Globetrotter: Connecting Languages by Connecting Images

Dídac Surís
Columbia University
didac.suris@columbia.edu

Dave Epstein
UC Berkeley
dave@eecs.berkeley.edu

Carl Vondrick
Columbia University
vondrick@cs.columbia.edu

Abstract

Machine translation between many languages at once is highly challenging, since training with ground truth requires supervision between all language pairs, which is difficult to obtain. Our key insight is that, while languages may vary drastically, the underlying visual appearance of the world remains consistent. We introduce a method that uses visual observations to bridge the gap between languages, rather than relying on parallel corpora or topological properties of the representations. We train a model that aligns segments of text from different languages if and only if the images associated with them are similar and each image in turn is well-aligned with its textual description. We train our model from scratch on a new dataset of text in over fifty languages with accompanying images. Experiments show that our method outperforms previous work on unsupervised word and sentence translation using retrieval. Code, models and data are available on globetrotter.cs.columbia.edu

1. Introduction

Researchers have been building machine translation models for over 60 years [20], converting input sentences in one language to equivalent ones in another. In recent years, sequence-to-sequence deep learning models have overtaken statistical methods as the state-of-the-art in this field, with widespread practical applications. However, these models require large supervised corpora of parallel text for all language pairs, which are expensive to collect and often impractical for uncommon pairs.

Rather than attempting to manually gather this ground truth, we use a source of supervision natural to the world: its consistent visual appearance. While language can take on many shapes and forms, visual observations are universal, as depicted in Fig. 1. This property can be freely leveraged to learn correspondences between the different languages of the world without any cross-lingual supervision.

Since we can learn how similar two images are to each other [12], and how compatible an image is with a textual description [36], we can introduce a transitive relation to estimate how similar two sentences are to each other: if (and only if) each sentence matches its image, and the two images match, then the two sentences should also match. We propose a multimodal contrastive approach to solve this problem, using vision to bridge between otherwise unrelated languages.

Figure 1. While each language represents a bicycle with a different word, the underlying visual representation remains consistent. A bicycle has similar appearance in the UK, France, Japan and India. We leverage this natural property to learn aligned multilingual representations for machine translation without paired training corpora.

In our experiments and visualizations, we show that the transitive relations through vision provide excellent self-supervision for learning machine translation. Although we train our approach without paired language data, our approach is able to translate between 52 different languages better than several baselines. While vision is necessary for our approach during learning, there is no dependence on vision during inference. After learning language representations, our approach can translate both individual words and full sentences using retrieval.

Our contribution is threefold. First, we propose a method that leverages cross-modal alignment between language and vision to train a multilingual translation system without any parallel corpora. Second, we show that our method outperforms previous work by a significant margin on both sentence and word translation, where we use retrieval to test translation. Finally, to evaluate and analyze our approach, we release a federated multimodal dataset spanning 52 different languages. Overall, our work shows that grounding language in vision yields models that are significantly more robust across languages, even in cases where ground truth parallel corpora are not available. Code, data, and pretrained models will be released.

2. Related work

Our unsupervised joint visual and multilingual model builds on recent progress in both the natural language processing and computer vision communities. We briefly summarize the prior work.

Unsupervised language translation has been studied as a word representation alignment problem in [34], where the distribution of word embeddings for two unpaired languages is aligned to minimize a statistical distance between them. [1, 32, 33, 35] build on top of this idea, and train an encoder-decoder structure to enforce cycle-consistency between language translations. This method achieves strong unsupervised word translation results, but does not scale beyond two languages. It also does not leverage visual information in learning, limiting performance.

Multi-language models are general language models that develop language-independent architectures that work equally well for any language [26]. [2, 15, 18, 32, 39, 42] share the same token embeddings across different languages, showing that this improves language modeling both for general downstream single-language NLP tasks and also for supervised language translation across multiple languages. [2, 15, 32] use a shared Byte Pair Encoding (BPE), which we use in our work. We loosely follow the architecture of [15] in that we train a transformer-based [50] masked language model with BPE.

Vision as multimodal bridge implies using vision as an interlingua between all languages. Using a third language as a pivot to translate between pairs of languages without source-target paired corpora has been studied extensively [*e.g.* 23, 24, 29]. [3, 27] use vision for the same purpose, operating directly on speech waveforms instead of text. [13] use images to help translate between languages in the text modality. Their model involves both generation and reinforcement learning, which makes optimization difficult, and they do not generalize to more than two languages. Sigurdsson *et al.* [46] also use vision as a pivot for unsupervised word translation. However, unlike their approach, our model is not limited by a reliance on extensive visual supervision for pre-training or inexpressive topological methods to relate concepts across languages. Further, our approach scales very naturally to multiple languages at once (instead of just two), models misalignment between vision and language, and crucially learns to translate at the sentence level rather than just words. Our experiments quantitatively compare the two approaches, showing that our approach performs better both in word and sentence translation.

