

Multi-Level Sensor Fusion with Deep Learning

Valentin Vielzeuf^{1,2}, Alexis Lechervy², Stéphane Pateux¹,
and Frédéric Jurie²

¹ Orange Labs, Rennes

² Normandie Univ., UNICAEN, ENSICAEN, CNRS

November 7, 2018

Abstract

In the context of deep learning, this article presents an original deep network, namely CentralNet, for the fusion of information coming from different sensors. This approach is designed to efficiently and automatically balance the trade-off between early and late fusion (*i.e.* between the fusion of low-level vs high-level information). More specifically, at each level of abstraction – the different levels of deep networks – unimodal representations of the data are fed to a central neural network which combines them into a common embedding. In addition, a multi-objective regularization is also introduced, helping to both optimize the central network and the unimodal networks. Experiments on four multimodal datasets not only show state-of-the-art performance, but also demonstrate that CentralNet can actually choose the best possible fusion strategy for a given problem.

1 Introduction and Related Work

In recent years, the multiplication of sensor technologies has resulted in a massive increase of multimodal data such as videos, RGB+D data, *etc.*. The fusion of modalities has been successful in various domains, from medical imaging to autonomous driving, including natural language processing or image analysis. The main motivation behind the use of multimodal data is to combine complementary information from different modalities, leading to better decisions than using only one. Many problems that benefit from multiple modalities can be found in the recent deep learning literature such as video classification [22], emotion recognition [16], or human activity recognition [3].

A first basic taxonomy of multimodal fusion [11] distinguishes between the approaches by the level of representation at which the fusion is done, as

deep neural networks process the data at different levels of abstraction, the different layers of the networks. A universal optimal level of fusion (early or late) does not exist, as it is task dependent. For example, Simonyan *et al.* [17]’s two stream convolutional neural network for human activity recognition is fusing modalities at prediction level (late fusion). In the same way, in audiovisual emotion recognition, better performance is observed using late fusion [10, 21]. In contrast, Arevalo *et al.* [2] propose a Gated Multimodal Unit, weighting the modalities features using relatively early fusion information on a textual-visual dataset. Finally, Chen *et al.* [4] follow an early fusion hard-gated approach for textual-visual sentiment analysis.

Nevertheless, the early versus late fusion paradigm can be insufficient to describe some approaches. Indeed, Neverova *et al.* [15] highlight that there are benefits to fuse similar modalities earlier. Another example is the multilayer approach of Yang *et al.* [23] performing the fusion on human activity videos with a boosting algorithm applied to all layers. Kang *et al.* [8] also use a multilayer framework where the layers’ representations are aggregated into a unique representation.

Finally, it is worth mentioning some approaches relying on regularization: Andrew *et al.* [1] propose the deep Canonical Correlation Analysis method aiming at maximizing the correlation between the multimodal representations, while Neverova *et al.* [14, 12] propose modDrop and modout regularizations, consisting in dropping modalities during training.

The proposed method borrows from both visions: it is a multilayer framework and uses a constrained multi-objective loss. It builds on existing deep convolutional neural networks designed to process each modality independently. We suggest connecting these networks, at different levels, using an additional central network dedicated to the projection of the features coming from the different modalities into the same common space. In addition, the global loss allows to back propagate some global constraints on each modality, coordinating their representations. The proposed approach aims at automatically identifying an optimal fusion strategy for a given task. The approach is multi-objective in the sense that it simultaneously tries to minimize per modality losses as well as the global loss defined on the joint space. This article is an extension of [20].

After presenting our approach in the next section, Section 3 gives an extended experimental validation of the approach showing the advantage of our method on 4 different multimodal tasks.

2 CentralNet

In this text we use the phrase *multimodal fusion* to refer to the combination of information provided by different media, under the form of their associated features or the intermediate decisions. More formally, if M^1 and M^2 denote

the media and D^1 and D^2 the decisions inferred respectively from M^1 and M^2 , the goal is to make a better decision $D^{1,2}$ using both M^1 and M^2 . More than 2 modalities can be used. This paper addresses the case of classification tasks, but any other task, *e.g.* regression, can be addressed in the same way.

This article focuses on the case of neural nets, for which the data is sequentially processed by a succession of layers. We assume having one neural net per modality, capable of inferring a decision from each modality taken in isolation, and want to combine them. One recurrent question with multimodal fusion is where the fusion has to be done: close to the data (early fusion as in Fig.1.a), at the decision level (late fusion as in Fig.1.b) or in between. In the case of neural networks, fusion can be performed at any level between the input and the output of the different unimodal networks.

