Learning Translations via Images with a Massively Multilingual Image Dataset

John Hewitt* **Daphne Ippolito*** **Brendan Callahan** **Reno Kriz**
Derry Wijaya **Chris Callison-Burch**
University of Pennsylvania
Computer and Information Science Department
{johnhew,daphnei,rekriz,derry,ccb}@seas.upenn.edu

Abstract

We conduct the most comprehensive study to date into translating words via images. To facilitate research on the task, we introduce a large-scale multilingual corpus of images, each labeled with the word it represents. Past datasets have been limited to only a few high-resource languages and unrealistically easy translation settings. In contrast, we have collected by far the largest available dataset for this task, with images for approximately 10,000 words in each of 100 languages. We run experiments on a dozen high resource languages and 20 low resources languages, demonstrating the effect of word concreteness and part-of-speech on translation quality. To improve image-based translation, we introduce a novel method of predicting word concreteness from images, which improves on a previous state-of-the-art unsupervised technique. This allows us to predict when image-based translation may be effective, enabling consistent improvements to a state-of-the-art text-based word translation system. Our code and the Massively Multilingual Image Dataset (MMID) are available at http://multilingual-images.org/.

1 Introduction

Learning the translations of words is important for machine translation and other tasks in natural language processing. Typically this learning is done using sentence-aligned bilingual parallel texts. However, for many languages, there are not sufficiently large parallel texts to effectively learn translations. In this paper, we explore the question of whether it is possible to learn translations with images. We systematically explore an idea originally proposed by Bergsma and Van Durme (2011): translations can be identified via images associated with words in different languages that have a high degree of visual similarity. This is illustrated in Figure 1.

Figure 1: Our dataset and approach allow translations to be discovered by comparing the images associated with foreign and English words. Shown here are five images for the Indonesian word *kucing*, a word with high predicted concreteness, along with its top 4 ranked translations using CNN features.

Most previous image datasets compiled for the task of learning translations were limited to the translation of nouns in a few high-resource languages. In this work, we present a new large-scale dataset that contains images for 100 languages, and is not restricted by part-of-speech. We collected images using Google Image Search for up to 10,000 words in each of 100 foreign languages, and their English translations. For each word, we collected up to 100 images and the text on images' corresponding web pages.

We conduct a broad range of experiments to evaluate the utility of image features across a number of factors:

*These authors contributed equally; listed alphabetically.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 2566–2576
Melbourne, Australia, July 15 - 20, 2018.

- We evaluate on 12 high-resource and 20 low-resource languages.
- We evaluate translation quality stratified by part-of-speech, finding that nouns and adjectives are translated with much higher accuracy than adverbs and verbs.
- We present a novel method for predicting word concreteness from image features that better correlates with human perception than existing methods. We show that choosing concrete subsets of words to translate results in higher accuracy.
- We augment a state-of-the-art text-based word translation system with image feature scores and find consistent improvements to the text-only system, ranging from 3.12% absolute top-1 accuracy improvement at 10% recall to 1.30% absolute improvement at 100% recall.

A further contribution of this paper is our dataset, which is the largest of its kind and should be a standard for future work in learning translations from images. The dataset may facilitate research into multilingual, multimodal models, and translation of low-resource languages.

2 Related Work

The task of learning translations without sentence-aligned bilingual parallel texts is often called bilingual lexicon induction (Rapp, 1999; Fung and Yee, 1998). Most work in bilingual lexicon induction has focused on text-based methods. Some researchers have used similar spellings across related languages to find potential translations (Koehn and Knight, 2002; Haghighi et al., 2008). Others have exploited temporal similarity of word frequencies to induce translation pairs (Schafer and Yarowsky, 2002; Klementiev and Roth, 2006). Irvine and Callison-Burch (2017) provide a systematic study of different text-based features used for bilingual lexicon induction. Recent work has focused on building joint distributional word embedding spaces for multiple languages, leveraging a range of levels of language supervision from bilingual dictionaries to comparable texts (Vulić and Korhonen, 2016; Wijaya et al., 2017).

The most closely related work to ours is research into bilingual lexicon induction using image similarity by Bergsma and Van Durme (2011) and Kiela et al. (2015). Their work differs from ours in that they focused more narrowly on the translation of nouns for a limited number of high resource languages. Bergsma and Van Durme (2011) compiled datasets for Dutch, English, French, German, Italian, and Spanish by downloading 20 images for up to 500 concrete nouns in each of the foreign languages, and 20,000 English words.

