GLOBETROTTER: UNSUPERVISED MULTILINGUAL TRANSLATION FROM VISUAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine translation in a multi-language scenario requires large amounts of parallel corpora for every language pair. While several unsupervised systems have been proposed, they are limited to word-level alignments and do not generalize to more than two languages. We introduce a framework that leverages visual similarity to align multiple languages, using images as the bridge between them. We estimate the cross-modal alignment between language and images, and use this estimate to guide the learning of cross-lingual representations. Our language representations are trained jointly in one model with a single stage. Experiments with fifty-two languages show that our method outperforms prior work on unsupervised word-level and sentence-level translation.

1 INTRODUCTION

Machine translation aims to learn a mapping between sentences of different languages while also maintaining the underlying intent. In the last few years, sequence-to-sequence models have emerged as remarkably powerful methods for this task, leading to widespread applications in robust language translation. However, sequence-to-sequence models also require large data sets of parallel corpora for learning, which is expensive to collect and often impractical for low-resource languages.

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

We propose to leverage the synchronization between language and vision in order to learn models for machine translation without parallel training corpora. Instead of learning a direct mapping between languages, we present a model that aligns them by first mapping through a visual representation. We show how vision creates a transitive closure across modalities, which we use to establish positive and negative pairs of sentences without supervision. Since the visual appearance of scenes and objects will remain relatively stable between different spoken languages, vision acts as a "bridge" between them. Our approach integrates these transitive relations into multi-modal contrastive learning.

Experiments and visualizations show that the transitive relations through vision provide excellent self-supervision for learning neural 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 the language representation, our approach can translate both individual words and full sentences either by retrieval or generation.

The contributions of this paper are three-fold. 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 multi-modal dataset spanning 52 different languages. Overall, our work shows that grounding language in vision helps developing language

processing tools that are robust across languages, even in cases where ground truth alignment across languages is not available. Code, data, and pre-trained 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 Lample et al. (2018b), where the distribution of word embeddings for two unpaired languages is aligned to minimize a statistical distance between them. Lample et al. (2018a); Artetxe et al. (2018); Lample et al. (2018c); Lample & Conneau (2019) build on top of this idea, and train an encoder-decoder structure to enforce cycle-consistency when translating from one language to another and back to the first one. This method achieves strong unsupervised word translation results, but does not scale beyond two languages. It also does not leverage visual information in learning.

Multi-language models are general language models that approach the cross-lingual challenge of developing general language-independent architectures that work equally well for any language (Gerz et al., 2018). Lample & Conneau (2019); Conneau et al. (2019); Artetxe & Schwenk (2019); Devlin et al. (2018) 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. Lample & Conneau (2019); Conneau et al. (2019); Artetxe & Schwenk (2019) use a shared Byte Pair Encoding (BPE), which we use in our work. We loosely follow the architecture of Conneau et al. (2019) in that we train a transformer-based (Vaswani et al., 2017) masked language model with BPE.

Vision as multi-modal 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 for the past few years (e.g. Firat et al., 2016; Johnson et al., 2017). Harwath et al. (2018); Azuh et al. (2019) use vision for the same purpose, and they work directly on the speech signal instead of text. Chen et al. (2018) 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. (2020) also use vision as a pivot for unsupervised translation. However, our approach works for multiple languages at once (instead of just two) and also obtains an explicit cross-lingual alignment. We share a single word embedding and language model for all languages, and use different training strategies. 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, and we refer readers to Specia et al. (2016) 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, in general leveraging visual-language correspondence, but without translation as a goal. Unlike this paper, they make use of ground truth language pairs (Wehrmann et al., 2019; Gella et al., 2017; Kim et al., 2020; Burns et al., 2020).

3 Method

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, 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

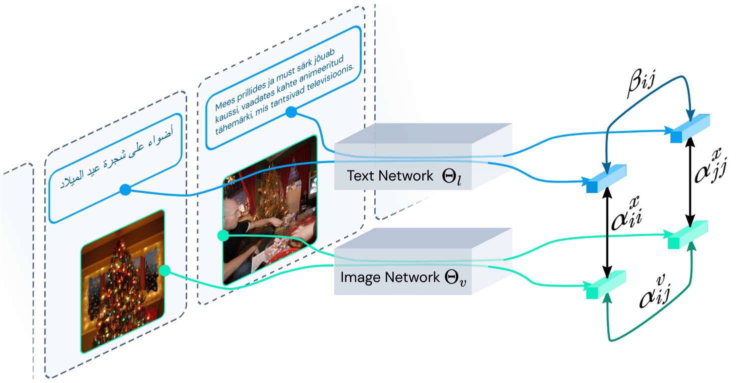

Figure 2: Our model learns an aligned embedding space for language translation by leveraging a transitive relation through vision. Cross-sentence similarity β_{ij} is estimated by the path through an image collection. See Section 3 for details.

meaning are mapped to similar positions in the embedding space despite coming from 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 z 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 contrastive learning, we need to define which pairs of examples should be close in the learned embedding space (the positives), and which pairs of examples 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 (Lample et al., 2018b; Lample & Conneau, 2019). However, these constraints provide a sparse 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 similarity as well as a cross-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:

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

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.