Other work views the input image as extra information for translation [*e.g.* 10, 48], and we refer readers to [47] for an extensive overview on this topic. Instead of using images as a bridge, paired data between languages is used. There has also been research on training multilingual language representations for downstream vision tasks, leveraging visual-linguistic correspondence, but without translation as a goal. Unlike this paper, they make use of ground truth language pairs [9, 25, 30, 52].

Translation by retrieval. We evaluate the representations using retrieval-based machine translation [5, 38], often used in the context of example-based machine translation [*e.g.* 6, 7, 8, 16, 21], analogy-based translation [*e.g.* 31, 41], or translation memories [*e.g.* 4, 11, 19, 51].

State-of-the-art cross-lingual retrieval approaches rely on supervised language pairs, and range from training the models in a standard contrastive learning setting [14] to more complex combinations of the language pairs such as cross-attention [40] or using custom fusion layers [22]. Our approach does not require supervised language pairs.

3. Approach

We present an approach that learns to map words and sentences from one language to semantically similar words and sentences from different languages, for a large number of languages simultaneously. Our approach does not require any paired data between languages, and instead only depends on image-language pairs. Fig. 2 provides an overview of our framework.

3.1. Sentence embedding

Our approach learns an aligned embedding space for sentences across languages. Let $z_i^l \in \mathbb{R}^D$ be the learned embedding of sentence i (l stands for *language*), obtained by processing the text through a language network Θ_l. Moreover, let β_{ij} be the similarity between sentences z_i^l and z_j^l, for example through the cosine similarity. Our goal is to learn the parameters of the embedding z such that sentences with the same meaning are mapped to similar positions in the embedding space despite being in different languages. After learning, we will have a sentence embedding z_i^l that we can use for a variety of tasks, such as retrieving or generating sentences in different languages.

We learn the parameters of the embedding space by optimizing the contrastive learning problem:

$$\mathcal{L}_t = -\sum_i \sum_{j \neq i} \alpha_{ij} \log \frac{\exp(\beta_{ij}/\tau)}{\sum_{k \neq i} \exp\left(\beta_{ik}/\tau\right)} \quad \text{with} \quad \beta_{ij} = \text{sim}\left(z_i^l, z_j^l\right) \tag{1}$$

In this framework, we need to define which pairs of examples should be close in the learned embedding space (the

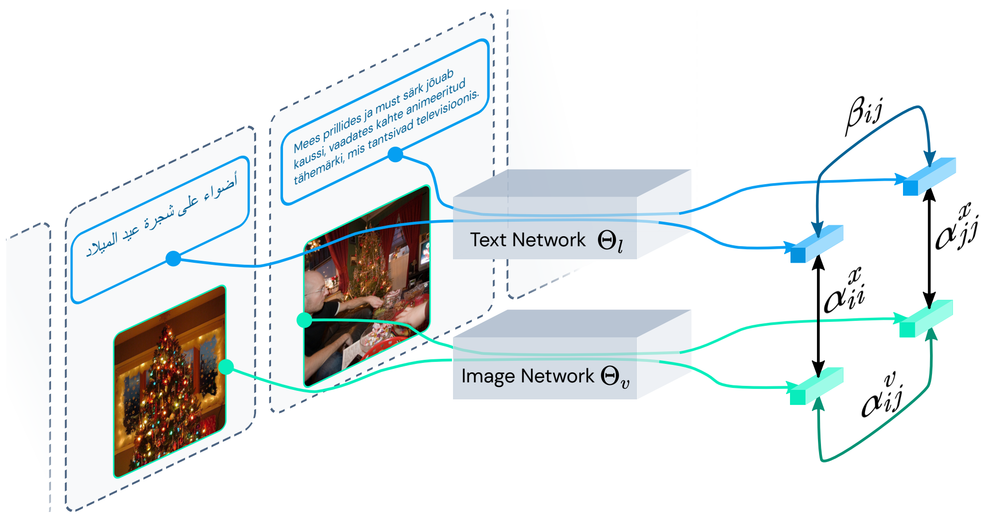

Figure 2. Our model learns an aligned embedding space for language translation by leveraging a transitive relation through vision. Cross-sentence similarity β_{ij} is estimated by the path through an image collection. See Section 3 for details.

positives), and which should not (the negatives). In the above formulation, the scalar $\alpha_{ij} \in [0, 1]$ indicates this assignment. However, since we are in an unsupervised translation setting, we do not have ground truth pairs. Our main idea, which we introduce in the next section, is that we can use the visual modality to discover these pairs.

3.2. Transitive relations

Estimating the similarity for sentences of different languages is challenging without labels. Unsupervised machine translation approaches typically rely on topological properties, such as distributional alignment or back-translation [32, 34]. However, these constraints provide a noisy gradient for learning, which makes large-scale optimization difficult.