Another dataset was generated by Vulic and Moens (2013) who collected images for 1,000 words in Spanish, Italian, and Dutch, along with the English translations for each. Their dataset also consists of only nouns, but includes abstract nouns. Our corpus will allow researchers to explore image similarity for bilingual lexicon induction on a much wider range of languages and parts of speech, which is especially desirable given the potential utility of the method for improving translation between languages with little parallel text.

The ability of images to usefully represent a word is strongly dependent on how concrete or abstract the word is. The terms *abstractness* and *concreteness* are used in the psycholinguistics and cognitive psychology literature. *Concrete* words directly reference a sense experience (Paivio et al., 1968), while abstract words can denote ideas, emotions, feelings, qualities or other abstract or intangible concepts. Concreteness ratings are closely correlated with *imagery* ratings, defined as the ease with which a word arouses a mental image (Gilhooly and Logie, 1980; Friendly et al., 1982). Intuitively, concrete words are easier to represent visually, so a measure of a word's concreteness ought to be able to predict the effectiveness of using images to translate the word.

Kiela et al. (2014) defines an unsupervised method called *image dispersion* that approximates a word's concreteness by taking the average pairwise cosine distance of a set of image representations of the word. Kiela et al. (2015) show that image dispersion helps predict the usefulness of image representations for translation. In this paper, we introduce novel supervised approaches for predicting word concreteness from image and textual features. We make use of a dataset created by Brysbaert et al. (2014) containing human evaluations of concreteness for 39,954 English words.

Concurrently with our work, Hartmann and Søgaard (2017) released an unpublished arXiv draft challenging the efficacy of using images for translation. Their work presents several difficulties of using image features for translation, difficulties which

our methods address. They find that image features are only useful in translating simple nouns. While we did indeed find that nouns perform better than other parts of speech, we do not find that images are only effective in translating simple words. Instead, we show a gradual degradation in performance as words become more abstract. Their dataset is restricted to six high-resource languages and a small vocabulary of 557 English words. In contrast, we present results for over 260,000 English words and 32 foreign languages.

Recent research in the NLP and computer vision communities has been enabled by large collections of images associated with words or longer texts. Object recognition has seen dramatic gains in part due to the ImageNet database (Deng et al., 2009), which contains 500-1000 images associated with 80,000 synsets in WordNet. Ferraro et al. (2015) surveys existing corpora that are used in vision and language research. Other NLP+Vision tasks that have been enabled by the availability of large datasets include caption generation for images, action recognition in videos, visual question answering, and others.

Most existing work on multilingual NLP+Vision relies on having a corpus of images manually annotated with captions in several languages, as in the Multi30K dataset (Elliott et al., 2016). Several works have proposed using image features to improve sentence level translations or to translate image captions (Gella et al., 2017; Hitschler and Riezler, 2016; Miyazaki and Shimizu, 2016). Funaki and Nakayama (2015) show that automatically scraped data from websites in English and Japanese can be used to effectively perform zero-shot learning for the task of cross-lingual document retrieval. Since collecting multilingual annotations is difficult at a large-scale or for low-resource languages, our approach relies only on data scraped automatically from the web.

3 Corpus Construction

We present a new dataset for image-based word translation that is more expansive than any previous ones, encompassing all parts-of-speech, the gamut of abstract to concrete, and both low- and high-resource languages.

3.1 Dictionaries

We collect images for words in 100 bilingual dictionaries created by Pavlick et al. (2014). They selected the 10,000 most frequent words on Wikipedia pages in the foreign language, and then collected their translations into English via crowdsourcing. We will denote these dictionaries as CROWDTRANS. The superset of English translations for all foreign words consists of 263,102 translations. The English portion of their data tends to be much noisier than the foreign portion due to its crowdsourced nature (e.g. misspellings, or definition included with translations.)

We computed part-of-speech for entries in each dictionary. We found that while nouns are the most common, other parts-of-speech are reasonably represented (Section 5.1).

3.2 Method

For each English and foreign word, we query Google Image Search to collect 100 images associated with the word. A potential criticism of our use of Google Image Search is that it may be using a bilingual dictionary to translate queries into English (or other high resource languages) and returning images associated with the translated queries (Kilgarriff, 2007). We take steps (Section 3.3) to filter out images that did not appear on pages written in the language that we are gathering images for. After assembling the collection of images associated with words, we construct low-dimensional vector representations of the images using convolutional neural networks (CNNs). We also save the text from each web page that an image appeared on. Further detail on our corpus construction pipeline can be found in Section 2 of the supplemental materials.