Since both $\alpha^x_{ii} \in [0, 1]$ and $\alpha^v_{ij} \in [0, 1]$, the final similarity is in the same 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 parameterize both with a neural network, and we train them to directly estimate the similarity also with contrastive learning.

Visual Similarity: We jointly learn a visual feature space using contrastive learning (Chen et al., 2020) in order 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, j) of images with representations z^v_i and z^v_j, we compute a contrastive loss, where all the other $2(N-1)$ images are the negatives. We use the loss function:

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

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, following Chen et al. (2020).

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

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

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 (Devlin et al., 2018) 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:

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

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 will be estimating an aligned multi-lingual representation β jointly with the transitive similarity α. As learning progresses, α_{ij} will form soft positive and negative pairs, which the model will use to learn the aligned multi-language representation. The quality of the multi-language representation will depend 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, which our experiments support.

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 aligned well 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 (Schönemann, 1966) 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 Taitelbaum et al. (2019), which extends the original algorithm to more than two distributions. To translate words, we then directly use 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: To extract visual features, we apply a convolutional network over the images, which we label Θ_v. We use a ResNet-18, initialized with ImageNet features (He et al., 2016; Deng et al., 2009), and we add a prediction head after the last hidden layer of the ResNet.

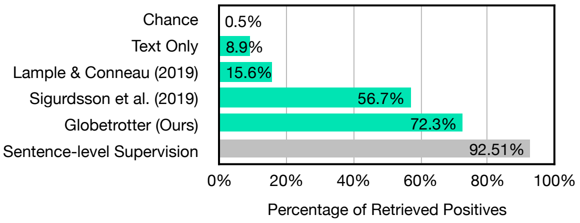

Figure 3: We evaluate our translations at the sentence-level. Our approach outperforms several unsupervised translation baselines. While unsupervised approaches are still no match for fully supervised methods, our approach uses significantly less supervision.

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 גג *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.

Text network: We use a neural network to embed a sentence, which we label Θ_l. We use a single encoder with shared word embeddings across all languages, which has been shown to scale well to the multilingual setting (Artetxe & Schwenk, 2019; Conneau et al., 2019). All languages share the same vocabulary created using Byte Pair Encoding (Sennrich et al., 2015), which improves the alignment of embedding spaces across languages that share the same alphabet (Lample et al., 2018a). We then use a transformer from Vaswani et al. (2017), 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 have collected a federated dataset of images and captions that span 52 different languages. The full list of languages is in the footnote.[1] We combined three captioning datasets and translated them using Amazon Translate from Amazon Web Services. We use captions and images from the Flickr30k (Young et al., 2014), MSCOCO (Lin et al., 2014) and Conceptual Captions (Sharma et al., 2018) datasets. The language in the federated dataset is diverse, covering both captions from human annotators and captions harvested from the web. 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 a train, validation, and testing set. 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 have ground truth alignment.

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**.

[1] Afrikaans, Albanian, Amharic, Arabic, Azerbaijani, Bengali, Bosnian, Bulgarian, Chinese, Croatian, Czech, Danish, Dari, Dutch, English, Estonian, Finnish, French, Georgian, German, Greek, Hausa, Hebrew, Hindi, Hungarian, Indoniesian, Italian, Japanese, Korean, Latvian, Malay, Norwegian, Persian, Pashto, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Somali, Spanish, Swahili, Swedish, Tagalog, Tamil, Thai, Turkish, Ukrainian, Urdu, Vietnamese.

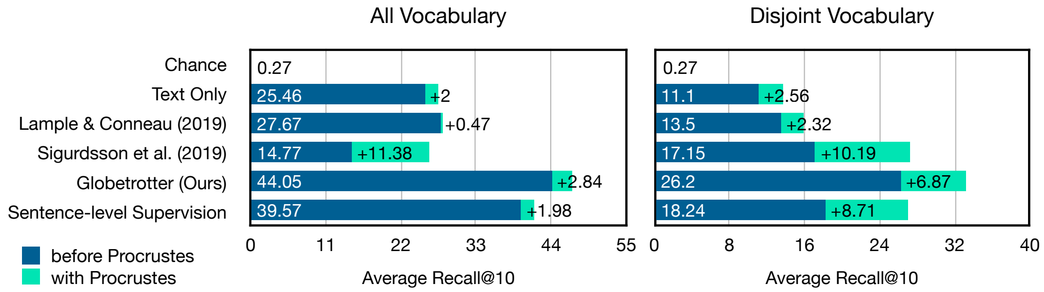

Figure 4: We also evaluate word-level translation. Although our approach is trained on sentence-level similarity, the word embeddings also learn to provide strong word-level translation. The results can be further refined with Procrustes.

