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.

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 dif-
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 a 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 \( \Theta_l \). Moreover, let \( \beta_{ij} \) 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:

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

In this framework, we need to define which pairs of examples should be close in the learned embedding space (the
We propose to take advantage of a transitive relation through the visual modality in order to estimate the similarity in language space \( \alpha_{ij} \). Given a dataset of images and their corresponding captions, we estimate both a cross-modal (sentence-image) similarity as well as a cross-image similarity in language space through the visual modality, along with the other modal augmentations, resulting in two different versions of features. This results in \( 2N \) feature maps. For every pair \((i_1, i_2)\) of images with representations \( z_{i_1}^v \) and \( z_{i_2}^v \), we compute a contrastive loss, where all the other \( 2(N-1) \) images are the negatives. We use the loss function:

\[
L_v = - \sum_{i_1,i_2} \log \frac{\exp(\alpha^v_{i_1i_2}/\tau)}{\sum_{j \neq i_1} \exp(\alpha^v_{i_1j}/\tau)} \\
\text{where } \alpha^v_{ij} = \text{sim}(z_i^v, z_j^v).
\]

\( z_i^v \) represents the learned features for image \( i \), obtained by processing the images through an image network \( \Theta_v \). 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 \( \alpha^x_{ij} \). The visual representation anchors inter-language alignment, and this similarity constrains the sentence embedding

\[
\alpha_{ij} = f\left( \left[ \alpha^x_{ii} \cdot \alpha^v_{ij} \cdot \alpha^x_{jj} \right]^{1/3} \right),
\]

where

\[
f(x) = \max(0, x - m)/(1 - m),
\]

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 \( \alpha^x_{ij} \) and \( \alpha^v_{ij} \). 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 \( \alpha^v_{ij} \). 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 \( z_{i_1}^v \) and \( z_{i_2}^v \), we compute a contrastive loss, where all the other \( 2(N-1) \) images are the negatives. We use the loss function:

\[
L_v = - \sum_{i_1,i_2} \log \frac{\exp(\alpha^v_{i_1i_2}/\tau)}{\sum_{j \neq i_1} \exp(\alpha^v_{i_1j}/\tau)}
\]

where

\[
\alpha^v_{ij} = \text{sim}(z_i^v, z_j^v).
\]

Cross-modal similarity: We also need to estimate the similarity between images and their corresponding captions \( \alpha^x_{ij} \). The visual representation anchors inter-language alignment, and this similarity constrains the sentence embedding...
for each language to share the same space as the image embedding. We learn this similarity metric through the contrasive objective:

\[
L_x = - \sum_i \left( \log \frac{\exp(\alpha_{ii}^x/\tau)}{\sum_j \exp(\alpha_{ij}^x/\tau)} + \log \frac{\exp(\alpha_{ij}^x/\tau)}{\sum_j \exp(\alpha_{ji}^x/\tau)} \right)
\]

with \( \alpha_{ij}^x = \text{sim}(z_i^x, z_j^x) \).

(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 \( L_c \).

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

\[
\min_\Theta L_t + \lambda_1 L_v + \lambda_2 L_x + \lambda_3 L_c
\]

(5)

where \( \Theta \) are the neural network parameters, and \( \lambda \) are scalar hyper-parameters to the balance the terms. Over the course of optimization, the model learns a cross-lingual similarity metric \( \beta \) jointly with the transitive similarities \( \alpha \). As learning progresses, \( \alpha_{ij} \) 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 \( \alpha_{ij} \) 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 \( \Theta_i \):** 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 \( \Theta_t \):** 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 \( ||z||_2 = 1 \).
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 $\alpha$ 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.
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 \( L_c \) (Eq. 3) or \( L_z \) (Eq. 4) implies breaking the transitive closure represented in Fig. 2, which results in a drastic decrease in performance. \( L_1 \) (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, \( L_c \) 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
Table 1

<table>
<thead>
<tr>
<th>Source: Spanish (English trans.)</th>
<th>Target: Russian (English trans.)</th>
<th>Target: Hebrew (English trans.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chica (girl)</td>
<td>девушка (girl)</td>
<td>אשה (wife)</td>
</tr>
<tr>
<td>tenis (tennis)</td>
<td>теннис (prefix for tennis)</td>
<td>(tennis)</td>
</tr>
<tr>
<td>personas (people)</td>
<td>людей (people)</td>
<td>люди (people)</td>
</tr>
<tr>
<td>aire (air)</td>
<td>воздух (air)</td>
<td>пред (background)</td>
</tr>
<tr>
<td>campo (field)</td>
<td>поле (field)</td>
<td>в (in the field)</td>
</tr>
<tr>
<td>béisbol (baseball)</td>
<td>бейсбол (baseball)</td>
<td>(baseball)</td>
</tr>
<tr>
<td>espect (prefix for show)</td>
<td>(prefix for show)</td>
<td>(event)</td>
</tr>
<tr>
<td>motocic (prefix for motorcycle)</td>
<td>мотоцикл (мотоцикл is motorcycle)</td>
<td>автобус (bus) (in the street)</td>
</tr>
<tr>
<td>camion (truck)</td>
<td>автобус (bus)</td>
<td>брат (brother)</td>
</tr>
<tr>
<td>sombrero (hat)</td>
<td>костюм (suit)</td>
<td>(in the street)</td>
</tr>
<tr>
<td>hombre (man)</td>
<td>жена (woman is man)</td>
<td>об (after the)</td>
</tr>
<tr>
<td>par (two, or prefix for couple)</td>
<td>пара (couple)</td>
<td>(the second)</td>
</tr>
<tr>
<td>calle (street)</td>
<td>улица (the outside)</td>
<td>(in the street)</td>
</tr>
<tr>
<td>camino (path)</td>
<td>пляж (beach)</td>
<td>(path)</td>
</tr>
</tbody>
</table>

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.

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 multilingual and multimodal representation.

5.5. Analysis

Visualizing transitive matches: Fig. 3 shows examples of estimated transitive similarity values. We show predicted \( \alpha^v \) (inter-image similarity), \( \alpha^x \) (cross-modal similarity), and \( \beta \) (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 \( \alpha \). If these pairs were to be used for training, they would be positives. The model correctly predicts a high \( \beta \) value between the two texts. Fig. 3c demonstrates the importance of using \( \alpha^x \) in addition to \( \alpha^v \) to create language pairs. In this case,
the visual content between the two images corresponds, and the model detects that correctly with a high $\alpha^v$ 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 $\alpha^x$ 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.

### 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


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


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