For simplicity, we assume that the extracted features (at the input of the fusion layers) have the same dimensionality. If it is not the case, the features can be projected, *e.g.* with 1x1 convolutional layers or zero padded to give them the same size. In practice, the last convolution layers or the first dense layers of separately trained unimodal networks can be used as features.

CentralNet Architecture The architecture of CentralNet [20] is a neural network which combines the features issued from different modalities, by taking as input of each one of its layers, a weighted sum of the layers of the corresponding unimodal networks and of its own previous layers. This is illustrated in Figure 1(c). Such fusion layers can be defined by the following equation:

$$h_{C_{i+1}} = \alpha_{C_i} h_{C_i} + \sum_{k=1}^n \alpha_{M_i^k} h_{M_i^k} \quad (1)$$

where n is the number of modalities, α are trainable scalar weights, $h_{M_i^k}$ is the hidden representation of each modality at layer i , and h_{C_i} is the central hidden representation. The resulting representation $h_{C_{i+1}}$ is then fed to an operating layer cell (which can be a convolutional or a dense layer followed by an activation function).

Regarding the first layer of the central network ($i = 0$), as we do not have any previous central hidden representation, we only weight and sum the representations of all modalities (M^i) issued from unimodal networks. At the output level, the last weighted sum is done between the unimodal predictions and the central prediction. Then, the output of the central net is used as the final prediction.

Training the CentralNet model All trainable weights of the unimodal networks, the ones of the CentralNet and the fusion parameters $\alpha_{M_i^k}$, are optimized jointly by applying a stochastic gradient descent using the Adam approach. The global loss is defined as: $loss = loss_C + \sum_{k=1}^n loss_{M^k}$ where $loss_C$ is the (classification) loss computed from the output of the central

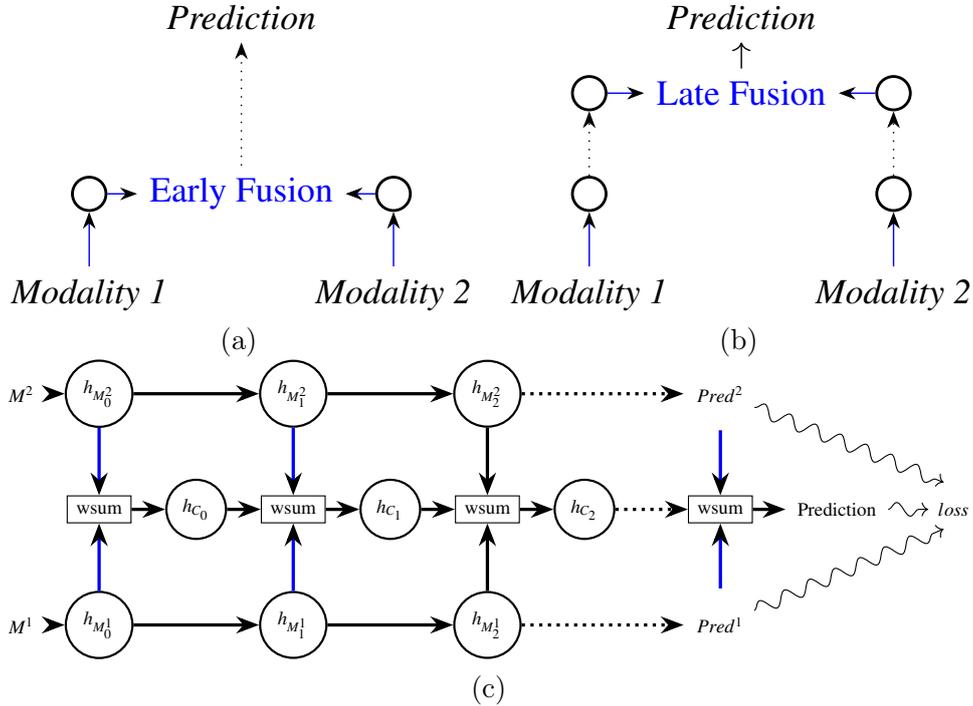


Figure 1: (a) Early fusion method, fusing the low-level representations of the modalities. (b) Late fusion method, fusing the high-level representations of the modalities (or even predictions) obtained for each unimodal neural network. (c) Our CentralNet fusion model, using both unimodal hidden representations and a central joint representation at each layer. The fusion of the unimodal representations is done in this case using a learned weighted sum. For the sake of simplicity, only the overall synoptic views of the architectures are represented. More details are provided in Section 2.

model and $loss_{M^k}$ the (classification) loss when using only modality k . As already observed by Neverova *et al.* [14], when dealing with multimodal fusion, it is crucial to maintain the performance of each unimodal neural networks. It is the reason why the global loss also includes unimodal losses. It helps generalization by acting as a multi-objective regularization (see Fig. 3).

