Lessons learned in multilingual grounded language learning

Ákos Kádár
Tilburg University
a.kadar@uvt.nl

Desmond Elliott*
University of Copenhagen
de@di.ku.dk

Marc-Alexandre Côté
Microsoft Research Montreal
macote@microsoft.com

Grzegorz Chrupała
Tilburg University
g.chrupala@uvt.nl

Afra Alishahi
Tilburg University
a.alishahi@uvt.nl

1 Introduction

Multimodal representation learning is largely motivated by evidence of perceptual grounding in human concept acquisition and representation (Barsalou et al., 2003). It has been shown that visually grounded word and sentence-representations (Kiela et al., 2014; Baroni, 2016; Elliott and Kádár, 2017; Kiela et al., 2017; Yoo et al., 2017) improve performance on the downstream tasks of paraphrase identification, semantic entailment, and multimodal machine translation (Dolan et al., 2004; Marelli et al., 2014; Specia et al., 2016). Multilingual sentence representations have also been successfully applied to many-languages-to-one character-level machine translation (Chung et al., 2016) and multilingual dependency parsing (Ammar et al., 2016).

Recently, Gella et al. (2017) proposed to learn both bilingual and multimodal sentence representations using images paired with captions independently collected in English and German. Their results show that bilingual training improves image-sentence ranking performance over a monolingual baseline, and it improves performance on semantic textual similarity benchmarks (Agirre et al., 2014, 2015). These findings suggest that it may be beneficial to consider another language as another *modality* in a monolingual grounded language learning model. In the grounded learning scenario, descriptions of an image in multiple languages can be considered as multiple views of the same or closely related data. These additional views can help overcome the problems of data sparsity, and have practical implications for efficiently collecting image-text datasets in different languages. In real-life applications, many tasks and domains can involve code switching (Barman et al., 2014), which is easier to deal with using a multilingual model. Furthermore, it is more convenient to maintain a single multilingual system than one system for each considered language. However, there is a need for a systematic exploration of the conditions under which it is useful to add additional views of the data. We investigate the impact of the following conditions on the performance of a multilingual grounded language learning model in sentence and image retrieval tasks:

Additional languages. Multilingual models have not been explored yet in a multimodal setting. We investigate the contribution of adding more than one language by performing bilingual experiments on English and German as well as adding French and Czech captioned images.

Data alignment: We assess the performance of a multilingual models trained using either captions that are translations of each other, or captions that are independently collected in different languages for the same set of images. Additionally we consider the setup when non-overlapping sets of images and their captions are collected in different languages. Such disjoint settings have been explored in pivot-based multimodal representation learning (Funaki and Nakayama, 2015; Rajendran et al., 2015) or zero-shot multi-modal machine translation (Nakayama and Nishida, 2017). We compare translated vs. independently collected captions and overlapping vs. disjoint images.

High-to-low resource transfer: We investigate whether low-resource languages benefit from jointly training on larger data sets from higher-resource languages. This type of transfer has previously been shown to be effective in machine translation (e.g., Zoph et al., 2016).

*Work carried out at the University of Edinburgh.

Training objective: In addition to learning to map images to sentences, we study the effect of also learning relationships between captions of the same image in different languages Gella et al. (2017). We assess the contribution of such a caption–caption ranking objective throughout our experiments.

Our results show that multilingual joint training improves upon bilingual joint training, and that grounded sentence representations for a low-resource language can be substantially improved with data from different high-resource languages. Our results suggest that independently-collected captions are more useful than translated captions, for the task of learning multilingual multimodal sentence embeddings. Finally, we recommend to collect captions for the same set of images in multiple languages, due to the benefits of the additional caption–caption ranking objective function.

2 Highlight Results

2.1 Translation vs. independent captions

Table 1the translations vs. comparable captions experiment with data in four languages. The Multi-translation models are trained on 29K images paired with a single caption in each language. These models perform better than their Monolingual counterparts, and the German, French, and Czech models are further improved with the c2c objective. The Multi-comparable models are trained by randomly sampling one English and one German caption from the *comparable* dataset, alongside the French and Czech translation pairs. These models perform as well as the Multi-translation models, and the c2c objective brings further improvements for all languages in this setting. These results clearly demonstrate the advantage of jointly training on more than two languages. Text-to-image retrieval performance increases by more than 11 R@10 points for each of the four languages in our experiment.

2.2 High-to-low resource transfer

We now examine whether the lower-resource French and Czech models benefit from training with the full complement of the higher-resource English and German comparable data. Therefore we train a joint model on the *translation* as well as *comparable* portions of Multi30K, and examine the performance on French and Czech.

	En	De	Fr	Cz
Monolingual	50.4	39.5	47.0	42.0
Multi-translation	58.7	51.2	57.0	51.0
+ c2c	56.3	52.2	55.0	51.6
Multi-comparable	59.2	49.6	57.2	50.8
+ c2c	**61.8**	**52.7**	**59.2**	**55.2**

Table 1: The Monolingual and joint Multi-translation models trained on *translation pairs*, and the Multi-comparable trained on the downsampled *comparable* set with one caption per image.

	French	Czech
Monolingual	47.0	42.0
Multilingual	56.3	51.3
+ Comparable	58.9	52.4
+ c2c	**61.6**	**57.2**

Table 2: Multilingual is trained on all *translation pairs*, + Comparable adds the *comparable* data set.

Table 2 shows the results of this experiment. We find that the French and Czech models improve by 8.8 and 5.5 R@10 points respectively when they are only trained on the multilingual translation pairs (compared to the monolingual version), and by another 2.2 and 2.8 points if trained on the extra 155K English and German *comparable* descriptions. We also find that the additional c2c objective improves the Czech model by a further 4.8 R@10 points (this improvement is likely caused by training the model on 46 possible caption pairs). Our results show the impact of jointly training with the larger English and German resources, which demonstrates the benefits of high-to-low resource transfer.

References

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., et al. (2015). Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau,

G., and Wiebe, J. (2014). Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *arXiv preprint arXiv:1602.01595*.

Barman, U., Das, A., Wagner, J., and Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., and Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91.

Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.

Elliott, D. and Kádár, A. (2017). Imagination improves multimodal translation. *arXiv preprint arXiv:1705.04350*.

Funaki, R. and Nakayama, H. (2015). Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 585–590.

Gella, S., Sennrich, R., Keller, F., and Lapata, M. (2017). Image pivoting for learning multilingual multimodal representations. *arXiv preprint arXiv:1707.07601*.

Kiela, D., Conneau, A., Jabri, A., and Nickel, M. (2017). Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*.

Kiela, D., Hill, F., Korhonen, A., and Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 835–841.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Nakayama, H. and Nishida, N. (2017). Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64.

Rajendran, J., Khapra, M. M., Chandar, S., and Ravindran, B. (2015). Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519*.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Yoo, K. M., Shin, Y., and Lee, S.-g. (2017). Improving visually grounded sentence representations with self-attention. *arXiv preprint arXiv:1712.00609*.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.