[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}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. 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.