Overall, CentralNet is easy to implement and can be built on top of existing architectures known to be efficient for each modality. We refer to [20] for implementation details. The resulting values of the α weights allow some interesting interpretations on where modalities are combined. For instance, getting $\alpha_{M^k_i}$ values close to 0 for $i > 0$ is equivalent to early fusion, while having all the α_{C_i} close to 0 up to the last weighted sum would be equivalent to late fusion.

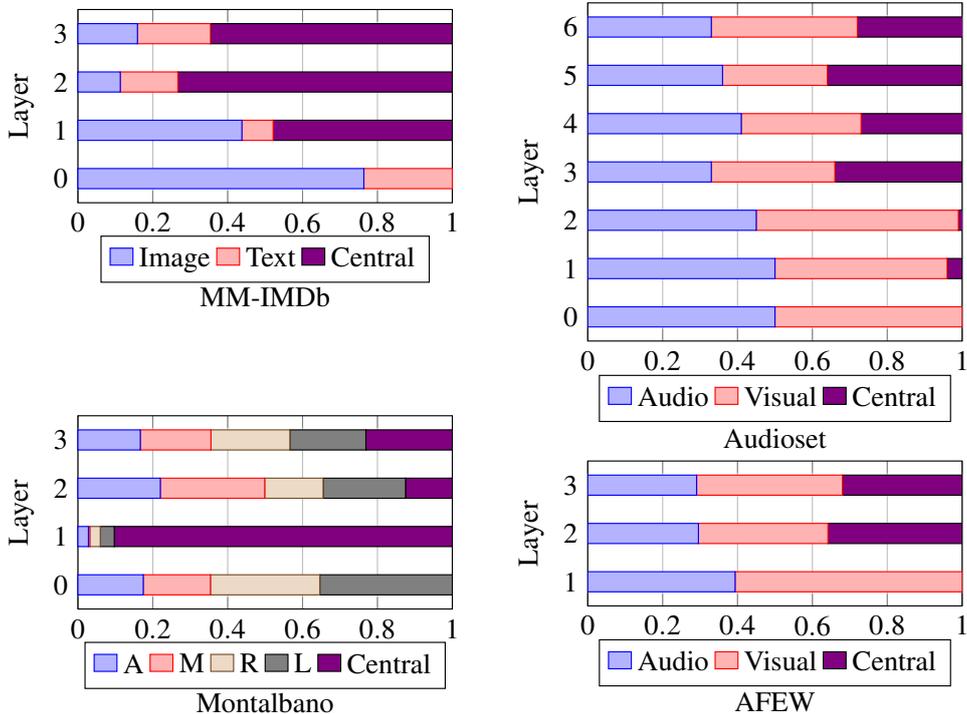


Figure 2: Visualization of the α_i weights after training. They are displayed as the percentage given to each modality and to the central hidden representations across layers and datasets. We observe that the learned fusion strategy is different for each dataset.

3 Experiments

This section provides extensive evaluations of CentralNet on four multimodal datasets, namely Montalbano [6], MM-IMDb [2], the 'animal' subset of Audioset [7] and AFEW [5]. For each dataset, CentralNet is compared with four different fusion approaches: *early* fusion (concatenation of low-level features), *late* fusion (concatenation of the unimodal scores), *ModDrop* (our own implementation of ModDrop, giving the same results as the one in Neverova *et al.* [14]) and *Gated Multimodal Unit* (GMU) (re-implemented following Arevalo *et al.* [2], our own implementation reaching slightly better performance than the original one). To assess the statistical significance of the results, 64 models with different random initialization are evaluated and we report the average performance and the standard deviation.

3.1 Experiments on Montalbano

Montalbano [6] gathers more than 14,000 samples of 20 Italian sign gesture categories. These videos were recorded with a Kinect sensor, capturing

Table 1: Overall performance on the different datasets.

