *Combining Vision and Language Representations for Patch-based Identification of Lexico-Semantic Relations*. In 30th ACM International Conference on Multimedia (ACM MM 2022), Lisbonne, Portugal, Octobre 2022. (Cité en pages 40, 41 et 140.)

[Jia 2020] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He et Yaoshuai Zhao. *GraphSleepNet : Adaptive Spatial-Temporal Graph Convolutional Networks for Sleep Stage Classification.* In IJCAI, pages 1324–1330, 2020. (Cité en page 33.)

[Jiang 2020] Junguang Jiang, Ximei Wang, Mingsheng Long et Jianmin Wang. *Resource Efficient Domain Adaptation*. In Proceedings of the 28th ACM International Conference on Multimedia, MM '20, pages 2220–2228, New York, NY, USA, Octobre 2020. Association for Computing Machinery. (Cité en page 27.)

[Jiang 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix et William El Sayed. *Mistral 7B*, Octobre 2023. (Cité en page 17.)

[Jousselme 2001] Anne-Laure Jousselme, Dominic Grenier et Éloi Bossé. *A New Distance between Two Bodies of Evidence*. Information Fusion, vol. 2, no. 2, pages 91–101, Juin 2001. (Cité en page 42.)

[Jung 2019] Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang et Changkyu Choi. *Learning to Quantize Deep Networks by Optimizing Quantization Intervals With Task Loss*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Juin 2019. (Cité en page 23.)

[Kalamkar 2023] Shrida Kalamkar et Geetha Mary A. *Multimodal Image Fusion : A Systematic Review*. Decision Analytics Journal, vol. 9, page 100327, Décembre 2023. (Cité en page 31.)

[Kang 2017] Miao Kang, Kefeng Ji, Xiangguang Leng et Zhao Lin. *Contextual Region-Based Convolutional Neural Network with Multilayer Fusion for SAR Ship Detection*. Remote Sensing, vol. 9, no. 8, page 860, 2017. (Cité en page 39.)

[Kaur 2021] Sukhpal Kaur, Himanshu Aggarwal et Rinkle Rani. *MR Image Synthesis Using Generative Adversarial Networks for Parkinson's Disease Classification*. In Poonam Bansal, Meena Tushir, Valentina Emilia Balas et Rajeev Srivastava, editeurs, Proceedings of International Conference on Artificial Intelligence and Applications, pages 317–327, Singapore, 2021. Springer. (Cité en page 12.)

[Kaymak 2019] Çağrı Kaymak et Ayşegül Uçar. *A Brief Survey and an Application of Semantic Image Segmentation for Autonomous Driving*. In Valentina Emilia Balas, Sanjiban Sekhar Roy, Dharmendra Sharma et Pijush Samui, editeurs, Handbook of Deep Learning Applications, pages 161–200. Springer International Publishing, Cham, 2019. (Cité en page 41.)

[Kendall 2015] Alex Kendall, Matthew Grimes et Roberto Cipolla. *Posenet : A Convolutional Network for Real-Time 6-Dof Camera Relocalization*. In Computer Vision (ICCV), 2015 IEEE International Conference On, pages 2938–2946, 2015. (Cité en pages 36 et 37.)

[Kendall 2016] Alex Kendall et Roberto Cipolla. *Modelling Uncertainty in Deep Learning for Camera Relocalization*. In Robotics and Automation (ICRA), 2016 IEEE International Conference On, pages 4762–4769, 2016. (Cité en pages 36 et 37.)

[Kervadec 2018] Corentin Kervadec, Valentin Vielzeuf, Stéphane Pateux, Alexis Lechervy et Frédéric Jurie. *CAKE : Compact and Accurate K-dimensional Representation of Emotion*. In Image Analysis for Human Facial and Activity Recognition (BMVC Workshop), Newcastle, United Kingdom, Septembre 2018. Dr. Zhaojie Ju. (Cité en pages 39, 40 et 141.)

[Kim 2017] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi et Byung Cheol Song. *Multi-Modal Emotion Recognition Using Semi-Supervised Learning and Multiple Neural Networks in the Wild*. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pages 529–535. ACM, 2017. (Cité en page 39.)

[Kim 2020] Jang-Hyun Kim, Wonho Choo et Hyun Oh Song. *Puzzle Mix : Exploiting Saliency and Local Statistics for Optimal Mixup*. In International Conference on Machine Learning, pages 5275–5285. PMLR, 2020. (Cité en page 12.)

[Kim 2021] Jang-Hyun Kim, Wonho Choo, Hosan Jeong et Hyun Oh Song. *Co-Mixup : Saliency Guided Joint Mixup with Supermodular Diversity*, Fèvrier 2021. (Cité en page 12.)

[Kirillov 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár et Ross Girshick. *Segment Anything*. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 3992–4003, Octobre 2023. (Cité en page 27.)

[Kiros 2015] Ryan Kiros, Yukun Zhu, Russ R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba et Sanja Fidler. *Skip-Thought Vectors*. Advances in neural information processing systems, vol. 28, 2015. (Cité en page 14.)

[Knyazev 2018] Boris Knyazev, Roman Shvetsov, Natalia Efremova et Artem Kuharenko. *Leveraging Large Face Recognition Data for Emotion Classification*. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 692–696, Mai 2018. (Cité en page 40.)

[Koestinger 2012] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth et Horst Bischof. *Large Scale Metric Learning from Equivalence Constraints*. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference On, pages 2288–2295. IEEE, 2012. (Cité en page 20.)

[Koley 2012] B. Koley et D. Dey. *An Ensemble System for Automatic Sleep Stage Classification Using Single Channel EEG Signal*. Computers in Biology and Medicine, vol. 42, no. 12, pages 1186–1195, Décembre 2012. (Cité en page 33.)

[Kontras 2024] Konstantinos Kontras, Christos Chatzichristos, Huy Phan, Johan Suykens et Maarten De Vos. *CoRe-Sleep : A Multimodal Fusion Framework for Time Series Robust to Imperfect Modalities*. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, pages 840–849, 2024. (Cité en page 33.)

[Kouris 2022] Alexandros Kouris, Stylianos I. Venieris, Stefanos Laskaridis et Nicholas Lane. *Multi-Exit Semantic Segmentation Networks*. In Computer Vision – ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI, pages 330–349, Berlin, Heidelberg, Octobre 2022. Springer-Verlag. (Cité en page 27.)

[Kumar 2022] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma et Percy Liang. *Fine-Tuning Can Distort Pretrained Features and Underperform Out-of-Distribution*. In International Conference on Learning Representations, 2022. (Cité en page 13.)

[Kupyn 2018] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin et Jiri Matas. *DeblurGAN : Blind Motion Deblurring Using Conditional Adversarial Networks*. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8183–8192, Juin 2018. (Cité en page 12.)

[Kwon 2022] Gihyun Kwon et Jong Chul Ye. *CLIPstyler : Image Style Transfer with a Single Text Condition*. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18041–18050, Juin 2022. (Cité en page 22.)

[Lake 2017] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum et Samuel J. Gershman. *Building Machines That Learn and Think like People*. Behavioral and Brain Sciences, vol. 40, page e253, Janvier 2017. (Cité en page 11.)

[lan 2018] xu lan, Xiatian Zhu et Shaogang Gong. *Knowledge Distillation by On-the-Fly Native Ensemble*. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi et R. Garnett, editeurs, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. (Cité en page 23.)