3.3 Filtering by Web Page Language

We used the following heuristic to filter images: if text could be extracted from an image's web page, and the expected language was in the top-3 most likely languages output by the CLD2[1] language detection system then we kept the image; otherwise it was discarded. This does not filter all images from webpages with English text; instead it acknowledges the presence of English in the multilingual web and keeps images from pages with some target-language presence. An average of approximately 42% of images for each foreign language remained after the language-filtering step.

[1] https://github.com/CLD2Owners/cld2

Language	Concreteness Ratings 1-2	2-3	3-4	4-5	Overall
English	.804	.814	.855	.913	.857
French	.622	.653	.706	.828	.721
Indonesian	.505	.569	.665	.785	.661
Uzbek	.568	.530	.594	.683	.601
All	.628	.649	.713	.810	.717
# Words	77	292	292	302	963

Table 1: The proportion of images determined to be good representations of their corresponding word. In columns 2-5, we bucket the results by the word's ground-truth concreteness, while column 6 shows the results over all words. The last row shows the number of words in each bucket of concreteness, and the number of words overall for each language.

3.4 Manual Evaluation of Images

By using a dataset scraped from the web, we expect some fraction of the images for each word to be incorrectly labeled. To confirm the overall quality of our dataset, we asked human evaluators on Amazon Mechanical Turk to label a subset of the images returned by queries in four languages: our target language, English; a representative high-resource language, French; and two low-resource languages, Indonesian and Uzbek. In total, we collected 36,050 judgments of whether the images returned by Google Image Search were a good match for the keyword. Details on the experimental setup can be found in Section 1 of the Supplemental Materials.

Table 1 shows the fraction of images that were judged to be good representations of the search word. It also demonstrates that as the concreteness of a word increases, the proportion of good images associated with that word increases as well. We further discuss the role of concreteness in Section 6.1. Overall, 85% of the English images, 72% of French, 66% of Indonesian, and 60% of Uzbek were judged to be good.

4 Finding Translations Using Images

Can images help us learn translations for low-resource languages? In this section we replicate prior work in high-resource languages, and then evaluate on a wide array of low-resource languages.

Although we scraped images and text for 100 languages, we have selected a representative set of 32 for evaluation. Kiela et al. (2015) established that CNN features are superior to the SIFT plus color histogram features used by Bergsma and Van Durme (2011), and so we restrict all analysis to the former.

4.1 Translation Prediction with AvgMax

To learn the English translation of each foreign word, we rank the English words as candidate translations based on their visual similarity with the foreign words. We take the cosine similarity score for each image i_f associated the foreign word w_f with each of image i_e for the English word w_e, and then compute the average maximum similarity as

$$\text{AvgMax}(w_f, w_e) = \frac{1}{|w_f|} \sum_{i_f \in w_f} \max_{i_e \in w_e} (cosine(i_f, i_e))$$

Each image is represented by a 4096-dimensional vector from the fully connected 7th (FC7) layer of a CNN trained on ImageNet (Krizhevsky et al., 2012). AvgMax is the best-performing method described by Bergsma and Van Durme (2011) on images created with SIFT and color histogram features. It was later validated on CNN features by Kiela et al. (2015).

The number of candidate English words is the number of entries in the bilingual dictionary after filtering out dictionary entries where the English word and foreign word are identical. In order to compare with Kiela et al. (2015), we evaluate the models' rankings using Mean Reciprocal Rank (MRR), top-1, top-5 and top-20 accuracy. We prefer the more interpretable top-k accuracy in our subsequent experiments. We choose to follow Wijaya et al. (2017) in standardizing to $k = 10$, and we report top-1 accuracy only when it is particularly informative.

4.2 Replication of Prior Work

We evaluate on the five languages–Dutch, French, German, Italian, and Spanish–which have been the focus of prior work. Table 2 shows the results reported by Kiela et al. (2015) on the BERGSMA500 dataset, along with results using our image crawl method (Section 3.2) on BERGSMA500's vocabulary.

On all five languages, our dataset performs better than that of Kiela et al. (2015). We attribute this to improvements in image search since they collected images. We additionally note that in the BERGSMA500 vocabularies, approximately 11% of the translation pairs are string-identical, like *film* ↔ *film*. In all subsequent experiments, we remove trivial translation pairs like this.

We also evaluate the identical model on our full data set, which contains 8,500 words, covering all parts of speech and the full range of concreteness ratings. The top-1 accuracy of the model is 23% on

our more realistic and challenging data set, versus 68% on the easier concrete nouns set.