We propose to take advantage of a transitive relation through the visual modality in order to estimate the similarity in language space α_{ij}. Given a dataset of images and their corresponding captions, we estimate both a cross-modal (sentence-image) similarity as well as a cross-image (image-image) similarity. Let α^x_{ii} be the cross-modal similarity, which indicates the alignment between image i and its corresponding caption i. We also let α^v_{ij} be the cross-image similarity, indicating the perceptual similarity between image i and another image j. This provides the transitive relation as the product of similarities

$$\alpha_{ij} = f\left(\left[\alpha^x_{ii} \cdot \alpha^v_{ij} \cdot \alpha^x_{jj}\right]^{1/3} \right), \quad \text{where} \quad f(x) = \max(0, x-m)/(1-m), \tag{2}$$

and m is a margin that we set to $m = 0.4$, which prevents pairs with low similarity from being used as positives. Note that $\alpha_{ij} = \alpha_{ji}$. The transitive similarity causes two sentences from different languages to be similar if they appear in similar visual contexts.

The final similarity is in the range $\alpha_{ij} \in [0, 1]$. Only when there is a strong alignment between an image and its caption, and there is also another image with close perceptual similarity, will a transitive relation be formed. In realistic scenes, the correspondence for some image and caption pairs may be difficult to establish in the presence of noise, which our formulation handles by breaking the transitive relation. In other words, we only consider paths with high total similarity as positives for the contrastive objective, and discard those paths with low total similarity, since their sentences likely do not match.

3.3. Learning

In order to optimize Equation 1, we need to estimate α^x_{ii} and α^v_{ij}. We parametrize both with neural networks and train them to directly estimate the similarity, also using contrastive learning [12].

Visual similarity: We jointly learn a visual feature space to estimate α^v_{ij}. For every image, we perform two random augmentations, resulting in two different versions of the same image. These two transformed images are run through the image network, along with the other $N-1$ pairs (in a batch of N samples). This results in $2N$ feature maps. For every pair (i_1, i_2) of images with representations $z^v_{i_1}$ and $z^v_{i_2}$, we compute a contrastive loss, where all the other $2(N-1)$ images are the negatives. We use the loss function:

$$\mathcal{L}_v = -\sum_{i_1, i_2} \log \frac{\exp\left(\alpha^v_{i_1 i_2}/\tau\right)}{\sum_{j \neq i_1} \exp\left(\alpha^v_{i_1 j}/\tau\right)} \quad \text{where} \quad \alpha^v_{ij} = \mathrm{sim}(z^v_i, z^v_j). \tag{3}$$

z^v_i represents the learned features for image i, obtained by processing the images through an image network Θ_v. We augment images using random image cropping, random Gaussian blurring, and random color distortions, as in [12].

Cross-modal similarity: We also need to estimate the similarity between images and their corresponding captions α^x_{ii}. The visual representation anchors inter-language alignment, and this similarity constrains the sentence embedding

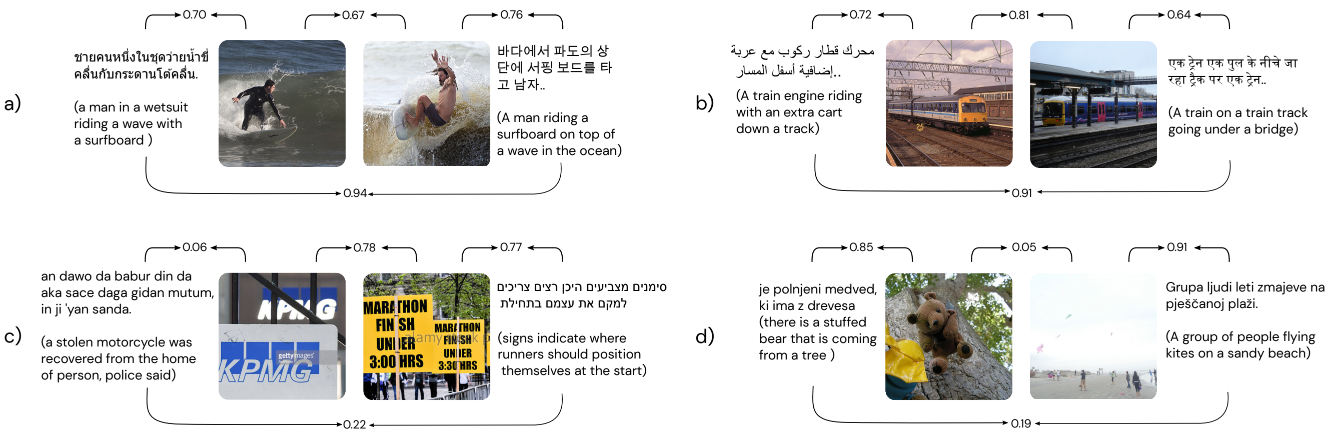

Figure 3. We show two examples of positive matches (top) and two examples of negative matches (bottom). Our model trains its text-to-text estimate (bottom) using the three scores on the top. At test time, it directly estimates similarity between text in different languages, without requiring visual input.

for each language to share the same space as the image embedding. We learn this similarity metric through the contrastive objective:

$$\mathcal{L}_x = -\sum_i \left(\log \frac{\exp\left(\alpha^x_{ii}/\tau\right)}{\sum_j \exp\left(\alpha^x_{ij}/\tau\right)} + \log \frac{\exp\left(\alpha^x_{ii}/\tau\right)}{\sum_j \exp\left(\alpha^x_{ji}/\tau\right)} \right)$$
$$\text{with} \quad \alpha^x_{ij} = \text{sim}(z^v_i, z^l_j). \tag{4}$$

Token cloze: We finally also train the model with a token cloze task in order to make the language representation contextual. We follow the same loss and objective as BERT [18] over the sentence input. We label this loss $\mathcal{L}_c$.