Fusion meth.	Montalbano [6] Macro accuracy	MM-IMDb [2] F1-score	AFEW [5] Accuracy	Animals [7] F1-score
Best moda.	88.00±0.30	60.20±0.2	52.5±0.7	55.2
Early	97.81±0.05	63.0±0.24	40.1±1.2	50.1
Late	97.54±0.02	63.5±0.14	55.3±1.2	58.3
ModDrop [14]	98.19±0.03	62.4±0.15	53.2±0.9	
GMU [2]	97.98±0.04	63.0±0.17	55.1±0.8	58.7
CentralNet	98.27± 0.03	63.9± 0.12	57.2± 0.4	59.1

audio, skeleton joints, RGB and depth (4 modalities). The task is to recognize the gestures from the video data. The performance is measured as the macro accuracy, which is the average of the accuracy for each class. We used the features given by [14]: audio features (size 350), motion capture of the skeleton (size 350), RGB+depth left/right-hand features (size 400). Features are zero-padded to give vectors of size 400, which is possible as done after convolution layers. Fusion architecture includes 1 multilayer perceptron per modality, each having 3 layers of size: 400×128 , 128×42 , 42×21 . CentralNet connects the 3 layers of the different modalities to the central network.

We use dropout (50% dropping) and batch normalization. The learning rate starts at 0.05 and is multiplied by 0.96 at each epoch, the batch size is of 42, containing two samples of each class, and the model is trained on 100 epochs for all experiments. For ModDrop, the modality drop probability is of 0.5.

The performance obtained with each modality varies from 46% (left hand) to 88% (motion capture of skeletons). The first column of Table 1 shows that both late fusion and early fusion give significant improvement compared to the best modality, thus suggesting some complementarity between the modalities. CentralNet outperforms other approaches. In addition, Fig. 2, gives more information regarding the learned fusion strategy by visualizing the weights proportion at each layer. At the first layer (layer 0), the weights reflect the dimensionality of the modalities. At the next layer, almost no information is taken from the modalities, while at layers 2 and 3, the weight given to each modality and to the central representation are relatively similar. This may be interpreted as a hybrid fusion strategy, mixing "early" and "late" fusions.

3.2 Experiments on MM-IMDb

The MM-IMDb dataset [2] contains 25,959 samples. One sample is composed of a movie plot, its associated poster and its genres (e.g. Drama, Fantasy). The task is to predict a movie's genre based on its plot and on its poster

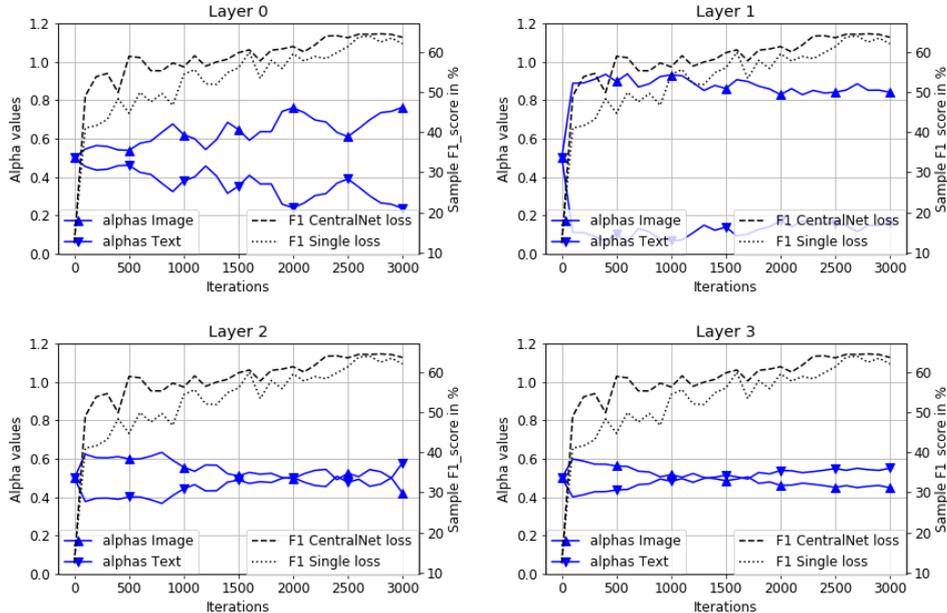


Figure 3: Evolution of α values (normalized to 1) and F1-scores (on the test set) during training on MM-IMDb

(2 modalities). One movie can belong to more than one of the 23 possible genres. The task hence has to be evaluated as a multilabel classification task. As in [2, 9], we measure performance with the micro F1 score. More details about other F1 score measures may be found in our workshop paper [20]. For these experiments, we use the features kindly provided by the authors of [2]. The visual features of size 4096 are extracted from the posters using the VGG-16 [18] network pre-trained on Imagenet. The 300-d textual ones are computed with a fine-tuned word2vec [13] encoder.