4.3 High- and Low-resource Languages

To determine whether image-based translation is effective for low resource languages, we sample 12 high-resource languages (HIGHRES), and 20 low-resource languages (LOWRES). Table 3 reports the top-10 accuracy across all 32 languages.

For each language, we predict a translation for each foreign word in the language's CROWDTRANS dictionary. This comes to approximately 7,000 to 10,000 foreign words per language. We find that high-resource languages' image features are more predictive of translation than those of low-resource languages. Top-10 accuracy is 29% averaged across high-resource languages, but only 16% for low-resource languages. This may be due to the quality of image search in each language, and the number of websites in each language indexed by Google, as suggested by Table 1. The difficulty of the translation task is dependent on the size of the English vocabulary used, as distinguishing between $5,000$ English candidates as in Slovak is not as difficult as distinguishing between $10,000$ words as in Tamil.

4.4 Large Target Vocabulary

How does increasing the number of candidate translations affect accuracy? Prior work used an English vocabulary of 500 or 1,000 words, where the correct English translation is guaranteed to appear. This is unrealistic for many tasks such as machine translation, where the target language vocabulary is likely to be large. To evaluate a more realistic scenario, we take the union of the English vocabulary of every dictionary in CROWDTRANS, and run the same translation experiments as before. We call this large common vocabulary LARGEENG.

Confirming our intuition, experiments with LARGEENG give significantly lower top-10 accuracies across parts of speech, but still provide discriminative power. We find .181 average top-10 accuracy using LARGEENG, whereas on the same languages, average accuracy on the CROWDTRANS vocabularies was .260. The full results for these experiments are reported in Table 4.

5 Evaluation by Part-of-speech

Can images be used to translate words other than nouns? This section presents our methods for de-

dataset	**BERGSMA500** Kiela et al. (2015)	**BERGSMA500** (ours)	**all** (ours)
# words	500	500	8,500
MRR	0.658	0.704	0.277
Top 1	0.567	0.679	0.229
Top 5	0.692	0.763	0.326
Top 20	0.774	0.811	0.385

Table 2: Our results are consistently better than those reported by Kiela et al. (2015), averaged over Dutch, French, German, Italian, and Spanish on a similar set of 500 concrete nouns. The rightmost column shows the added challenge with our larger, more realistic dataset.

HIGHRES	**All**	**VB**	**RB**	**JJ**	**NN**	**#**
Spanish	.417	.144	.157	.329	.593	9.9k
French	.366	.104	.107	.315	.520	10.5k
Dutch	.365	.085	.064	.262	.511	10.5k
Italian	.323	.086	.085	.233	.487	8.9k
German	.307	.071	.098	.164	.463	10.1k
Swedish	.283	.048	.048	.146	.328	9.6k
Turkish	.263	.035	.143	.233	.346	10.2k
Romanian	.255	.029	.080	.150	.301	9.1k
Hungarian	.240	.030	.082	.193	.352	10.9k
Bulgarian	.236	.024	.106	.116	.372	8.6k
Arabic	.223	.036	.084	.149	.344	10.2k
Serbian	.218	.023	.111	.090	.315	8.3k
Average	*.291*	*.059*	*.097*	*.198*	*.411*	*9.7k*
LOWRES						
Thai	.367	.139	.143	.264	.440	5.6k
Indonesian	.306	.103	.041	.238	.404	10.3k
Vietnamese	.303	.079	.058	.106	.271	6.6k
Bosnian	.212	.035	.084	.103	.277	7.5k
Slovak	.195	.024	.042	.095	.259	6.5k
Ukrainian	.194	.024	.131	.070	.273	5.0k
Latvian	.194	.028	.058	.114	.266	7.1k
Hindi	.163	.024	.068	.057	.231	9.4k
Cebuano	.153	.014	.070	.098	.180	7.7k
Azerbaijani	.150	.016	.031	.113	.174	6.2k
Welsh	.138	.007	.025	.033	.062	7.6k
Albanian	.127	.013	.017	.080	.154	6.0k
Bengali	.120	.026	.050	.063	.173	12.5k
Tamil	.089	.006	.013	.030	.140	9.9k
Uzbek	.082	.093	.066	.114	.077	12.4k
Urdu	.073	.005	.017	.032	.108	11.1k
Telugu	.065	.002	.018	.010	.095	9.6k
Nepali	.059	.002	.039	.018	.089	11.6k
Gujarati	.039	.004	.016	.012	.056	12.0k
Average	*.159*	*.034*	*.052*	*.087*	*.196*	*8.7k*

Table 3: Top-10 accuracy on 12 high-resource languages and 20 low-resource languages. The parts of speech Noun, Adjective, Adverb, and Verb are referred to as NN, JJ, RB, VB, respectively. The "all" column reports accuracy on the entire dictionary. The "#" column reports the size of the English vocabulary used for each experiment.

