



AFRL-RI-RS-TR-2022-120

SEMI-SUPERVISED LEARNING OF MULTIMODAL REPRESENTATIONS

UNIVERSITY OF PENNSYLVANIA

AUGUST 2022

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

**AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (<http://www.dtic.mil>).

AFRL-RI-RS-TR-2022-120 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /

ALEKSEY PANASYUK
Work Unit Manager

/ S /

SCOTT PATRICK
Deputy Chief
Intelligence Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE AUGUST 2022	2. REPORT TYPE FINAL TECHNICAL REPORT	3. DATES COVERED START DATE JUNE 2019 END DATE JUNE 2022	
4. TITLE AND SUBTITLE Semi-supervised Learning of Multimodal Representations			
5a. CONTRACT NUMBER FA8750-19-2-3334	5b. GRANT NUMBER N/A	5c. PROGRAM ELEMENT NUMBER 623038	
5d. PROJECT NUMBER	5e. TASK NUMBER	5f. WORK UNIT NUMBER R2XX	
6. AUTHOR(S) Chris Callison-Burch Ph.D., Derry Wijaya Ph.D.			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) UNIVERSITY OF PENNSYLVANIA 3451 Walnut Street Philadelphia PA 19104-6205			8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RIEA 525 Brooks Road Rome NY 13441-4505		10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RI	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RI-RS-TR-2022-120
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.			
13. SUPPLEMENTARY NOTES			
14. ABSTRACT The goal of this project was to create better representations of words by improving vector space language models with multimodal and multilingual information. A large-scale dataset of multilingual images, called MMID, was assembled, it associates images with words for 98 different languages (up to 10K words for each language, with 100 images per word). This dataset let us perform a comprehensive analysis of whether visual similarity could be used to identify translations, and the extent to which this is affected by linguistic factors like part of speech and concreteness. We studied whether MMID could be used to mitigate the geographical bias in image classification datasets like ImageNet (for example wedding is visually distinct in different regions of the world). The extent to which geography impacts the translatability across pairs of languages was investigated; factors such as shared language families, ethnic groups or shared religions have a larger impact than geography on the visual similarity and therefore translatability via images. We also collected a dataset from Wikipedia by aggregating shared images with multilingual captions giving us full sentences rather than the individual words in MMID.			
15. SUBJECT TERMS Multilingual images, MMID, Natural Language Processing (NLP), Workshop on Machine Translation (WMT)			
16. SECURITY CLASSIFICATION OF:		17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR 74
19a. NAME OF RESPONSIBLE PERSON ALEKSEY PANASYUK			19b. PHONE NUMBER (Include area code) N/A

TABLE OF CONTENTS

List of Figures	iii
List of Tables	v
1.0 SUMMARY	1
2.0 INTRODUCTION	2
2.1 Related Work.....	10
2.1.1 Learning Translations via Images.....	10
2.1.2 Language Characteristics in Translation.....	11
2.1.3 Background on Cross Lingual Tasks	12
2.1.4 Cross-lingual NLP Tasks Using Minimal Supervision.....	13
3.0 METHODS, ASSUMPTIONS, AND PROCEDURES.....	14
3.1 Study 1: Learning Translations with a Massively Multilingual Image Dataset	14
3.1.1 Corpus Construction	14
3.1.2 Finding Translations Using Images	15
3.1.3 Translation Prediction with AVGMAX.....	15
3.2 Study 2: Mitigating Geographic Bias of Image Classifiers with Multilingual Image Data	18
3.2.1 Data Expansion, Cleaning, and Selection	19
3.2.2 Model and Training.....	20
3.3 Study 3: Cultural and Geographical Influences on Image Translatability of Words across Languages	23
3.3.1 Dataset collection and preparation.....	23
3.3.2 Methodology	24
3.4 Study 4: Weakly Supervised Neural Translation Tailored to Cross-Lingual Tasks	27
3.4.1 Learning Translation from Wikipedia	27
3.4.2 Bilingual Dictionary Extraction and Cross-Lingual Word Embeddings	29
3.4.3 Mining Parallel Sentences.....	29
3.4.4 Leveraging Similar Languages	29
3.4.5 Translation Model	30
3.4.6 Cross-Lingual Tasks	31
4.0 RESULTS AND DISCUSSION	33
4.1 Study 1: Learning Translations with a Massively Multilingual Image Dataset	33
4.1.1 Evaluation by Part-of-speech.....	33
4.1.2 Evaluation by Concreteness	35
4.1.3 Predicting Word Concreteness.....	36
4.1.4 Translation with Images and Text.....	37

4.2	Study 2: Mitigating Geographic Bias of Image Classifiers with Multilingual Image Data	40
4.3	Study 3: Cultural and Geographical Influences on Image Translatability of Words across Languages	42
4.4	Study 4: Weakly Supervised Neural Translation Tailored to Cross-Lingual Tasks	48
4.4.1	Datasets and Settings	48
4.4.2	Pretraining.....	48
4.4.3	Translation Results.....	49
4.4.4	Image Captioning Results	50
4.4.5	Dependency Parsing Results.....	51
5.0	CONCLUSIONS.....	53
6.0	RECOMMENDATIONS.....	54
7.0	REFERENCES	55
	APPENDIX A.....	65
7.1	Publications from this project:	65
7.2	Dataset and Code:.....	65
	LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS.....	66

LIST OF FIGURES

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, along with its top 4 ranked translations.	2
Figure 2: Fraction of ImageNet images from each country. Over half of the data comes from the US and Great Britain Source: Shankar et al (2017).....	4
Figure 3: Wedding photographs labeled by a classifier trained on the Open Images dataset, from Google AI.....	4
Figure 4: A photo of an Indian wedding selected from MMID using the Hindi translation of the English word wedding.	5
Figure 5: A pair of Wikipedia documents in Arabic and English, along with the same image with two captions.	9
Figure 6: An example of annotation projection for which the source on top is a translation of the Romanian target via our “wikily” translation model. The supervised source tree is projected using intersected word alignments.....	13
Figure 7: Photos for “athlete”, “children”, and “farmer”. The photos on top are drawn from the English words in the MMID, and the photos on the bottom are drawn from their translations into a language spoken in India.	19
Figure 8: Photos for “police”, “soldier”, and “wedding”. The photos on top are drawn from the English words in the MMID, and the photos on the bottom are drawn from their translations into a language spoken in India.	19
Figure 9: Examples of noisy images that were discarded because they were illustrations rather than photos, or they contained text, or they did not correctly represent one of our six categories.	20
Figure 10: Training curves for logistic regression.....	20
Figure 11: Training curves for CNN.....	21
Figure 12: Training curves for Resnet	22
Figure 13: A brief description of the training pipeline for our weakly supervised cross lingual bootstrapping approach.	28
Figure 14: An image from MS-Coco (Chen et al., 2015) with gold-standard English captions, and Arabic translations from our Wikily translation model.	31
Figure 15: Shown here are five images for the abstract Indonesian word konsep, along with its top 4 ranked translations using CNN features. The actual translation, konsep, was ranked 3,465.	34
Figure 16: Strawberry is an example of a word with a high concreteness value in psycholinguistic studies. This is reflected with a consistent visual representation.....	35

Figure 17: Hope is an example of a word with a low concreteness value in psycholinguistic studies. This is reflected with inconsistent visual representations, both within a single language and across languages.....	35
Figure 18: 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.	36
Figure 19: 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.	37
Figure 20: Reranking top-1 and top-10 accuracies of our image+text combination systems compared to the text-only BPR system. The X-axis shows percent of foreign words evaluated on, sorted by decreasing predicted concreteness.	39
Figure 21: Confusion matrices for combined, network trained on English.....	41
Figure 22: Confusion matrices for combined, network trained on combined	42
Figure 23: (left) PCA plot of image embeddings of the word universe in Afrikaans (heelal) and in Dutch (universum) showing their tightly overlapping image clusters, (center) images of heelal in Afrikaans, (right) images of universum in Dutch	45
Figure 24: (left) PCA plot of image embeddings of the word dance in Afrikaans (dans) and in Swahili (kucheza) showing their tight but disjoint image clusters, (center) images of dans in Afrikaans, (right) images of kucheza in Swahili	46
Figure 25: Images of the word park in (left) Afrikaans, (center) Dutch, (right) Swahili.	46
Figure 26: PCA plots for image embeddings of (left) park in Afrikaans to park in Dutch and (right) park in Afrikaans to hifadhi in Swahili.....	47
Figure 27: An example of a Gujarati sentence and its outputs from different models, as well as Google Translate.....	50
Figure 28: An example of different outputs in our captioning experiments both for English and Arabic, as well as Arabic translations of English outputs on the Arabic Flickr dataset (ElJundi. et al., 2020).	51

LIST OF TABLES

Table 1: The 19 language pairs we explore in this work and the nature of their similarity: Geographical or Cultural: the same Language family, Ethnicity or Religion.	7
Table 2: 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.	15
Table 3: Our results are consistently better than those reported by Kiela et al. (2015), on a similar set of 500 concrete nouns. The rightmost column shows the added challenge with our larger, more realistic dataset.....	16
Table 4: Top-10 accuracy on 12 high-resource languages and 20 low-resource languages. The “all” column reports accuracy on the entire dictionary. The “#” column reports the size of the English vocabulary used for each experiment.	18
Table 5: 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 averages results from Table 4 across the same 10 languages.	18
Table 6: Comparison of models trained on combined English/non-English dataset.....	22
Table 7: Comparison of Resnet models.....	22
Table 8: The 19 language pairs we explore in this work and the nature of their similarity: Geographical or Cultural: the same Language family, Ethnicity or Religion.	22
Table 9: 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.	39
Table 10: Language word pairs and their and translatability ratios, for all and concrete word pairs. Bold indicates pairs with the same source language whose normalized scores are significantly different and whose ratio differences are high.....	43
Table 11: p-values of differences between normalized score distributions of language pairs that share the same source language.	45
Table 12: Data sizes for different pairs. We use a sample of English sentences with similar sizes to each data.	48
Table 13: BLEU scores for different models. Reference results are from *: Our implementation, 1: Kim et al. (2020), 2: Li et al. (2020), 3: Liu et al. (2020) (supervised), 4: Tran et al. (2020) (unsupervised with mined parallel data).	49
Table 14: Image captioning results evaluated on the Arabic Flickr dataset (ElJundi. et al., 2020) using SacreBLEU (Post, 2018). “pretrained” indicates initializing our captioning model with our translation parameters.....	51
Table 15: Dependency parsing results on the Universal Dependencies dataset (Zeman et al., 2020). Previous work has used different sub-versions of the Universal Dependencies data in which slight differences are expected.....	52

1.0 SUMMARY

This report summarizes the main findings from the project entitled “Semi-supervised Learning of Multimodal Representations”. The principal investigators for this project are Prof. Chris Callison-Burch at the University of Pennsylvania and Prof. Derry Wijaya from Boston University. The period of performance was 6/26/19 to 6/14/22. This project was broadly relevant to other projects that DARPA Program Manager Boyan Onyshkevych was managing, including the DARPA AIDA program, but it was a separate effort. This project was funded through DARPA’s OpenBAA call. The total budget for this project was \$428,000. The bulk of that budget was used for salaries for researchers, and for computing expenses. Prof. Chris Callison-Burch received 2 months of support, Prof. Derry Wijaya received a similar amount of support, and 17 months support for a post-doctoral researcher Mohammad Sadegh Rasooli.

The goal of this project was to create better representations of words by improving vector space language models with multimodal and multilingual information. This final report will detail four research studies:

1. In preparation for this project, we assembled a large-scale dataset of multilingual images, called MMID, which associates images with words for 98 different languages. MMID contains up to 10,000 vocabulary items for each language, with 100 images per word. Additionally, we have 100 images for each of 250K English translations. This dataset let us perform a comprehensive analysis of whether visual similarity could be used to identify translations, and the extent to which this is affected by linguistic factors like part of speech and concreteness.
2. We conducted a study to see if MMID could be used to mitigate the geographical bias in image classification datasets like ImageNet. Because of the Western bias in ImageNet, classifiers trained using it often fail to recognize objects and events that are visually distinct in other regions of the world (for instance, Indian weddings).
3. We investigated the extent to which geography impacts the translatability across pairs of languages, and compared it to the impact of cultural factors. We find that factors such as shared language families, ethnic groups or shared religions have a larger impact than geography on the visual similarity and therefore translatability via images.
4. We collect a second dataset from Wikipedia by aggregating shared images with multilingual captions. This dataset gives us full sentences rather than the individual words in MMID. We performed a set of experiments that showed how our Wikipedia data could be used as a source of weakly supervised data to perform several multilingual tasks that traditionally required bilingual sentence-aligned parallel corpora.

2.0 INTRODUCTION

This project focused on the developing method for learning the translations of words across languages without the use of traditional training resources. 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, along with its top 4 ranked translations.

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.

In our first study, we conducted a broad range of experiments to evaluate the utility of image features across a number of factors:

- We evaluated 12 high-resource and 20 low resource languages.
- We evaluated translation quality stratified by part-of-speech, finding that nouns and adjectives are translated with much higher accuracy than adverbs and verbs.
- We presented 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.

In our second study, we investigated whether MMID could be used to help mitigate the biases that have been observed in computer vision systems that are trained on datasets that were collected in a geographically biased fashion. Advances in computer vision driven by deep learning have greatly improved the performance of image algorithms on benchmark datasets and in domains like autonomous driving, but the lack of geodiversity in canonical datasets like ImageNet significantly limits the use of pretrained models in certain geographical regions, especially in the Global South. Initial approaches to remedy geographic bias have not explicitly charted a path forward on building more representative datasets.

The creation of publicly-available, large-scale datasets for image classification like ImageNet, which consists of around 1.2 million images with human-verified labels for 1000 categories, has facilitated the advancement of deep learning methods. Pretrained image classifiers are commonly used in practice. However, the lack of geo-representative training data leads to stark differences in model performance across location. Researchers at Google Brain found that ImageNet and Open Images “appear to exhibit an observable amero-centric and euro-centric representation bias.” Reliable country geolocation for a large subset of ImageNet found that 45% of the data comes from the United States (4% of the population), but only 1% and 2.1% come from China and India respectively (36% of the population). Images of grooms crowdsourced in Hyderabad were much less likely to be correctly classified than ImageNet images. Lastly, saliency maps like SmoothGrad activated around faces but predicted classes related to attire (Shankar et al, 2017). After identifying the problems of ImageNet and Open Images, Google researchers introduced the Inclusive Images Competition, which asked participants to develop algorithms to train models with geographically skewed data that would be evaluated on more balanced datasets. The winning team had a relatively straightforward solution, using an ensemble of fine-tuned convolutional neural networks. We explore a more data-driven approach to remedy the bias of classifiers trained on existing and predominantly Western large-scale image datasets, using images searched in many languages.

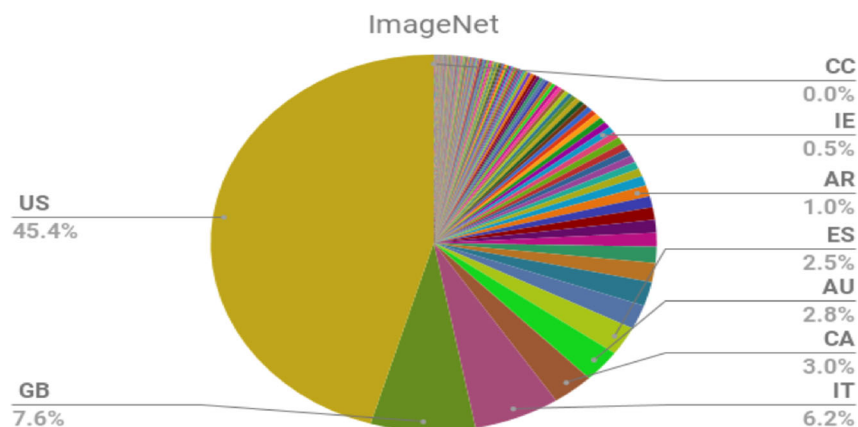


Figure 2: Fraction of ImageNet images from each country. Over half of the data comes from the US and Great Britain Source: Shankar et al (2017)

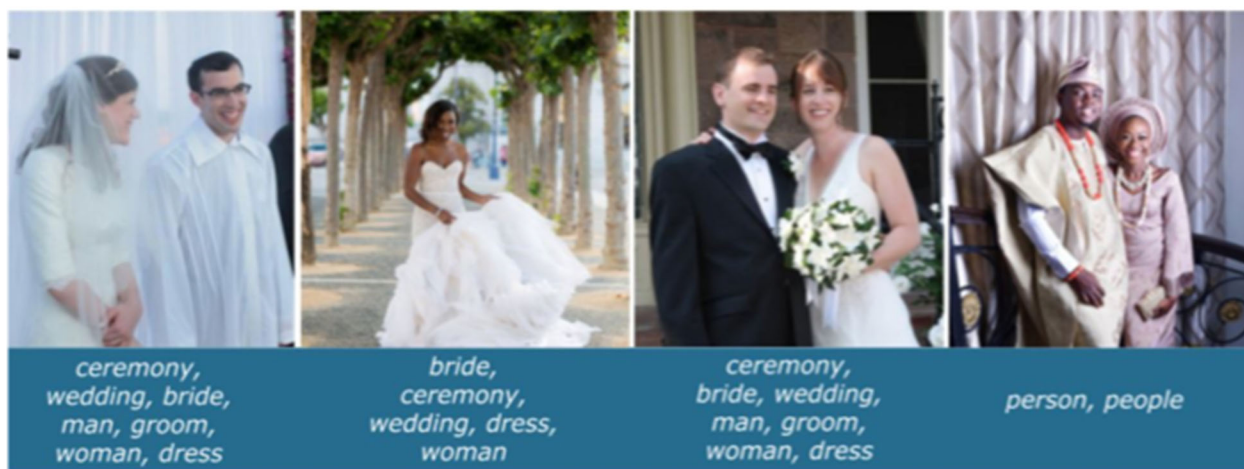


Figure 3: Wedding photographs labeled by a classifier trained on the Open Images dataset, from Google AI

Indian weddings are visually distinct from Western weddings – which are the only kind that are well-represented in ImageNet. We performed a study to investigate whether it may be possible to mitigate this using MMID, which contains millions of images for words in 100 different languages. We created a more diverse set of images by translating words like *wedding* into foreign languages like Bengali, Bishnupriya-Manipuri, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and Telugu, which are spoken in India. By doing so, we include images of Indian weddings like the one below.