[Langheim 2011] Frederick J. P. Langheim, Michael Murphy, Brady A. Riedner et Giulio Tononi. *Functional Connectivity in Slow-Wave Sleep : Identification of Synchronous Cortical Activity during Wakefulness and Sleep Using Time Series Analysis of Electroencephalographic Data*. Journal of Sleep Research, vol. 20, no. 4, pages 496–505, 2011. (Cité en page 34.)

[Laskaridis 2021] Stefanos Laskaridis, Alexandros Kouris et Nicholas D. Lane. *Adaptive Inference through Early-Exit Networks : Design, Challenges and Directions*. In Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning, EMDL'21, pages 1–6, New York, NY, USA, 2021. Association for Computing Machinery. (Cité en page 23.)

[Lechervy 2010a] Alexis Lechervy, Philippe-Henri Gosselin et Frédéric Precioso. *Active Boosting for Interactive Object Retrieval.* In International Conference on Pattern Recognition, page 1, Istanbul, Turkey, Août 2010. (Cité en pages 4 et 141.)

[Lechervy 2010b] Alexis Lechervy, Philippe-Henri Gosselin et Frédéric Precioso. *Boosting actif pour la recherche interactive d'images*. In Reconnaissance des Formes et Intelligence Artificielle (RFIA2010), page 1, Caen, France, Janvier 2010. (Cité en pages 4 et 142.)

[Lechervy 2012a] Alexis Lechervy, Philippe-Henri Gosselin et Frédéric Precioso. *Boosting Kernel Combination for Multi-Class Image Categorization*. In 2012 IEEE International Conference on Image Processing (ICIP), page 4, Orlando, United States, Septembre 2012. IEEE. (Cité en pages 6 et 141.)

[Lechervy 2012b] Alexis Lechervy, Philippe-Henri Gosselin et Frédéric Precioso. *Linear Kernel Combination Using Boosting*. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, page 6, Bruges, Belgium, Avril 2012. (Cité en pages 6 et 141.)

[Lechervy 2014] Alexis Lechervy, Philippe-Henri Gosselin et Frédéric Precioso. *Boosted Kernel for Image Categorization*. Multimedia Tools and Applications, vol. 69, no. 2, pages 471–490, Mars 2014. (Cité en pages 6 et 139.)

[Ledaguenel 2024] Arthur Ledaguenel, Céline Hudelot et Mostepha Khouadjia. *Improving Neural-based Classification with Logical Background Knowledge*. In Workshop on Composite AI (CompAI), ECAI, Santiago de Compostela, Spain, Fèvrier 2024. arXiv. (Cité en pages 16 et 49.)

[Lee 2013] Dong-Hyun Lee. *Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks*. In Workshop on Challenges in Representation Learning, ICML, volume 3, page 896. Atlanta, 2013. (Cité en page 12.)

[Lee 2020] Jin-Ha Lee, Muhammad Zaigham Zaheer, Marcella Astrid et Seung-Ik Lee. *SmoothMix : A Simple Yet Effective Data Augmentation to Train Robust Classifiers*. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 3264–3274, Juin 2020. (Cité en page 12.)

[Lee 2023] Hojung Lee et Jong-Seok Lee. *Rethinking Online Knowledge Distillation with Multi-exits*. In Lei Wang, Juergen Gall, Tat-Jun Chin, Imari Sato et Rama Chellappa, editeurs, Computer Vision – ACCV 2022, volume 13846, pages 408–424. Springer Nature Switzerland, 2023. (Cité en page 24.)

[Lemaire 2024a] R. Lemaire, C. Raboutet, T. Leleu, C. Jaudet, L. Dessoude, F. Missohou, Y. Poirier, P. Y. Deslandes, A. Lechervy, J. Lacroix, I. Moummad, S. Bardet, J. Thariat, D. Stefan et A. Corroyer-Dulmont. *Artificial Intelligence Solution to Accelerate the Acquisition of MRI Images : Impact on the Therapeutic Care in Oncology in Radiology and Radiotherapy Departments*. Cancer/Radiothérapie, vol. 28, no. 3, pages 251–257, Juin 2024. (Cité en page 139.)

[Lemaire 2024b] Raphaëlle Karine Lemaire, Alexis Lechervy et Youssef Chahir. *Appariement d'images d'oeuvres d'Art avec descriptions textuelles variées*. In Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP), Lille, France, Juillet 2024. (Cité en page 141.)

[Lester 2021] Brian Lester, Rami Al-Rfou et Noah Constant. *The Power of Scale for Parameter-Efficient Prompt Tuning*. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia et Scott Wen-tau Yih, editeurs, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3045–3059, Online and Punta Cana, Dominican Republic, Novembre 2021. Association for Computational Linguistics. (Cité en page 17.)

[Lewis 2020a] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov et Luke Zettlemoyer. *BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. In Dan Jurafsky, Joyce Chai, Natalie Schluter et Joel Tetreault, editeurs, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online, Juillet 2020. Association for Computational Linguistics. (Cité en pages 14 et 16.)

[Lewis 2020b] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih et Tim Rocktäschel. *Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks*. Advances in Neural Information Processing Systems, vol. 33, pages 9459–9474, 2020. (Cité en pages 17 et 49.)

[Li 2017] Shan Li, Weihong Deng et JunPing Du. *Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild*. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2584–2593, Juillet 2017. (Cité en page 40.)

[Li 2018a] Haoliang Li, Sinno Jialin Pan, Shiqi Wang et Alex C. Kot. *Domain Generalization with Adversarial Feature Learning*. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5400–5409, Juin 2018. (Cité en page 22.)

[Li 2018b] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang et Jan Kautz. *A Closed-Form Solution to Photorealistic Image Stylization*. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu et Yair Weiss, editeurs, Computer Vision – ECCV 2018, pages 468–483, Cham, 2018. Springer International Publishing. (Cité en page 12.)

[Li 2019a] Hao Li, Hong Zhang, Xiaojuan Qi, Ruigang Yang et Gao Huang. *Improved Techniques for Training Adaptive Deep Networks*. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1891–1900. arXiv, Août 2019. (Cité en pages 23 et 25.)

[Li 2019b] Yingming Li, Ming Yang et Zhongfei Zhang. *A Survey of Multi-View Representation Learning*. IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 10, pages 1863–1883, Octobre 2019. (Cité en page 14.)

[Li 2021a] Xiang Lisa Li et Percy Liang. *Prefix-Tuning : Optimizing Continuous Prompts for Generation*. In Chengqing Zong, Fei Xia, Wenjie Li et Roberto Navigli, editeurs, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers), pages 4582–4597, Online, Août 2021. Association for Computational Linguistics. (Cité en page 17.)

[Li 2021b] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart et Anima Anandkumar. *Fourier Neural Operator for Parametric Partial Differential Equations*. In International Conference on Learning Representations, 2021. (Cité en pages 16 et 49.)

[Li 2022a] Junnan Li, Dongxu Li, Caiming Xiong et Steven Hoi. *BLIP : Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. In Proceedings of the 39th International Conference on Machine Learning, pages 12888–12900. PMLR, Juin 2022. (Cité en pages 22 et 27.)

[Li 2022b] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang et Jian Ren. *Efficientformer : Vision Transformers at Mobilenet Speed*. Advances in Neural Information Processing Systems, vol. 35, pages 12934–12949, 2022. (Cité en page 26.)