Language	All	VB	RB	JJ	NN
Arabic	.149	.015	.053	.078	.219
Bengali	.066	.009	.042	.025	.084
Dutch	.265	.042	.039	.164	.350
French	.268	.051	.092	.196	.368
German	.220	.035	.040	.080	.321
Indonesian	.211	.050	.035	.156	.257
Italian	.233	.046	.028	.139	.350
Spanish	.320	.068	.076	.207	.449
Turkish	.171	.011	.086	.139	.201
Uzbek	.057	.121	.075	.104	.045
LARGEENG *Avg*	*.181*	*.041*	*.055*	*.118*	*.244*
SMALL *Avg*	*.260*	*.089*	*.078*	*.210*	*.392*

Table 4: Top-10 accuracy on the expanded English dictionary task. For each experiment, 263,102 English words were used as candidate translations for each foreign word. The SMALL average is given for reference, averaging the results from Table 3 across the same 10 languages.

Figure 2: Shown here are five images for the abstract Indonesian word *konsep*, along with its top 4 ranked translations using CNN features. The actual translation, *concept*, was ranked 3,465.

termining part-of-speech for foreign words even in low-resource languages, and presents our image-based translation results across part-of-speech.

5.1 Assigning POS Labels

To show the performance of our translation method for each particular POS, we first assign a POS tag to each foreign word. Since we evaluate on high- and low-resource languages, many of which do not have POS taggers, we POS tag English words, and transfer the tag to their translations. We scraped the text on the web pages associated with the images of each English word, and collected the sentences that contained each query (English) word. We chose to tag words in sentential context, rather than simply collecting parts of speech from a dictionary, because many words have multiple senses, often with different parts of speech.

We assign universal POS tags (Petrov et al., 2012) using spaCy[2], giving each word its majority tag. We gathered part-of-speech tags for 42% of the English words in our translations. Of the remaining untagged English entries, 40% were multi-word expressions, and 18% were not found in the text of the web pages that we scraped.

When transferring POS tags to foreign words, we only considered foreign words where every English translation had the same POS. Across all 32 languages, on average, we found that, after filtering, 65% of foreign words were nouns, 14% were verbs, 14% were adjectives, 3% were adverbs, and 3% were other (i.e. they were labeled a different POS).

[2]https://spacy.io

5.2 Accuracy by Part-of-speech

As we see in the results in Table 3, the highest translation performance is obtained for nouns, which confirms the observation by Hartmann and Søgaard (2017). However, we see considerable signal in translating adjectives as well, with top-10 accuracies roughly half that of nouns. This trend extends to low-resource languages. We also see that translation quality is relatively poor for adverbs and verbs. There is higher variation in our performance on adverbs across languages, because there were relatively few adverbs (3% of all words.) From these results, it is clear that one can achieve higher accuracy by choosing to translate only nouns and adjectives.

Analysis by part-of-speech only indirectly addresses the question of when translation with images is useful. For example, Figure 2 shows that nouns like *concept* translate incorrectly because of a lack of consistent visual representation. However, verbs like *walk* may have concrete visual representation. Thus, one might perform better overall at translation on *concrete* words, regardless of part-of-speech.

6 Evaluation by Concreteness

Can we effectively predict the concreteness of words in a variety of languages? If so, can these predictions be used to determine when translation via images is helpful? In this section, we answer both of these questions in the affirmative.

6.1 Predicting Word Concreteness

Previous work has used *image dispersion* as a measure of word concreteness (Kiela et al., 2014). We

introduce a novel supervised method for predicting word concreteness that more strongly correlates with human judgements of concreteness.

To train our model, we took Brysbaert et al. (2014)'s dataset, which provides human judgments for about 40k words, each with a 1-5 abstractness-to-concreteness score, and scraped 100 images from English Google Image Search for each word. We then trained a two-layer perceptron with one hidden layer of 32 units, to predict word concreteness. The inputs to the network were the element-wise mean and standard deviation (concatenated into a 8094-dimensional vector)of the CNN features for each of the images corresponding to a word. To better assess this image-only approach, we also experimented with using the distributional word embeddings of Salle et al. (2016) as input. We used these 300-dimensional vectors either seperately or concatentated with the image-based features. Our final network was trained with a cross-entropy loss, although an L2 loss performed nearly as well. We randomly selected 39,000 words as our training set. Results on the remaining held-out validation set are visualized in Figure 3.