Figure 4: A photo of an Indian wedding selected from MMID using the Hindi translation of the English word wedding.

In our second study, we create a 6-way classifier between for people centric images that depict *athletes, children, farmers, police, soldiers, and weddings*. Training the image classifier only on images collected via English labels results in average accuracy of 77% when tested on images that come from a mixture of English and Indian language labels. Augmenting the dataset with images sampled from MMID by translating the English labels into several Indian languages results in an improved accuracy of 91%, suggesting that we are able to mitigate the geographic bias in ImageNet that was identified by Google researchers. But is the problem that the image datasets are geographically unrepresentative, or is the problem that they do not represent a sufficiently diverse set of cultures?

In a third study, we investigate the role of culture versus geography in visual similarity of concepts across languages. We define several quantifiable measures of *translatability* across two languages using images associated with the words in each of the languages. Previous research on using images for translation typically accepts that images are useful due to their language invariance (Rotman et al., 2018). Since words such as *horse* denote concepts that exist independently of any language, images that ground their meanings should also be invariant to the language.

However, to the best of our knowledge, this conjecture on image-language invariance has never been tested. As images’ usefulness for translation has only been shown to be marginal (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018), it is important to study this conjecture in relation to the characteristics of languages to understand *when* and to what extent images can aid translation. An alternative view would be that images may be different to some extent in different languages since they reflect the ways different people interact with these concepts; this may depend on *where* they live and the *communities* they live in (Evans and Levinson, 2009). For example, images of the word *breakfast* in different languages may reflect the different cuisines of the communities that speak the languages.

Similarly to the geographic bias in ImageNet, most multimodal MT datasets are limited to a small set of European languages that come from the same language family, and are spoken by communities that are culturally and geographically close. We leverage our MMID, which we constructed

specifically to facilitate large-scale multilingual research in translating words via images. Specifically, we explored if translatability of words between two languages via images is influenced by the cultural similarity and geographical proximity of their communities. A recent study, (Thompson et al., 2020), has observed such correlations of culture and geography to semantic alignment of word meanings between languages that are measured through similarities in the word embeddings. We hypothesize that the same is true for images, and that the alignment of meanings conveyed via images coincides with culture and geography.

In our study, we primarily define culture as the set of “Language, Norms and Beliefs” of a community (Heather Griffiths, 2015). These elements form our interpretation of cultural closeness between languages, which consist of their common linguistic, ethnic, and religious properties. Our goal is to intrinsically evaluate to what extent images can aid in word translation, for words in languages close in a variety of these characteristics. Assuming that each word is associated with a number of images that convey its meaning, we measure the degree to which images of words that are translations of each other in different languages have similar representations (thus will help in translation). We call this measure *image translatability* or the capacity of word meaning to be transferred from one language to another via images. If images are indeed language invariant, we should observe similar image translatability across different language pairs.

We identify how close word translations are in terms of their image representations (embeddings). Our findings suggest that languages with cultural similarity (defined as a combination of linguistic, ethnicity, or religious similarity of the communities at the cultural centers of the languages by Glottolog (Hammarström et al., 2020)) coincides with their translatability via images, and that the translatability of languages with cultural similarity outperforms that in those with geographical proximity.

We define a translatability ratio for a language pair as:

$$RATIO = \frac{\# \text{ of word pairs that are good translations}}{\# \text{ of word pairs for the language pair}} \quad (1)$$

We then compare that ratio for a carefully selected set of language pairs that are geographically close (e.g. the are spoken in countries that share a border), versus languages that come from a shared language family, are spoken by similar ethnic groups, or by peoples who share a religion. Table 1 shows our language pairs and their categorization. We find that cultural factors –shared language family, ethnicity or religion– result in a higher translatability ratio than similar geography. This suggests that AI systems that attempt to automatically understand images through classification, captioning or translation need to take different cultural signifiers into account.

Table 1: The 19 language pairs we explore in this work and the nature of their similarity: Geographical or Cultural: the same Language family, Ethnicity or Religion.

Language Pair	Geography	Family	Ethnicity	Religion	Ratio
Azerbaijani Turkish	✓	✓	✓	✓	.31
Azerbaijani Russian	✓	✗	✗	✗	.17
Korean Chinese	✓	✗	✓	✓	.18
Korean Japanese	✗	✗	✓	✓	.69
Chinese Japanese	✗	✗	✓	✓	.56
Chinese Korean	✓	✗	✓	✓	.36
Japanese Chinese	✗	✗	✓	✓	.39
Japanese Korean	✗	✗	✓	✓	.67
Arabic Urdu	✗	✗	✗	✓	.39
Arabic Farsi	✓	✗	✗	✓	.50
Arabic Hebrew	✓	✓	✗	✗	.69
Urdu Arabic	✗	✗	✗	✓	.39
Urdu Hindi	✓	✓	✓	✗	.12
Spanish French	✓	✓	✓	✓	.45
Spanish Portuguese	✓	✓	✓	✓	.40
Finnish Hungarian	✗	✓	✓	✓	.29
Finnish Norwegian	✓	✗	✗	✓	.17
Afrikaans Dutch	✗	✓	✓	✓	.39
Afrikaans Swahili	✗	✗	✗	✓	.25

In our fourth and final study, present a simple but effective approach for leveraging the kind of data that we collected in MMID. Although our data is different from the bilingual parallel data that is typically used in machine translation, and although it does not have labels for traditional NLP tasks like dependency parsing, we show that this kind of “weak supervision” can be for cross-lingual NLP tasks. We collect a second multilingual image dataset by collecting shared images and captions from Wikipedia articles written in different languages. We also gather the first sentences and titles of the articles as non-parallel approximate translations. We use this data to bootstrap training data for several multilingual NLP tasks: neural machine translation, cross-lingual image captioning, and cross-lingual dependency parsing without using any direct supervision from external parallel data or supervised models in the target language.

Developing machine translation models without using bilingual parallel text is an intriguing research problem with real applications: obtaining a large volume of parallel text for many languages is hard if not impossible. Moreover, translation models could be used in downstream cross-lingual

tasks in which annotated data does not exist for some languages. There has recently been a great deal of interest in unsupervised neural machine translation (e.g. Artetxe et al. (2018a); Lample et al. (2018a,c); Conneau and Lample (2019); Song et al. (2019a); Kim et al. (2020); Tae et al. (2020)). Unsupervised neural machine translation models often perform nearly as well as supervised models when translating between similar languages, but they fail to perform well in low-resource or distant languages (Kim et al., 2020) or out-of-domain monolingual data (Marchisio et al., 2020). In practice, the highest need for unsupervised models is to expand beyond high resource, similar European language pairs.

There are two key goals in this paper: Our first goal is developing accurate translation models for low-resource distant languages *without* any supervision from a supervised model or gold-standard parallel data. Our second goal is to show that our machine translation models can be directly tailored to downstream natural language processing tasks. In this paper, we showcase our claim in cross-lingual image captioning and cross-lingual transfer of dependency parsers, but this idea is applicable to a wide variety of tasks.

We present a fast and accurate approach for learning translation models using Wikipedia. Unlike *unsupervised machine translation* that solely relies on raw monolingual data, we believe that we should not neglect the availability of incidental supervisions from online resources such as Wikipedia. Wikipedia contains articles in nearly 300 languages and more languages might be added in the future, including indigenous languages and dialects of different regions in the world. Different from similar recent work (Schwenk et al., 2019a), we do not rely on any supervision from supervised translation models. Instead, we leverage the fact that many first sentences in linked Wikipedia pages are rough translations, and furthermore, many captions of the same images are similar sentences, sometimes translations. Figure 5 shows a real example of a pair of linked Wikipedia pages in Arabic and English in which the titles, first sentences and also the image captions are rough translations of each other. Our method learns a seed bilingual dictionary from a small collection of first sentence pairs, titles and captions and then learns cross-lingual word embeddings. We make use of cross-lingual word embeddings to extract parallel sentences from Wikipedia. Our experiments show that our approach improves over strong unsupervised translation models for low-resource languages: we improve the BLEU score of English to Gujarati from 0.6 to 15.2, and English to Kazakh from 0.8 to 12.1.

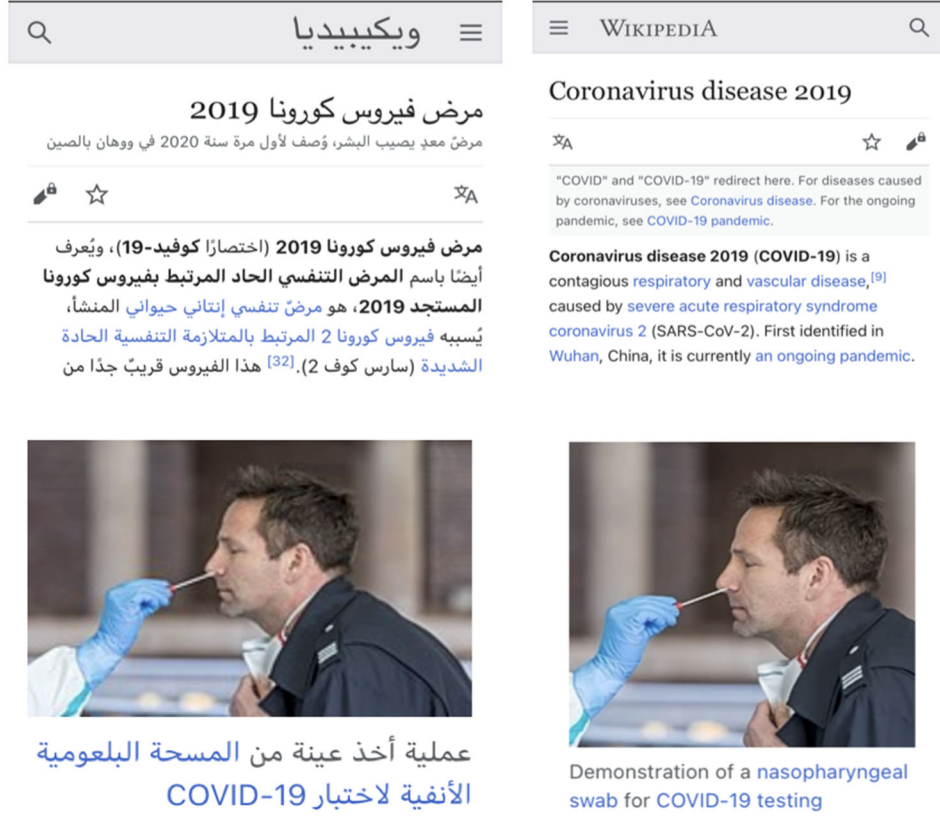


Figure 5: A pair of Wikipedia documents in Arabic and English, along with the same image with two captions.

For downstream tasks, we show that we can easily use our translation models to generate high-quality translations of the MS-COCO dataset (Chen et al, 2015) and the Flickr dataset (Hodosh et al, 2013), and that we can train a cross-lingual image captioning model in a multi-task pipeline paired with machine translation in which the model is initialized by the parameters from our translation model. Our results on Arabic captioning show a BLEU score of 5.72 that is slightly better than a supervised captioning model with a BLEU score of 5.22. We also investigate another task, *dependency parsing*. We first translate a large amount of monolingual data using our translation models and then apply transfer using the *annotation projection* method (Yarowsky et al, 2001, Hwa et al 2005). Our results show that our approach performs similarly compared to using gold-standard parallel text in high-resource scenarios, and significantly better in low-resource languages.

Our final study made several contributions:

1. We developed a simple, fast and effective approach towards using the Wikipedia monolingual data for machine translation without any explicit supervision. Our mining algorithm easily scales on large comparable data using limited computational resources. We achieved very high BLEU scores for distant languages, especially those in which current unsupervised methods perform very poorly.
2. We proposed novel methods for leveraging our current translation models in image captioning. We showed how a combination of translating caption training data, and multi-

task learning with English captioning as well as translation improves the performance. Our results on Arabic shows results slightly superior to that of a supervised captioning model trained on gold-standard datasets.

3. We proposed a novel modification to the annotation projection method to be able to leverage our translation models. Our results on dependency parsing performs better than previous work in most cases, and performs similarly to using gold-standard parallel datasets.

2.1 Related Work

2.1.1 Learning Translations via Images

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 (Vulic' 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.

2.1.2 Language Characteristics in Translation

The claim that the concepts of language, culture and their geographical affiliations are interdependent, constantly and dynamically evolving and defining each other, has been widely discussed and is well established in the literature. Culture is considered an indistinguishable part of languages when translating from one language to another. The importance of cultural literacy of the translator and his/her awareness of cultural factors, views and tradition, apart from word meaning, for producing high quality translations is indisputable (Nida, 1945; Wakabayashi, 1991; Janfaza et al., 2012).

Despite the importance of language, culture and geography in translation, and findings that parallel data from similar, higher resource languages can help improve MT of low resource languages (Kocmi and Bojar, 2018), no previous work has studied how language similarity may influence translation via images. The most notable recent work in this area, that is most similar to ours, is

that of [Thompson et al. \(2020\)](#). The authors predict semantic similarity of words in 41 languages from the NEL dataset ([Dellert et al., 2020](#)) and examine the relationships between word semantic similarity (measured via word embeddings) with the cultural, historical and geographical aspects of the communities speaking the language. Their findings, that the role of cultural similarities to this prediction is superior to that of geographical ones, align with ours. However, their methods differ from ours in many aspects. They use word-only embeddings to measure semantic alignment of words and only a small and publicly available set of images ([Duñabeitia et al., 2018](#)), for validation of the predicted scores, in a supervised manner, and for a small subset of 6 languages in the Indo-European family.

2.1.3 Background on Cross Lingual Tasks

2.1.3.1 Supervised neural machine translation

Supervised machine translation uses a parallel text. Neural machine translation uses sequence-to-sequence models with attention ([Cho et al., 2014](#); [Bahdanau et al., 2015](#); [Vaswani et al., 2017](#)) for which the likelihood of training data is maximized by maximizing the log likelihood of predicting each target word given its previous predicted words and source sequence:

$$\mathcal{L}(\mathcal{P}) = \sum_{i=1}^n \sum_{j=1}^{|t_i|} \log p(t_{i,j} | t_{i,k < j}, s_i; \theta) \quad (2)$$

where θ is a collection of parameters to be learned.

2.1.3.2 Unsupervised neural machine translation

Unsupervised neural machine translation does not have access to any parallel data. Instead, it tailors monolingual datasets M1 and M2 for learning multilingual language models. These language models usually mask parts of every input sentence, and try to uncover the masked words ([Devlin et al., 2019](#)). The monolingual language models are used along with iterative back-translation ([Huang et al., 2018](#)) to learn unsupervised translation. An input sentence s is translated to t' using current model θ , then the model assumes that (t', s) is a gold standard translation, and uses the same training objective as of supervised translation.

Dependency parsing Dependency parsing algorithms capture the best scoring dependency trees for sentences among an exponential number of possible dependency trees. A valid dependency tree for a sentence $s = s_1, \dots, s_n$ assigns heads h_i for each word s_i where $1 \leq i \leq n$, $0 \leq h_i \leq n$ and $h_i \neq i$. The zeroth word represents a dummy root token as an indicator for the root of the sentence. For more details about efficient parsing algorithms, we encourage the reader to see [Kübler et al. \(2009\)](#).

Annotation projection

Annotation projection is an effective method for transferring supervised annotation from a rich-resource language to a low-resource language through translated text ([Yarowsky et al., 2001](#)). Having a parallel data $P = \{(s_i, t_i)\}_{i=1}^n$, and supervised source annotations for source sentences s_i , we transfer those annotations through word translation links $0 \leq a_{(j)} \leq |t_i|$ for $1 \leq j \leq |s_i|$ where $a_{(j)} = 0$ shows a null alignment. The alignment links are learned in an unsupervised fashion using unsupervised word alignment algorithms ([Och and Ney, 2003a](#)). In dependency parsing, if $h_i = j$ and $a_{(j)} = k$ and $a_{(i)} = m$, we project a dependency $k \rightarrow m$ (i.e. $h_m = k$) to the target side. Previous work

(Rasooli and Collins, 2017, 2019) has shown that annotation projection only works when a large amount of translation data exists. In the absence of parallel data, we create artificial parallel data using our translation models. Figure 6 shows an example of annotation projection using translated text.

Figure 6: An example of annotation projection for which the source on top is a translation of the Romanian target via our “wikily” translation model. The supervised source tree is projected using intersected word alignments.

Kim et al. (2020) has shown that unsupervised translation models often fail to provide good translation systems for distant languages. Our work solves this problem by leveraging the Wikipedia data. Using pivot languages has been used in previous work (Al-Shedivat and Parikh, 2019), as well as using related languages (Zoph et al., 2016; Nguyen and Chiang, 2017). Our work only explores a simple idea of adding one similar language pair. Most likely, adding more language pairs and using ideas from recent work might improve the performance.

Cross-lingual and unsupervised image captioning has been studied in previous work (Gu et al., 2018; Feng et al., 2019; Song et al., 2019b; Gu et al., 2019; Gao et al., 2020; Burns et al., 2020). Unlike previous work, we do not have a supervised translation model. Cross-lingual transfer for dependency parsers has a long history. We encourage the reader to read a recent survey on this topic (Das and Sarkar, 2020). Our work does not use gold standard parallel data or even supervised translation models to apply annotation projection.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Study 1: Learning Translations with a Massively Multilingual Image Dataset

3.1.1 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.

