Limitations of cross-lingual learning from image search

Mareike Hartmann
Department of Computer Science
University of Copenhagen
Denmark
hartmann@di.ku.dk

Anders Søgaard
Department of Computer Science
University of Copenhagen
Denmark
soegaard@di.ku.dk

Abstract

Cross-lingual representation learning is an important step in making NLP scale to all the world's languages. Previous work on bilingual lexicon induction suggests that it is possible to learn cross-lingual representations of words based on similarities between images associated with these words. However, that work focused (almost exclusively) on the translation of nouns only. Here, we investigate whether the meaning of other parts-of-speech (POS), in particular adjectives and verbs, can be learned in the same way. Our experiments across five language pairs indicate that previous work does not scale to the problem of learning cross-lingual representations beyond simple nouns.

1 Introduction

Typically, cross-lingual word representations are learned from word alignments, sentence alignments, from aligned, comparable documents (Levy et al., 2017), or from monolingual corpora using seed dictionaries (Ammar et al., 2016).[1] However, for many languages such resources are not available.

Bergsma and Van Durme (2011) introduced an alternative idea, namely to learn bilingual representations from image data collected via web image search. The idea behind their approach is to represent words in a visual space and find valid translations between words based on similarities between their visual representations. Representations of words in the visual space are built by representing a word by a set of images that are associated with that word, i.e., the word is a semantic tag for the images in the set.

Kiela et al. (2015) improve performance for the same task using a feature representation extracted from convolutional networks. However, both works only consider nouns, leaving open the question of whether learning cross-lingual representations for other POS from images is possible.[2]

In order to evaluate whether this work scales to verbs and adjectives, we compile wordlists containing these POS in several languages. We collect image sets for each image word and represent all words in a visual space. Then, we rank translations computing similarities between image sets and evaluate performance on this task.

Another field of research that exploits image data for NLP applications is the induction of multi-modal embeddings, i.e. semantic representations that are learned from textual and visual information jointly (Kiela et al., 2014; Hill and Korhonen, 2014; Kiela and Bottou, 2014; Lazaridou et al., 2015; Silberer et al., 2017; Kiela et al., 2016; Vulić et al., 2016). The work presented in our paper differs from these approaches, in that we do not use image data to improve semantic representations, but use images as a resource to learn cross-lingual representations. Even though lexicon induction from text resources might be more promising in terms of performance, we think that lexicon induction from visual data is worth exploring as it might give insights in the way that language is grounded in visual context.

[1] Recent work by Lample et al. (2018) introduces unsupervised bilingual lexicon induction from monolingual corpora, however, it was shown that this approach has important limitations (Søgaard et al., 2018).

[2] Kiela et al. (2016) induce English-Italian word translations from image data for the Simlex-999 dataset which contains adjectives and verbs, but they do not evaluate the performance for these POS compared to nouns.

Proceedings of the 3rd Workshop on Representation Learning for NLP, pages 159–163
Melbourne, Australia, July 20, 2018.

1.1 Contributions

We evaluate the approaches by Bergsma and Van Durme (2011) and Kiela et al. (2015) on an extended data set, which apart from nouns includes both adjectives and verbs. Our results suggest that none of the approaches involving image data are directly applicable to learning cross-lingual representations for adjectives and verbs.

2 Data

Wordlists We combined 3 data sets of English words to compile the wordlists for our experiments: the original wordlist used by Kiela et al. (2015), the Simlex-999 data set of English word pairs (Hill et al., 2014) and the MEN data set (Bruni et al., 2014). Whereas the first wordlist contains only nouns, the latter two datasets contain words of three POS classes (nouns, adjectives and verbs). We collect all distinct words and translate the final wordlist into 5 languages (German, French, Russian, Italian, Spanish) using the Google translation API[3], choosing the most frequent translation with the respective POS tag. Table 1 shows the POS distribution in the datasets.

	MEN	Simlex	Bergsma	Combined
N	656	751	500	1406
V	38	170	0	206
A	57	107	0	159

Table 1: Distribution of POS tags in the datasets used to compile the final wordlist.

Image Data Sets We use the Google Custom Search API[4] to represent each word in a wordlist by a set of images. We collect the first 50 jpeg images returned by the search engine when querying the words specifying the target language.[5] This way, we compile image data sets for 6 languages.[6] Figure 1 shows examples for images associated with a word in two languages.

[3] https://translate.google.com/

[4] https://developers.google.com/custom-search/

[5] Even though we get the search results for the first 50 images, some of them cannot be downloaded. On average, we collect 42 images for each image word.