Full objective: The final objective we optimize is the combination of all four losses defined above:

$$\min_{\Theta} \ \mathcal{L}_t + \lambda_1 \mathcal{L}_v + \lambda_2 \mathcal{L}_x + \lambda_3 \mathcal{L}_c \tag{5}$$

where Θ are the neural network parameters, and λ are scalar hyper-parameters to the balance the terms. Over the course of optimization, the model learns a cross-lingual similarity metric β jointly with the transitive similarities α. As learning progresses, α_{ij} forms soft positive and negative pairs, which the model uses to learn aligned multi-language representations. The quality of the multi-language representation depends on the quality of transitive alignments α_{ij} our model discovers. However, since the contrastive objective relies on statistical patterns over a large dataset, our approach is fairly robust to noise, as supported by our experiments.

3.4. Refining word-level alignment

Our approach learns a common embedding space between vision and sentences in multiple languages, which our experiments will show provides a robust representation for unsupervised machine translation. This representation is trained to be well-aligned at the sentence level. We can further refine the representation by aligning them along words as well.

To obtain word-level alignment, we use the Procrustes algorithm [43] on the learned word embeddings: we find a linear transformation from the word embeddings of one language to the word embeddings of another language. To estimate the linear transformation, we follow standard practice and identify the anchor points by finding the $k = 5$ mutual nearest neighbors between the word embeddings across languages. We then proceed with the Procrustes approach from [49], which extends the original algorithm to more than two distributions. To translate words, we then directly retrieve using the transformed word embeddings.

3.5. Architecture

Our method uses a two-branch architecture, which extracts text and image features that share the same semantic embedding space. We briefly describe the network architecture choices below. We refer readers to the supplemental material for complete details.

Image network Θ_v: To extract visual features, we apply a convolutional network over the images. We use a ResNet-18, initialized with ImageNet features [17, 28], and we add a prediction head after the last hidden layer of the ResNet.

Text network Θ_l: We use a neural network to embed a sentence. We use a single encoder with shared word embeddings across all languages, which has been shown to scale well to the multilingual setting [2, 15]. All languages share the same vocabulary created using Byte Pair Encoding [44], which improves the alignment of embedding spaces across languages that share the same alphabet [33]. We then use a transformer from [50], shared by all the languages.

To produce outputs, we add a prediction head, and normalize the outputs so that $||z||_2 = 1$.

Figure 4. We show some examples of our dataset. See the supplementary material for the English translations. See Section 4 for details.

4. The Globetrotter dataset

In order to train and evaluate our approach, we collect a federated dataset of images and captions that span 52 different languages. The full list of languages is in Supplementary Material. We combine three captioning datasets and translate them using Amazon Translate from Amazon Web Services. We use captions and images from the Flickr30k [53], MSCOCO [37], and Conceptual Captions [45] datasets. The language in the federated dataset is diverse, covering both captions from human annotators and captions harvested from the web. We show some examples in Fig. 4. The dataset contains a total of 4.1M image-caption pairs, with an English sentence mean length of 10.4 words. We will publicly release this dataset.

We split our dataset into train, validation, and testing sets. We make the partition ensuring that they each contain a disjoint set of images and sentences. We use 3.15M unique text-image pairs for training, 787k for validation, and 78.7k for testing. The training and validation splits contain samples corresponding to all languages, and each image only has one language associated with it. The testing set is translated to all languages (the same samples), to obtain ground truth alignment for evaluation. We further collect a test set of 200 English captions translated by fluent speakers to 11 different languages (see Supplementary Material), for a total of 2200 human-generated translations.

5. Experimental evaluation

Our experiments analyze the language translation capabilities of our model, and quantify the impact of vision on the learning process. We call our model **Globetrotter**.

5.1. Baselines

Sigurdsson *et al.* [46]: The closest approach to ours is [46], which is a state-of-the-art approach for unsupervised word translation using cross-modal information. Their original model is trained to translate between just two languages, and our experiments work with more than fifty languages. We therefore extended their method to multiple languages by creating a different word embedding and adapting layer for each language, which we use as the baseline. We use the same vocabulary as in our method, but train separate word embeddings for different languages.