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.

5.1 Baselines

Sigurdsson et al. (2020): The closest approach to ours is Sigurdsson et al. (2020), 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 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.

Lample & Conneau (2019): 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 Lample et al. (2018b), but this performed worse than Lample & Conneau (2019).

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 Devlin et al. (2018). 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.

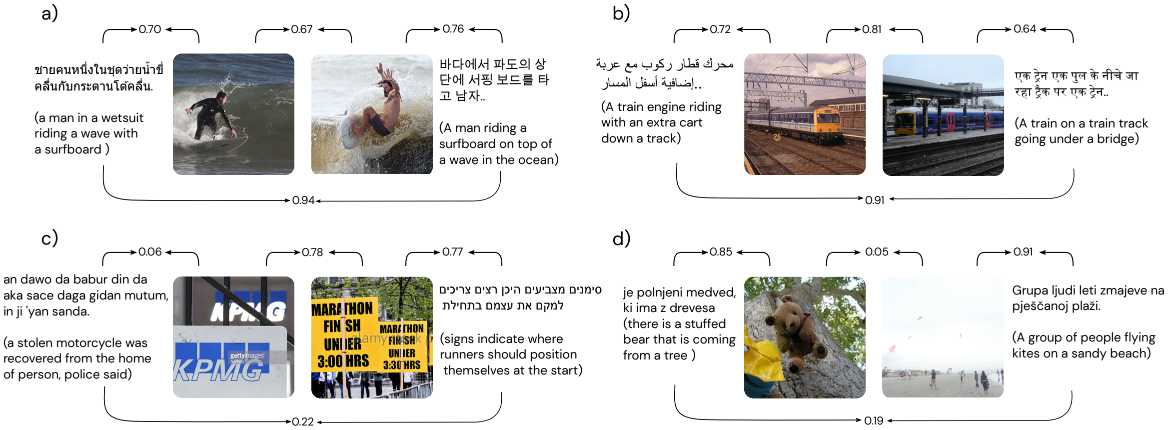

Figure 5: Qualitative results. We show two examples of positive matches (top) and two examples of negative matches (bottom). We refer the reader to Section 5.4 for further analysis.

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. 3. Our method outperforms all baselines by a significant margin, underscoring the utility of transitive relations across modalities.

We show some examples of our sentence translations in Tab. 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.

5.3 Word-level translation

Following the evaluation in Sigurdsson et al. (2020), we also evaluate the word-level translation. Since we lack ground truth translation at this level, 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 Appendix B.2. Then we test each pair of languages separately. For every translation, we evaluate retrieval in both directions. Fig. 4 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 Tab. 2.

5.4 Analysis

Visualizing transitive matches: Fig. 5 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. 5a and 5b 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. 5c demonstrates the importance of using α^x in addition to α^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 α^v value. However, because web data is not always clean, the caption in 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

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. Please see the end of Section 5.4 for discussion.

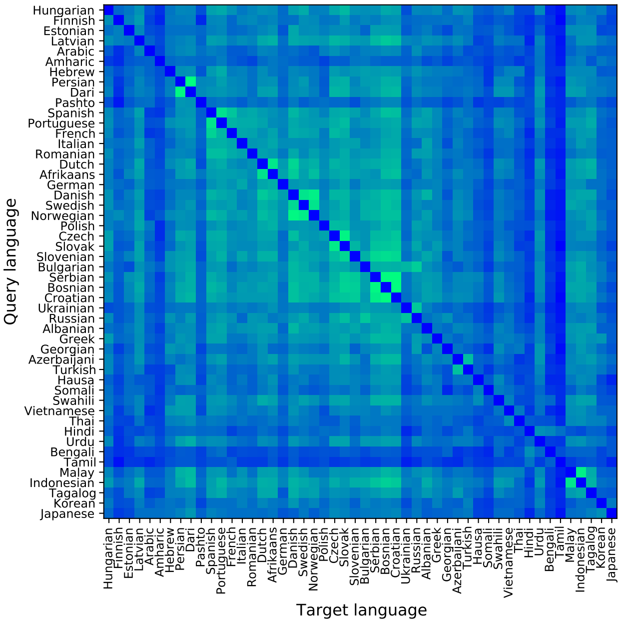

Figure 6: We show sentence-level translation accuracy by query-target language pair. In the figure, the languages are sorted by family (Romance, Baltic, etc.). The block-diagonal structure shows that languages from the same family are easier to translate between. We also find that language isolates in our dataset perform worse overall (e.g. Tamil, the only Dravidian language).