[6] The wordlists and image datasets are available at https://github.com/coastalcph/cldi_from_image_search/

3 Approach

The assumption underlying the approach is that semantically similar words in two languages are associated with similar images. Hence, in order to find the translation of a word, e.g. from English to German, we compare the images representing the English word with all the images representing German words, and pick as translation the German word represented by the most similar images. To compute similarities between images, we compute cosine similarities between their feature representations.

3.1 Convolutional Neural Network Feature Representations

Following Kiela et al. (2015), we compute convolutional neural network (CNN) feature representations using a model pre-trained on the ImageNet classification task (Russakovsky et al., 2015). For each image, we extract the pre-softmax layer representation of the CNN. Instead of an AlexNet (Krizhevsky et al., 2012) as used by Kiela et al. (2015), we use the Keras implementation of the VGG19 model as described in Simonyan and Zisserman (2014), which was shown to achieve similar performance for word representation tasks by Kiela et al. (2016). Using this model, we represent each image by a 4069-dimensional feature vector.

Similarities Between Individual Images Bergsma and Van Durme (2011) determine similarities between image sets based on similarities between all individual images. For each image in image set 1, the maximum similarity score for any image in image set 2 is computed. These maximum similarity scores are then either averaged (AVGMAX) or their maximum is taken (MAXMAX).

Similarities Between Aggregated Representations In addition to the above described methods, Kiela et al. (2015) generate an aggregated representation for each image set and then compute the similarity between image sets by computing the similarity between the aggregated representations. Aggregated representations for image sets are generated by either taking the component-wise average (CNN-MEAN) or the component-wise maximum (CNN-MAX) of all images in the set.

K-Nearest Neighbor For each image in an image set in language 1, we compute its nearest

(a) Images associated with the English noun *cow* (left) and the German translation *Kuh* (right).

(b) Images associated with the English verb *discuss* (left) and the German translation *diskutieren* (right).

(c) Images associated with the English adjective *sad* (left) and the German translation *traurig* (right).

Figure 1: Examples for images associated with equivalent words in two languages (English and German).

neighbor across all image sets in language 2. Then, we find the image set in language 2 that contains the highest number of nearest neighbors. The image word is translated into the image word that is associated with that image 2 set. Ties between image sets containing an equivalent number of nearest neighbors are broken by computing the average distance between all members. We refer to the method as KNN. Whereas the other approaches described above provide a ranking of translations, this method determines only the one translation that is associated with the most similar image set.

Clustering Image Sets As we expect the retrieved image sets for a word to contain images associated with different senses of the word, we first cluster images into k clusters. This way, we hope to group images representing different word senses. Then, we apply the KNN method as described above. We refer to this method as KNN-C.

3.2 Evaluation Metrics

Ranking performance is evaluated by computing the Mean Reciprocal Rank (MRR) as $MRR = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{rank(w_s, w_t)}$ M is the number of words to be translated and $rank(w_s, w_t)$ is the position the correct translation w_t for source word w_s is ranked on.

In addition to MRR, we also evaluate the cross-lingual representations by means of precision at k (P@k).

4 Experiments and Results

We run experiments for 5 language pairs English–German, English–Spanish, English–French, English–Russian and English–Italian. We evaluate the representations computed from image data and compare the different methods for similarity computation described in 3. For each English word, we rank all the words in the corresponding target languages based on similarities between image sets and evaluate the models' ability to identify correct translations, i.e. to rank the correct translation on a position near the top. We compare 4 settings that differ in the set of English words that are translated. In the setting ALL, all English words in the wordlist are translated. NN, VB and ADJ refer to the settings where only nouns, verbs and adjectives are translated.

4.1 Results

Comparison of similarity computation methods for visual representations Table 2 displays results averaged over all language pairs.[7] First, comparing the different methods to compute similarities between image sets, AVGMAX outperforms the other methods in almost all cases. Most importantly, we witness a very significant drop in performance when moving from nouns to verbs and adjectives. For verbs, we rarely pick the right translation based on the image-based word representations. This behavior applies across all methods for similarity computation. Further, we see small improvements if we cluster the image sets prior to applying the KNN method, which might indicate that the clustering helps in finding translations for polysemous words.

4.2 Analysis

If we try to learn translations from images, integrating verbs and adjectives into the dataset worsens results compared to a dataset that contains only nouns. One possible explanation is that images associated with verbs and adjectives are less suited to represent the meaning of a concept than images associated with nouns.

Kiela et al. (2015) suppose that lexicon induction via image similarity performs worse for

[7] We also evaluate our visual representations on the set of 500 nouns used by Kiela et al. (2015), which results in P@1=0.6 and MRR=0.63 averaged over 5 language pairs for the AVGMAX method.