[Li 2023a] Junnan Li, Dongxu Li, Silvio Savarese et Steven Hoi. *BLIP-2 : Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. In Proceedings of the 40th International Conference on Machine Learning, pages 19730–19742. PMLR, Juillet 2023. (Cité en pages 22 et 27.)

[Li 2023b] Xiaoxu Li, Xiaochen Yang, Zhanyu Ma et Jing-Hao Xue. *Deep Metric Learning for Few-Shot Image Classification : A Review of Recent Developments*. Pattern Recognition, vol. 138, page 109381, Juin 2023. (Cité en page 13.)

[Liang 2012] Sheng-Fu Liang, Chin-En Kuo, Yu-Han Hu et Yu-Shian Cheng. *A Rule-Based Automatic Sleep Staging Method*. Journal of Neuroscience Methods, vol. 205, no. 1, pages 169–176, Mars 2012. (Cité en page 33.)

[Liang 2013] Jason Zhi Liang, Nicholas Corso, Eric Turner et Avideh Zakhor. *Image Based Localization in Indoor Environments*. In 2013 Fourth International Conference on Computing for Geospatial Research and Application, pages 70–75, Juillet 2013. (Cité en page 36.)

[Liu 2017] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan et Changshui Zhang. *Learning Efficient Convolutional Networks Through Network Slimming*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Octobre 2017. (Cité en page 23.)

[Liu 2023] Haotian Liu, Chunyuan Li, Qingyang Wu et Yong Jae Lee. *Visual Instruction Tuning*. Advances in Neural Information Processing Systems, vol. 36, pages 34892–34916, Décembre 2023. (Cité en page 27.)

[Liu 2024] Fan Liu, Tianshu Zhang, Wenwen Dai, Chuanyi Zhang, Wenwen Cai, Xiaocong Zhou et Delong Chen. *Few-Shot Adaptation of Multi-Modal Foundation Models : A Survey*. Artificial Intelligence Review, vol. 57, no. 10, page 268, Août 2024. (Cité en page 16.)

[Lorre 2020] Guillaume Lorre, Jaonary Rabariosa, Astrid Orcesi, Samia Ainouz et Stephane Canu. *Temporal Contrastive Pretraining for Video Action Recognition*. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 651–659, Mars 2020. (Cité en page 14.)

[Lotte 2007] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche et B. Arnaldi. *A Review of Classification Algorithms for EEG-based Brain–Computer Interfaces*. Journal of Neural Engineering, vol. 4, no. 2, page R1, Janvier 2007. (Cité en page 34.)

[Lou 2018] Zhongyu Lou, Fares Alnajar, Jose M. Alvarez, Ninghang Hu et Theo Gevers. *Expression-Invariant Age Estimation Using Structured Learning*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 2, pages 365–375, Fèvrier 2018. (Cité en page 18.)

[Lowe 1999] David G Lowe. *Object Recognition from Local Scale-Invariant Features*. In ICCV, 1999. (Cité en pages 36 et 37.)

[Luo 2017] Zelun Luo, Yuliang Zou, Judy Hoffman et Li F Fei-Fei. *Label Efficient Learning of Transferable Representations Acrosss Domains and Tasks*. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. (Cité en page 16.)

[Madan 2024] Neelu Madan, Andreas Moegelmose, Rajat Modi, Yogesh S. Rawat et Thomas B. Moeslund. *Foundation Models for Video Understanding : A Survey*, Mai 2024. (Cité en page 16.)

[Martin 2012] Arnaud Martin. *About Conflict in the Theory of Belief Functions*. In Thierry Denoeux et Marie-Hélène Masson, editeurs, Belief Functions : Theory and Applications, pages 161–168, Berlin, Heidelberg, 2012. Springer. (Cité en page 42.)

[Merkle 2017] Nina Merkle, Wenjie Luo, Stefan Auer, Rupert Müller et Raquel Urtasun. *Exploiting Deep Matching and SAR Data for the Geo-Localization Accuracy Improvement of Optical Satellite Images*. Remote Sensing, vol. 9, no. 6, page 586, 2017. (Cité en page 38.)

[Meronen 2024] Lassi Meronen, Martin Trapp, Andrea Pilzer, Le Yang et Arno Solin. *Fixing Overconfidence in Dynamic Neural Networks*. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 2668–2678, Janvier 2024. (Cité en pages 23, 24 et 26.)

[Mignon 2012] Alexis Mignon et Frédéric Jurie. *PCCA : A New Approach for Distance Learning from Sparse Pairwise Constraints*. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 2666–2672, France, 2012. (Cité en pages 18 et 20.)

[Misra 2020] Ishan Misra et Laurens van der Maaten. *Self-Supervised Learning of Pretext-Invariant Representations*. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6706–6716, Juin 2020. (Cité en pages 14 et 15.)

[Morago 2016] Brittany Morago, Giang Bui et Ye Duan. *2D Matching Using Repetitive and Salient Features in Architectural Images*. IEEE Transactions on Image Processing, vol. 25, no. 10, pages 4888–4899, Octobre 2016. (Cité en page 36.)

[Moreau 2025] Noémie N. Moreau, Samuel Valable, Cyril Jaudet, Loïse Dessoude, Romain Herault, Romain Modzelewski, Dinu Stefan, Juliette Thariat, Alexis Lechervy et Aurélien Corroyer-Dulmont. *Early Characterization and Prediction of Glioblastoma and Brain Metastases Treatment Efficacy Using Medical Imaging-Based Radiomics and Artificial Intelligence Algorithms*. Frontiers in Oncology, vol. 15, Janvier 2025. (Cité en page 139.)

[Motiian 2017] Saeid Motiian, Quinn Jones, Seyed Iranmanesh et Gianfranco Doretto. *Few-Shot Adversarial Domain Adaptation*. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. (Cité en page 15.)

[Moummad 2021] I Moummad, C Jaudet, A Lechervy, S Valable, C Raboutet, J Lacroix, A Batalla et A Corroyer-Dulmont. *Impact d'un Algorithme de Machine Learning de Resampling Sur Les Radiomiques En IRM*. In Journée Thématique Santé et Sciences Du Numérique, Caen, France, Juin 2021. (Cité en page 142.)

[Moummad 2022] Ilyass Moummad, Cyril Jaudet, Alexis Lechervy, Samuel Valable, Charlotte Raboutet, Zamila Soilihi, Juliette Thariat, Nadia Falzone, Joëlle Lacroix, Alain Batalla et Aurélien Corroyer-Dulmont. *The Impact of Resampling and Denoising Deep Learning Algorithms on Radiomics in Brain Metastases MRI*. Cancers, vol. 14, no. 1, page 36, Janvier 2022. (Cité en page 139.)

[Mumuni 2022] Alhassan Mumuni et Fuseini Mumuni. *Data Augmentation : A Comprehensive Survey of Modern Approaches*. Array, vol. 16, page 100258, Décembre 2022. (Cité en page 12.)

[Nagrani 2021] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid et Chen Sun. *Attention Bottlenecks for Multimodal Fusion*. In Proceedings of the 35th International Conference on Neural Information Processing Systems, volume 34 of *NIPS '21*, pages 14200–14213, Red Hook, NY, USA, 2021. Curran Associates Inc. (Cité en pages 32, 43 et 44.)