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. 5d shows an example where both sentences accurately describe their corresponding image, but the images do not match. As expected, this would result in a negative pair.

Failure cases: We show three prototypical examples of failure cases in Tab. 3. In the first example, the caption is not related to any visual concept, causing our model to translate it incorrectly. The second example shows how some words are related to incorrect concepts 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 overreliant on context to translate instead.

Translation difficulty by language: We itemize the performance of sentence-level translation by language in Fig. 6. Languages from the same family are often easier to translate between. The most difficult language is Tamil, the only Dravidian language in our dataset.

6 Conclusion

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.

References

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.

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

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

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A Plummer. Learning to scale multilingual representations for vision-language tasks. *arXiv preprint arXiv:2004.04312*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

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.

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. *arXiv preprint arXiv:1911.02116*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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*, pp. 268–277, 2016.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. Image pivoting for learning multilingual multimodal representations. *arXiv preprint arXiv:1707.07601*, 2017.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, 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*, pp. 316–327, 2018.

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)*, pp. 4969–4973. IEEE, 2018.

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*, pp. 770–778, 2016.

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.

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.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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, 2018a. URL `https://openreview.net/forum?id=rkYTTf-AZ`.

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, 2018b. URL `https://openreview.net/forum?id=H196sainb`.

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*, pp. 5039–5049, 2018c.

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*, pp. 740–755. Springer, 2014.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

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.

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 *CVPR*, 2020. URL `https://arxiv.org/abs/2003.05078`.

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*, pp. 543–553, 2016.

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)*, pp. 3551–3556, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

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*, pp. 5804–5813, 2019.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2: 67–78, 2014.

APPENDIX

We divide the appendix in two sections. In Section A we show more results, and in Section B we provide more information about the implementation of our method.

A ADDITIONAL RESULTS

A.1 FEATURE GENERALIZATION

Training a language model, as opposed to a text representation only designed for image retrieval, has the crucial advantage that it can be finetuned to perform downstream NLP tasks. In this work we are interested in evaluating how well the representations generalize across languages, after training on a downstream task. We evaluate our model on sentence correspondence: we split sentences in two, and half of the times we swap the second half of the sentences with other sentences of the same language. The model has to determine whether or not a sentence is coherent and the beginning of the sentence corresponds to the end of the sentence. We control for uppercase, word breaks, length of sentences etc. so that the model cannot find an easy shortcut (cheat), and has to rely on the semantic and syntactic structure of the sentence. We show examples of the test in Tab. 4 for English.

We train all the models for one epoch on half of the languages in the testing split (first half in alphabetical order), and test on both held-out samples from that half, and *on the languages from the other half* (new languages the sentence correspondence downstream task has not seen). We train a single transformer layer on top of our representation, with one head. For Sigurdsson et al. (2020), we do not apply the max-pooling over words in order to have a representation for each word. We show results on Tab. 5. The results show that methods trained with language models are much better at performing language tasks. It also shows that our method, trained with alignment, not only performs better on the languages the downstream task has been trained on, but it also generalizes better to other languages the sentence correspondence task has never seen, indicating that the model has a very aligned representation across languages. The relative decrease in accuracy is computed as the percentage decrease of the difference between the accuracy and the chance accuracy.

A.2 ADAPTATION TO A NEW LANGUAGE

We test how well our framework can adapt to incoming languages. For this purpose, we test on English and Chinese (separately), which were held out during training. To do so, we precompute features for images and texts from the languages we used during training, and finetune the model for the new language using the same losses as before. We train for one epoch.

After finetuning for English and Chinese, we repeat the same experiments performed for the other languages, showing that our system is able to adapt to new languages without losing the multilingual alignment. See Tab. 6 for translation results, and Tab. 7 for sentence correspondence results. For the sentence correspondence test, we use the head we trained before (without finetuning on the new languages).

A.3 MORE RESULTS ON TRANSLATION DIFFICULTY PER LANGUAGE

Similarly to Fig. 6, we show in Fig. 7 the *word* translation accuracy matrix for every pair of languages. As expected, languages that share an important part of their vocabulary are the ones with highest similarity scores. Specifically, there is a very high similarity between Bosnian, Croatian and

Sentence	**Corresponds**
A piece of cake sitting next to pastries on a white plate with red and yellow sauce	Yes
Seamless pattern with white bugs on a black background	Yes
A big tower with a big tv genre and a common language	No
A hand holding a smartphone with of a picnic by a lake	No

Table 4: Sentence correspondence task examples. See Appendix A.1.