	ALL			NN			VB			ADJ		
	MRR	P@1	P@10	MRR	P@1	P@10	MRR	P@1	P@10	MRR	P@1	P@10
AVGMAX	0.53	0.49	0.60	0.60	0.56	0.67	0.20	0.15	0.30	0.28	0.22	0.37
MAXMAX	0.44	0.38	0.54	0.49	0.43	0.61	0.19	0.15	0.24	0.23	0.18	0.31
CNNMEAN	0.49	0.44	0.57	0.56	0.52	0.64	0.15	0.10	0.26	0.24	0.20	0.32
CNNMAX	0.47	0.43	0.55	0.55	0.50	0.63	0.15	0.10	0.24	0.19	0.15	0.27
KNN	–	0.42	–	–	0.50	–	–	0.06	–	–	0.13	–
KNN-C ($k = 3$)	–	0.47	–	–	0.56	–	–	0.10	–	–	0.16	–

Table 2: Results for translation ranking with images represented by CNN features averaged over 5 language pairs. KNN and KNN-C do not produce a ranking, hence we only provide P@1 values.

datasets containing words that are more abstract. In order to approximate the degree of abstractness of a concept, they compute the *image dispersion* d for a word w as the average cosine distance between all image pairs in the image set $\{i_j, \ldots, i_n\}$ associated with word w according to

$$d(w) = \frac{2}{n(n-1)} \sum_{k<j\leq n} 1 - \frac{i_j \cdot i_k}{|i_j||i_k|}$$

In their analysis, Kiela et al. (2015) find that their model performs worse on datasets with a higher average image dispersion. Kiela et al. (2014) introduce a dispersion-based filtering approach for learning multi-modal representations of nouns. They show that the quality of their representations with respect to a monolingual word-similarity prediction task improves, if they include visual information only in cases where the dispersion of the visual data is low.

Computing the average image dispersion for our data across languages shows that image sets associated with verbs and adjectives have a higher average image dispersion than image sets associated with nouns (nouns: $d = 0.60$, verbs: $d = 0.68$, adjectives: $d = 0.66$).

Table 3 shows the image words associated with the image sets that have the highest and lowest dispersion values in the English image data. For nouns and adjectives, we observe that the words with lowest dispersion values express concrete concepts, whereas the words with highest dispersion values express more abstract concepts that can be displayed in many variants. Manually inspecting the dataset, we find e.g. that the images associated with the noun *animal* display many different animals, such as birds, dogs, etc, whereas the images for *mug* all show a prototypical mug.

Besides the dispersion values, we also analyze the number of word senses per POS using WordNet[8]. We find that the verbs in our dataset have a higher average number of word senses ($n = 9.18$) than the adjectives ($n = 6.88$) and the nouns ($n = 5.08$). That we get worst results for the words with highest number of different word senses is in agreement with Gerz et al. (2016), who find that in a monolingual word similarity prediction task, models perform worse for verbs with more different senses than for less polysemous verbs.

	Lowest dispersion		Highest dispersion	
	Word	d	Word	d
NN	mug	0.31	animal	0.78
	oscilloscope	0.32	companion	0.78
	padlock	0.33	mammal	0.78
VB	vanish	0.43	differ	0.76
	shed	0.43	hang	0.76
	divide	0.47	arrange	0.75
ADJ	yellow	0.39	huge	0.79
	white	0.40	large	0.79
	fragile	0.43	big	0.78

Table 3: English image words associated with the image sets with highest and lowest dispersion scores d.

5 Conclusion

We showed that existing work on learning cross-lingual word representations from images obtained via web image search does not scale to other POS than nouns. It is possible that training convolutional networks on different resources than ImageNet data will provide better features represent-

[8] https://wordnet.princeton.edu/

ing verbs and adjectives. Finally, it would be interesting to extend the approach to multi-modal input, combining images and texts, e.g. from comparable corpora with images such as Wikipedia.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR* abs/1602.01925.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49(1):1–47.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *Proceedings of EMNLP*.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of EMNLP*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *CoRR* abs/1408.3456. http://arxiv.org/abs/1408.3456.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*. Baltimore, Maryland.

Douwe Kiela, Anita Lilla Verő, and Stephen Clark. 2016. Comparing data sources and architectures for deep visual representation learning in semantics. In *Proceedings of EMNLP*.

Douwe Kiela, Ivan Vulic, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of EMNLP*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of NIPS*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018. Word translation without parallel data. In *Proceedings of ICLR*.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL HLT*.

Omer Levy, Yoav Goldberg, and Anders Søgaard. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of EACL*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3).

Shane Bergsma and Benjamin Van Durme. 2011. Learning Bilingual Lexicons using the Visual Similarity of Labeled Web Images. In *Proceedings of IJCAI*. Barcelona, Spain, IJCAI '11.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2017. Visually grounded meaning representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.

K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL*.

Ivan Vulić, Douwe Kiela, S. Clark, and M.F Moens. 2016. Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of ACL*.