3.1.1.1 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.

3.1.1.2 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 (Kilgariff, 2007). We take steps 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 appears on.

3.1.1.3 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 CLD21 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.

3.1.1.4 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.

Table 2: 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.

Language	Concreteness Ratings				Overall
	1-2	2-3	3-4	4-5	
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 2 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. Overall, 85% of the English images, 72% of French, 66% of Indonesian, and 60% of Uzbek were judged to be good.

3.1.2 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.

3.1.3 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:

$$\text{AVGMAX}(w_f, w_e) = \frac{1}{|w_f|} \sum_{i_f \in w_f} \max_{i_e \in w_e} (\text{cosine}(i_f, i_e)) \quad (3)$$

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.

3.1.3.1 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 3 shows the results reported by Kiela et al. (2015) on the BERGMA500 dataset, along with results using our image crawl method on BERGMA500’s vocabulary.

Table 3: Our results are consistently better than those reported by Kiela et al. (2015), on a similar set of 500 concrete nouns. The rightmost column shows the added challenge with our larger, more realistic dataset.

dataset	BERGMA500 Kiela et al. (2015)	BERGMA500 (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

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 BERGMA500 vocabularies, approximately 11% of the translation pairs are string-identical, like *film* $\leftrightarrow$ *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.

3.1.3.2 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 4 reports the top-10 accuracy across all 32 languages.

For each language, we predict a translation for each foreign word in the language’s CROWD-TRANS 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% is not as difficult as distinguishing between 10,

000 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 2. 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 words as in Tamil.

3.1.3.3 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 LARGE-ENG.

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

Table 5.

Table 4: Top-10 accuracy on 12 high-resource languages and 20 low-resource languages. The “all” column reports accuracy on the entire dictionary. The “#” column reports the size of the English vocabulary used for each experiment.

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	.0	.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
Language						
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 5: Top-10 expanded English each experiment, words were used as for each foreign word. results from Table 4 languages.

accuracy on the Dictionary task. For 263,102 English candidate translations The SMALL averages across the same 10

3.2 Study 2: Mitigating Geographic Bias of Image Classifiers with Multilingual Image Data

We use the MMID as our source of multilingual images. This study targets on 10 languages spoken in India (Bengali, Bishnupriya-Manipuri, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu), as well as English. We focused on a set of 6 people labels (“athlete”, “children”, “farmer”, “police”, “soldier”, “wedding”) To expand the dataset, the 10 closest synonyms found by using cosine similarity of Glove embeddings were used as queries.

3.2.1 Data Expansion, Cleaning, and Selection

Noise in MMID, through inaccurate translations or bad Google Image data, was a major issue in our initial dataset. We deleted folders of completely-noisy images (likely due to bad translations) and then sub-selected images from both English and non-English languages equally. Images with cartoons, text, irrelevant subject matter, or other noise were discarded.



Figure 7: Photos for “athlete”, “children”, and “farmer”. The photos on top are drawn from the English words in the MMID, and the photos on the bottom are drawn from their translations into a language spoken in India.



Figure 8: Photos for “police”, “soldier”, and “wedding”. The photos on top are drawn from the English words in the MMID, and the photos on the bottom are drawn from their translations into a language spoken in India.



Figure 9: Examples of noisy images that were discarded because they were illustrations rather than photos, or they contained text, or they did not correctly represent one of our six categories.

3.2.2 Model and Training

3.2.2.1 Training Details

The dataset that we used consisted of 120-150 images from each class. We used three different models to classify the data: logistic regression as a baseline, a convolutional neural network (CNN), and a residual CNN (Resnet-50). Because the pretrained Resnet in PyTorch expects 224x224 RGB images, we resized all input images to that size using bilinear interpolation. We tuned our batch size to 50 and used random horizontal flips and color jitter to combat overfitting. The Hindi and English datasets consisted of 750 and 799 images respectively. As a result, the validation sets for the English and Hindi datasets are 150 and 160 images respectively and for the combined dataset, 310 images.

Logistic Regression

The training curves for logistic regression on the combined dataset can be seen above. While the training accuracy reached around 90%, the validation set accuracy reached only around 45%. This is likely due to the simplistic assumption of linearity inherent to logistic regression. This resulted in the model being unable to generalize well over the entire data and overfitting on the training set.

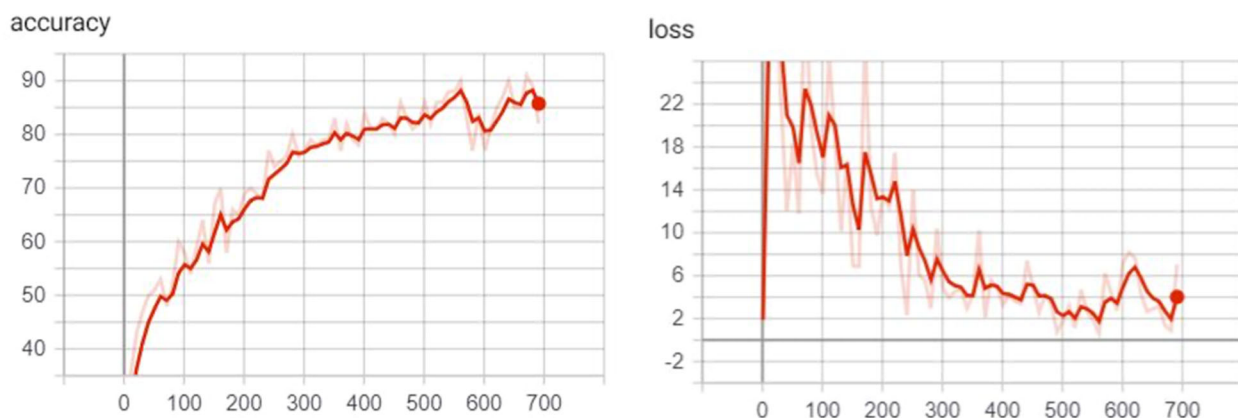


Figure 10: Training curves for logistic regression

CNN

The CNN that we built consisted of 6 convolutional layers followed by two fully connected layers to lower the dimensionality down to the 6 output channels. Each convolutional layer is followed by a ReLU activation function and batch normalization. The first two convolutional layers are also followed by a max pool layer to reduce the size of the images and offset any interpolations of the resize transformation mentioned earlier. Every convolution operation uses a 3x3 filter. There are dropout layers between convolutions 3 to 5 to prevent overfitting. The fully connected layers first reduce the number of outputs to 100 and then to 6. These 6 outputs are then put through a softmax layer to determine the probability that each image belongs to a specific class. The learning curves for the CNN on the combined dataset can be seen at the top right. While the training set did not reach the same accuracy as the logistic regression model, the validation set achieved an accuracy of around 70%, meaning the model did not overfit nearly as much.

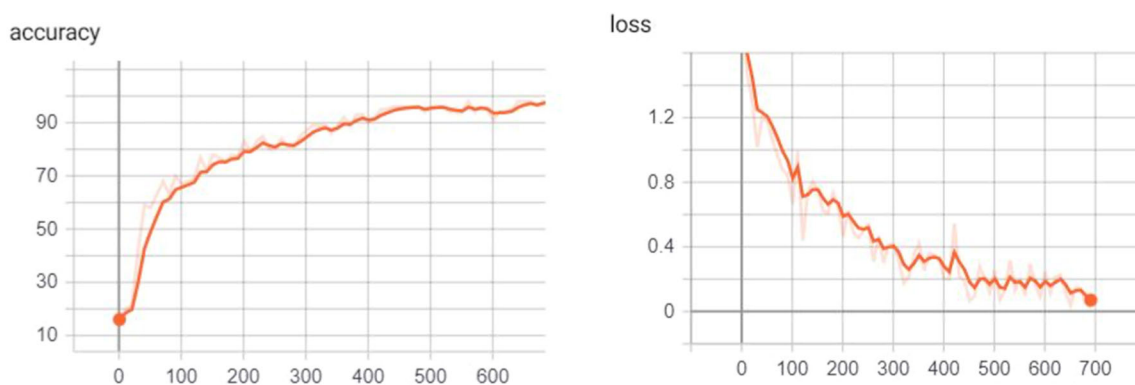


Figure 11: Training curves for CNN

Resnet

Resnet-50 outperformed the other two methods on the combined dataset, reaching a training accuracy of around 95% and validation accuracy of around 92%. The architecture we used was pre-trained on ImageNet, and we only added a fully connected layer at the end, once again outputting the network to 6 channels. Using a pretrained model was helpful given the limited size of our dataset, as it allowed us to use a model that already learned to identify edges and other features. Also, using a Resnet is often better than a CNN due to the residual connections that allow back-propagation between non-consecutive layers, even to the beginning of the network.

The results are summarized in the tables above. The

Table 76 shows the results of the combined dataset, while the Table 7 shows the results of training on the English dataset and validating on the combined dataset.

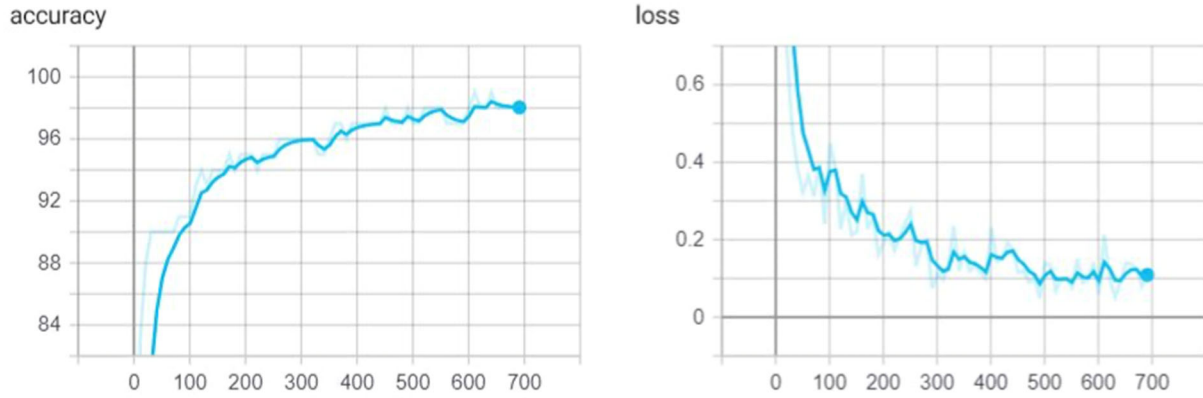


Figure 12: Training curves for Resnet

Table 6: Comparison of models trained on combined English/non-English dataset

Method	Train Set	Test Set	Test Accuracy
LogReg	Combined	Combined	0.452
CNN	Combined	Combined	0.698
Resnet	Combined	Combined	0.919

Table 7: Comparison of Resnet models

Method	Train Set	Test Set	Test Accuracy
Resnet	English	Combined	0.741
Resnet	Combined	Combined	0.919

Table 8: The 19 language pairs we explore in this work and the nature of their similarity: Geographical or Cultural: the same Language family, Ethnicity or Religion.

Language Pair	Geography	Family	Ethnicity	Religion
Azerbaijani Turkish	✓	✓	✓	✓
Azerbaijani Russian	✓	✗	✗	✗
Korean Chinese	✓	✗	✓	✓
Korean Japanese	✗	✗	✓	✓
Chinese Japanese	✗	✗	✓	✓
Chinese Korean	✓	✗	✓	✓
Japanese Chinese	✗	✗	✓	✓
Japanese Korean	✗	✗	✓	✓
Arabic Urdu	✗	✗	✗	✓
Arabic Farsi	✓	✗	✗	✓

Arabic	Hebrew	✓	✓	✗	✗
Urdu	Arabic	✗	✗	✗	✓
Urdu	Hindi	✓	✓	✓	✗
Spanish	French	✓	✓	✓	✓
Spanish	Portuguese	✓	✓	✓	✓
Finnish	Hungarian	✗	✓	✓	✓
Finnish	Norwegian	✓	✗	✗	✓
Afrikaans	Dutch	✗	✓	✓	✓
Afrikaans	Swahili	✗	✗	✗	✓

3.3 Study 3: Cultural and Geographical Influences on Image Translatability of Words across Languages

We again use the MMID. We choose to examine specific language pairs so that for each source language there are two or more target languages whose shared characteristics with the source language differ in zero or more aspects. The shared characteristics between the source and target language include shared culture (i.e., either they are from the same language family¹ or the communities at their cultural centers have the same major ethnic group² or major religion³) or shared geography (i.e., the countries at their cultural centers share land borders). For example, for Finnish, we include two target languages: one that has geographical proximity (Norwegian) and another that has ethnolinguistic similarity (Hungarian). In this way, we intend to examine for each source language, which of these groups of characteristics (culture or geography) are more important in image aided word translation, and whether culture or geography dominate one another. We form language pairs from the following 20 languages: Afrikaans (af), Arabic (ar), Azerbaijani (az), Chinese (zh), Dutch (nl), Finnish (fi), French (fr), Hebrew (he), Hindi (hi), Hungarian (hu), Japanese (ja), Korean (ko), Norwegian (no), Persian (fa), Portuguese (pt), Russian (ru), Spanish (es), Swahili (sw), Turkish (tr) and Urdu (ur). We summarize the language pairs and their shared characteristics in

Table 8.

3.3.1 Dataset collection and preparation

We download MMID images for all the source and target languages in our language pairs. In order to get vector embeddings for the images, we scale the images to 224 x 224 pixels, normalize and

¹ https://en.wikipedia.org/wiki/List_of_language_families

² https://en.wikipedia.org/wiki/Ethnic_group

³ <https://www.pewresearch.org/religion/2015/04/02/religious-projection-table/>

feed them as input into the ResNet-50 network (He et al., 2015), using network weights pre-trained on ImageNet. We obtain image embeddings from the last average pooling layer of ResNet-50, which gives us a 2048 dimensional vector embedding for each image. For each word, we call the embeddings of the associated images the word’s image embedding. Because cosine similarity, which underlies parts of this work and previous works for bilingual lexicon induction via images (Bergsma et al., 2011; Kiela et al., 2015; Hewitt et al., 2018), is non-invariant to translation (Korenienius et al., 2007) we treat all vectors with respect to the origin rather than some mean center for each image cluster.⁴

Since the MTurk word translations that come with MMID (Pavlick et al., 2014) are limited in coverage and quality i.e., they contain only translations *to* English and the coverage and quality are high ($\geq 70\%$ accuracy) only for a small set (13) of European and Indian languages where many MTurk workers are located; we create translation dictionaries for each of our language pairs using Google Translate, translating all words in the source language to the target language. We compute the translatability of words whose translations have associated images in MMID. If a word in the source is translated to a phrase, we use the last word in the phrase to find associated images in the dataset. This heuristic applies to only 10% of the words in the dataset and the Google translations with the majority (80%) of the first word in the phrase translations being indicative of functional words: shared and appearing more than 50 times in the dataset.

3.3.2 Methodology

Given images that are associated with the word w_s in the source language s , and w_t in the target language t , we define two measures that determine how well a word can be translated by its images. The first measures whether w_s and w_t have overlapping or disjoint image embeddings. The second measures whether the spread of the image embedding for w_s , and, similarly, for w_t , is tight or loose: such a measure of image dispersion has been found to help predict the usefulness of image representations for translation (Kiela et al., 2014, 2015).

Specifically, when images of w_s and w_t are tight and overlapping in the embedding space, it shows that the images have little diversity (low dispersion) and are similar between w_s and w_t , indicating potentially good translation between them. Conversely, if the images are either spread out or disjoint, it means that the images have greater diversity (high dispersion) or differ between w_s and w_t , indicating potentially poor translation between them. We refer to the degree of overlap between two clusters of images associated with w_s and w_t respectively as their *inter-cluster similarity*, and to the degree of tightness or looseness of the images in each cluster as their *intra-cluster similarity*.

Our conjecture is that this is equivalent to representing image embeddings as samples from some generator distribution G . We can call the generator distribution for a given source word G_s and a generator distribution for a given target word G_t . Two words are translations of each other when $G_s = G_t$, and conversely two words are poor translations when $G_s \neq G_t$. Thus, *inter-cluster similarity* checks to see if an image embedding from G_s could have been produced by G_t . Note that this is a necessary condition but it is not sufficient to say $G_s = G_t$ if inter-cluster similarity is high, because an image embedding from G_s can also be produced by some random image embedding generator

⁴ Our code is available at <https://github.com/nikzadkhani/MMID-CNN-Analysis>

G_r with $G_s \neq G_r$. *Intra-cluster similarity* is a measure of how similar samples from a single generator are to each other. This will ensure that G_t and G_s are not random generators and are accurate representations of the word they are generating image embeddings for. In other words, having a high intra-cluster similarity implies that $G_s \neq G_r$ and $G_t \neq G_r$. Thus, we have sufficient conditions when we have high inter- and intra-cluster similarities to say $G_s = G_t$.

3.3.2.1 Inter-Cluster Similarity

To measure the degree of overlap (*inter-cluster similarity*) between images associated with the word w_s in the source language, and those associated with the word w_t in the target language, we first cluster their image embeddings with a k -means clustering algorithm ($k = 2$). Then, we measure the degree of overlap between images of the two words by the homogeneity score of the resulting clusters, $h_{w_s, w_t} \in [0, 1]$ (Rosenberg and Hirschberg, 2007), calculated given the words w_s and w_t as image labels. A homogeneity score of 0 signals that all the image embeddings come from the distribution of a single class, hence represent the same word or concept ($w_s = w_t$). In this case, we say that the images of the two words have high inter-cluster similarity. A score of 1 means that the k -means clustering was able to identify two mutually exclusive clusters of images indicative that the images come from two different generators ($G_s \neq G_t$). In other words, the image embeddings were sampled from two different words or senses ($w_s \neq w_t$).