Although the concatenated image and word embedding features performed the best, we do not expect to have high-quality word embeddings for words in low-resource languages. Therefore, for the evaluation in Section 6.2, we used the image-embeddings-only model to predict concreteness for every English and foreign word in our dataset.

6.2 Accuracy by Predicted Concreteness

It has already been shown that the images of more abstract words provide a weaker signal for translation (Kiela et al., 2015). Using our method for predicting concreteness, we determine which images sets are most concrete, and thereby estimate the likelihood that we will obtain a high quality translation.

Figure 4 shows the reduction in translation accuracy as increasingly abstract words are included in the set. The concreteness model can be used to establish recall thresholds. For the 25% of foreign words we predict to be most concrete, (25% recall,) AVGMAX achieves top-10 accuracy of 47.0% for high-resource languages and 32.8% for low-resource languages. At a 50% most-concrete recall treshold, top-10 translation accuracies are 25.0% and 37.8% for low- and high-resource languages respectively, compared to 18.6% and 29.3% at 100% recall.

7 Translation with Images and Text

Translation via image features performs worse than state-of-the-art distributional similarity-based methods. For example, Wijaya et al. (2017) demonstrate top-10 accuracies in range of above 85% on the VULIC1000 a 1,000-word dataset, whereas with only image features, Kiela et al. (2015) report top-10 accuracies below 60%. However, there may be utility in combining the two methods, as it is likely that visual and textual distributional representations are contributing different information, and fail in different cases.

We test this intuition by combining image scores with the current state-of-the-art system of Wijaya et al. (2017), which uses Bayesian Personalized Ranking (BPR). In their arXiv draft, Hartmann and Søgaard (2017) presented a negative result when directly combining image representations with distributional representations into a single system. Here, we present a positive result by formalizing the problem as a reranking task. Our intuition is that we hope to guide BPR, clearly the stronger system, with aid from image features and a predicted concreteness value, instead of joining them as equals and potentially washing out the stronger signal.

7.1 Reranking Model

For each foreign word w_f and each English word w_e, we have multiple scores for the pair $p_{f,e} = (w_f, w_e)$, used to rank w_e against all other $w_{e'} \in E$, where E is the English dictionary used in the experiment. Specifically, we have $\textsc{Txt}(p_{f,e})$ and $\textsc{Image}(p_{f,e})$ for all pairs. For each foreign word, we also have the concreteness score, $\textsc{Cnc}(w_f)$, predicted from its image set by the method described in Section 6.1.

We use a small bilingual dictionary, taking all pairs $p_{f,e}$ and labeling them $\{\pm 1\}$, with 1 denoting the words are translations. We construct training data out of the dictionary, treating each labeled pair as an independent observation. We then train a 2-layer perceptron (MLP), with 1 hidden layer of 4 units, to predict translations from the individual scores, minimizing the squared loss.[3]

$$\textsc{MLP}(p_{f,e}) = \textsc{MLP}\big([\textsc{Txt}(p_{f,e}); \textsc{Image}(p_{f,e}); \textsc{Cnc}(w_f)]\big) = \{\pm 1\}$$

[3] We use DyNet (Neubig et al., 2017) for constructing and training our network with the Adam optimization method (Kingma and Ba, 2014).

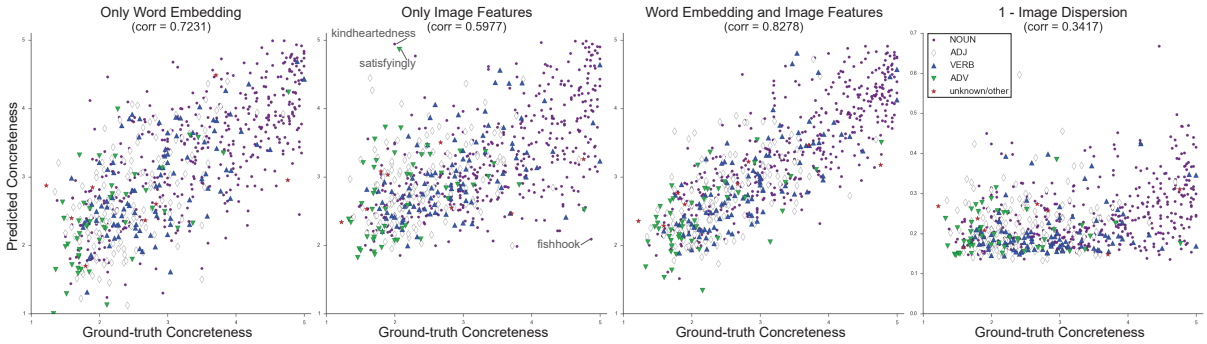

Figure 3: Plots visualizing the distribution of concreteness predictions on the validation set for our three trained models and for image dispersion. Spearman correlation coefficients are shown. For the model trained only on images, the three worst failure cases are annotated. False positives tend to occur when one concrete meaning of an abstract word dominates the search results (i.e. many photos of "satisfyingly" show food). False negatives often stem from related proper nouns or an overabundance of clipart, as is the case for "fishhook."