Conneau & Lample [32]: We also compare to the state-of-the-art unsupervised translation approach that does not use visual information. We experimented with several baselines, and chose the one that performs the best. This baseline uses a cycle-consistency (or back-translation) loss between pairs of languages. We train their method on our dataset, for all M languages simultaneously. We originally experimented with adding cycle-consistency constraints for all M^2 language pairs, but this resulted in poor performance. We randomly select a total of $5M$ pairs, where each language appears five times as the source and five times as the target. We also experimented with [34], but this performed worse than [32].

Text-only model: To quantify the impact of vision, we also train a version of our model where all images and image-related losses are removed, as in [18]. This model is capable of learning some basic cross-lingual concepts by having different languages using the same tokens.

Fully supervised: To understand the gap between unsupervised and supervised approaches, we train our method with paired language corpora. We use our same framework, except we set the values of α to 1 for paired sentences, and 0 for unpaired sentences.

Common evaluation setup: Throughout our experiments, we adopt a common evaluation setup to evaluate all models. We train all models for 200 epochs and select the best model on the held-out validation set. In all cases, vision is not used during testing.

Source: Spanish	**Target: Russian**	**Target: Hebrew**
Una vista aérea durante su remodelación *An aerial view during its redevelopment*	Вид на город с бара на крыше *View of the city from rooftop bar*	נוף ממרפסת ג *View from a roof terrace*
Actor asiste al estreno de los angeles celebrado *Actor attends the los angeles premiere held*	Актер посещает премьеру сезона *Actor attends the season premiere*	אדם מגיע לבכורה *Person arrives at the premiere*
Ilustración de la niña de dibujos animados en color negro sobre el fondo blanco *Illustration of cartoon girl in black color on the white background*	Нарисованный эскиз с мягким классическим диваном и подушками на белом фоне *Hand drawn sketch with soft classic couch and pillows on the white background*	קריקטורה של קבוצה של נערות *Cartoon of a group of teenage girls*

Table 1. We show some examples of sentence-level translations obtained by our approach. English is only shown for visualization purposes.

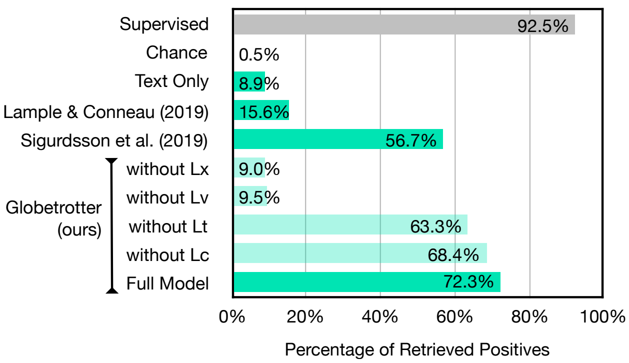

Figure 5. We evaluate our translations at the sentence-level. Our approach outperforms several unsupervised translation baselines. While unsupervised approaches are still no match for fully supervised methods, our approach uses significantly less supervision.

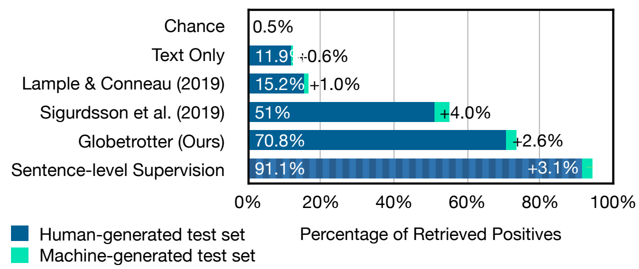

Figure 6. We evaluate our translations at the sentence-level with a human-generated test set. Fluent speakers for 11 of the languages manually annotated translations in the test set. Our approach outperforms several unsupervised translation baselines on this test set as well.

5.2. Sentence-level translation

We evaluate sentence translation using held-out data that contains a set of sentences translated to all languages. We produce translations by retrieving the nearest examples given a query. From the test set, we randomly select 200 captions, for all M languages, with a total of $200M$ sentences. Each one of these sentences is used as a query during test, and it has $M-1$ positives (same sentence in different languages). The metric we report is the percentage of positives the model ranks in the top $M-1$, among all the $200M-1$ possible options. In order to rank target sentences, we compute the similarity between them and the query sentence, and rank them according to this value. We show results in Fig. 5. Our method outperforms all baselines by a significant margin, underscoring the utility of transitive relations across modalities.

Fig. 5 also reports ablations of our framework when not training with each one of the four losses in Eq. 5. Training without losses $\mathcal{L}_v$ (Eq. 3) or $\mathcal{L}_x$ (Eq. 4) implies breaking the transitive closure represented in Fig. 2, which results in a drastic decrease in performance. $\mathcal{L}_t$ (Eq. 1) is the loss that makes the cross-lingual alignment explicit, but importantly it is not required to close the transitive relation through the visual modality. Training without it represents a considerable drop in accuracy, but the results are still better than baselines. Finally, $\mathcal{L}_c$ also contributes to the final performance, consistently with prior work [32, 39].