However, if images are highly dispersed (have high diversity, loose clusters), then the inter-cluster similarity may be deceptively high (i.e., low homogeneity) since loose clusters may overlap to some extent. Thus, the homogeneity score is only an effective measure of how good an image-aided translation is on the condition that the clusters are sufficiently tight (i.e., have high intra-cluster similarity).

3.3.2.2 Intra-Cluster Similarity

Images of a given word have low intra-cluster similarity when the images have high dispersion, which may be due to the word being abstract (e.g., words like *concept* whose images might be very diverse) or when the word has many different senses (e.g., words like *bug* whose images might represent the different senses of the word). On the other hand, when the intra-cluster similarity is high, it indicates that there is a general consensus on the meaning of the word as represented by the images, which makes for an easier transfer of the word meaning via images (i.e., better image translatability).

The metric we choose for the intra-cluster similarity of a word w is Median Max Cosine Similarity, which, given the set of images associated with the word, I_w , is:

$$\text{MEDMAX}_w = \text{median} \max_{i \in I_w} \max_{j \in I_w} \begin{cases} i \neq j : \text{cosine}(i, j) \\ i = j : 0 \end{cases} \quad (4)$$

This is a variation of the Average Maximum Cosine Similarity in Bergsma and Van Durme (2011), using the median to reduce the effect of outliers. Additionally, note that the worst case of this

metric giving an undesirable outcome is when we have 50 random pairs of image embeddings for a given word cluster. This will result in a high *intra-cluster similarity* despite the randomness of the overall cluster. However, in our findings this scenario is extremely unlikely. As words have dominant senses, the effect of outliers is mitigated due to the use of the median.

However, intra-cluster similarity on its own is not enough to indicate if the word in the target language w_t is a good translation of the word in the source language w_s . For example, the word *train* may be represented with images of locomotives in one language and with images of people exercising in another, if its meaning differs across languages. Both of the words' images will have high intra-cluster similarity but low inter-cluster similarity, indicating poor translatability via images. Thus, intra-cluster similarity is only an effective measure of how good an image-aided translation is, on the condition that the inter-cluster similarity is sufficiently high.

3.3.2.3 Inter-Cluster Similarity Threshold

We define a homogeneity score threshold to determine if two words w_s and w_t have sufficiently high overlap in their image embeddings to indicate a good translation. For each language pair, we compute this threshold by taking the average homogeneity score of clusters of images of 10 randomly chosen word pairs from the source s and the target language t . We take the average of these scores since we want to first be able to compare the threshold with other homogeneity scores and second to be able to capture the skew in negative thresholds as well. These pairs serve as negative examples of translation and we expect their image embeddings to be disjoint. Hence, a word pair with homogeneity score *lower* than this threshold means that the word pair has a good overlap in their image embeddings (i.e., a high inter-cluster similarity), which indicates a good translation.

3.3.2.4 Intra-Cluster Similarity Threshold

Similarly, we define an intra-cluster similarity threshold to determine if an image cluster associated with a word w is sufficiently tight. Since intra-cluster similarity is computed for each word (and not word pair), we compute this threshold $MEDMAX_{thld}$ for each language l by constructing a negative example for the language i.e., an image cluster with a high dispersion. We create this negative example by taking five random words from the language and for each word a random sample of 20 images to build a cluster of 100 images (mimicking the typical image cluster size for a word in our dataset). We set the Median Max Cosine Similarity of this image cluster as the intra-cluster similarity threshold. A word that has an intra-cluster similarity *higher* than this threshold would mean that this word has a tight image cluster, a consistent meaning as represented in its images' representations.

3.3.2.5 Normalized Score

We define a normalized score $NORM_{w_s, w_t}$ to combine intra- and inter-cluster similarity scores for a word w_s in the source language and its translation in the target language w_t .

Given the intra-cluster similarity of word w_s ($MEDMAX_{w_s}$) and that of word w_t ($MEDMAX_{w_t}$); and the maximum and minimum intra-cluster similarities for the source language ($MEDMAX_{max}$ and $MEDMAX_{min}$), and those of the target language ($MEDMAX_{max}$ and $MEDMAX_{min}$); as well as the homogeneity score of the words h_{w_s, w_t} and the maximum and minimum homogeneity scores of words in the language pair (h_{max} and h_{min}), we compute the normalized score $NORM_{w_s, w_t}$ as:

$$\text{NORM}_{w_s, w_t} = \text{NORM MEDMAX}_{w_s} + \text{NORM MEDMAX}_{w_t} - \text{NORM } h_{w_s, w_t}$$

Where:

$$\text{NORM MEDMAX}_{w_s} = \frac{\text{MEDMAX}_{w_s} - \text{MEDMAX}_s^{\min}}{2(\text{MEDMAX}_s^{\max} - \text{MEDMAX}_s^{\min})} \quad (5)$$

$$\text{NORM MEDMAX}_{w_t} = \frac{\text{MEDMAX}_{w_t} - \text{MEDMAX}_t^{\min}}{2(\text{MEDMAX}_t^{\max} - \text{MEDMAX}_t^{\min})}$$

$$\text{NORM } h_{w_s, w_t} = \frac{h_{w_s, w_t} - h_{s, t}^{\min}}{h_{s, t}^{\max} - h_{s, t}^{\min}}$$

For each language pair, we also define a threshold on this normalized score (i.e., $\text{NORM}_{\text{thld}}$) by substituting MEDMAX_{w_s} , MEDMAX_{w_t} , and h with $\text{MEDMAX}_{\text{thld}}$, $\text{MEDMAX}_{\text{thld}}$, and $h_{\text{thld } w_s, w_t \text{ s t s, t}}$ respectively in the equation above.

3.4 Study 4: Weakly Supervised Neural Translation Tailored to Cross-Lingual Tasks

3.4.1 Learning Translation from Wikipedia

The key component of our approach is to leverage the multilingual cues from linked Wikipedia pages across languages. Wikipedia is a great comparable data in which many of its pages explain entities in the world in different languages. In most cases, first sentences define or introduce the mentioned entity in that page (e.g. Figure 5). Therefore, we observe that many first sentence pairs in linked Wikipedia documents are rough translations of each other. Moreover, captions of images in different languages are usually similar but not necessarily direct translations of each other. We leverage this information to extract many parallel sentences from Wikipedia without using any external supervision. In this section, we describe our algorithm which is briefly shown in Figure 13.

Definitions: 1) e is English, f is the foreign language, and g is a language similar to f , 2) `learn_dict` (P) extracts a bilingual dictionary from parallel data P , 3) `t` ($x|m$) translates input x given model m , 4) `pretrain` (x) pretrains on monolingual data x using MASS (Song et al., 2019a), 5) `train` ($P|m$) trains on parallel data P initialized by model m , 6) `bt_train` ($x_1, x_2|m$) trains iterative back-translation on monolingual data $x_1 \in e$ and $x_2 \in f$ initialized by model m .

Inputs: 1) Wikipedia documents $w^{(e)}$, $w^{(f)}$, and $w^{(g)}$, 2) Monolingual word embedding vectors v_e and v_f , 3) Set of linked pages from Wikipedia `COMP`, their aligned titles $\mathcal{T}$, and their first sentence pairs $\mathcal{F}$, 4) Set of paired image captions $\mathcal{C}$, and 5) Gold-standard parallel data $\mathcal{P}^{(e,g)}$.

Algorithm:

→ **Learn bilingual dictionary and embeddings**

$\mathcal{S} = \mathcal{F} \cup \mathcal{C} \cup \mathcal{T}$

$\mathcal{D}^{(f,e)} = \text{learn_dict}(\mathcal{S})$

$\mathcal{D}^{(g,e)} = \text{learn_dict}(\mathcal{P}^{(e,g)})$ ▷ Related language

Learn $v_e \rightarrow v'_e$ and $v_f \rightarrow v'_f$ using $\mathcal{D}^{(f,e)} \cup \mathcal{D}^{(g,e)}$

→ **Mine parallel data**

Extract comparable sentences $\mathcal{Z}$ from `COMP`

Extract $\mathcal{P}^{(f,e)}$ from $\mathcal{Z}$.

$\mathcal{P}^{(f,e)} = \mathcal{P}^{(f,e)} \cup \mathcal{T}$ ▷ Mined Data

→ **Train MT with pretraining and back-translation**

$\theta_0 = \text{pretrain}(w^{(e)} \cup w^{(f)} \cup w^{(g)})$ ▷ MASS Training

$\theta_{\rightleftharpoons} = \text{train}(\mathcal{P}^{(f,e)} \cup \mathcal{P}^{(g,e)} | \theta_0)$ ▷ NMT Training

$\mathcal{P}^{(e \rightarrow f)} = (\text{t}(w^{(f)} | \theta_{\rightleftharpoons}), w^{(f)})$

$\mathcal{P}^{(f \rightarrow e)} = (\text{t}(w^{(e)} | \theta_{\rightleftharpoons}), w^{(e)})$

$\mathcal{P}'^{(f,e)} = \mathcal{P}^{(e \rightarrow f)} \cup \mathcal{P}^{(f \rightarrow e)} \cup \mathcal{P}^{(f,e)}$

$\theta'_{\rightleftharpoons} = \text{train}(\mathcal{P}'^{(f,e)} | \theta_0)$

$\theta_{\rightleftharpoons}^* = \text{bt_train}(w^{(e)}, w^{(f)} | \theta'_{\rightleftharpoons})$

Output: $\theta_{\rightleftharpoons}^*$

Figure 13: A brief description of the training pipeline for our weakly supervised cross lingual bootstrapping approach.

3.4.1.1 Data Definitions

For languages e and f in which e is English and f is a low-resource target language of interest, there are Wikipedia documents $w_e = \{w_{1(e)} \dots w_{n(e)}\}$ and $w_f = \{w_{1(f)} \dots w_{n(f)}\}$. We refer to $w_{(l)(i,j)}$ as the j th sentence in the i th document for language l . A subset of these documents are aligned (using Wikipedia *languages links*). Thus we have an aligned set of document pairs in which we can easily extract many sentence pairs that are potentially translations of each other. A smaller subset F is the set of first sentences in Wikipedia ($w_{e(i,1)}$, $w_{f(i',1)}$) in which documents i and i' are linked and

their first sentence lengths are in a similar range. In addition to text content, Wikipedia has a large set of images. Each image comes along with one or more captions, sometimes in different languages. A small subset of these images have captions both in English and the target language. We refer to this set as C . We use the set of all caption pairs (C), title pairs (T), and first sentences (F) as the seed parallel data: $S=FUCUT$.

3.4.2 Bilingual Dictionary Extraction and Cross-Lingual Word Embeddings

Having the seed parallel data S , we run unsupervised word alignment (Dyer et al., 2013) in both English-to-target, and target-to-English directions. We use the intersected alignments to extract highly confident word-to-word connections. Finally, we pick the most frequently aligned word for each word in English as translation. This set serves as a bilingual dictionary D .

Given two monolingual trained word embeddings $v_e \in \mathbb{R}_{N_e \times d}$ and $v_f \in \mathbb{R}_{N_f \times d}$, and the extracted bilingual dictionary D , we use the method of Faruqui and Dyer (2014) to project these two embedding vectors to a shared cross-lingual space. This method uses a bilingual dictionary along with canonical correlation analysis (CCA) to learn two projection matrices to map each embedding vector to a shared space.

3.4.3 Mining Parallel Sentences

We use cross-lingual embedding vectors for calculating cosine similarity between pairs of words. Moreover, we use the extracted bilingual dictionary to boost the accuracy of the scoring function. For a pair of sentences (s, t) after filtering sentence pairs with different numerical values (e.g. sentences containing 2019 in the source and 1987 in the target), we use a modified version of cosine similarity between words:

$$sim(s_i, t_j) = \begin{cases} 1.0, & \text{if } (s_i, t_j) \in \mathcal{D} \\ cos(s_i, t_j), & \text{otherwise} \end{cases} \quad (6)$$

Using the above definition of word similarity, we use the average-maximum similarity between pairs of sentences.

$$score(s, t) = \frac{\sum_{i=1}^n \max_{j=1}^m sim(s_i, t_j)}{n} \quad (7)$$

From a pool of candidates, we pick those pairs that have the highest score in both directions.

3.4.4 Leveraging Similar Languages

In many low-resource scenarios, the number of paired documents is very small, leading to a small number and often noisy extracted parallel sentences.

To alleviate this problem to some extent, we assume to have another language g in which g has a large lexical overlap with the target language f (such as g =Russian and f =Kazakh). We assume that a parallel data exists between language g and English, and we can use it both as an auxiliary parallel data in training, and also for extracting extra lexical entries for the bilingual dictionaries: as shown in Figure 13, we supplement the extracted bilingual dictionary from seed parallel data with the bilingual dictionary extracted from related language parallel data.

3.4.5 Translation Model

We use a standard sequence-to-sequence transformer-based translation model (Vaswani et al., 2017) with a six-layer BERT-based (Devlin et al., 2019) encoder-decoder architecture from HuggingFace (Wolf et al., 2019) and Pytorch (Paszke et al., 2019) with a shared SentencePiece (Kudo and Richardson, 2018) vocabulary. All input and output token embeddings are summed up with the language id embedding. First tokens of every input and output sentence are shown by the language ID. Our training pipeline assumes that the encoder and decoder are shared across different languages, except that we use a separate output layer for each language in order to prevent input copying (Artetxe et al., 2018b; Sen et al., 2019). We pretrain the model on a tuple of three Wikipedia datasets for the three languages g , f , and e using the MASS model (Song et al., 2019a). The MASS model masks a contiguous span of input tokens, and recovers that span in the output sequence.

To facilitate multi-task learning with image captioning, our model has an image encoder that is used in cases of image captioning (more details in §4.1). In other words, the decoder is shared between the translation and captioning tasks. We use the pretrained ResNet-152 model (He et al., 2016) from Pytorch to encode every input image. We extract the final layer as a 7×7 grid vector and project it to a new space by a linear transformation and then add location embeddings by using entry-wise addition. Afterwards, we assume that the 49 vectors are encoded text representations as if a sentence with 49 words occurs. This is similar to but not exactly the same as the Virtex model (Desai and Johnson, 2021).

3.4.5.1 Back-Translation: One-shot and Iterative

Finally, we use the back-translation technique to improve the quality of our models. Back Translation is done by translating a large amount of monolingual text to and from the target language. The translated texts serve as noisy input text along with the monolingual data as the silver standard translations. Previous work (Sennrich et al., 2016b; Edunov et al., 2018) has shown that back-translation is a very simple but effective technique to improve the quality of translation models. Henceforth, we refer to this method as one-shot back-translation. Another approach is to use iterative back-translation (Hoang et al., 2018), the most popular approach in unsupervised translation (Artetxe et al., 2018b; Conneau and Lample, 2019; Song et al., 2019a). The main difference from one-shot translation is that the model uses an online approach, and updates its parameters in every batch.

We empirically find one-shot back-translation faster to train but with much less potential to reach a high translation accuracy. A simple and effective way to have both a reliable and accurate model is to first initialize a model with one-shot back-translation, and then apply iterative back translation. The model that is initialized with a more accurate model reaches a higher accuracy.

3.4.6 Cross-Lingual Tasks

In this section, we describe our approaches for tailoring our translation models to cross-lingual tasks. Note that henceforth we assume that our translations model training is finished, and we have access to trained translation models for cross-lingual tasks.

3.4.6.1 Cross-Lingual Image Captioning

Having gold-standard image captioning training data $I = \{(I_i, c_i)\}_{i=1}^n$ where I_i is the image as pixel values, and as the textual description with k_i words, our goal is to learn a captioning model that is able to describe new (unseen) images. We use a transformer decoder from our translation model and a ResNet image encoder (He et al., 2016) for our image captioning pipeline. Unfortunately, annotated image captioning datasets do not exist in many languages. Having our translation model parameter θ , we can use its translation functionality to translate each caption c_i to $c'_i = \text{translate}(c_i|\theta)$. Afterwards, we will have a translated annotated dataset $I' = \{(I_i, c'_i)\}_{i=1}^n$ in which the textual descriptions are not gold-standard but translations from the English captions. Figure 14 shows a real example from MS-Coco (Chen et al., 2015) in which Arabic translations are provided by our translation model. Furthermore, to augment our learning capability, we initialize our decoder with decoding parameters of θ , and also continue training with both English captioning and translation.

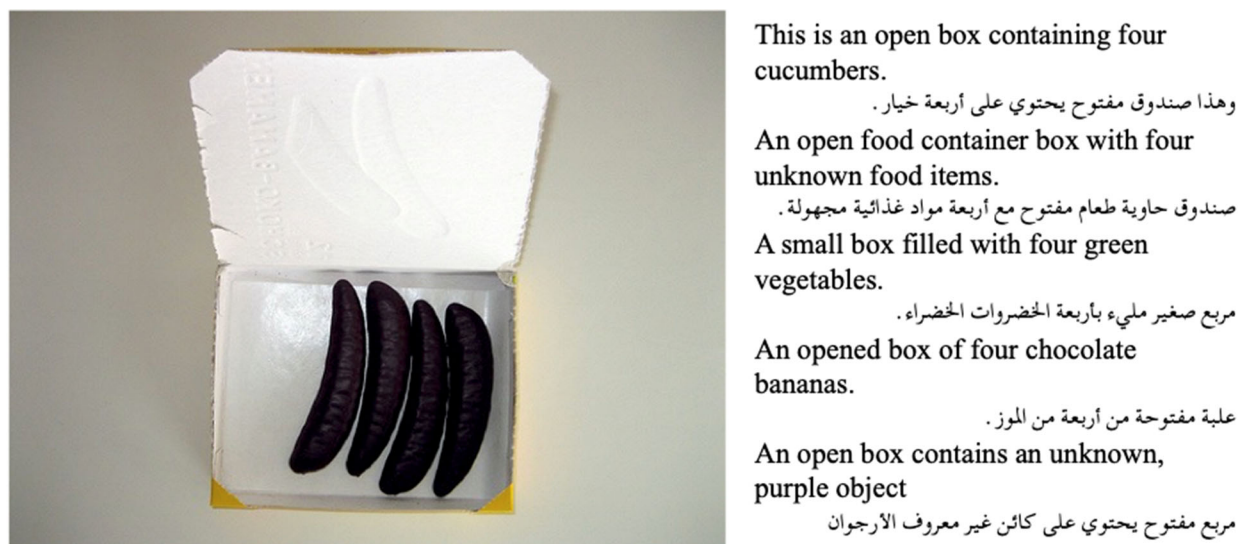


Figure 14: An image from MS-Coco (Chen et al., 2015) with gold-standard English captions, and Arabic translations from our Wikily translation model.