Serbian, since the three of them are standardized varieties of the Serbo-Croatian language. Also, Indonesian is very close to Malay, as the former is a standardized variety of the latter. A final example is the Czech and Slovak pair: the two of them are languages from the Czech–Slovak group. This shows the importance of cognates across languages. We can find similar patterns for languages that are not as close, but that share the same family or alphabet.

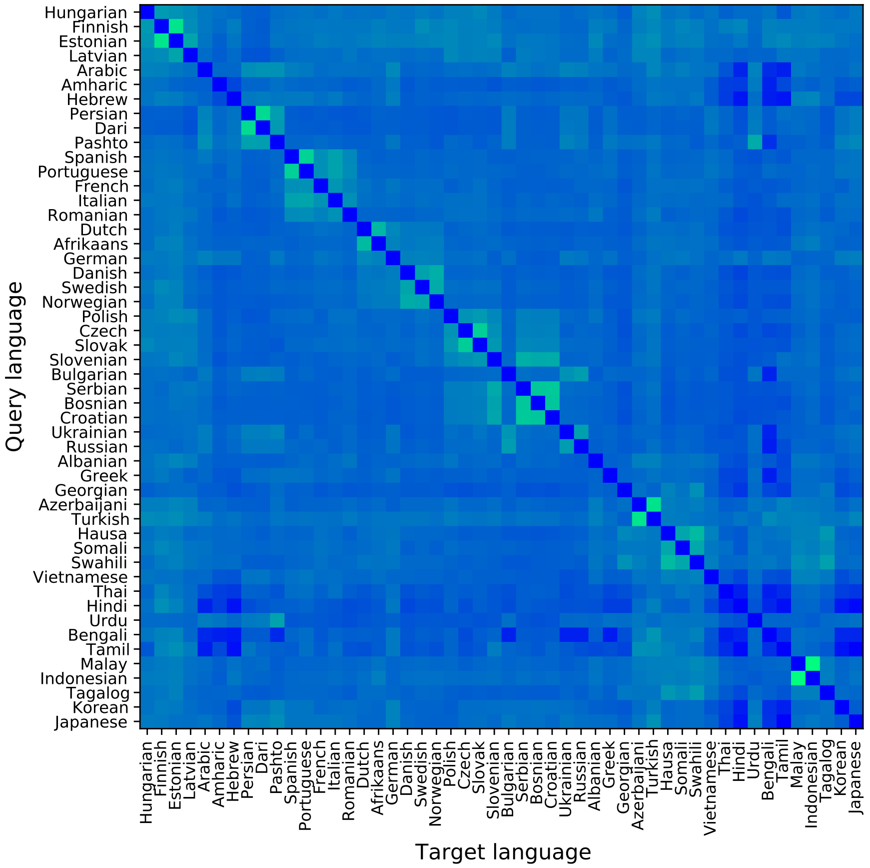

Figure 7: Word-level similarity across languages. See Appendix A.3 for more information.

We also show in Fig. 8 the sentence-level translation values from Fig. 6, but now we plot $A - A^T$. Instead of illustrating which language pairs are close, or are easier to work with, it shows which language pairs are asymmetric in the difficulty of the translation. Rarer languages —e.g. languages that are far from the others in the linguistic tree such as Somali, Tamil or Hindi— are easier to translate from than to translate to.

	Seen accuracy (%)	**Unseen accuracy (%)**	**Relative decrease (%)**
Chance	50	50	-
Text only	71.54	64.94	30.64
Lample & Conneau (2019)	72.41	68.22	18.70
Sigurdsson et al. (2020)	53.25	52.89	11.07
Globetrotter (Ours)	**75.95**	**74.54**	**5.43**
Supervised	75.64	68.73	26.95

Table 5: Sentence correspondence results. See Appendix A.1.

	English retrieved positives (%)	**Chinese retrieved positives (%)**
Chance	0.48	0.48
Text only	19.27	12.98
Sigurdsson et al. (2020)	59.18	37.96
Globetrotter (Ours)	**75.67**	**62.81**
Supervised	94.87	92.77

Table 6: Sentence translation results for finetuning. See Appendix A.2.

	English accuracy (%)	**Chinese accuracy (%)**
Chance	50	50
Text only	65.97	55.75
Sigurdsson et al. (2020)	50.2	50.5
Globetrotter (Ours)	**73.27**	**67.17**
Supervised	69.17	62.14

Table 7: Sentence correspondence results for finetuning. See Appendix A.2.