[Nanthini 2023] K. Nanthini, D. Sivabalaselvamani, K. Chitra, P. Gokul, S. KavinKumar et S. Kishore. *A Survey on Data Augmentation Techniques*. In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), pages 913–920, Fèvrier 2023. (Cité en page 11.)

[Negrel 2015] Romain Negrel, Alexis Lechervy et Frédéric Jurie. *Boosted Metric Learning for Efficient Identity-Based Face Retrieval*. In 26th British Machine Vision Conference, volume 13, pages 1007–1036, Swansea, United Kingdom, Septembre 2015. (Cité en pages 19, 21 et 141.)

[Negrel 2016] Romain Negrel, Alexis Lechervy et Frederic Jurie. *MLBoost Revisited : A Faster Metric Learning Algorithm for Identity-Based Face Retrieval*. In Proceedings of the British Machine Vision Conference (BMVC), pages 103.1–103.13, York, UK, Septembre 2016. British Machine Vision Association. (Cité en pages 19, 20, 21 et 141.)

[Nistér 2004] David Nistér. *An Efficient Solution to the Five-Point Relative Pose Problem*. IEEE transactions on pattern analysis and machine intelligence, vol. 26, no. 6, pages 756–770, 2004. (Cité en page 37.)

[Oliva 2001] Aude Oliva et Antonio Torralba. *Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope*. International Journal of Computer Vision, vol. 42, no. 3, pages 145–175, Mai 2001. (Cité en page 36.)

[OpenAI 2024] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat,

Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth

Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk et Barret Zoph. *GPT-4 Technical Report*, Mars 2024. (Cité en page 17.)

[Ouali 2021] Yassine Ouali, Céline Hudelot et Myriam Tami. *Spatial Contrastive Learning for Few-Shot Classification*. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read et Jose A. Lozano, editeurs, Machine Learning and Knowledge Discovery in Databases. Research Track, pages 671–686, Cham, 2021. Springer International Publishing. (Cité en page 14.)

[Park 2020] Taesung Park, Alexei A. Efros, Richard Zhang et Jun-Yan Zhu. *Contrastive Learning for Unpaired Image-to-Image Translation*. In Andrea Vedaldi, Horst Bischof, Thomas Brox et Jan-Michael Frahm, editeurs, Computer Vision – ECCV 2020, pages 319–345, Cham, 2020. Springer International Publishing. (Cité en page 15.)

[Parnami 2022] Archit Parnami et Minwoo Lee. *Learning from Few Examples : A Summary of Approaches to Few-Shot Learning*, Mars 2022. (Cité en page 11.)

[Pedreschi 2019] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri et Franco Turini. *Meaningful Explanations of Black Box AI Decision Systems*. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, volume 33 of *AAAI'19/IAAI'19/EAAI'19*, pages 9780–9784, Honolulu, Hawaii, USA, Janvier 2019. AAAI Press. (Cité en page 28.)

[Petit 2024] Grégoire Petit, Michael Soumm, Eva Feillet, Adrian Popescu, Bertrand Delezoide, David Picard et Céline Hudelot. *An Analysis of Initial Training Strategies for Exemplar-Free Class-Incremental Learning*. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1826–1836, Janvier 2024. (Cité en page 22.)

[Phan 2022] Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins et Maarten De Vos. *SleepTransformer : Automatic Sleep Staging with Interpretability and Uncertainty Quantification*. IEEE Transactions on Biomedical Engineering, vol. 69, no. 8, pages 2456–2467, 2022. (Cité en page 33.)

[Phuong 2019] Mary Phuong et Christoph Lampert. *Distillation-Based Training for Multi-Exit Architectures*. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1355–1364, Seoul, Korea (South), Octobre 2019. IEEE. (Cité en pages 23 et 25.)

[Piasco 2018] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux et Valérie Gouet-Brunet. *A Survey on Visual-Based Localization : On the Benefit of Heterogeneous Data*. Pattern Recognition, vol. 74, pages 90–109, Fèvrier 2018. (Cité en page 36.)

[Piasco 2019] Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet et Cédric Demonceaux. *Learning Scene Geometry for Visual Localization in Challenging Conditions*. In 2019 International Conference on Robotics and Automation (ICRA), pages 9094–9100, Mai 2019. (Cité en page 37.)

[Piczak 2015] Karol J Piczak. *ESC : Dataset for Environmental Sound Classification*. In Proceedings of the 23rd ACM International Conference on Multimedia, pages 1015–1018. ACM, 2015. (Cité en page 39.)

[Pourpanah 2022] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang et QM Jonathan Wu. *A Review of Generalized Zero-Shot Learning Methods*. IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 4, pages 4051–4070, 2022. (Cité en page 16.)

[Prabhavalkar 2023] Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter et Shinji Watanabe. *End-to-End Speech Recognition : A Survey*, Mars 2023. (Cité en page 23.)

[Prakash 2021] Aditya Prakash, Kashyap Chitta et Andreas Geiger. *Multi-Modal Fusion Transformer for End-to-End Autonomous Driving*. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7073–7083, Juin 2021. (Cité en page 32.)

[Qian 2021] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie et Yin Cui. *Spatiotemporal Contrastive Video Representation Learning*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6960–6970, Juin 2021. (Cité en page 14.)

[Quan 2023] Shengjiang Quan, Masahiro Hirano et Yuji Yamakawa. *Semantic Information in Contrastive Learning*. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5686–5696, 2023. (Cité en page 15.)

[Quéau 2019] Yvain Quéau, Florian Leporcq, Alexis Lechervy et Ayman Alfalou. *Learning to Classify Materials Using Mueller Imaging Polarimetry*. In Fourteenth International Conference on Quality Control by Artificial Vision (QCAV), Mulhouse, France, Mai 2019. (Cité en page 140.)

[Radford 2018] Alec Radford, Karthik Narasimhan, Tim Salimans et Ilya Sutskever. *Improving Language Understanding by Generative Pre-Training*. 2018. (Cité en page 14.)

[Radford 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger et Ilya Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. In International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, Fèvrier 2021. (Cité en pages 16, 22, 27, 41, 44 et 48.)

[Raffel 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li et Peter J. Liu. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of machine learning research, vol. 21, no. 140, pages 1–67, 2020. (Cité en page 16.)

[Rahmath P 2024] Haseena Rahmath P, Vishal Srivastava, Kuldeep Chaurasia, Roberto G. Pacheco et Rodrigo S. Couto. *Early-Exit Deep Neural Network - A Comprehensive Survey*. ACM Comput. Surv., vol. 57, no. 3, pages 75 :1–75 :37, Novembre 2024. (Cité en page 25.)

[Raissi 2019] M. Raissi, P. Perdikaris et G. E. Karniadakis. *Physics-Informed Neural Networks : A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations*. Journal of Computational Physics, vol. 378, pages 686–707, Fèvrier 2019. (Cité en pages 16 et 49.)

[Ramalingam 2011] Srikumar Ramalingam, Sofien Bouaziz et Peter Sturm. *Pose Estimation Using Both Points and Lines for Geo-Localization*. In 2011 IEEE International Conference on Robotics and Automation, pages 4716–4723, Mai 2011. (Cité en page 36.)

[Rane 2021] Chinmay Rane, Gaël Dias, Alexis Lechervy et Asif Ekbal. *Improving Neural Text Style Transfer by Introducing Loss Function Sequentiality*. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, pages 2197–2201, New York, NY, USA, Juillet 2021. Association for Computing Machinery. (Cité en page 140.)