We show some examples of our sentence translations in Table 1. Our approach works on all language pairs and we simply select a few for visualization purposes. These examples show how our method aligns languages following their visual semantics.

Despite training on machine-generated translations, our method generalizes with minimal degradation to natural human language. To demonstrate this, we evaluate all methods on the human-translated subset of the Globetrotter dataset. We report results in Fig. 6, where we show the accuracy values both for human-translated and machine-translated texts. We use the same metric as before, now for $M = 11$. While all methods experience a minimal decrease in performance, our approach also outperforms the unsupervised baselines on the human-generated test.

5.3. Word-level translation

Following the evaluation in [46], we also evaluate word-level translation. Since dictionaries are not readily available for most language pairs, we obtain ground truth for evaluation by automatically matching words across languages. For every language pair, we find which words co-occur frequently in a sentence between the two languages. See supplementary materials. Then we test each pair of languages

Source: Spanish (English trans.)	Target: Russian (English trans.)	Target: Hebrew (English trans.)
chica (girl)	девушка (girl)	אישה (wife)
tenis (tennis)	тенни (prefix for tennis)	טניס (tennis)
personas (people)	людей (people)	אנשים (people)
aire (air)	воздух (air)	רקע (background)
campo (field)	поле (field)	בשדה (in the field)
béisbol (baseball)	бейсбол (baseball)	בייסבול (baseball)
espect (prefix for show)	шоу (show)	אירוע (event)
motocic (prefix for motorcycle)	мотоцик (мотоцик is motorcycle)	אופן (**אופנוע** is motorcycle)
camion (truck)	автобус (bus)	ברחוב (in the street)
sombrero (hat)	костюм (suit)	ולצה (חולצה is shirt)
hombre (man)	жчина (мужчина is man)	אדם (man)
mientras (while)	когда (when)	לאחר (after the)
par (two, or prefix for couple)	пара (couple)	השני (the second)
calle (street)	улица (the outside)	ברחוב (in the street)
camino (path)	пляже (beach)	דרך (path)

Table 2. We show examples of Spanish-Russian and Spanish-Hebrew word-level translations.

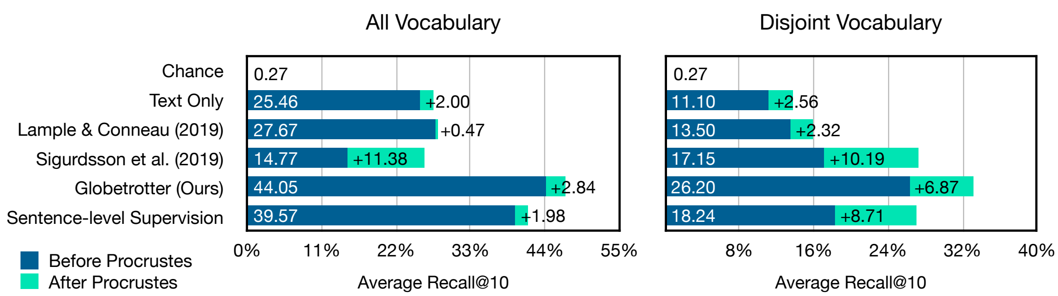

Figure 7. We also evaluate word-level translation. Although our approach is trained on sentence-level similarity, the word embeddings also learn to provide strong word-level translation. The results can be further refined with Procrustes.

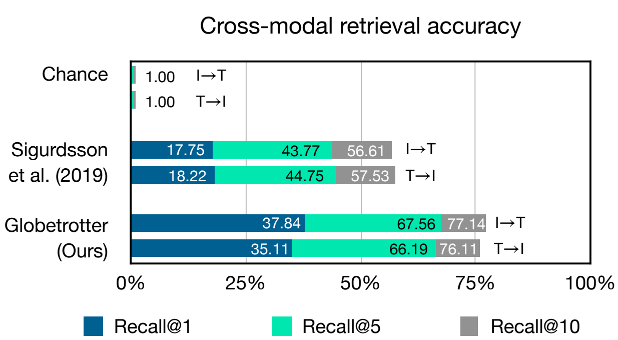

Figure 8. Cross-modal retrieval results. We show Recall at [1,5,10] for text-to-image (T→I) and image-to-text (I→T). We compare all the models that use images to perform the translation.

separately. For every translation, we evaluate retrieval in both directions. Fig. 7 reports the average Recall@10 for all pairs of translations and all pairs of languages. In the right column, we exclude from the list of pairs those where the token is the same in the two languages. Even the model trained with text only – which performs poorly on sentence-level translation – obtains strong results, highlighting the importance of using a shared vocabulary. We show some examples of word translation in Table 2.

5.4. Cross-modal retrieval

Alignment between image and text representations is crucial for our model to perform properly. We analyze this cross-modal alignment by performing retrieval from one modality to the other. Fig. 8 shows recall both for our model and for Sigurdsson et al. [46]. For each language, we select $1,000$ text-image pairs and compute Recall@K results for each one of the pairs, using the other pairs as negatives. We compute these values both from image to text and from text to image, and use $K = 1, 5, 10$. We report the average for all languages. Our model performs significantly better than the baselines, showing our approach learns a strong multi-lingual and multimodal representation.