3.4.6.2 Cross-Lingual Dependency Parsing

Assuming that we have a large body of monolingual text, we translate that monolingual text to create artificial parallel data. We run unsupervised word alignments on the artificial parallel text. Following previous work (Rasooli and Collins, 2015; Ma and Xia, 2014), we run Giza++ (Och and Ney, 2003b) alignments on both source-to-target and target-to-source directions, and extract intersected alignments to keep high-precision one-to-one alignments. We run a supervised dependency parser of English as our rich-resource language. Then, we project dependencies to the target language sentences via word alignment links. Inspired by previous work (Rasooli and Collins, 2015), to remove noisy projections, we keep those sentences that at least 50% of words or 5 consecutive words in the target side have projected dependencies.

4.0 RESULTS AND DISCUSSION

4.1 Study 1: Learning Translations with a Massively Multilingual Image Dataset

4.1.1 Evaluation by Part-of-speech

Can images be used to translate words other than nouns? This section presents our methods for determining part-of-speech for foreign words even in low-resource languages, and presents our image based translation results across part-of-speech.

4.1.1.1 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 a 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 spaCy², 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).



Figure 15: 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.

4.1.1.2 Accuracy by Part-of-speech

As we see in the results in Table 4, 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 15 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.

Figure 16 and Figure 17 illustrate the effect of concreteness for visual representations of nouns.

4.1.2 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.



Figure 16: Strawberry is an example of a word with a high concreteness value in psycholinguistic studies. This is reflected with a consistent visual representation.

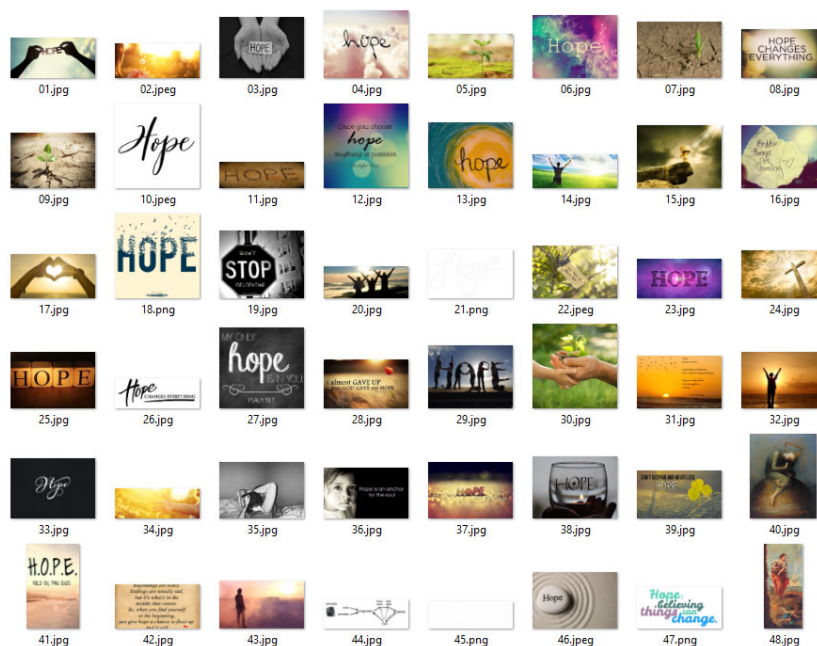


Figure 17: Hope is an example of a word with a low concreteness value in psycholinguistic studies. This is reflected with inconsistent visual representations, both within a single language and across languages.

4.1.3 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 separately or concatenated 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 18.

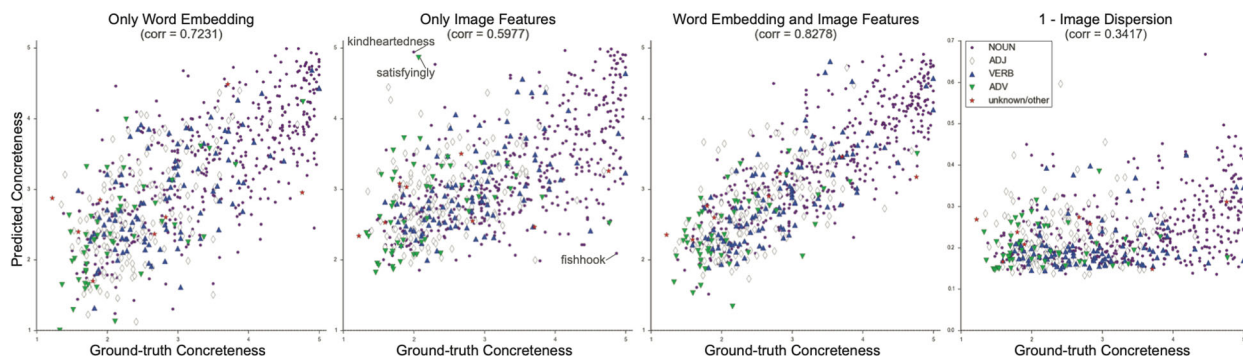


Figure 18: Plots visualizing the distribution of concreteness predictions on the validation set for our three trained models and for image dispersion. Spearman correlation coefficients are shown.

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, in our evaluation we used the image-embeddings-only model to predict concreteness for every English and foreign word in our dataset.

4.1.3.1 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 image sets are most concrete, and thereby estimate the likelihood that we will obtain a high quality translation.

Figure 19 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

threshold, 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.

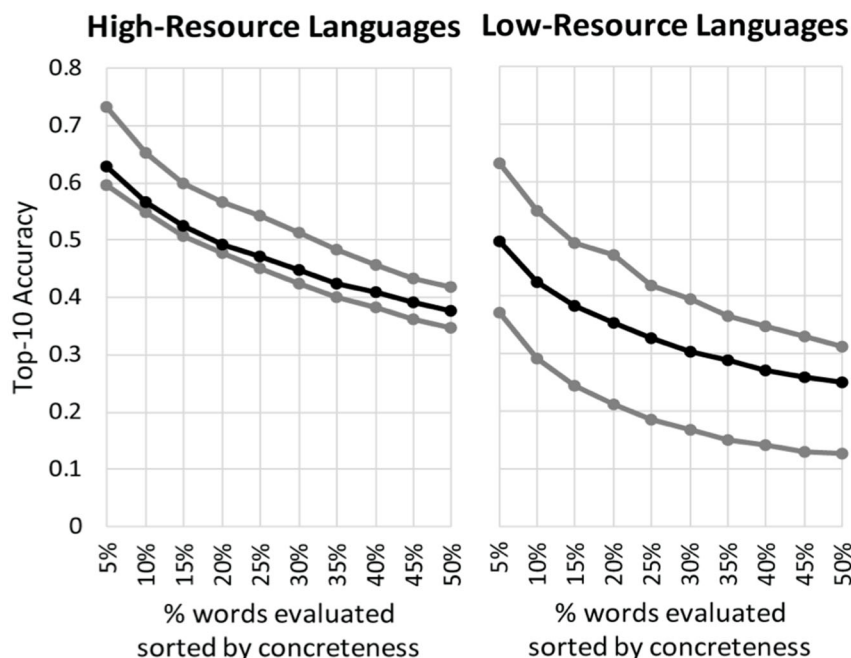


Figure 19: The black curve shows mean top-10 accuracy over the HIGHRES and LOWRES sets sorted by predicted concreteness. The gray curves show the 25th and 75th percentiles.

4.1.4 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.

4.1.4.1 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 $\text{TXT}(p_{f,e})$ and $\text{IMAGE}(p_{f,e})$ for all pairs. For each foreign word, we also have the concreteness score, $\text{CNC}(w_f)$, predicted from its image set.

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.

$$\begin{aligned} \text{MLP}(p_{f,e}) = \\ \text{MLP}([\text{TXT}(p_{f,e}); \text{IMAGE}(p_{f,e}); \text{CNC}(w_f)]) = \{\pm 1\} \end{aligned} \quad (8)$$

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 $\text{MLP}(p_{f,e})$. Using these scores, we sort E for each w_f .

4.1.4.2 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 percentages of testing data sorted by predicted concreteness. As shown in Figure 20, both image augmented methods beat TXT across concreteness thresholds on the top-1 accuracy metric.

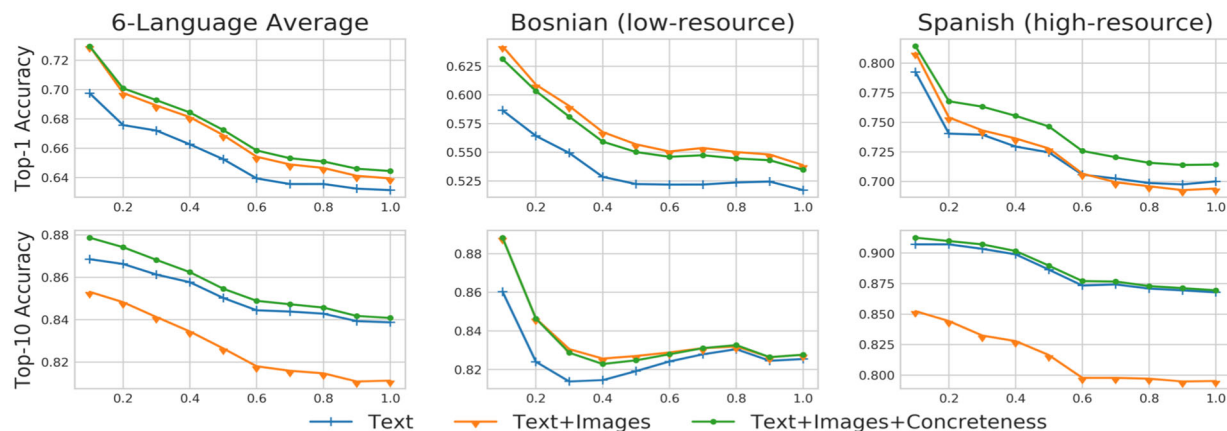


Figure 20: Reranking top-1 and top-10 accuracies of our image+text combination systems compared to the text-only BPR system. The X-axis shows percent of foreign words evaluated on, sorted by decreasing predicted concreteness.

Results across the 6 languages are reported in Table 69. Confirming our intuition, images are useful at high concreteness, improving the state-of-the-art text based method 3.21% at 10% recall. At 100% recall our method with images still improves the state-of-the-art by 1.3%. For example, the text-only system translates the Bosnian word *kos'arkas'ki* incorrectly as *football*, whereas the image+text system produces the correct *basketball*.

Table 9: 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.

	Method	% words evaluated		
		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

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.

4.2 Study 2: Mitigating Geographic Bias of Image Classifiers with Multilingual Image Data

Confusion matrices are useful for determining which classes a classifier is performing relatively poorly on and relatively well on. We first created a confusion matrix to summarize the results of training on an English dataset and testing on a combined validation set. The confusion matrix can be seen above. Interestingly enough, our classifier did a very good job of learning most classes. However, it mostly classified images of weddings as children. This is likely due to the fact that the Western images of children often are filled with a wide variety of colors, as are Indian, but not Western, weddings.

We created the confusion matrix shown in the top right by training and testing on the combined dataset. The classifier performed very well, and seemed to learn to classify Indian weddings properly. However, certain classes became confused more often, such as police and soldiers. This is likely due to the fact that in India, police wear brown and sometimes camouflage outfits, which could be confused with military.

Another interesting result was that training and classifying on Indian language images resulted in a validation accuracy of around 90%, while training and classifying on English images resulted in a validation accuracy of around 96%. In the future, a model that could identify what language an image is from and put the image through a trained Resnet model of that language might outperform our model.

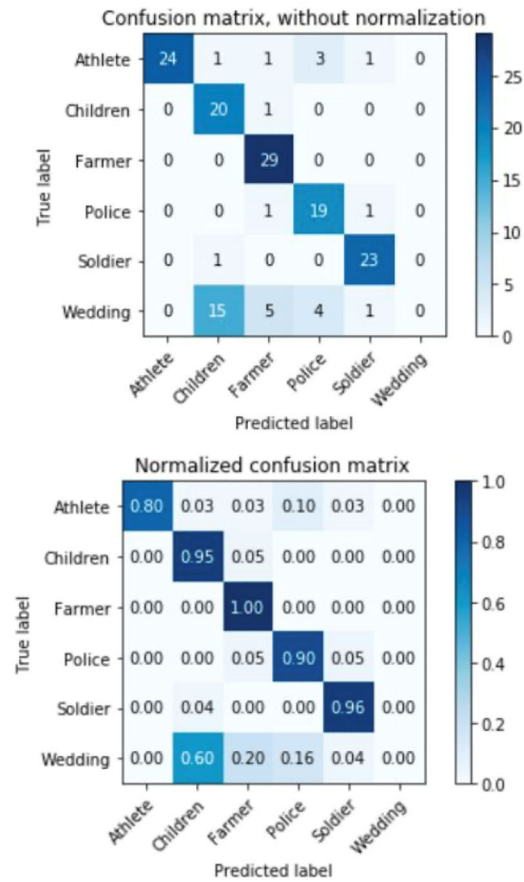


Figure 21: Confusion matrices for combined, network trained on English

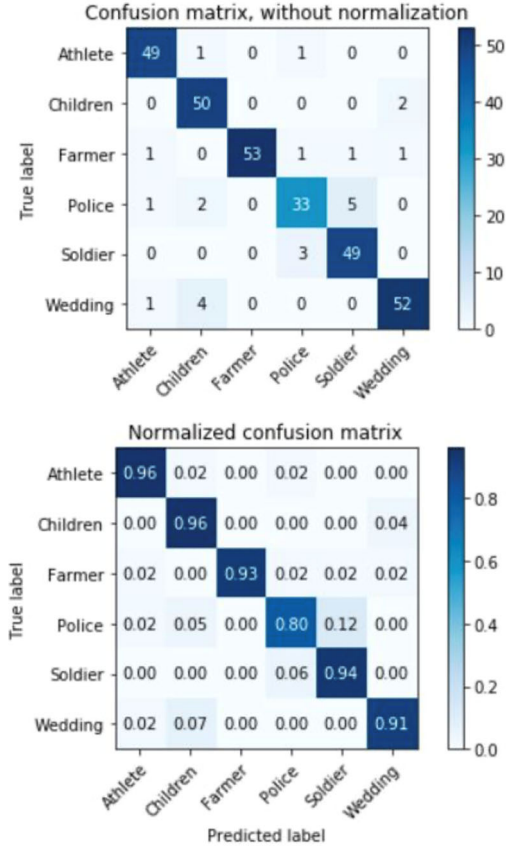


Figure 22: Confusion matrices for combined, network trained on combined

4.3 Study 3: Cultural and Geographical Influences on Image Translatability of Words across Languages

In order to compare the image translatability of two language pairs with different characteristics, we compare the ratio of the number of word pairs that are good translations divided by the total number of word pairs in each language pair using equation (1).

A word pair w_s and w_t has a good translation via images (or good image translatability) if its homogeneity score h_{w_s, w_t} is lower than the homogeneity threshold h_{thld} and its Median Max Cosine $s_{s,t}$ Similarities i.e., $MEDMAX_{w_s}$ and $MEDMAX_{w_t}$ are higher than the thresholds $MEDMAX_{thld}$ and $MEDMAX_{thld}$, respectively.

The higher this ratio, the more translatable the language pair is via images. When we compare two language pairs that have the same source language but different target languages (with different shared characteristics with the source), we can distinguish how different characteristics such as cultural similarity or geographical proximity affect image translatability. In Table 10, we show image translatability ratios of language pairs with the same source but different target languages side-by-side.

To understand the role of concreteness in translation via images, we also compute how many concrete words have good translations according to our image translatability measures. We consider a word pair, w_s and w_t , to be concrete if w_s has a concreteness score greater than 3. Source word concreteness is taken to be the pair concreteness, considering translation directionality. The ratio of how many concrete words in each language pair have good translations is also shown in Table 10.

Table 10: Language word pairs and their and translatability ratios, for all and concrete word pairs. Bold indicates pairs with the same source language whose normalized scores are significantly different and whose ratio differences are high.

Language Pair		Number of Words		Ratio	
		All	Concrete	All	Concrete
az	tr	4538	3470	0.31	0.37
az	ru	5380	2953	0.17	0.22
ko	zh	338	214	0.18	0.22
ko	ja	748	499	0.69	0.72
zh	ja	367	212	0.56	0.58
zh	ko	310	197	0.36	0.44
ja	zh	212	137	0.39	0.44
ja	ko	741	488	0.67	0.70
ar	ur	4916	3226	0.39	0.44
ar	fa	448	318	0.50	0.55
ar	he	2887	1874	0.69	0.73
ur	ar	4243	2466	0.39	0.45
ur	hi	4588	2817	0.12	0.15
es	fr	6392	3506	0.45	0.58
es	pt	7116	3920	0.40	0.53
fi	hu	5615	3190	0.29	0.40
fi	no	5336	3033	0.17	0.26
af	nl	5436	3247	0.39	0.50
af	sw	4553	2611	0.25	0.31

To test whether the difference in image translatability between language pairs that share the same source language (e.g., Finnish to Norwegian vs. Finnish to Hungarian) is statistically significant, we conduct a simple t-test between their normalized score distributions. The resulting p-values signal the difference between their distributions. Low p-values (< 0.05) indicate statistical significance and high variation between the distributions, while higher values suggest low variation and large similarities between the language pairs (Table 11).