[Rastegari 2016] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon et Ali Farhadi. *XNOR-Net : ImageNet Classification Using Binary Convolutional Neural Networks*. In Bastian Leibe, Jiri Matas, Nicu Sebe et Max Welling, editeurs, Computer Vision – ECCV 2016, pages 525–542, Cham, 2016. Springer International Publishing. (Cité en page 23.)

[Ravaut 2024] Mathieu Ravaut, Aixin Sun, Nancy Chen et Shafiq Joty. *On Context Utilization in Summarization with Large Language Models*. In Lun-Wei Ku, Andre Martins et Vivek Srikumar, editeurs, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), pages 2764–2781, Bangkok, Thailand, Août 2024. Association for Computational Linguistics. (Cité en page 16.)

[Ravi 2017] Sachin Ravi et Hugo Larochelle. *Optimization as a Model for Few-Shot Learning*. In International Conference on Learning Representations, Fèvrier 2017. (Cité en page 17.)

[Ren 2021] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen et Xin Wang. *A Comprehensive Survey of Neural Architecture Search : Challenges and Solutions*, Mars 2021. (Cité en page 23.)

[Ricanek 2006] K. Ricanek et T. Tesafaye. *MORPH : A Longitudinal Image Database of Normal Adult Age-Progression*. In 7th International Conference on Automatic Face and Gesture Recognition (FGR06), pages 341–345, Avril 2006. (Cité en page 19.)

[Rodríguez Bribiesca 2021] Isaac Rodríguez Bribiesca, Adrián Pastor López Monroy et Manuel Montes-y-Gómez. *Multimodal Weighted Fusion of Transformers for Movie Genre Classification*. In Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, Candace Ross, Ruslan Salakhutdinov, Soujanya Poria, Erik Cambria et Kelly Shi, editeurs, Proceedings of the Third Workshop on Multimodal Artificial Intelligence, pages 1–5, Mexico City, Mexico, Juin 2021. Association for Computational Linguistics. (Cité en pages 32 et 43.)

[Rublee 2011] Ethan Rublee, Vincent Rabaud, Kurt Konolige et Gary Bradski. *ORB : An Efficient Alternative to SIFT or SURF*. In Computer Vision (ICCV), 2011 IEEE International Conference On, pages 2564–2571, 2011. (Cité en pages 36 et 37.)

[Ruder 2017] Sebastian Ruder. *An Overview of Multi-Task Learning in Deep Neural Networks*. CoRR, vol. abs/1706.05098, 2017. (Cité en page 15.)

[Ruffino 2022] Cyprien Ruffino, Rachel Blin, Samia Ainouz, Gilles Gasso, Romain Hérault, Fabrice Meriaudeau et Stéphane Canu. *Physically-Admissible Polarimetric Data Augmentation for Road-Scene Analysis*. Computer Vision and Image Understanding, vol. 222, page 103495, Septembre 2022. (Cité en page 12.)

[Russell 2011] Bryan C. Russell, Josef Sivic, Jean Ponce et Hélène Dessales. *Automatic Alignment of Paintings and Photographs Depicting a 3D Scene*. In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pages 545–552, Novembre 2011. (Cité en page 36.)

[Sarah 2022] Anthony Sarah, Daniel Cummings, Sharath Nittur Sridhar, Sairam Sundaresan, Maciej Szankin, Tristan James Webb et Juan Pablo Muñoz. *A Hardware-Aware System for Accelerating Deep Neural Network Optimization*. ArXiv, vol. abs/2202.12954, 2022. (Cité en page 23.)

[Saurer 2016] Olivier Saurer, Georges Baatz, Kevin Köser, L'ubor Ladický et Marc Pollefeys. *Image Based Geo-localization in the Alps*. International Journal of Computer Vision, vol. 116, no. 3, pages 213–225, Fèvrier 2016. (Cité en page 36.)

[Schuhmann 2022] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk et Jenia Jitsev. *LAION-5B : An Open Large-Scale Dataset for Training next Generation Image-Text Models*. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, pages 25278–25294, Red Hook, NY, USA, Novembre 2022. Curran Associates Inc. (Cité en page 22.)

[Sensoy 2018] M. Sensoy, L. Kaplan et M. Kandemir. *Evidential Deep Learning to Quantify Classification Uncertainty*. Advances in Neural Information Processing Systems, vol. 31, pages 3179–3189, 2018. (Cité en page 42.)

[Seo 2020] Hogeon Seo, Seunghyeok Back, Seongju Lee, Deokhwan Park, Tae Kim et Kyoobin Lee. *Intra- and Inter-Epoch Temporal Context Network (IITNet) Using Sub-Epoch Features for Auto-*

matic Sleep Scoring on Raw Single-Channel EEG. Biomedical Signal Processing and Control, vol. 61, page 102037, 2020. (Cité en page 33.)

[Seraphim 2023a] Mathieu Seraphim, Paul Dequidt, Alexis Lechervy, Florian Yger, Luc Brun et Olivier Etard. *Analyse automatique de l'état de sommeil sur données EEG par utilisation de Transformers et de matrices de covariance*. In 19ème Colloque ORASIS (ORASIS 2023) :, journées francophones des jeunes chercheurs en vision par ordinateur,, Mai 2023. (Cité en pages 33, 34, 43 et 142.)

[Seraphim 2023b] Mathieu Seraphim, Paul Dequidt, Alexis Lechervy, Florian Yger, Luc Brun et Olivier Etard. *Temporal Sequences of EEG Covariance Matrices for Automated Sleep Stage Scoring with Attention Mechanisms*. In Nicolas Tsapatsoulis, Andreas Lanitis, Marios Pattichis, Constantinos Pattichis, Christos Kyrkou, Efthyvoulos Kyriacou, Zenonas Theodosiou et Andreas Panayides, editeurs, Computer Analysis of Images and Patterns, pages 67–76, Cham, 2023. Springer Nature Switzerland. (Cité en pages 33, 34, 43 et 140.)

[Seraphim 2024a] Mathieu Seraphim, Alexis Lechervy, Florian Yger, Luc Brun et Olivier Etard. *Automatic Classification of Sleep Stages from EEG Signals Using Riemannian Metrics and Transformer Networks*. SN Computer Science, 2024. (Cité en pages 33, 34, 43 et 139.)

[Seraphim 2024b] Mathieu Seraphim, Alexis Lechervy, Florian Yger, Luc Brun et Olivier Etard. *Structure-Preserving Transformers for Sequences of SPD Matrices*. In European Signal Processing Conference (EUSIPCO) 2024, Août 2024. (Cité en pages 33, 34, 43 et 140.)

[Sermanet 2018] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine et Google Brain. *Time-Contrastive Networks : Self-Supervised Learning from Video*. In 2018 IEEE International Conference on Robotics and Automation (ICRA), pages 1134–1141, Mai 2018. (Cité en page 14.)

[Shafer 1976] G. Shafer. A mathematical theory of evidence. Princeton University Press, 1976. (Cité en page 42.)

[Sharma 2019] Monika Sharma, Abhishek Verma et Lovekesh Vig. *Learning to Clean : A GAN Perspective*. In Gustavo Carneiro et Shaodi You, editeurs, Computer Vision – ACCV 2018 Workshops, pages 174–185, Cham, 2019. Springer International Publishing. (Cité en page 12.)