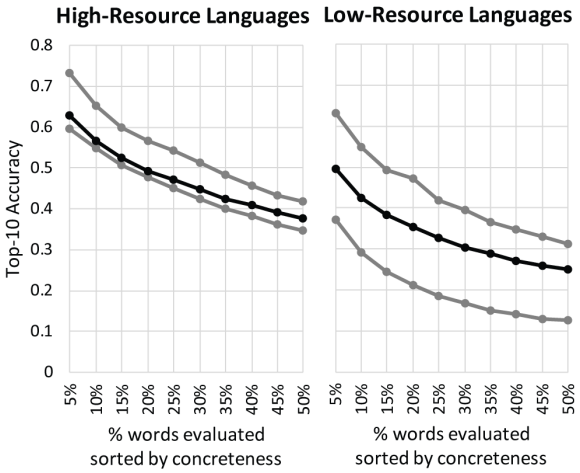

Figure 4: The black curve shows mean top-10 accuracy over the HIGHRES and LOWRES sets sorted by predicted concreteness. The gray curves show the 25th and 75th percentiles.

Once the model is trained, we fix each foreign word w_f, and score all pairs $(w_f, w_{e'})$ for all $e' \in E$, using the learned model MLP$(p_{f,e'})$. Using these scores, we sort E for each w_f.

7.2 Evaluation

We evaluate our text-based and image-based combination method by translating Bosnian, Dutch, French, Indonesian, Italian, and Spanish into English. For each language, we split our bilingual dictionary (of 8,673 entries, on average) into 2,000 entries for a testing set, 20% for training the text-based BPR system, 35% for training the reranking MLP, and the rest for a development set. We filtered out multi-word phrases, and translations where w_f and w_e are string identical.

We compare three models: TXT is Wijaya et al. (2017)'s text-based state-of-the-art model. TXT+IMG is our MLP-learned combination of the two features. TXT+IMG+CNC uses our predicted concreteness of the foreign word as well. We evaluate all models on varying percents of testing data sorted by predicted concreteness, as in Section 6.2. As shown in Figure 5, both image-augmented methods beat TXT across concreteness thresholds on the top-1 accuracy metric.

Results across the 6 languages are reported in Table 5. Confirming our intuition, images are useful at high concreteness, improving the SOA text-based method 3.21% at 10% recall. At 100% recall our method with images still improves the SOA by 1.3%. For example, the text-only system translates the Bosnian word *košarkaški* incorrectly as *football*, whereas the image+text system produces the correct *basketball*.

Further, gains are more pronounced for low-resource languages than for high-resource languages. Concreteness scores are useful for high-resource languages, for example Spanish, where TXT+IMG falls below TXT alone on more abstract words, but TXT+IMG+CNC remains an improvement. Finally, we note that the text-only system also performs better on concrete words than abstract words, indicating a general trend of ease in translating concrete words regardless of method.

8 Summary

We have introduced a large-scale multilingual image resource, and used it to conduct the most comprehensive study to date on using images to learn translations. Our Massively Multilingual Image Dataset will serve as a standard for future work in image-based translation due to its size and generality, covering 100 languages, hundreds of thousands