From the t-test, we find that the difference in distributions of pairs that share the same source language is almost in all statistically significant (p-value < 0.05 , Table 11) except for Japanese to Chinese vs. Japanese to Korean.

Of other pairs whose normalized score differences are statistically significant and whose differences in translatability ratios are high (boldfaced, Table 10), we observe that the language pair with the higher image translatability ratio (i.e., Azerbaijani to Turkish, Korean to Japanese, Chinese to Japanese, Arabic to Hebrew, Urdu to Arabic, Finnish to Hungarian, and Afrikaans to Dutch) is always the pair that shares cultural similarity (i.e., either similar language family, similar major ethnicity, or similar major religion) even when they have little to no geographical proximity. For example, between Arabic and Hindi, Urdu's words are more translatable via images to Arabic (whose speakers share the same major religion as speakers of Urdu), despite Pakistan's geographic proximity to India. Similarly, Finnish words are more translatable via images to Hungarian (whose speakers belong to the same ethnolinguistic group as speakers of Finnish) than to Norwegian, despite Hungary not sharing any land border with Finland. In addition, there may be other language relatedness factors that result in better image translatability between languages, such as the similar writing system of Chinese and Japanese, or the similar grammatical structure of Korean and Japanese; despite their different language families.

From Table 10 we can see that Spanish shares similar characteristics with both French and Portuguese. Similarly, Japanese shares similar attributes with both Chinese and Korean (matching ethnicity and religion, different geography and language family). In such cases, where the two language pairs do not differ in characteristics, we observe that the difference in their translatability ratios is either small (in the case of Spanish to French and Spanish to Portuguese ratios in Table 10) or insignificant (in the case of Japanese to Chinese and Japanese to Korean p-value in Table 11).

On the contrary, Korean to Japanese and Korean to Chinese pairs, and Chinese to Korean and Chinese to Japanese pairs, have at least one difference in their attributes, accounting for the pairs' results' statistical significance.

In addition, we observe that concreteness of words largely affects the quality of translations due to the low diversity in its image representations, which facilitate translation between words. On average, across language pairs, 62.4% of words with normalized scores above the threshold are concrete, while only 37.6% are abstract. At the same time, in Table 10 we see that the translatability ratio is considerably higher for concrete word pairs than all pairs. This supports our idea, and other previous works, that concrete words are better represented visually and, so, more likely to have good image aided translations, compared to abstract ones.

Table 11: p-values of differences between normalized score distributions of language pairs that share the same source language.

Language Pair I		Language Pair II		p-value
az	tr	az	ru	2.26×10^{-25}
ko	zh	ko	ja	2.2×10^{-4}
zh	ja	zh	ko	10.6×10^{-5}
ja	zh	ja	ko	0.43
ar	ur	ar	fa	9.44×10^{-29}
ar	fa	ar	he	2.39×10^{-13}
ar	he	ar	ur	16.35×10^{-5}
es	fr	es	pt	4.3×10^{-47}
fi	hu	fi	no	2.66×10^{-104}
af	nl	af	sw	10^{-100}



Figure 23: (left) PCA plot of image embeddings of the word universe in Afrikaans (heelal) and in Dutch (universum) showing their tightly overlapping image clusters, (center) images of heelal in Afrikaans, (right) images of universum in Dutch



Figure 24: (left) PCA plot of image embeddings of the word dance in Afrikaans (dans) and in Swahili (kucheza) showing their tight but disjoint image clusters, (center) images of dans in Afrikaans, (right) images of kucheza in Swahili



Figure 25: Images of the word park in (left) Afrikaans, (center) Dutch, (right) Swahili.

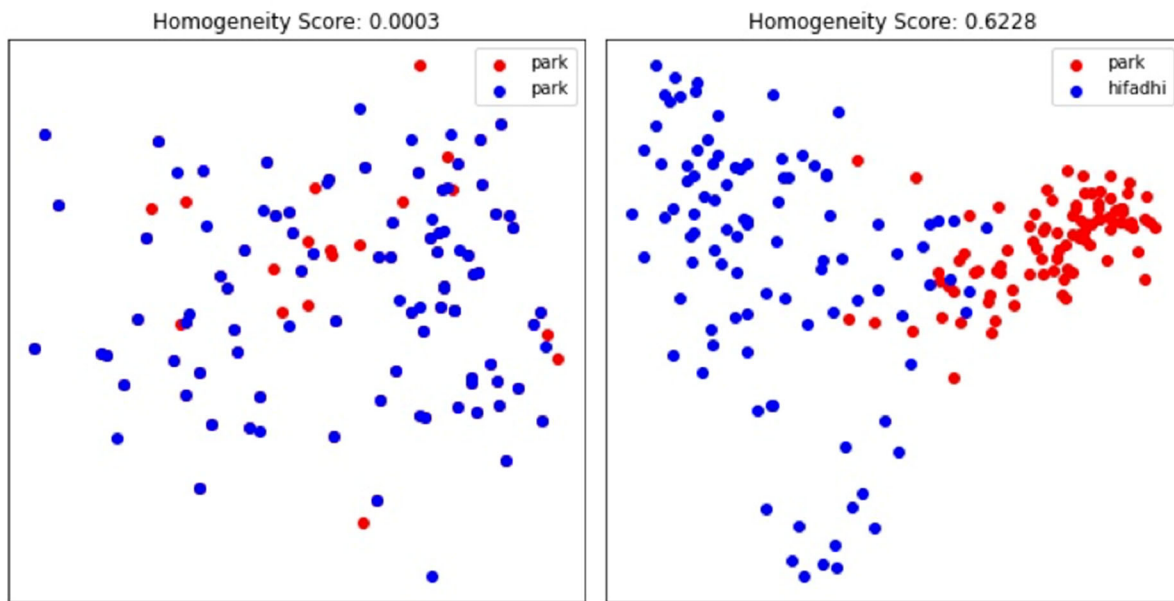


Figure 26: PCA plots for image embeddings of (left) park in Afrikaans to park in Dutch and (right) park in Afrikaans to hifadhi in Swahili.

Our work has identified that language relatedness affects word translations via images. We observe that languages with cultural relatedness have better image translatability, suggesting that cultural relatedness should be taken into account when using images to aid translation. The image translatability measures we have defined can be used to identify a potentially good or poor translation or discover a cultural similarity or disconnect between words in two languages. For example, a word pair that has a high intra-cluster similarity and a high inter-cluster similarity in their image representations indicates that the image clusters are tight and overlapping, signaling a good translation between them. For example, the word *heelal* in Afrikaans and the word *universum* in Dutch have tight and overlapping image clusters (i.e., low homogeneity score) as can be seen in the PCA plot of their image embeddings and in their images (Figure 23).

On the other hand, when a word pair has tight but disjoint image clusters, it can mean that their images express different meanings of the word. For example, the word *dance*: *dans* in Afrikaans and the word *kucheza* in Swahili have tight but disjoint image clusters (i.e., high homogeneity score) as can be seen in the PCA plot of their image embeddings and in the images (Figure 24), as *kucheza* means both to *play* and to *dance* in Swahili.

We observe that Afrikaans, for example, has a higher image translatability to Dutch due to their cultural (ethnolinguistic) similarity, than to Swahili, despite the relative distances of their cultural centers—South Africa is more distant geographically to the Netherlands than to Tanzania, defined in Glottolog as being the cultural center of the Swahili language. We observe higher visual similarities between words that are translations of each other in Afrikaans and Dutch than in Afrikaans and Swahili. For example, images of the word *park* in Afrikaans are more visually similar to images of the word *park* in Dutch than to images of the word *park* in Swahili (*hifadhi*) (Figure 25). The images of *park* in Afrikaans and in Dutch refer to a Western style park, while its images in

Swahili refer more to a wildlife reservation, a culturally different representation of the word *park* that is potentially influenced by how speakers of the different languages interact with the concept of the word. Interestingly, such connotations that are apparent in images may not be apparent in word embeddings, since *hifadhi* is used similarly with *park* in the texts of the language.

4.4 Study 4: Weakly Supervised Neural Translation Tailored to Cross-Lingual Tasks

In this section, we provide details about our experimental settings and results for translation, captioning, and dependency parsing.

4.4.1 Datasets and Settings

Languages: We focus on four language pairs: Arabic-English, Gujarati-English, Kazakh-English, and Romanian-English. We choose these pairs to provide enough evidence that our model works in distant languages, morphologically-rich languages, as well as similar languages. As for similar languages, we use Persian for Arabic (written with very similar scripts and have many words in common), Hindi for Gujarati (similar languages), Russian for Kazakh (written with the same script), and Italian for Romanian (Romance languages).

Monolingual and Translation Datasets: We use a shared SentencePiece vocabulary (Kudo and Richardson, 2018) with size 60K. Table 12 shows the sizes of Wikipedia data in different languages. For evaluation, we use the Arabic-English UN data (Ziems et al., 2016), WMT 2019 data (Barrault et al., 2019) for Gujarati-English and Kazakh-English, and WMT 2016 shared task data (Bojar et al., 2016) for Romanian-English. Following previous work (Sennrich et al., 2016a), diacritics are removed from the Romanian data.

Table 12: Data sizes for different pairs. We use a sample of English sentences with similar sizes to each data.

Direction	ar $\leftrightarrow$ en	gu $\leftrightarrow$ en	kk $\leftrightarrow$ en	ro $\leftrightarrow$ en
Foreign docs	1.0m	28k	230k	400k
Paired docs	745k	7.3k	80k	270k
First sents.	205k	3.2k	52k	78k
Captions	92k	2.2k	1.9k	35k
Comparable pairs	0.1b	14m	32m	64m
Mined sents.	1.7m	49k	183k	675k
BT	2.1m	1.5m	2.2m	2.1m
Iterative BT	4.0m	3.8m	4.0m	6.1m

4.4.2 Pretraining

We pretrain four models on 3-tuples of languages via a single NVIDIA Geforce RTX 2080 TI with 11GB of memory. We create batches of 4K words, run pretraining for two million iterations where we alternate between language batches, and accumulate gradients for 8 steps. We use the apex library³ to use FP-16 tensors. This whole process takes four weeks in a single GPU. We use the Adam optimizer (Kingma and Ba, 2015) with inverse square root and learning rate of 10^{-4} , 4000 warm-up steps, and dropout probability of 0.1.

4.4.3 Translation Results

Training Table 12 shows the sizes of different types of datasets in our experiments. We pick comparable candidates for sentence pairs whose lengths are within a range of half to twice of each other. As we see, the final size of mined datasets heavily depends on the number of paired English-target language Wikipedia documents. We train our translation models initialized by pretrained models. All of our evaluations are conducted using SacreBLEU (Post, 2018) except for en↔ro in which we use BLEU score (Papineni et al., 2002) from Moses decoder scripts (Koehn et al., 2007) for the sake of comparison to previous work.

Table 13 shows the results of different settings in addition to baseline and state-of-the-art results. We see that Arabic as a clear exception needs more rounds of training: we train our Arabic model once again on mined data by initializing it by our back-translation model. We have not seen further improvement by back-translation. To have a fair comparison, we list the best supervised models for all language pairs (to the best of our knowledge). In low-resource settings, we outperform strong supervised models that are boosted by backtranslation. In high-resource settings, our Arabic models achieve very high performance but regarding the fact that the parallel data for Arabic has 18M sentences, it is quite impossible to reach that level of accuracy.

Table 13: BLEU scores for different models. Reference results are from *: Our implementation, 1: Kim et al. (2020), 2: Li et al. (2020), 3: Liu et al. (2020) (supervised), 4: Tran et al. (2020) (unsupervised with mined parallel data).

Model	ar→en	en→ar	gu→en	en→gu	kk→en	en→kk	ro→en	en→ro
UNMT	Conneau and Lample (2019)	—	—	—	—	—	31.8	33.3
	Song et al. (2019a) (MASS; 8 GPUs)	—	—	—	—	—	33.1	35.2
	Best published results	11.0*	9.4*	0.6 ¹	0.6 ¹	2.0 ¹	0.8 ¹	37.6 ⁴
Wikily UNMT	First sentences + captions + titles	6.1	3.1	0.7	1.1	2.3	1.0	2.0
	Mined Corpora	23.1	19.7	4.2	4.9	2.8	1.6	22.1
	+ Related Language	—	—	9.1	7.8	7.3	2.3	23.2
	+ One-shot back-translation (bt-beam=4)	23.0	18.8	13.8	13.9	7.0	12.1	25.2
	+ Iterative back-translation (bt-beam=1)	24.4	18.9	13.3	15.2	9.0	10.8	32.5
	+ Retrain on mined data	30.6	23.4	—	—	—	—	—
	(Semi-)Supervised	48.9*	40.6*	14.2 ¹	4.0 ¹	12.5 ¹	3.1 ¹	39.9 ³

Figure 27 shows a randomly chosen example from the Gujarati-English development data. As depicted, we see that the model after back-translation reaches to somewhat the core meaning of the sentence with a bit of divergence from exactly matching the reference. The final iterative back translation output almost catches a correct translation. We also see that the use of the word “creative” is seen in Google Translate output, a model that is most likely trained on much larger parallel data than what is currently available for public use. In general, unsupervised translation performs very poorly compared to our approach in all directions.

Input	અર્થાત આપણે પહેલા તુલનાએ વધુ રચનાત્મક બનવું પડશે.
Outputs	Unsupervised First sentences + captions + titles Mined Corpora + Related Language + One-shot back-translation + Iterative back-translation
	Ut numerousત્રીlit the mother, onwards, in theover અધિકૃશexualit theotherit theIN રોS 19 A view of the universe from the present to the present day. For example, if the ghazal is more popular than ghazal. We need to become more creative than before. For example, we must become more creative than before. Meanwhile, we'll have to become more constructive than before.
Google Translate	That means we have to be more creative than before.
Reference	That means we have to be more constructive than before.

Figure 27: An example of a Gujarati sentence and its outputs from different models, as well as Google Translate.

4.4.4 Image Captioning Results

We use the Flickr (Hodosh et al., 2013) and MS-Coco (Chen et al., 2015) datasets for English, and the gold-standard Arabic Flickr dataset (ElJundi. et al., 2020) for evaluation. The Arabic test set has 1000 images with 3 captions per image. We translate all the training datasets to Arabic for having translated caption data. The final training data contains 620K captions for about 125K unique images. Throughout experiments, we use the pretrained Resnet-152 models (He et al., 2016) from Pytorch (Paszke et al., 2019), and let it fine-tune during our training pipeline. Each training batch contains 20 images. We accumulate gradients for 16 steps, and use a dropout of 0.1 for the projected image output representations. Other training parameters are the same as our translation training. To make our pipeline fully unsupervised, we use translated development sets to pick the best checkpoint during training.

Table 14 shows the final results on the Arabic test set using the SacreBLEU measure (Post, 2018). First, we should note that similar to ElJundi. et al. (2020), we see lower scales of BLEU scores due to morphological richness in Arabic. We see that if we initialize our model with the translation model and multitask it with translation and also English captioning, we achieve much higher performance. It is interesting to observe that translating the English output on the test data to Arabic achieves a much lower result. This is a strong indicator of the strength of our approach. We also see that supervised translation fails to perform well. This might be due to the UN translation training dataset which has a different domain from the caption dataset. Furthermore, we see that our model outperforms Google Translate which is a strong machine translation system, and that is actually what is being used as seed data for manual revision in the Arabic dataset. Finally, it is interesting to see that our model outperforms supervised captioning. Multi-tasking makes translation performance slightly worse.

Table 14: Image captioning results evaluated on the Arabic Flickr dataset (ElJundi. et al., 2020) using SacreBLEU (Post, 2018). “pretrained” indicates initializing our captioning model with our translation parameters.

	Supervision	Pretrained	Multi-task		BLEU	
			EN	MT	@1	@4
Translate train data	wikily	✗	✗	✗	33.1	4.57
	wikily	✓	✗	✗	32.9	5.28
	wikily	✓	✓	✗	32.8	4.37
	wikily	✓	✗	✓	33.3	5.72
	wikily	✓	✓	✓	36.8	5.60
	supervised	✓	✗	✗	17.7	1.26
Translate test	English test performance→				68.7	20.42
	wikily	✓	✗	✗	30.6	4.20
	supervised	✓	✗	✗	15.8	0.92
	Google	✓	✗	✗	31.8	5.56
	Gold	✓	✗	✗	33.7	3.76
		✓	✓	✗	37.9	5.22

Figure 28 shows a randomly picked example with different model outputs. We see that the two outputs from our approach with multi-tasking are roughly the same but one of them as more syntactic order overlap with the reference while both orders are correct in Arabic as a free-word order language.

	English gold	A child on a red slide. A little boy sits on a slide on the playground. A little boy slides down a bright red corkscrew slide. A little boy slides down a red slide. a young boy wearing a blue outfit sliding down a red slide.
	English supervised	A boy is sitting on a red slide.
	En- supervised translate	صبي صبي يجلس على شاحنة خفيفة .
	En- unsupervised translate	الطفل يجلس على شريحة حمراء.
	En- Google translate	صبي يجلس على شريحة حمراء.
	Supervised MT	صبي صبي على شظية
	Unsupervised (mt + ar + en)	يجلس صبي صغير على شريحة برتقالية.
	Unsupervised (mt + ar)	صبي صغير يجلس على شريحة حمراء.
	Supervised	صبي في قميص أزرق يقفز في الهواء
	Arabic Gold	طفل على منزلقة حمراء صبي صغير يجلس على زلاجة في الملعب ينزل صبي صغير أسفل منزلقة حمراء

Figure 28: An example of different outputs in our captioning experiments both for English and Arabic, as well as Arabic translations of English outputs on the Arabic Flickr dataset (ElJundi. et al., 2020).

In general, we observe that although superficially the BLEU scores for Arabic is low, it is mostly due to its lexical diversity, free-word order, and morphological complexity.