[Shen 2012] Chunhua Shen, Junae Kim, Lei Wang et Anton Van Den Hengel. *Positive Semidefinite Metric Learning Using Boosting-like Algorithms*. The Journal of Machine Learning Research, vol. 13, no. 1, pages 1007–1036, 2012. (Cité en page 20.)

[Simonyan 2015] Karen Simonyan et Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. In 3rd International Conference on Learning Representations (ICLR), 2015. (Cité en page 41.)

[Song 2011] Zheng Song, Bingbing Ni, Dong Guo, Terence Sim et Shuicheng Yan. *Learning Universal Multi-View Age Estimator Using Video Context*. In 2011 International Conference on Computer Vision, pages 241–248, Novembre 2011. (Cité en page 18.)

[Song 2023] Yisheng Song, Ting Wang, Puyu Cai, Subrota K. Mondal et Jyoti Prakash Sahoo. *A Comprehensive Survey of Few-shot Learning : Evolution, Applications, Challenges, and Opportunities*. ACM Comput. Surv., vol. 55, no. 13s, pages 271 :1–271 :40, Juillet 2023. (Cité en page 11.)

[Sun 2016] Baochen Sun, Jiashi Feng et Kate Saenko. *Return of Frustratingly Easy Domain Adaptation*. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, pages 2058–2065, Phoenix, Arizona, Fèvrier 2016. AAAI Press. (Cité en page 22.)

[Supratak 2017] Akara Supratak, Hao Dong, Chao Wu et Yike Guo. *DeepSleepNet : A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG*. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 11, pages 1998–2008, Novembre 2017. (Cité en page 33.)

[Tagliazucchi 2012] Enzo Tagliazucchi, Frederic von Wegner, Astrid Morzelewski, Sergey Borisov, Kolja Jahnke et Helmut Laufs. *Automatic Sleep Staging Using fMRI Functional Connectivity Data*. NeuroImage, vol. 63, no. 1, pages 63–72, Octobre 2012. (Cité en page 34.)

[Tamaazousti 2020] Youssef Tamaazousti, Herve Le Borgne, Celine Hudelot, Mohamed-El-Amine Seddik et Mohamed Tamaazousti. *Learning More Universal Representations for Transfer-Learning*. IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 9, pages 2212–2224, Septembre 2020. (Cité en page 22.)

[Teerapittayanon 2016] Surat Teerapittayanon, Bradley McDanel et Hsiang-Tsung Kung. *Branchynet : Fast Inference via Early Exiting from Deep Neural Networks*. In 2016 23rd International Conference on Pattern Recognition (ICPR), pages 2464–2469. IEEE, 2016. (Cité en pages 23 et 24.)

[Thukral 2012] Pavleen Thukral, Kaushik Mitra et Rama Chellappa. *A Hierarchical Approach for Human Age Estimation*. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1529–1532, Mars 2012. (Cité en page 18.)

[Thung 2018] Kim-Han Thung et Chong-Yaw Wee. *A Brief Review on Multi-Task Learning*. Multimedia Tools and Applications, vol. 77, no. 22, pages 29705–29725, Novembre 2018. (Cité en page 15.)

[Tian 2020a] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid et Phillip Isola. *What Makes for Good Views for Contrastive Learning ?* In Advances in Neural Information Processing Systems, volume 33, pages 6827–6839. Curran Associates, Inc., 2020. (Cité en page 14.)

[Tian 2020b] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum et Phillip Isola. *Rethinking Few-Shot Image Classification : A Good Embedding Is All You Need ?* In Andrea Vedaldi, Horst Bischof, Thomas Brox et Jan-Michael Frahm, editeurs, Computer Vision – ECCV 2020, pages 266–282, Cham, 2020. Springer International Publishing. (Cité en page 13.)

[Tian 2021] Yuandong Tian, Xinlei Chen et Surya Ganguli. *Understanding Self-Supervised Learning Dynamics without Contrastive Pairs*. In Proceedings of the 38th International Conference on Machine Learning, pages 10268–10278. PMLR, Juillet 2021. (Cité en page 15.)

[Tian 2024] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning et Prayag Tiwari. *A Survey on Few-Shot Class-Incremental Learning*. Neural Networks, vol. 169, pages 307–324, Janvier 2024. (Cité en page 11.)

[Tong 2021a] Zheng Tong, Philippe Xu et Thierry Denœux. *An Evidential Classifier Based on Dempster-Shafer Theory and Deep Learning*. Neurocomputing, vol. 450, pages 275–293, Août 2021. (Cité en page 32.)

[Tong 2021b] Zheng Tong, Philippe Xu et Thierry Denœux. *Evidential Fully Convolutional Network for Semantic Segmentation*. Applied Intelligence, vol. 51, no. 9, pages 6376–6399, Septembre 2021. (Cité en page 42.)

[Touvron 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave et Guillaume Lample. *LLaMA : Open and Efficient Foundation Language Models*, Fèvrier 2023. (Cité en page 17.)

[Troester 2023] Matthew M. Troester, Stuart F. Quan, American Academy of Sleep Medicine et Richard B. Berry. The AASM Manual for the Scoring of Sleep and Associated Events, Version 3. American Academy Of Sleep Medicine, Juin 2023. (Cité en page 33.)

[van den Oord 2019] Aaron van den Oord, Yazhe Li et Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*, Janvier 2019. (Cité en pages 14 et 15.)

[Van Der Donckt 2023] Jeroen Van Der Donckt, Jonas Van Der Donckt, Emiel Deprost, Nicolas Vandenbussche, Michael Rademaker, Gilles Vandewiele et Sofie Van Hoecke. *Do Not Sleep on Traditional Machine Learning : Simple and Interpretable Techniques Are Competitive to Deep Learning for Sleep Scoring*. Biomedical Signal Processing and Control, vol. 81, page 104429, Mars 2023. (Cité en page 33.)

[Vaswani 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser et Illia Polosukhin. *Attention Is All You Need*. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, editeurs, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. (Cité en pages 26 et 34.)

[Vidit 2023] Vidit Vidit, Martin Engilberge et Mathieu Salzmann. *CLIP the Gap : A Single Domain Generalization Approach for Object Detection*. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3219–3229, Juin 2023. (Cité en page 22.)

[Vielzeuf 2017] Valentin Vielzeuf, Stéphane Pateux et Frédéric Jurie. *Temporal Multimodal Fusion for Video Emotion Classification in the Wild*. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pages 569–576. ACM, 2017. (Cité en page 39.)

[Vielzeuf 2018a] Valentin Vielzeuf, Corentin Kervadec, Stéphane Pateux, Alexis Lechervy et Frédéric Jurie. *An Occam's Razor View on Learning Audiovisual Emotion Recognition with Small Trai-*

ning Sets. In ICMI (EmotiW) 2018, Boulder, Colorado, United States, Octobre 2018. (Cité en pages 39, 40 et 141.)

[Vielzeuf 2018b] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux et Frédéric Jurie. *CentralNet : A Multilayer Approach for Multimodal Fusion*. In European Conference on Computer Vision Workshops : Multimodal Learning and Applications, pages 575–589, Munich, Germany, Septembre 2018. (Cité en pages 38, 39, 40, 41, 43 et 141.)

[Vielzeuf 2019] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux et Frédéric Jurie. *Multilevel Sensor Fusion With Deep Learning*. IEEE Sensors Letters, vol. 3, no. 1, pages 1–4, Janvier 2019. (Cité en pages 39, 40 et 139.)