We build a multilayer perceptron on top of the features of each modality. For both modalities, the network has 3 layers of size $input_size \times 4096$, 4096×512 and 512×23 . The CentralNet architecture is the same, taking 4096-d vectors as inputs, zero-padding the textual features to reach the visual features' size.

We use dropout (50% dropping) and batch normalization. The learning rate is 0.01 and the batch size is 128. For Moddrop, the modality drop probability is of 0.25. The loss of the models is a weighted cross entropy, with a weight of 2.0 on the positive terms to balance precision and recall. More formally, the loss is:

$$loss = -y \log(2\sigma(pred)) - (1 - y) \log(1 - \sigma(pred)) \quad (2)$$

with $\sigma(pred)$ the sigmoid activation of the last output of the network and y

the multiclass label. As recommended by Arevalo *et al.* [2], we also use early stopping on the validation set.

Table 1 reports the performance measured on the test set. First of all, the worst std. dev. we observe is very small. Second, the best modality is text (60.2) and it outperforms image (47.8) by a large margin. Third, we note that the early and late fusion methods bring a large improvement and are not far from the CentralNet, which gives the best performance. Fig. 2 shows that CentralNet gives more weight to the first layers, indicating that an "earlier fusion" strategy is privileged in this case, even if the two modalities contribute significantly at all levels. Fig. 3 shows the evolution of the (normalised to 1) α s during training. We can see that the α s converge to stable values, favoring images at the first 2 layers (early fusion) and slightly favoring text in the last 2 layers (late fusion). We also see that the CentralNet loss makes the convergence faster in addition to making the performance slightly better.

3.3 Experiments on AFEW

AFEW [5] is a dataset of 1,156 video clips extracted from movies recorded by a standard video-camera. The task is to predict the emotion of a person, among seven classes. The performance is measured as the average of the accuracy on 64 runs (and standard deviation is also provided).

The audio modality is encoded as 1,583-d feature vector, obtained with the OpenSmile toolkit, while the visual modality is following the same approach as [19] and is a 512-d feature vector. CentralNet is made of a multi-layer perceptron with one hidden layer [1583 -> 128 -> 7]. To perform the first weighted sum, the visual modality is zero-padded to the size of the audio modality. The choice of this architecture with a small number of parameters can be explained by the great risk of over-fitting on such a small dataset.

The third column of Table 1 reports the performance measured on the validation set during the different experiments, as the test set is not available. The visual modality achieves an accuracy of 52.5%, while the audio accuracy is of 34.2%. Thus, the visual modality is performing best by a large margin. Early fusion brings no benefits compared to the visual modality, while all other methods are better and more stable (smaller standard deviations). CentralNet brings the largest improvement and prefers a relatively late fusion strategy, as shown in Fig. 2 where the central weights are small in the first layers.

3.4 Experiments on Audioset

Audioset [7] is a large audiovisual dataset containing YouTube videos (therefore recorded by different types of video-cameras). We only focus on a subset containing 19,842 videos of 10 seconds of different pets (named "domestic

animals, pets" in the ontology of audio set). These videos are annotated with 211 different class labels, which are not only relative to animals but also to context (e.g. "ringtone", "meow"). Therefore, the classification task is multilabel. To take this issue into account, we propose to use the micro F1-score.

Modalities are a spectrogram of the audio data, resized to 512x512 and the video sequence subsampled to 16x256x256 images. An inflated Resnet-50 is applied on the image sequence, in the same fashion as in [3]. For audio inputs, we use a model containing six convolutional layers (kernel size is 4 and stride is 2), an average two-dimensional pooling layer (7 by 7) and a fully connected layer. The first convolutional layer is outputting 64 filters, and the number of filters is doubled at each layer until having 2048 filters. The associated 8x8 maps are pooled to 1x1 and the obtained 2048 features are fed to the fully connected layer. Between each layer, we use batch normalization and reLU activation.

The CentralNet architecture is identical to the audio CNN with the first two convolutional layers removed (beginning with 256 filters). The weighted sum of the features, at a given layer, is the average of the features of the inflated Resnet-50 alongside the temporal dimension.