5.5. Analysis

Visualizing transitive matches: Fig. 3 shows examples of estimated transitive similarity values. We show predicted α^v (inter-image similarity), α^x (cross-modal similarity), and β (inter-sentence similarity). Fig. 3a and 3b show examples where both the similarity between images and the cross-modal similarity are high, resulting in a large α. If these pairs were to be used for training, they would be positives. The model correctly predicts a high β value between the two texts. Fig. 3c demonstrates the importance of using α^x in addition to α^v to create language pairs. In this case,

Source: Spanish	**Target: Russian**	**Target: Hebrew**
Si escuchas, el silencio de una persona te ayudará a entender de maneras que las palabras simplemente no pueden *If you listen, a person's silence will help you to understand in ways that words simply can not*	Праздник, написанный на листе бумаги, на деревянном фоне *Holiday written on piece of paper, on a wood background*	אם אחתוך אותך זה בגתת שנלל לי אח המספדיים *If I cut you off it's because you gave me the scissors*
Un vistazo a un nuevo concepto *A glimpse at new concept*	Заднее изображение модели автомобиля в пальто *Rear image of automobile model in coat*	רכישת מכונית חדשה? הנה כמה טכנולוגיות לחפש *Purchasing a new car? here are some technologies to look out for*
Un tabby gris manchado se encuentra entre plantas verdes. *A spotted gray tabby sits among green plants*	Кролик ждет на переднем плане для обычной проверки *A rabbit waits in the foreground for a routine check*	שועל אדום בשדה *Red fox in a field*

Table 3. We illustrate some failure cases. See the end of Section 5.5 for discussion.

the visual content between the two images corresponds, and the model detects that correctly with a high α^v value. However, because web data is not always clean, the caption on the left does not correspond to the visual content. This is correctly captured in the small α^x value. If we were using this pair for training, it would be considered a negative example despite significant visual similarity. Thus, the misalignment noise is not propagated to the cross-lingual loss. Finally, Fig. 3d shows an example where both sentences accurately describe their corresponding image, but the images do not match. As expected, this results in a negative pair.

Translation difficulty by language: We itemize the performance of sentence-level translation by language in Fig. 9. Languages from the same family are often easier to translate between. The most difficult language is Tamil, the only Dravidian language in our dataset.

Limitations: We show three representative failure cases in Table 3. In the first, the caption is not related to any visual concept, causing our model to translate it incorrectly. The second example shows some words incorrectly translated due to spurious correlations in the training set. In this specific case, the phrase "new concept" is strongly associated to cars, since it appears in training in the context of "concept cars", i.e. vehicles from car companies to explore new designs. Therefore, the model retrieves sentences referring to cars, even though they do not have any relation to the phrase "new concept". Finally, the third failure case shows a sentence with a new word ("tabby"), where the model is overly reliant on context to translate instead.

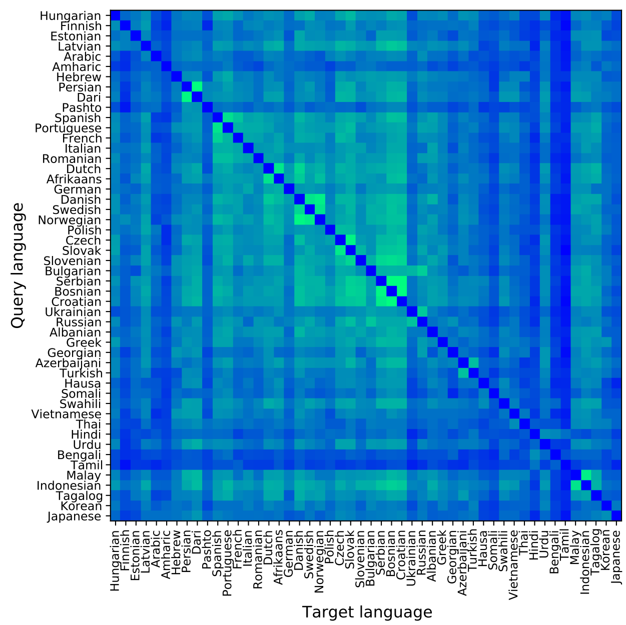

Figure 9. We show sentence-level translation accuracy by query-target language pair. In the figure, the languages are sorted by family (Romance, Baltic, etc.). The block-diagonal structure shows that languages from the same family are easier to translate between. We also find that language isolates in our dataset perform worse overall (*e.g*. Tamil, the only Dravidian language). Green implies high accuracy, blue implies low accuracy.

6. Conclusions

Leveraging a transitive relation between language and vision, our experiments show our framework learns a strong representation for both sentence-level and word-level machine translation without parallel corpora. We believe vision will continue to be valuable for learning robust language models.