[Vivone 2023] Gemine Vivone. *Multispectral and Hyperspectral Image Fusion in Remote Sensing : A Survey*. Information Fusion, vol. 89, pages 405–417, Janvier 2023. (Cité en page 31.)

[Vo 2017] Nam Vo, Nathan Jacobs et James Hays. *Revisiting IM2GPS in the Deep Learning Era*. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2640–2649, Octobre 2017. (Cité en page 36.)

[Walawalkar 2020] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu et Marios Savvides. *Attentive Cutmix : An Enhanced Data Augmentation Approach for Deep Learning Based Image Classification*. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3642–3646, Mai 2020. (Cité en page 12.)

[Walch 2017] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck et Daniel Cremers. *Image-Based Localization Using LSTMs for Structured Feature Correlation*. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, volume 1, pages 627–637. IEEE Computer Society, 2017. (Cité en pages 36 et 37.)

[Wang 2020a] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo et Qinghua Hu. *ECA-Net : Efficient Channel Attention for Deep Convolutional Neural Networks*. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11531–11539, Juin 2020. (Cité en page 32.)

[Wang 2020b] Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang et Gao Huang. *Glance and Focus : A Dynamic Approach to Reducing Spatial Redundancy in Image Classification*. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pages 2432–2444. Curran Associates, Inc., 2020. (Cité en pages 23 et 25.)

[Wang 2021a] Yaqing Wang, Quanming Yao, James T. Kwok et Lionel M. Ni. *Generalizing from a Few Examples : A Survey on Few-shot Learning*. ACM Computing Surveys, vol. 53, no. 3, pages 1–34, Mai 2021. (Cité en page 11.)

[Wang 2021b] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang et Gao Huang. *Not All Images Are Worth 16x16 Words : Dynamic Transformers for Efficient Image Recognition*. In Advances in Neural

Information Processing Systems, volume 34, pages 11960–11973. Curran Associates, Inc., 2021. (Cité en page 24.)

[Wang 2024] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei et Yuanchun Zhou. *A Comprehensive Survey on Data Augmentation*, Mai 2024. (Cité en page 11.)

[Weinberger 2009] Kilian Q. Weinberger et Lawrence K. Saul. *Distance Metric Learning for Large Margin Nearest Neighbor Classification*. Journal of Machine Learning Research, vol. 10, pages 207–244, 2009. (Cité en page 20.)

[Wilson 2020] Garrett Wilson et Diane J. Cook. *A Survey of Unsupervised Deep Domain Adaptation*. ACM Trans. Intell. Syst. Technol., vol. 11, no. 5, pages 51 :1–51 :46, Juillet 2020. (Cité en page 22.)

[Wu 2012] Changwei W. Wu, Po-Yu Liu, Pei-Jung Tsai, Yu-Chin Wu, Ching-Sui Hung, Yu-Che Tsai, Kuan-Hung Cho, Bharat B. Biswal, Chia-Ju Chen et Ching-Po Lin. *Variations in Connectivity in the Sensorimotor and Default-Mode Networks During the First Nocturnal Sleep Cycle*. Brain Connectivity, vol. 2, no. 4, pages 177–190, Août 2012. (Cité en page 34.)

[Wu 2018] Zhirong Wu, Yuanjun Xiong, Stella X. Yu et Dahua Lin. *Unsupervised Feature Learning via Non-parametric Instance Discrimination*. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3733–3742, Juin 2018. (Cité en page 15.)

[Xiao 2024] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu et Lu Yuan. *Florence-2 : Advancing a Unified Representation for a Variety of Vision Tasks*. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4818–4829. IEEE Computer Society, Juin 2024. (Cité en pages 22, 27, 32 et 44.)

[Xie 2016] Junyuan Xie, Ross Girshick et Ali Farhadi. *Unsupervised Deep Embedding for Clustering Analysis*. In Proceedings of The 33rd International Conference on Machine Learning, pages 478–487. PMLR, Juin 2016. (Cité en page 15.)

[Xin 2021] Ji Xin, Raphael Tang, Yaoliang Yu et Jimmy Lin. *BERxiT : Early Exiting for BERT with Better Fine-Tuning and Extension to Regression*. In Paola Merlo, Jorg Tiedemann et Reut Tsarfaty, editeurs, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume, pages 91–104, Online, Avril 2021. Association for Computational Linguistics. (Cité en page 27.)

[Xiong 2014] Fei Xiong, Mengran Gou, Octavia Camps et Mario Sznaier. *Person Re-Identification Using Kernel-Based Metric Learning Methods*. In ECCV 2014, pages 1–16. Springer, Octobre 2014. (Cité en page 20.)

[Xu 2023] Mingle Xu, Sook Yoon, Alvaro Fuentes et Dong Sun Park. *A Comprehensive Survey of Image Augmentation Techniques for Deep Learning*. Pattern Recognition, vol. 137, page 109347, Mai 2023. (Cité en page 12.)

[Yan 2015] Wang Yan, Jordan Yap et Greg Mori. *Multi-Task Transfer Methods to Improve One-Shot Learning for Multimedia Event Detection.* In BMVC, pages 37–1, 2015. (Cité en page 16.)

[Yang 2016] Xiaodong Yang, Pavlo Molchanov et Jan Kautz. *Multilayer and Multimodal Fusion of Deep Neural Networks for Video Classification*. In Proceedings of the 2016 ACM on Multimedia Conference, pages 978–987. ACM, 2016. (Cité en page 39.)

[Yang 2020] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai et Gao Huang. *Resolution Adaptive Networks for Efficient Inference*. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2366–2375, Seattle, WA, USA, Juin 2020. IEEE. (Cité en pages 23, 24 et 25.)

[Yang 2021] Le Yang, Haojun Jiang, Ruojin Cai, Yulin Wang, Shiji Song, Gao Huang et Qi Tian. *CondenseNet V2 : Sparse Feature Reactivation for Deep Networks*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3569–3578, Juin 2021. (Cité en page 23.)

[Yi 2016] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit et Pascal Fua. *Lift : Learned Invariant Feature Transform*. In European Conference on Computer Vision, pages 467–483. Springer, 2016. (Cité en pages 36 et 37.)

[Yi 2018] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann et Pascal Fua. *Learning to Find Good Correspondences*. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), volume 3, 2018. (Cité en page 37.)

[Yosinski 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio et Hod Lipson. *How Transferable Are Features in Deep Neural Networks ?* In Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014. (Cité en pages 16 et 28.)

[Yu 2022] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini et Yonghui Wu. *CoCa : Contrastive Captioners Are Image-Text Foundation Models*, Juin 2022. (Cité en page 22.)

[Yu 2023] Haichao Yu, Haoxiang Li, Gang Hua, Gao Huang et Humphrey Shi. *Boosted Dynamic Neural Networks*. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 10989–10997. AAAI Press, 2023. (Cité en page 24.)

[Yun 2019] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo et Junsuk Choe. *CutMix : Regularization Strategy to Train Strong Classifiers With Localizable Features*. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6022–6031. IEEE Computer Society, Octobre 2019. (Cité en page 12.)

[Zagoruyko 2015] Sergey Zagoruyko et Nikos Komodakis. *Learning to Compare Image Patches via Convolutional Neural Networks*. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference On, pages 4353–4361, 2015. (Cité en page 38.)