As training is time consuming, we made only one run. The 4th column of Table 1 reports the performance measured on this dataset. ModDrop score is not provided as the authors were not able to make it work on this dataset (convergence issues). Audio and visual models alone respectively achieve F1-scores of 55.2% and 54.8%. Late fusion is outperformed by GMU, itself outperformed by CentralNet. We note that early fusion did not bring any improvements here, mainly because the combined low-level features may not be relevant to each other. This is in line with the Fig. 2, where we observe that the central contribution is ignored in low-level layers, and progressively taken into account when getting closer to prediction level.

3.5 Comparison with other fusion methods

Table 1 shows that the proposed method outperforms 5 alternative fusion models (including the very recent approaches of [14] and [2]) on 4 different datasets. The difference is statically significant according the two sample z test, with z values between 12 and 47 for 64 runs. Additional computational cost is marginal, as the number of parameters of the central network is marginal compared to the number of parameters of the unimodal networks, and does not require any additional iterations at training time.

4 Conclusions

This article, presenting CentralNet and experimentally validating it on 4 datasets, consolidates the idea that the best fusion strategy heavily depends

on datasets and tasks. It is underlined by the fact that CentralNet is performing best on all of the 4 datasets. This is explainable by its capacity to automatically find at which layers the multimodal fusion should be performed. Moreover, the interpretation of the different fusion strategies is in line with the observations made in the literature on optimal depths for fusion. Future works will explore more flexible frameworks, using other operations than weighting sums and allowing more than one by one layer combinations.

References

- [1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Int. Conference on Machine Learning*, pages 1247–1255, 2013.
- [2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *ICLR Workshop*, 2017.
- [3] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. *Computer Vision and Pattern Recognition (CVPR)(To appear)*, 3, 2018.
- [4] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and 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.
- [5] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012.
- [6] Sergio Escalera, Xavier Baró, Jordi Gonzalez, Miguel A Bautista, Meysam Madadi, Miguel Reyes, Víctor Ponce-López, Hugo J Escalante, Jamie Shotton, and Isabelle Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *Workshop at the European Conference on Computer Vision*, pages 459–473. Springer, 2014.
- [7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 776–780. IEEE, 2017.

- [8] Miao Kang, Kefeng Ji, Xiangguang Leng, and Zhao Lin. Contextual region-based convolutional neural network with multilayer fusion for sar ship detection. *Remote Sensing*, 9(8):860, 2017.
- [9] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. *CoRR*, abs/1802.02892, 2018.
- [10] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In *International Conference on Multimodal Interaction*, pages 529–535. ACM, 2017.
- [11] Dana Lahat, Tülay Adalı, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. of the IEEE*, 103(9):1449–1477, 2015.
- [12] Fan Li, Natalia Neverova, Christian Wolf, and Graham Taylor. Modout: Learning to fuse modalities via stochastic regularization. *Journal of Computational Vision and Imaging Systems*, 2(1), 2016.
- [13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [14] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.
- [15] Natalia Neverova, Christian Wolf, Graham W Taylor, and Florian Nebout. Multi-scale deep learning for gesture detection and localization. In *Workshop at the European conference on computer vision*, pages 474–490. Springer, 2014.
- [16] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, and Maja Pantic. Summary for avec 2017: Real-life depression and affect challenge and workshop. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1963–1964, 2017.
- [17] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
- [18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

- [19] Valentin Vielzeuf, Corentin Kervadec, Stéphane Pateux, Alexis Lechervy, and Frédéric Jurie. An occam’s razor view on learning audiovisual emotion recognition with small training sets. In *International Conference on Multimodal Interaction*, pages 589–593. ACM, 2018.
- [20] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *ECCV Workshops*, 2018.
- [21] Valentin Vielzeuf, Stéphane Pateux, and 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.
- [22] Zhe Wang, Kingsley Kuan, Mathieu Ravaut, Gaurav Manek, Sibong Song, Fang Yuan, Kim Seokhwan, Nancy Chen, Luis Fernando D’Haro-Enríquez, Anh Tuan Luu, Hongyuan Zhu, Zeng Zeng, Ngai-Man Cheung, Georgios Piliouras, Jie Lin, and Vijay Chandrasekhar. Truly multi-modal youtube-8m video classification with video, audio, and text. *CoRR*, abs/1706.05461, 2017.
- [23] Xiaodong Yang, Pavlo Molchanov, and 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.