A.4 Clustering in the representation space

In this experiment, we show how differently the representation space is clustered when we train with and without visual alignment. We extract features for the test set examples both for the full model and the text-only model, and cluster these features using k-means, with $k = 50$ clusters. In Fig. 9 we show three sentences belonging to each one of the first three clusters (the selection of both the sentences and the clusters is arbitrary). When training with visual alignment the clusters have a semantic meaning, and when training without it the clusters are language-specific, proving that cross-modal alignment is necessary to obtain good semantic representations.

A.5 Generated translations

The learned representations are not only good to do translation by retrieval, but also to generate translations. In order to do so, we use a GPT-2 decoder (small version) from Radford et al., pre-trained on English. Next, we finetune it on English sentences from our dataset, and after that we finetune it yet again but conditioning it on feature vectors from the English finetuned model from Appendix A.2. To do this we use an extra linear layer at the input, and we concatenate the results with the input word embeddings. After that, we obtain a GPT-2 model that generates sentences in English based on the input representation. We then test it for translation by inputting representations obtained from other languages, and generating English translations for them. The sentences we used in the test were not used for any of the GPT-2 finetuning stages. We show results in Fig. 10. We selected the first 10 translations that were generated, without any cherry-picking. Interestingly, while our framework is not able to do an accurate literal translation, it does base the translation on the contextual knowledge provided by vision.

B Implementation details

B.1 Training and architecture details

We train a transformer network with 4 attention heads and $M = 4$ hidden layers, with a hidden size of $d = 512$. The size of the embeddings at the output of the heads (where the contrastive losses are computed) is $D = 128$. We use a batch size of 800. We set all the λ values in Eq 5 to $\lambda = 0.2$. We train with an Adam optimizer and a learning rate of $1e - 4$.

As mentioned in Section 3.5, we normalize the feature values z so that $||z||_2 = 1$. Then the similarity value is computed with a dot product, resulting in the cosine similarity. After that, we scale the value so that the range of the similarity is in $[0, 1]$, instead of $[-1, 1]$.

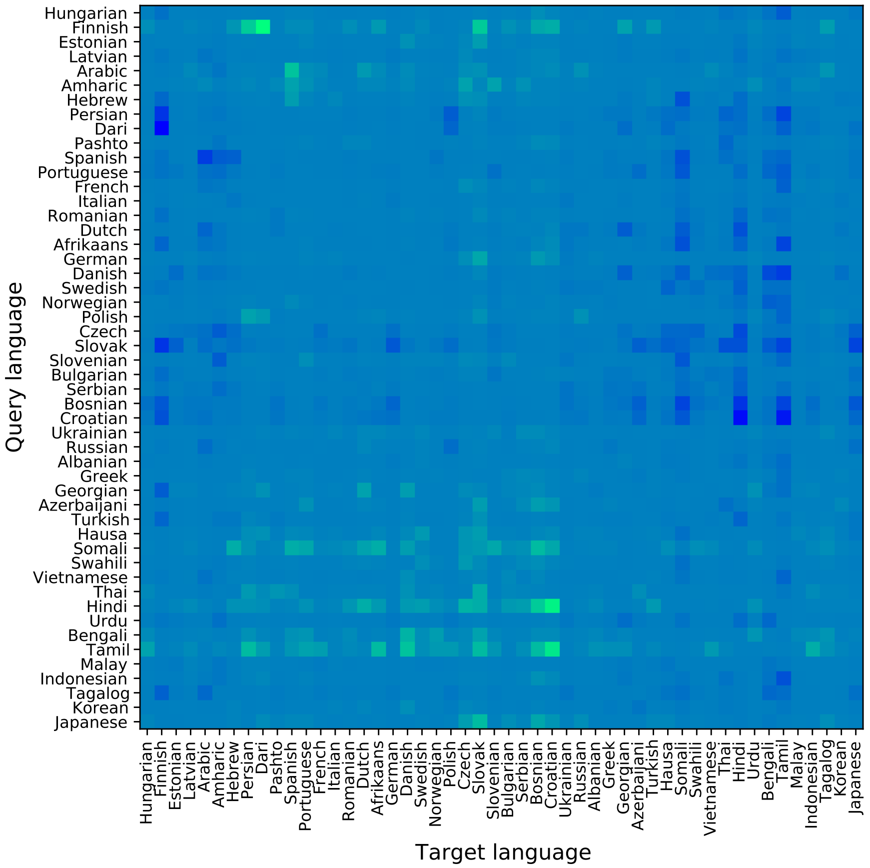

Figure 8: Asymmetry in the direction of the sentence-level translation. See Appendix A.3.