[Zamir 2016] Amir R. Zamir, Asaad Hakeem, Luc Van Gool, Mubarak Shah et Richard Szeliski, editeurs. Large-Scale Visual Geo-Localization. Advances in Computer Vision and Pattern Recognition. Springer International Publishing, Cham, 2016. (Cité en page 36.)

[Zbontar 2021] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun et Stéphane Deny. *Barlow Twins : Self-supervised Learning via Redundancy Reduction*. In International Conference on Machine Learning, pages 12310–12320. PMLR, 2021. (Cité en page 15.)

[Zhang 2018a] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin et David Lopez-Paz. *Mixup : Beyond Empirical Risk Minimization*. In International Conference on Learning Representations, 2018. (Cité en page 12.)

[Zhang 2018b] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin et Jian Sun. *ShuffleNet : An Extremely Efficient Convolutional Neural Network for Mobile Devices*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Juin 2018. (Cité en page 23.)

[Zhang 2018c] Yabin Zhang, Hui Tang et Kui Jia. *Fine-Grained Visual Categorization Using Meta-Learning Optimization with Sample Selection of Auxiliary Data*. In Proceedings of the European Conference on Computer Vision (ECCV), pages 233–248, 2018. (Cité en page 15.)

[Zhang 2018d] Yu Zhang et Qiang Yang. *An Overview of Multi-Task Learning*. National Science Review, vol. 5, no. 1, pages 30–43, 2018. (Cité en page 15.)

[Zhang 2019] Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo et Désiré Sidibé. *Exploration of Deep Learning-Based Multimodal Fusion for Semantic Road Scene Segmentation*. In VISAPP 2019 14Th International Conference on Computer Vision Theory and Applications, 2019. (Cité en page 32.)

[Zhang 2021a] Yifei Zhang, Désiré Sidibé, Olivier Morel et Fabrice Mériaudeau. *Deep Multimodal Fusion for Semantic Image Segmentation : A Survey*. Image and Vision Computing, vol. 105, page 104042, Janvier 2021. (Cité en page 31.)

[Zhang 2021b] Yu Zhang et Qiang Yang. *A Survey on Multi-Task Learning*. IEEE transactions on knowledge and data engineering, vol. 34, no. 12, pages 5586–5609, 2021. (Cité en page 15.)

[Zhang 2022] Lei Zhang, Na Jiang, Qishuai Diao, Zhong Zhou et Wei Wu. *Person Re-identification with Pose Variation Aware Data Augmentation*. Neural Computing and Applications, vol. 34, no. 14, pages 11817–11830, Juillet 2022. (Cité en page 12.)

[Zhang 2023a] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang et Rainer Stiefelhagen. *Delivering Arbitrary-Modal Semantic Segmentation*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1136–1147, 2023. (Cité en page 32.)

[Zhang 2023b] Yuxiao Zhang, Alexander Carballo, Hanting Yang et Kazuya Takeda. *Perception and Sensing for Autonomous Vehicles under Adverse Weather Conditions : A Survey*. ISPRS Jour-

nal of Photogrammetry and Remote Sensing, vol. 196, pages 146–177, Fèvrier 2023. (Cité en page 41.)

[Zhang 2024] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu et Changqing Zhang. *Multimodal Fusion on Low-quality Data : A Comprehensive Survey*, Mai 2024. (Cité en pages 31 et 32.)

[Zhao 2024] Fei Zhao, Chengcui Zhang et Baocheng Geng. *Deep Multimodal Data Fusion*. ACM Comput. Surv., vol. 56, no. 9, pages 216 :1–216 :36, Avril 2024. (Cité en pages 31 et 32.)

[Zhou 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy et Ziwei Liu. *Conditional Prompt Learning for Vision-Language Models*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16816–16825, 2022. (Cité en page 22.)

[Zhou 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy et Ziwei Liu. *Learning to Prompt for Vision-Language Models*. International Journal of Computer Vision, vol. 130, no. 9, pages 2337–2348, Septembre 2022. (Cité en page 22.)

[Zhou 2023] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang et Chen Change Loy. *Domain Generalization : A Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pages 4396–4415, Avril 2023. (Cité en page 22.)

[Zhou 2024] Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang et Yuan Wu. *A Survey on Data Augmentation in Large Model Era*, Mars 2024. (Cité en page 11.)

[Zhuang 2021] Liu Zhuang, Lin Wayne, Shi Ya et Zhao Jun. *A Robustly Optimized BERT Pre-training Approach with Post-training*. In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He et Gaoqi Rao, editeurs, Proceedings of the 20th Chinese National Conference on Computational Linguistics, pages 1218–1227, Huhhot, China, Août 2021. Chinese Information Processing Society of China. (Cité en page 16.)

[Zong 2024] Yongshuo Zong, Oisin Mac Aodha et Timothy Hospedales. *Self-Supervised Multimodal Learning : A Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–20, 2024. (Cité en page 31.)

Travaux sur l'apprentissage frugale et la fusion de modalités

Résumé :

L'intelligence artificielle est en plein essor et s'intègre progressivement dans notre quotidien. Son utilisation croissante entraîne un besoin toujours plus grand en données et en ressources de calcul pour les traiter. Dans ce document, j'aborde deux défis majeurs que cela engendre : l'apprentissage frugal et la fusion de modalités.

La première partie explore les stratégies développées pour entraîner des modèles performants malgré des données limitées ou des ressources computationnelles restreintes. Elle aborde notamment l'apprentissage cross-domaine, l'apprentissage de métriques et les architectures multi-sorties efficaces. Ces approches visent à optimiser l'utilisation des ressources disponibles tout en maintenant, voire en améliorant, la performance des modèles.

La seconde partie se concentre sur la capacité à intégrer et exploiter des informations issues de sources hétérogènes, telles que des capteurs variés, du texte et des images. Elle étudie différentes approches de fusion – précoce, tardive, hybride – pour améliorer la robustesse et la performance des systèmes. Ces méthodes trouvent des applications dans divers domaines, allant de l'analyse de scènes à l'estimation du sommeil, en passant par la reconnaissance d'activités humaines.

Ce document fait la synthèse de mes activités de recherche dans ces domaines et les replace dans le contexte de la littérature récente. Il met également en lumière les avancées actuelles et les perspectives futures possibles.

Mots clés : Apprentissage multimodal, Apprentissage frugale, Apprentissage avec peu de ressources, Apprentissage profond, Vision par ordinateur.

Works on frugal learning and modality fusion

Abstract :

Artificial intelligence is rapidly advancing and gradually integrating into our daily lives. Its increasing use leads to an ever-growing need for data and computational resources to process it. In this document, I address two major challenges that arise from this : frugal learning and modality fusion.

The first part explores strategies developed to train high-performing models despite limited data or restricted computational resources. It covers cross-domain learning, metric learning, and efficient multi-output architectures. These approaches aim to optimize the use of available resources while maintaining or even improving model performance.

The second part focuses on the ability to integrate and exploit information from heterogeneous sources, such as various sensors, text, and images. It studies different fusion approaches – early, late, hybrid – to enhance the robustness and performance of systems. These methods find applications in various fields, ranging from scene analysis to sleep estimation and human activity recognition.

This document synthesizes my research activities in these areas and places them in the context of recent literature. It also highlights current advancements and possible future perspectives.

Keywords : Multimodal Learning, Frugal Learning, Resource-efficient Learning, Deep Learning, Computer Vision.