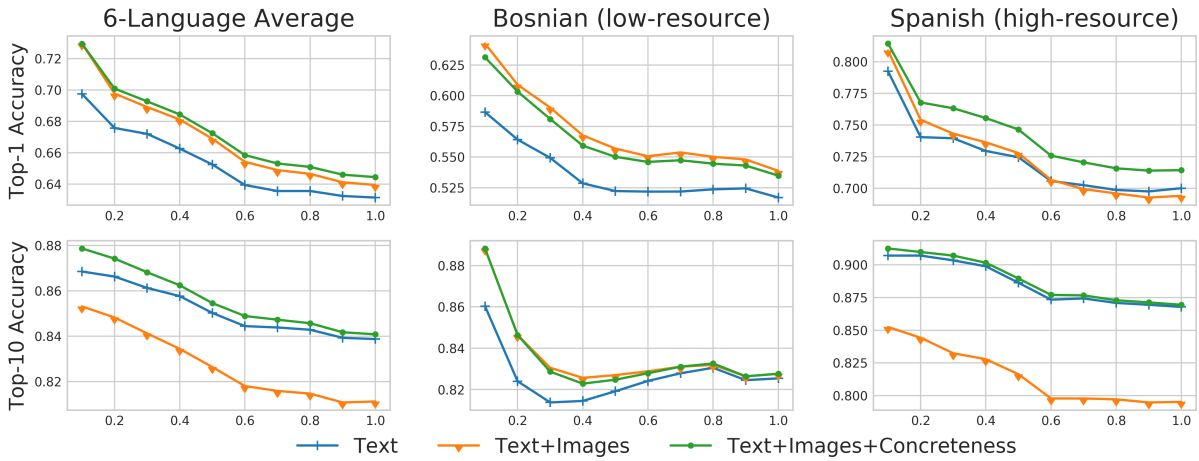

Figure 5: Reranking top-1 and top-10 accuracies of our image+text combination sytems compared to the text-only Bayesian Personalized Ranking system. The X-axis shows percent of foreign words evaluated on, sorted by decreasing predicted concreteness.

		% words evaluated		
	Method	**10%**	**50%**	**100%**
High-Res	TXT	.746	.696	.673
	TXT+IMG	**.771**	.708	.678
	TXT+IMG+Cnc	**.773**	**.714**	**.685**
Low-Res	TXT	.601	.565	.549
	TXT+IMG	**.646**	**.590**	**.562**
	TXT+IMG+Cnc	**.643**	**.589**	**.563**

Table 5: Top-1 accuracy results across high-resource (Dutch, French, Italian, Spanish) and low-resource (Bosnian, Indonesian) languages. Words evaluated on are again sorted by concreteness for the sake of analysis. The best result on each % of test data is bolded.

of words, and a broad range of parts of speech. Using this corpus, we demonstrated the substantial utility in supervised prediction of word concreteness when using image features, improving over the unsupervised state-of-the-art and finding that image-based translation is much more accurate for concrete words. Because of the text we collected with our corpus, we were also able to collect part-of-speech information and demonstrate that image features are useful in translating adjectives and nouns. Finally, we demonstrate a promising path forward, showing that incorporating images can improve a state-of-the-art text-based word translation system.

9 Dataset and Code

The MMID will be distributed both in raw form and for a subset of languages in memory compact featurized versions from http://multilingual-images.org along with code we used in our experiments. Additional details are given in our Supplemental Materials document, which also describes our manual image annotation setup, and gives numerous illustrative examples of our system's predictions.

Acknowledgements

We gratefully acknowledge Amazon for its support of this research through the Amazon Research Awards program and through AWS Research Credits.

This material is based in part on research sponsored by DARPA under grant number HR0011-15-C-0115 (the LORELEI program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA and the U.S. Government.

References

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

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

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *CoRR*, abs/1605.00459.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 207–213, Lisbon, Portugal. Association for Computational Linguistics.

Michael Friendly, Patricia E. Franklin, David Hoffman, and David C. Rubin. 1982. The Toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, 14(4):375–399.

Ruka Funaki and Hideki Nakayama. 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. Association for Computational Linguistics.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international Conference on Computational Linguistics*, volume 1, pages 414–420. Association for Computational Linguistics.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark. Association for Computational Linguistics.

K. J. Gilhooly and R. H. Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, volume 2008, pages 771–779.

Mareike Hartmann and Anders Søgaard. 2017. Limitations of cross-lingual learning from image search. *CoRR*, abs/1709.05914.

Julian Hitschler and Stefan Riezler. 2016. Multimodal pivots for image caption translation. *CoRR*, abs/1601.03916.

Ann Irvine and Chris Callison-Burch. 2017. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310.

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 the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 835–841, Baltimore, Maryland. Association for Computational Linguistics.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Lisbon, Portugal. Association for Computational Linguistics.

Adam Kilgarriff. 2007. Googleology is bad science. *Computational linguistics*, 33(1):147–151.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 817–824. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, volume 9, pages 9–16. Association for Computational Linguistics.

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

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1780–1790.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. In *Journal of Experimental Psychology*, volume 76, pages 207–213. American Psychological Association.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1115.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 419–424. Association for Computational Linguistics.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany. Association for Computational Linguistics.

Ivan Vulic and Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 106–116. ACL.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1453–1464, Copenhagen, Denmark. Association for Computational Linguistics.