Clusters in full model

Cluster 1: Savannah animals
(Arabic): یه گورخر که داره به یه گورخر دیگه نگاه میکنه پایین یه مسیر خاکي
(Croatian): popodne provedeno igrajući se sa slonovima
(Georgian): ფართო გასროლა, ჟირაფები სავანას გავლით

Cluster 2: Wedding
(Bengali):উইন্ডোতে নববধূ এবং বর
(Slovenian): nevesta v meri obleko, ki ima roza šopek
(Urdu): شخص آپ کی شادی کے دن خواب سچ ہو بنانے دو!

Cluster 3: Bicycle/Motorcycle
(Swedish): en cykel kastad ner i sanden på en strand.
(Japanese): 砂地の隣にモーターバイクが駐車しています。
(Tamil): உடற்பயிற்சி பைக் மீது பெண்.
.

Clusters in text-only model

Cluster 1: French
un grand éléphant se tient près d'une clôture
motif circulaire sur fond rouge
homme silhouette à la plage

Cluster 2: Hindi
हाथ का एक सेट - डिजाइन के लिए प्यारा फल खींचा.
एक मॉडल घटना के दौरान फैशन शो में रनवे चलता
एक पतली परत पिज्जा तिमाही टुकड़ों में विभाजित।

Cluster 3: Greek
ποταμός είναι ένα δημοφιλές σημείο για κανό.
παλιά πόρτα σε ένα ξεχασμένο κήπο
πράσινα ψάρια στο γύρο ενυδρείο.

Figure 9: Clustering in the representation space. When trained without visual alignment the clusters are language-specific, and when trained with visual correspondence the clusters have a semantic meaning.

B.2 GROUND TRUTH FOR WORD TRANSLATION

In order to generate the ground truth translations at the token level, we use the split of the dataset that is translated to all the languages. We then create ground truth token translations for every language

Original sentence	Generated English translation
(Russian) кошка отдыхает на обочине в солнечный летний день (cat resting on the curb in sunny summer day)	cat lying on the grass
(German) Hardrock-Künstler treten während des Musikfestivals auf (hard rock artists perform during music festival)	artist performs on stage at festival.
(Croatian) Nekoliko snowboardera koji su poletjeli niz snijeg prekriveno brdo. (A few snowboarders taking off down a snow covered hill.)	some people skiing in the snow
(Arabic) الأوراق النقدية على خلفية سوداء (banknotes on a black background)	silver coin on the black background
(German) Porträt auf dem blauen Himmel Hintergrund (portrait on the blue sky background)	bald eagle on the green background
(Georgian) დამატებითი ფოტო ქონების ჩამონათვალი (additional photo for property listing)	photo of the front porch
(Swedish) utsikt över sjön från rutten (view over lake from the route)	picture of the beach on a sunny day
(Hungarian) légi kilátás a strand a legtöbb fehér és tiszta homok (aerial view of the beach with the most white and clean sand)	photo of the mountain lake in winter
(Croatian) šetnja po kiši. (a walk in the rain .)	photo of the rain : walking along the streets
(Afrikaans) akteur woon die spesiale geleentheid by (actor attends the special event)	some person attends los angeles premiere

Figure 10: Translation by generation. See Appendix A.5 for more information.

pair separately. In order to do that, we follow the tf-idf algorithm. We exploit the fact that we have alignments of languages at the group-of-words (sentence) level. The idea is that if the word "car" appears in an English sentence every time that the word "voiture" (car in French) appears in its French translation, they probably mean the same. In the following explanation, assume we are looking for the translation of a specific token t_i^A from language A into some token t_j^B from language B. We just redefine the concept of "document" in the classical tf-idf algorithm to be the collection of all the words (with repetition) in language B that appear in the same (translated) sentence as t_i^A. We call this collection (document) d.

First, we create a count of tokens in language B that appear in the document d, and compute the *term frequency* (tf) using this count:

$$\text{tf}_{j,d} = \frac{f_{j,d}}{\sum_{j' \in d} f_{j',d}}, \tag{6}$$

where $f_{j,d}$ is the count of the token t_j^B in a document d. Second, we compute the inverse document frequency, that takes into account how usual a token is in general, for all D documents:

$$\text{idf}_j = \log \frac{|D|}{|d \in D : t_j^B \in d|}. \tag{7}$$

Multiplying the tf and idf terms we get a value for each (i, j) pairs of tokens (the value is not symmetric). We store tokens t_i^A and t_j^B as ground truth translation if and only if t_j^B is in the top 5 for the tf-idf value of (i, j), for all j, *and* t_i^A is in the top 5 for the tf-idf value of (j, i), for all i.