4.4.5 Dependency Parsing Results

We use the Universal Dependencies v2.7 collection (Zeman et al., 2020) for Arabic, Kazakh, and Romanian. We use the Stanza (Qi et al., 2020) pretrained supervised models for getting supervised parse trees for Arabic and Romanian, and use the UDPipe (Straka et al., 2016) pretrained model for Kazakh. We translate about 2 million sentences from each language to English, and also 2 million English sentences to Arabic. We use a simple modification to Stanza to facilitate training on partially projected trees by masking dependency and label assignments for words with missing

dependencies. All of our training on projected dependencies is blindly conducted with 100k training steps with default parameters of Stanza (Qi et al., 2020). As for gold-standard parallel data, we use our supervised translation training data for Romanian-English and KazakhEnglish and use a sample of 2 million sentences from the UN Arabic-English data due to its large size that causes word alignment significant slowdown. For Kazakh *wikily* projections, due to low supervised POS accuracy, we use the projected POS tags for projected words and supervised tags for unprojected words. We observe a two percent increase in performance by using projected tags.

Table 15 shows the results for dependency parsing experiments. We see that our model performs very high in Romanian with a unlabeled accuracy score of 74 which is much higher than that of Ahmad et al. (2019) and slightly lower than that of Rasooli and Collins (2019) which uses a combination of multi-source annotation projection and direct model transfer. Our work on Arabic outperforms all previous work and performs even better than using gold-standard parallel data. One clear highlight is our result in Kazakh. As mentioned before, by projecting the part-of-speech tags, we achieve roughly 2 percent absolute improvement. Our final results on Kazakh are significantly higher than that of using gold-standard parallel text (7K sentences).

Table 15: Dependency parsing results on the Universal Dependencies dataset (Zeman et al., 2020). Previous work has used different sub-versions of the Universal Dependencies data in which slight differences are expected.

Method		Version	Token and POS	Arabic			Kazakh			Romanian		
				UAS	LAS	BLEX	UAS	LAS	BLEX	UAS	LAS	BLEX
Previous	Rasooli and Collins (2019)	2.0	gold/supervised	61.2	48.8	–	–	–	–	76.3	64.3	–
	Ahmad et al. (2019)	2.2	gold	38.1	28.0	–	–	–	–	65.1	54.1	–
	Kurniawan et al. (2021)	2.2	gold	48.3	29.9	–	–	–	–	–	–	–
Projection	Wikily translation	2.7	gold	62.5	50.7	46.3	46.8	28.5	25.0	74.1	57.7	52.6
	supervised		60.2	48.7	42.1	46.2	27.8	14.1	73.6	57.4	50.9	
	Gold-standard Parallel data		gold	61.5	47.3	42.4	22.2	9.3	7.9	75.9	62.4	57.3
			supervised	59.1	45.3	38.5	21.8	9.2	3.8	75.6	62.0	55.6
Supervised			supervised	84.2	79.8	72.7	48.0	29.8	13.7	90.8	86.0	80.0

5.0 CONCLUSIONS

In our first study, we 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 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.

In our second study, we demonstrated that it may be possible to remedy the geographic bias of image classification datasets using the Massively Multilingual Image Dataset. We validate our inductive bias that a pretrained Resnet has learned general image features and trains better than other models. We also backed up our data hypothesis, that a model trained only on English people images would be insufficient to capture cultural differences in people from places that speak other languages. The ultimate goal of this work is to reduce bias in image datasets and classifiers, and as such, future work with MMID would benefit from cleaning data with crowds of workers from locales of interest.

In our third study, we investigated when images may be useful for translating words between two languages from the perspective of their cultural and geographical relatedness. We observe that translatability of words via images vary in different language pairs, with language pairs sharing cultural similarities having better image translatability. In the future, it will be interesting to study image translatability of more language pairs and their characteristics, including those outside MMID, as well as extend our work to sentence-level image aided translation. It will also be of great value to study if adding considerations of cultural relatedness to image-aided MT can further improve its performance. Additionally, using a different metric for intra-cluster similarity that does not calculate similarity with respect to the origin may be more accurate depending on the application. As many similarity functions, aside from cosine similarity, have been used in the computer vision literature, improving this function could be fruitful future work.

In our final study, we described a fast and effective algorithm for learning translation systems using Wikipedia. We show that by wisely choosing what to use as seed data, we can have very good seed parallel data to mine more parallel text from Wikipedia. We have also shown that our translation models can be used in downstream cross-lingual natural language processing tasks. In the future, we plan to extend our approach beyond Wikipedia to other comparable datasets like the BBC World Service. A clear extension of this work is to try our approach on other cross-lingual tasks. Moreover, as many captions of the same images in Wikipedia are similar sentences and sometimes translations, multimodal machine translation (Specia et al., 2016; Caglayan et al., 2019; Hewitt et al., 2018; Yao and Wan, 2020) based on this data or the analysis of the data, such as whether more similar languages may share more similar captions (Khani et al., 2021) are other interesting avenues.

6.0 RECOMMENDATIONS

This research demonstrated the value of using unconventional data sources to perform multilingual natural language processing. By combining text and image representations, our NLP algorithms arrived at better representations of the meanings of words. Furthermore, we illustrated the benefit of having large multilingual image datasets for overcoming some of the limitations of current image datasets. In particular, we demonstrated the value of having geographically and culturally diverse data, and showed that having diverse data leads to better image classifiers and better representations.

Since the outset of this project, the field of NLP has made several breakthroughs with ever larger neural language models such as GPT-3, and text+image systems such as CLIP. These allow for zero-shot or few-shot translation. However, these findings are mainly for high-resource languages. We believe that there are still many improvements that can be made to AI systems based on deep learning that benefit from our findings. Can we design algorithms that incorporate a diverse set of images in order to arrive at better multimodal representations? Can we facilitate better machine translation and cross-lingual applications through neural language models through multimodal data?

Finally, we believe that there are many ways of extending our research into more subtle questions. For instance, what information is present in a Wikipedia article in one language that is not present in the English language version, or vice versa? What are the cultural influences of information mismatches across languages? Is there a way of automatically identifying what these information mismatches are, and then characterizing them? This points towards the possibility of very interesting research that combines NLP with sociolinguistics and other social sciences.

7.0 REFERENCES

- Burns, Andrea, Donghyun Kim, D. Wijaya, Kate Saenko, Bryan A. Plummer. 'Learning to Scale Multilingual Representations for Vision-Language Tasks'. ArXiv abs/2004.04312 (2020).
- Bhatia, Vibhu, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, 'OpenFraming: Open-sourced Tool for Computational Framing Analysis of Multilingual Data'. EMNLP, 2021.
- Funaki, Ruka, Hideki Nakayama. 'Image-Mediated Learning for Zero-Shot Cross-Lingual Document Retrieval'. EMNLP, 2015.
- Conneau, Alexis, Guillaume Lample, Marc'aurelio Ranzato, Ludovic Denoyer, Herv'e J'egou. 'Word Translation Without Parallel Data'. ArXiv abs/1710.04087 (2018).
- Song, Yuqing, Shizhe Chen, Yida Zhao, Qin Jin. 'Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards'. Proceedings of the 27th ACM International Conference on Multimedia, 2019.
- Tae, Yunwon, Cheonbok Park, Taehee Kim, Soyoung Yang, Mohammad Azam Khan, Eunjeong Park, Tao Qin, Jaegul Choo. 'Meta-Learning for Low-Resource Unsupervised Neural MachineTranslation'. ArXiv abs/2010.09046 (2020).
- Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. NeurIPS, 2019.
- He, Kaiming, X. Zhang, Shaoqing Ren, Jian Sun. 'Deep Residual Learning for Image Recognition'. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770–78.
- Koehn, Philipp, Kevin Knight. 'Learning a Translation Lexicon from Monolingual Corpora'. ACL 2002, 2002.
- Hartmann, Mareike, Anders Søgaard. 'Limitations of Cross-Lingual Learning from Image Search'. Rep4NLP@ACL, 2018.
- ElJundi, Obeida, Mohamad Dhaybi, Kotaiba Mokadam, Hazem M. Hajj, Daniel C. Asmar. 'Resources and End-to-End Neural Network Models for Arabic Image Captioning'. VISIGRAPP, 2020.
- Abbasi, Gelavizh, Arezoo Assemi, Siamak Saadat Dehghan, Elenaz Janfaza. 'Language, Translation, and Culture', 2012.
- Li, Shen, João Graça, Ben Taskar. 'Wiki-ly Supervised Part-of-Speech Tagging'. EMNLP, 2012.
- Gu, Jiuxiang, Shafiq R. Joty, Jianfei Cai, Handong Zhao, Xu Yang, G. Wang. 'Unpaired Image Captioning via Scene Graph Alignments'. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 10322–31.
- Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, 'HuggingFace's Transformers: State-of-the-art Natural Language Processing'. ArXiv abs/1910.03771 (2019).
- Neubig, Graham, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, 'DyNet: The Dynamic Neural Network Toolkit'. ArXiv abs/1701.03980 (2017).

Ruiter, Dana, Cristina España-Bonet, Josef van Genabith. 'Self-Supervised Neural Machine Translation'. ACL, 2019.

Koehn, Philipp. 'Europarl: A Parallel Corpus for Statistical Machine Translation'. MTSUMMIT, 2005.

Barrault, Loïc, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, Stella Frank. 'Findings of the Third Shared Task on Multimodal Machine Translation'. WMT, 2018.

Krizhevsky, Alex, Ilya Sutskever, Geoffrey E. Hinton. 'ImageNet classification with deep convolutional neural networks'. Communications of the ACM 60 (2012): 84–90.

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

Tufis, Dan, Radu Ion, Stefan Daniel Dumitrescu, Dan C. Stefanescu. 'Wikipedia as an SMT Training Corpus'. RANLP, 2013.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. ArXiv abs/1810.04805 (2019).

Singhal, K., Karthik Raman, Balder ten Cate. 'Learning Multilingual Word Embeddings Using Image-Text Data'. ArXiv abs/1905.12260 (2019).

Elliott, Desmond, Stella Frank, Loïc Barrault, Fethi Bougares, Lucia Specia. 'Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description'. WMT, 2017.

Cho, Kyunghyun, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio. 'Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation'. EMNLP, 2014.

Schafer, Charles, David Yarowsky. 'Inducing Translation Lexicons via Diverse Similarity Measures and Bridge Languages'. CoNLL, 2002.

Och, Franz Josef, Hermann Ney. 'A Systematic Comparison of Various Statistical Alignment Models'. Computational Linguistics 29 (2003): 19–51.

Singh, Amanpreet, Vedanuj Goswami, Devi Parikh. 'Are we pretraining it right? Digging deeper into visio-linguistic pretraining'. ArXiv abs/2004.08744 (2020).

K'ad'ar, 'akos, Grzegorz Chrupała, A. Alishahi, Desmond Elliott. 'Bootstrapping Disjoint Datasets for Multilingual Multimodal Representation Learning'. ArXiv abs/1911.03678 (2019).

Vulic, Ivan, Anna Korhonen. 'On the Role of Seed Lexicons in Learning Bilingual Word Embeddings'. ACL, 2016.

Ziemski, Michal, Marcin Junczys-Dowmunt, Bruno Pouliquen. 'The United Nations Parallel Corpus v1.0'. LREC, 2016.

Koehn, Philipp, Hieu T. Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, 'Moses: Open Source Toolkit for Statistical Machine Translation'. ACL, 2007.

Nida, Eugene Albert. 'Linguistics and Ethnology in Translation-Problems'. WORD 1 (1945): 194–208.

Kiela, Douwe, Ivan Vulic, Stephen Clark. 'Visual Bilingual Lexicon Induction with Transferred ConvNet Features'. EMNLP, 2015.

Kim, Yunsu, Miguel Graça, Hermann Ney. 'When and Why is Unsupervised Neural Machine Translation Useless?' EAMT, 2020.

Calixto, Iacer, Qun Liu, Nick Campbell. 'Multilingual Multi-modal Embeddings for Natural Language Processing'. ArXiv abs/1702.01101 (2017).

Guo, Mandy, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Matthew Cer, Gustavo Hernández Ábrego, Keith Stevens, 'Effective Parallel Corpus Mining using Bilingual Sentence Embeddings'. WMT, 2018.

Zoph, Barret, Deniz Yuret, Jonathan May, Kevin Knight. 'Transfer Learning for Low-Resource Neural Machine Translation'. EMNLP, 2016.

Lample, Guillaume, Myle Ott, Alexis Conneau, Ludovic Denoyer, Marc'aurelio Ranzato. 'Phrase-Based & Neural Unsupervised Machine Translation'. EMNLP, 2018.

Chen, Shizhe, Qin Jin, Alexander Hauptmann. 'Unsupervised Bilingual Lexicon Induction from Mono-lingual Multimodal Data'. ArXiv abs/1906.00378 (2019).

Haghighi, Aria, Percy Liang, Taylor Berg-Kirkpatrick, Dan Klein. 'Learning Bilingual Lexicons from Monolingual Corpora'. ACL, 2008.

Duñabeitia, Jon Andoni, Davide Crepaldi, Antje S. Meyer, Boris New, Christos Pliatsikas, Eva Smolka, Marc Brysbaert. 'MultiPic: A standardized set of 750 drawings with norms for six European languages'. Quarterly Journal of Experimental Psychology 71 (2018): 808–16.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer. 'Multilingual Denoising Pre-training for Neural Machine Translation'. Transactions of the Association for Computational Linguistics 8 (2020): 726–42.

Sen, Sukanta, Kamal Kumar Gupta, Asif Ekbal, Pushpak Bhattacharyya. 'Multilingual Unsupervised NMT using Shared Encoder and Language-Specific Decoders'. ACL, 2019.

Yarowsky, David, Grace Ngai, Richard H. Wicentowski. 'Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora'. HLT, 2001.

Kim, Donghyun, Kuniaki Saito, Kate Saenko, Stan Sclaroff, Bryan A. Plummer. 'MULE: Multimodal Universal Language Embedding'. AAAI, 2020.

Kuwanto, Garry, Afra Feyza Akyurek, Isidora Chara Tourni, Siyang Li, D. Wijaya. 'Low-Resource Machine Translation for Low-Resource Languages: Leveraging Comparable Data, Code-Switching and Compute Resources'. ArXiv abs/2103.13272 (2021).

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning. 'Stanza: A Python Natural Language Processing Toolkit for Many Human Languages'. ACL, 2020.

Wang, Josiah, Pranava Swaroop Madhyastha, Josiel Maimoni de Figueiredo, Chiraag Lala, Lucia Specia. 'MultiSubs: A Large-scale Multimodal and Multilingual Dataset'. ArXiv abs/2103.01910 (2021).

Nguyen, Toan Q., David Chiang. 'Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation'. IJCNLP, 2017.

Aggarwal, Pranav, Ritiz Tambi, Ajinkya Kale. 'Towards Zero-shot Cross-lingual Image Retrieval and Tagging'. ArXiv abs/2109.07622 (2021).

Straka, Milan, Jan Hajic, Jana Straková. 'UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing'. LREC, 2016.

Marchisio, Kelly, Kevin Duh, Philipp Koehn. 'When Does Unsupervised Machine Translation Work?' WMT, 2020.

Hitschler, Julian, Shigehiko Schamoni, Stefan Riezler. 'Multimodal Pivots for Image Caption Translation'. ArXiv abs/1601.03916 (2016).

Kocmi, Tom, Ondrej Bojar. 'Trivial Transfer Learning for Low-Resource Neural Machine Translation'. WMT, 2018.

Green, Nathan. 'Dependency Parsing', 2011.

Thompson, Bill, Seán G. Roberts, Gary Lopyan. 'Cultural influences on word meanings revealed through large-scale semantic alignment'. Nature Human Behaviour, 2020, 1–10.

Dellert, Johannes, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, 'NorthEuraLex: a wide-coverage lexical database of Northern Eurasia'. Language Resources and Evaluation 54 (2020): 273–301.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, Tomas Mikolov. 'Learning Word Vectors for 157 Languages'. ArXiv abs/1802.06893 (2018).

Artetxe, Mikel, Gorka Labaka, Eneko Agirre. 'Unsupervised Statistical Machine Translation'. ArXiv abs/1809.01272 (2018).

Caglayan, Ozan, Pranava Swaroop Madhyastha, Lucia Specia, Loïc Barrault. 'Probing the Need for Visual Context in Multimodal Machine Translation'. NAACL, 2019.

Specia, Lucia, Stella Frank, K. Sima'an, Desmond Elliott. 'A Shared Task on Multimodal Machine Translation and Crosslingual Image Description'. WMT, 2016.

Vulic, Ivan, Marie-Francine Moens. 'Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses'. NAACL, 2013.

Gella, Spandana, Rico Sennrich, Frank Keller, Mirella Lapata. 'Image Pivoting for Learning Multilingual Multimodal Representations'. EMNLP, 2017.

Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 'Attention is All you Need'. ArXiv abs/1706.03762 (2017).

Ma, Xuezhe, F. Xia. 'Unsupervised Dependency Parsing with Transferring Distribution via Parallel Guidance and Entropy Regularization'. ACL, 2014.

Joshi, Pratik M., Sebastin Santy, Amarjit Budhiraja, Kalika Bali, Monojit Choudhury. 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World'. ACL, 2020.

Hodosh, Micah, Peter Young, J. Hockenmaier. 'Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics (Extended Abstract)'. J. Artif. Intell. Res. 47 (2013): 853–99.