Societal impact: traditional NMT approaches focus on languages with large amounts of parallel corpora, naturally biasing progress toward languages with many speakers and a robust online presence. By leveraging vision, our model provides a promising avenue for transferring NLP models to lower-resource languages. As with all deep learning systems, our model may inherit biases present in the image-text datasets used to train it.

References

[1] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018*, 2018. 2

[2] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. 2, 4

[3] Emmanuel Azuh, David Harwath, and James Glass. Towards bilingual lexicon discovery from visually grounded speech audio. *Proc. Interspeech 2019*, pages 276–280, 2019. 2

[4] Timothy Baldwin. Low-cost, high-performance translation retrieval: Dumber is better. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 18–25, Toulouse, France, July 2001. Association for Computational Linguistics. 2

[5] Timothy Baldwin and Hozumi Tanaka. The effects of word order and segmentation on translation retrieval performance. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, 2000. 2

[6] R. Brown. Transfer-rule induction for example-based translation. In *Proceedings of the MT Summit VIII Workshop on Example-Based Machine Translation*, 2001. 2

[7] Ralf D Brown. Example-based machine translation in the pangloss system. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. 2

[8] Ralf D. Brown. Automated dictionary extraction for "knowledge-free" example-based translation. In *In Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 111–118, 1997. 2

[9] Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. Learning to scale multilingual representations for vision-language tasks. *European Conference in Computer Vision*, 2020. 2

[10] Iacer Calixto and Qun Liu. Sentence-level multilingual multi-modal embedding for natural language processing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 139–148, Varna, Bulgaria, Sept. 2017. INCOMA Ltd. 2

[11] Konstantinos Chatzitheodorou. Improving translation memory fuzzy matching by paraphrasing. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 24–30, Hissar, Bulgaria, Sept. 2015. Association for Computational Linguistics. 2

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Everest Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 1, 3

[13] Yun Chen, Yang Liu, and Victor OK Li. Zero-resource neural machine translation with multi-agent communication game. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[14] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. ArXiv, July 2020. 2

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. 2, 4

[16] Lambros Cranias, Harris Papageorgiou, and Stelios Piperdis. A matching technique in example-based machine translation. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, 1994. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 4, 5

[19] Meiping Dong, Yong Cheng, Yang Liu, Jia Xu, Maosong Sun, Tatsuya Izuha, and Jie Hao. Query lattice for translation retrieval. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2031–2041, 2014. 2

[20] Leon E Dostert. The georgetown-ibm experiment. *1955). Machine translation of languages. John Wiley & Sons, New York*, pages 124–135, 1955. 1

[21] T El-Shishtawy and A El-Sammak. The best templates match technique for example based machine translation. *arXiv preprint arXiv:1406.1241*, 2014. 2

[22] Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. Filter: An enhanced fusion method for cross-lingual language understanding, 2020. 2

[23] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, 2016. 2

[24] Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur Parikh. A multilingual view of unsupervised machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3160–3170, Online, Nov. 2020. Association for Computational Linguistics. 2

[25] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 2

[26] Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Re-

ichart, and Anna Korhonen. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, 2018. 2

[27] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4969–4973. IEEE, 2018. 2

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[29] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. 2

[30] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A Plummer. Mule: Multimodal universal language embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2

[31] T. Kimura, J. Matsuoka, Y. Nishikawa, and Y. Lepage. Analogy-based machine translation using secability. In *2014 International Conference on Computational Science and Computational Intelligence*, volume 2, pages 297–298, 2014. 2

[32] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3, 5, 6

[33] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2, 4

[34] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2, 3, 5

[35] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, 2018. 2

[36] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017. 1

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[38] Chunyang Liu, Qi Liu, Yang Liu, and Maosong Sun. Thutr: A translation retrieval system. In *Proceedings of COLING 2012: Demonstration Papers*, pages 321–328, 2012. 2

[39] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. 2, 6

[40] Fuli Luo, W. Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and L. Si. Veco: Variable encoder-decoder pre-training for cross-lingual understanding and generation. *ArXiv*, abs/2010.16046, 2020. 2

[41] Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and human intelligence*, pages 351–354, 1984. 2

[42] Jason Phang, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Iacer Calixto, Katharina Kann, and Samuel R. Bowman. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 9th International Joint Conference on Natural Language Processing*, 2020. 2

[43] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 4

[44] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 4

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 5

[46] Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020. 2, 5, 6, 7

[47] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, 2016. 2

[48] Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2019. 2

[49] Hagai Taitelbaum, Gal Chechik, and Jacob Goldberger. A multi-pairwise extension of procrustes analysis for multilingual word translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3551–3556, 2019. 4

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-

reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 4

[51] Katharina Wäschle and Stefan Riezler. Integrating a large, monolingual corpus as translation memory into statistical machine translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 169–176, 2015. 2

[52] Jonatas Wehrmann, Douglas M Souza, Mauricio A Lopes, and Rodrigo C Barros. Language-agnostic visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5804–5813, 2019. 2

[53] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5