The following are some examples of translations we obtain between Spanish and English: (electr, electr), (fotograf, ograph), (ción, ction), (grande, lar), (atas, jam), (pare, couple), (decor, decor), (ventana, window), (deportivo, team), (1950, 1950), (form, form), (30, 30), (casa, hom), (lave, key), (1960, 1960), (del, the), (libro, ok), (kara, kara), (ola, surfer), (fan, fan), (viol, viol), (%, %), (dar, standard), (segundo, sec), (equipo, sports), (rojo, red), (árbol, tree), (hierba, gras), (durante, dur), (bron, ze), (mani, demonstr), (pequeño, sm), (tí, typ), (turística, attra), (corre, run), (mus, muse), (atrac, tour), (baño, bat), (mam, mom), (una, on), (element, element), (ijo, son), (ant, ol), (mural, mural), (chocola, chocola), (iste, sad), (cinta, bon), (carro, cart), (edif, bu), (planta, plant), (óc, broccoli), (prim, st), (camina, runway), (cerca, close), (pop, artist), (nacional, nation), (ustr, alian),

	Flickr30k	**MSCOCO**	**Conceptual Captions**	**Total**
Image/language pairs per language	$3.1k$	$11.9k$	$63.8k$	$78.7k$
Total image/language pairs	$159k$	$616k$	$3.3M$	$4.1M$

Table 8: Dataset statistics. There are a total of 52 languages.

(vest, dress), (motocic, motorc), (perro, dog), (largo, ong), (+, +), (ates, tom), (fram, rasp), (camina, wal), (inta, inta).

B.3 Text network details

The input to the text network is a sequence of tokens $\{[SEQ], w_1, \ldots, w_i\}$ that represent a sentence in any language (Devlin et al., 2018). Before inputting tokens to the transformer, we encode them with a fixed-length vector representation. To embed input tokens, we use a $\mathcal{V} \times d$ word embedding matrix ϕ_w, where $\mathcal{V}$ is the size of the vocabulary considered by the tokenizer. We use $\mathcal{V} = 30,000$. We augment the input encoding with positional information (word index), translating the encoding by a learned vector: $\phi_{\text{txt}}(w_i) = \phi_w^T w_i + \phi_{\text{pos}}(w_i)$ where ϕ_{pos} encodes the word position of w_i.

We then input the augmented tokens to the transformer. A transformer block (Vaswani et al., 2017) consists of a multi-headed self-attention layer followed by a linear layer, that outputs a hidden representation for every token in the input sequence. These transformer blocks are concatenated in series to get deeper representations. Let $H^m \in \mathbb{R}^{d \times j}$ be the d dimensional hidden vectors at layer m. The transformer first computes vectors for queries $Q = W_q^m H^m$, keys $K = W_k^m H^m$, and values $V = W_v^t H^m$ where each $W_* \in \mathbb{R}^{d \times d}$ is a matrix of learned parameters. Using these queries, keys, and values, the transformer computes the next layer representation by attending to all elements in the previous layer:

$$H^{m+1} = SV \quad \text{where} \quad S = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right). \tag{8}$$

In practice, the transformer uses multi-head attention, which repeats Equation 8 once for each head, and concatenates the results. The network produces a final representation $\{h^M_{[SEQ]}, h^M_1 \ldots, h^M_i\}$ for a stack of M transformer blocks.

As mentioned in Section 3.5, we also add a prediction head. This head takes as input the final hidden representation for the $[SEQ]$ token, $h^M_{[SEQ]}$.

B.4 Dataset details

To collect the dataset, we used captions from the Flickr30k (Young et al., 2014), MSCOCO (Lin et al., 2014) and Conceptual Captions (Sharma et al., 2018) datasets. Flickr30k and MSCOCO are image captioning datasets that have been carefully curated and annotated in a controlled setting, so the text descriptions are accurate and thorough. However, most of the images in our datasets come from Conceptual Captions, which consists of captions harvested from the web, so the visual-language alignment is more noisy.

We randomly split each dataset into 52 equally sized parts, one for each language supported by the machine translation service we use. Each split is assigned a unique language, and splits with the same language across datasets are combined. The split which is assigned the English language is set aside and translated into all 51 other languages, and only used in testing. We also set aside the split translated into Chinese for fine tuning experiments. The remaining 50 splits have their original English captions discarded, and are then split 80%-20% into training and validation data. All experiments shown in Section 5 are run on the reserved test data.

Note that there is no overlap at all (visual or linguistic) between the different splits, except for the test split. Please see Table 8 for more details about the dataset.