- . 'Density-Driven Cross-Lingual Transfer of Dependency Parsers'. EMNLP, 2015.
- Jones, Alex, D. Wijaya. 'Majority Voting with Bidirectional Pre-translation For Bitext Retrieval'. BUCC, 2021.
- Friendly, Michael, Patricia E. Franklin, David Hoffman, David C. Rubin. 'The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words'. *Behavior Research Methods & Instrumentation* 14 (1982): 375–99.
- Gao, Jiahui, Yi Zhou, P. Yu, Jiuxiang Gu. 'Unsupervised Cross-lingual Image Captioning'. ArXiv abs/2010.01288 (2020).
- Post, Matt. 'A Call for Clarity in Reporting BLEU Scores'. ArXiv abs/1804.08771 (2018).
- Bojar, Ondrej, Vojtech Diatka, P. Rychlý, Pavel Stranák, Vít Suchomel, Ales Tamchyna, Daniel Zeman. 'HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation'. LREC, 2014.
- Kastner, Marc A., Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, Hiroshi Murase. 'Estimating the imageability of words by mining visual characteristics from crawled image data'. *Multimedia Tools and Applications* 79 (2020): 18167–99.
- Wakabayashi, J. F. 'Translation Between Unrelated Languages and Cultures, as Illustrated by Japanese-English Translation'. *Meta: Translators' Journal* 36 (1991): 414–23.
- Zhou, Mingyang, Luwei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, Jingjing Liu. 'UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training'. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 4153–63.
- Vulic, Ivan, Douwe Kiela, Stephen Clark, Marie-Francine Moens. 'Multi-Modal Representations for Improved Bilingual Lexicon Learning'. ACL, 2016.
- Rasooli, Mohammad Sadegh, Chris Callison-Burch, D. Wijaya. "“Wikily” Supervised Neural Translation Tailored to Cross-Lingual Tasks'. EMNLP, 2021.
- Resnik, Philip. 'Parallel strands: a preliminary investigation into mining the Web for bilingual text'. ArXiv cmp-lg/9808003 (1998).
- Rapp, Reinhard. 'Automatic Identification of Word Translations from Unrelated English and German Corpora'. ACL, 1999.
- . 'Low-Resource Syntactic Transfer with Unsupervised Source Reordering'. NAACL, 2019.
- Lample, Guillaume, Alexis Conneau. 'Cross-lingual Language Model Pretraining'. NeurIPS, 2019.
- Sharma, Piyush, Nan Ding, Sebastian Goodman, Radu Soricut. 'Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning'. ACL, 2018.
- Gu, Jiuxiang, Shafiq R. Joty, Jianfei Cai, G. Wang. 'Unpaired Image Captioning by Language Pivoting'. ArXiv abs/1803.05526 (2018).
- Desai, Karan, Justin Johnson. 'VirTex: Learning Visual Representations from Textual Annotations'. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, 11157–68.

———. 'Improving Neural Machine Translation Models with Monolingual Data'. ArXiv abs/1511.06709 (2016).

Patry, Alexandre, Philippe Langlais. 'Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia'. BUCC@ACL, 2011.

Resnik, Philip, Noah A. Smith. 'The Web as a Parallel Corpus'. Computational Linguistics 29 (2003): 349–80.

Evans, Nicholas, Stephen C. Levinson. 'The myth of language universals: language diversity and its importance for cognitive science'. The Behavioral and brain sciences 32 5 (2009): 429–48; discussion 448–494.

Griffiths, Heather, Susan Cody-Rydzewski, Nathan Keirns, Eric Strayer. 'Introduction to Sociology 2e', 2014.

Kilgarrieff, Adam. 'Googleology is Bad Science', 2007.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara I. Cabezas, Okan Kolak. 'Bootstrapping parsers via syntactic projection across parallel texts'. Natural Language Engineering 11 (2005): 311–25.

Salle, Alexandre, Aline Villavicencio, Marco A. P. Idiart. 'Matrix Factorization using Window Sampling and Negative Sampling for Improved Word Representations'. ArXiv abs/1606.00819 (2016).

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu. 'MASS: Masked Sequence to Sequence Pre-training for Language Generation'. ICML, 2019.

Faruqui, Manaal, Chris Dyer. 'Improving Vector Space Word Representations Using Multilingual Correlation'. EACL, 2014.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin. 'CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web'. ACL, 2021.

Li, Z., Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, Hai Zhao. 'Data-dependent Gaussian Prior Objective for Language Generation'. ICLR, 2020.

Bergsma, Shane, David Yarowsky, Kenneth Ward Church. 'Using Large Monolingual and Bilingual Corpora to Improve Coordination Disambiguation'. ACL, 2011.

Pavlick, Ellie, Matt Post, Ann Irvine, Dmitry Kachaev, Chris Callison-Burch. 'The Language Demographics of Amazon Mechanical Turk'. Transactions of the Association for Computational Linguistics 2 (2014): 79–92.

Rosenberg, Andrew, Julia Hirschberg. 'V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure'. EMNLP, 2007.

Dinh, Tuan, Jy-Yong Sohn, Shashank Rajput, Timothy Ossowski, Yifei Ming, Junjie Hu, Dimitris Papailiopoulos, Kangwook Lee. 'Utilizing Language-Image Pretraining for Efficient and Robust Bilingual Word Alignment'. ArXiv abs/2205.11616 (2022).

Ferraro, Francis, N. Mostafazadeh, Ting-Hao 'kenneth'Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell. 'A Survey of Current Datasets for Vision and Language Research'. EMNLP, 2015.

Yao, Shaowei, Xiaojun Wan. 'Multimodal Transformer for Multimodal Machine Translation'. ACL, 2020.

Ahmad, Wasi Uddin, Zhisong Zhang, Xuezhe Ma, Eduard H. Hovy, Kai-Wei Chang, Nanyun Peng. 'On Difficulties of Cross-Lingual Transfer with Order Differences: A Case Study on Dependency Parsing'. NAACL, 2019.

Lample, Guillaume, Ludovic Denoyer, Marc'aurelio Ranzato. 'Unsupervised Machine Translation Using Monolingual Corpora Only'. ArXiv abs/1711.00043 (2018).

Aggarwal, Pranav, Ajinkya Kale. 'Towards Zero-shot Cross-lingual Image Retrieval'. ArXiv abs/2012.05107 (2020).

Miyazaki, Takashi, N. Shimizu. 'Cross-Lingual Image Caption Generation'. ACL, 2016.

Madaan, Aman, Shruti Rijhwani, Antonios Anastasopoulos, Yiming Yang, Graham Neubig. 'Practical Comparable Data Collection for Low-Resource Languages via Images'. ArXiv abs/2004.11954 (2020).

Kingma, Diederik P., Jimmy Ba. 'Adam: A Method for Stochastic Optimization'. CoRR abs/1412.6980 (2015).

Esplà-Gomis, M., Mikel L. Forcada, Gema Ramírez-Sánchez, Hieu T. Hoang. 'ParaCrawl: Web-scale parallel corpora for the languages of the EU'. MTSummit, 2019.

Artetxe, Mikel, Holger Schwenk. 'Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings'. ACL, 2019.

Rotman, Guy, Ivan Vulic, Roi Reichart. 'Bridging Languages through Images with Deep Partial Canonical Correlation Analysis'. ACL, 2018.

Orife, Iroko, Julia Kreutzer, Blessing K. Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, 'Masakhane - Machine Translation For Africa'. ArXiv abs/2003.11529 (2020).

Irvine, Ann, Chris Callison-Burch. 'A Comprehensive Analysis of Bilingual Lexicon Induction'. Computational Linguistics 43 (2017): 273–310.

———. 'Neural Machine Translation'. ArXiv abs/1709.07809 (2017).

Yang, Ziyang, Leticia Pinto-Alva, Franck Dernoncourt, Vicente Ordonez. 'Backpropagation-Based Decoding for Multimodal Machine Translation'. Frontiers in Artificial Intelligence 4 (2021).

Shankar, Shreya, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, D. Sculley. 'No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World'. arXiv: Machine Learning, 2017.

Sennrich, Rico, Barry Haddow, Alexandra Birch. 'Edinburgh Neural Machine Translation Systems for WMT 16'. ArXiv abs/1606.02891 (2016).

Kiela, Douwe, Felix Hill, Anna Korhonen, Stephen Clark. 'Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More'. ACL, 2014.

Feng, Yang, Lin Ma, Wei Liu, Jiebo Luo. 'Unsupervised Image Captioning'. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 4120–29.

- Al-Shedivat, Maruan, Ankur P. Parikh. 'Consistency by Agreement in Zero-Shot Neural Machine Translation'. NAACL, 2019.
- Tran, C., Y. Tang, Xian Li, Jiatao Gu. 'Cross-lingual Retrieval for Iterative Self-Supervised Training'. ArXiv abs/2006.09526 (2020).
- Dabre, Raj, Tetsuji Nakagawa, Hideto Kazawa. 'An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation'. PACLIC, 2017.
- Fung, Pascale, Yuen Yee Lo. 'An IR Approach for Translating New Words from Nonparallel, Comparable Texts'. ACL, 1998.
- Wibowo, Haryo Akbarianto, Made Nindyatama Nityasya, Afra Feyza Akyürek, Suci Fitriany, Alham Fikri Aji, Radityo Eko Prasajo, D. Wijaya. 'IndoCollex: A Testbed for Morphological Transformation of Indonesian Word Colloquialism'. FINDINGS, 2021.
- . 'Last Words: Googleology is Bad Science'. Computational Linguistics 33 (2007): 147–51.
- Bahdanau, Dzmitry, Kyunghyun Cho, Yoshua Bengio. 'Neural Machine Translation by Jointly Learning to Align and Translate'. CoRR abs/1409.0473 (2015).
- Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, K. Li, Li Fei-Fei. 'ImageNet: A large-scale hierarchical image database'. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, 248–55.
- Tourni, Isidora Chara, Lei Guo, Taufiq Husada Daryanto, Fabian Zhafransyah, Edward Edberg Halim, Mona Jalal, Boqi Chen, 'Detecting Frames in News Headlines and Lead Images in U.S. Gun Violence Coverage'. EMNLP, 2021.
- Brysbaert, Marc, Amy Beth Warriner, Victor Kuperman. 'Concreteness ratings for 40 thousand generally known English word lemmas'. Behavior Research Methods 46 (2014): 904–11.
- Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, Francisco Guzmán. 'WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia'. EACL, 2021.
- Wijaya, D., Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, Chris Callison-Burch. 'Learning Translations via Matrix Completion'. EMNLP, 2017.
- Petrov, Slav, Dipanjan Das, Ryan T. McDonald. 'A Universal Part-of-Speech Tagset'. ArXiv abs/1104.2086 (2012).
- Bergsma, Shane, Benjamin Van Durme. 'Learning Bilingual Lexicons Using the Visual Similarity of Labeled Web Images'. IJCAI, 2011.
- Barrón-Cedeño, Alberto, Cristina España-Bonet, Josu Boldoba, Lluís Màrquez i Villodre. 'A Factory of Comparable Corpora from Wikipedia'. BUCC@ACL/IJCNLP, 2015.
- Chen, Xinlei, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, C. Lawrence Zitnick. 'Microsoft COCO Captions: Data Collection and Evaluation Server'. ArXiv abs/1504.00325 (2015).
- Korenius, Tuomo, Jorma Laurikkala, Martti Juhola. 'On principal component analysis, cosine and Euclidean measures in information retrieval'. Inf. Sci. 177 (2007): 4893–4905.

Hangya, Viktor, Alexander M. Fraser. 'Unsupervised Parallel Sentence Extraction with Parallel Segment Detection Helps Machine Translation'. ACL, 2019.

Dyer, Chris, Victor Chahuneau, Noah A. Smith. 'A Simple, Fast, and Effective Reparameterization of IBM Model 2'. NAACL, 2013.

Kunchukuttan, Anoop, Pratik Mehta, Pushpak Bhattacharyya. 'The IIT Bombay English-Hindi Parallel Corpus'. ArXiv abs/1710.02855 (2018).

Kudo, Taku, John Richardson. 'SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing'. EMNLP, 2018.

Khani, Nikzad, Isidora Chara Tourni, Mohammad Sadegh Rasooli, Chris Callison-Burch, D. Wijaya. 'Cultural and Geographical Influences on Image Translatability of Words across Languages'. NAACL, 2021.

Kashefi, Omid. 'MIZAN: A Large Persian-English Parallel Corpus'. ArXiv abs/1801.02107 (2018).

Das, Ayan, Sudeshna Sarkar. 'A Survey of the Model Transfer Approaches to Cross-Lingual Dependency Parsing'. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19 (2020): 1–60.

Keung, Phillip, Julian Salazar, Y. Lu, Noah A. Smith. 'Unsupervised Bitext Mining and Translation via Self-Trained Contextual Embeddings'. Transactions of the Association for Computational Linguistics 8 (2021): 828–41.

Rasooli, Mohammad Sadegh, Michael Collins. 'Cross-Lingual Syntactic Transfer with Limited Resources'. Transactions of the Association for Computational Linguistics 5 (2017): 279–93.

Bojar, Ondrej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, 'Findings of the 2016 Conference on Machine Translation'. WMT, 2016.

Artetxe, Mikel, Gorka Labaka, Eneko Agirre, Kyunghyun Cho. 'Unsupervised Neural Machine Translation'. ArXiv abs/1710.11041 (2018).

Hewitt, John, Daphne Ippolito, Brendan Callahan, Reno Kriz, D. Wijaya, Chris Callison-Burch. 'Learning Translations via Images with a Massively Multilingual Image Dataset'. ACL, 2018.

Wehrmann, Jonatas, Mauricio A. Lopes, Douglas M. Souza, Rodrigo C. Barros. 'Language-Agnostic Visual-Semantic Embeddings'. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 5803–12.

Paivio, Allan, John C. Yuille, Stephen A. Madigan. 'Concreteness, imagery, and meaningfulness values for 925 nouns'. Journal of experimental psychology 76 1 (1968): Suppl:1-25.

Edunov, Sergey, Myle Ott, Michael Auli, David Grangier. 'Understanding Back-Translation at Scale'. EMNLP, 2018.

Papineni, Kishore, Salim Roukos, Todd Ward, Wei-Jing Zhu. 'Bleu: a Method for Automatic Evaluation of Machine Translation'. ACL, 2002.

Srinivasan, Krishna, Karthik Raman, Jiecao Chen, Michael Bendersky, Marc Najork. 'WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning'. Proceedings of

the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.

Lin, Wen-Pin, Matthew G. Snover, Heng Ji. 'Unsupervised Language-Independent Name Translation Mining from Wikipedia Infoboxes'. ULNLP@EMNLP, 2011.

Klementiev, A., Dan Roth. 'Weakly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora'. ACL, 2006.

Kurniawan, Kemal, Lea Frermann, Philip Schulz, Trevor Cohn. 'PPT: Parsimonious Parser Transfer for Unsupervised Cross-Lingual Adaptation'. EACL, 2021.

APPENDIX A

7.1 Publications from this project:

- [LAL: Linguistically Aware Learning for Scene Text Recognition](#). Yi Zheng, Wenda Qin, Derry Wijaya, Margrit Betke. Proc. of ACM Multimedia 2020
- [Learning to Scale Multilingual Representations for Vision-Language Tasks](#). Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, Bryan A. Plummer. Proc. of European Conference on Computer Vision, 2020
- [Multi-label and Multilingual News Framing Analysis](#). Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Margrit Betke, Prakash Ishwar, Derry T. Wijaya. Proc. of ACL 2020
- [Cultural and Geographical Influences on Image Translatability of Words across Languages](#). Nikzad Khani, Isidora Chara Tourni, Mohammad Sadegh Rasooli, Chris Callison-Burch and Derry Tanti Wijaya. NAACL 2021.
- [IndoCollex: A Testbed for Morphological Transformation of Indonesian Colloquial Words](#). Haryo Akbarianto Wibowo, Made Nindyatama Nityasya, Afra Feyza Akyürek, Suci Fitriany, Alham Fikri Aji, Radityo Eko Prasajo, and Derry Tanti Wijaya. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3170–3183, Online. Association for Computational Linguistics. 2021.
- [“Wikily” Supervised Neural Translation Tailored to Cross-Lingual Tasks](#). Mohammad Sadegh Rasooli, Chris Callison-Burch, Derry Tanti Wijaya. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

7.2 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.

The code for our analysis of the relative effects of geography versus cultural on translatability is available on github at <https://github.com/nikzadkhani/MMID-CNN-Analysis>

The code for our “wikily” supervised cross-lingual experiments is available on github at: <https://github.com/rasoolims/ImageTranslate>

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

- ACL - Association for Computational Linguistics
- ACM - Association for Computing Machinery
- AI - Artificial intelligence
- AIDA - Active Interpretation of Disparate Alternatives
- AVGMAX - Average over maximum scores for paired word vectors or image vector
- BBC - datasets derived from the publications of the British Broadcasting Company
- BERT - Bidirectional Encoder Representations from Transformers
- BLEU - BiLingual Evaluation Understudy (evaluation metric for machine translation)
- BPR - Bayesian Personalized Ranking
- CNN - Convolutional Neural Networks
- CROWDTRANS - Name of translation dictionaries created via crowdsourcing
- CVPR - Computer Vision and Pattern Recognition conference
- DARPA - Defense Advanced Research Projects Agency
- E - Symbol to denote an English sentence or word
- F - Symbol to denote a foreign language sentence or word
- GPU - Graphics Processing Unit
- HIGHRES - High Resource language or language pair
- LOWRES - Low Resource language or language pair
- MASS - MAsked Sequence to Sequence pre-training method
- MIPR - Military Interdepartmental Purchase Request
- MMID - Massively Multilingual Image Dataset
- MT - Machine Translation
- NAACL - North American chapter of the Association for Computational Linguistics
- NEL - NorthEuraLex: a wide-coverage lexical database of Northern Eurasia
- NeurIPS - Conference and Workshop on Neural Information Processing Systems
- NLP - Natural Language Processing
- NORM - Normalized score
- P - Probability
- PCA - Principal Component Analysis
- POS - Part of Speech
- RGB - Red Green Blue color channels in an image
- SIFT - Scale Invariant Feature Transformations
- TXT - denotes a text-based model
- UN - datasets derived from translated publications of the United Nations
- US - United States
- WMT - Workshop on Machine Translation