

# Using Visual Feature Space as a Pivot Across Languages

Ziyan Yang<sup>1</sup> Leticia Pinto-Alva<sup>1,2</sup> Franck Dernoncourt<sup>3</sup> Vicente Ordonez<sup>1</sup>

<sup>1</sup>University of Virginia, <sup>2</sup>San Pablo Catholic University, <sup>3</sup>Adobe Research  
{zy3cx, lp2rv, vicente}@virginia.edu, dernonco@adobe.com

## Abstract

Our work aims to leverage visual feature space to pass information across languages. We show that models trained to generate textual captions in more than one language conditioned on an input image can leverage their jointly trained feature space during inference to pivot across languages. We particularly demonstrate improved quality on a caption generated from an input image, by leveraging a caption in a second language. More importantly, we demonstrate that even without conditioning on any visual input, the model demonstrates to have learned implicitly to perform to some extent machine translation from one language to another in their shared visual feature space. We show results in German-English, and Japanese-English language pairs that pave the way for using the visual world to learn a common representation for language.

## 1 Introduction

There has been great interest in learning visual representations from images paired with natural language annotations. While tasks such as image caption generation e.g. (Young et al., 2014; Lin et al., 2014) have focused mostly on English text, there is a growing body of work extending to a larger set of languages (Calixto et al., 2012; Elliott et al., 2015, 2016). Images annotated in multiple languages offer the possibility of studying grounded models of languages along with their commonalities and intrinsics in direct connection with the visual world.

We focus in the multilingual image description generation setting where we train an image encoder with soft-attention (Xu et al., 2015) and multiple text decoders for each target language. Then, we demonstrate that information from one language can be transferred to another language using energy based inference (LeCun et al., 2006) in an iterative fashion by leveraging the backpropagation algorithm at test time. Effectively, we demonstrate that

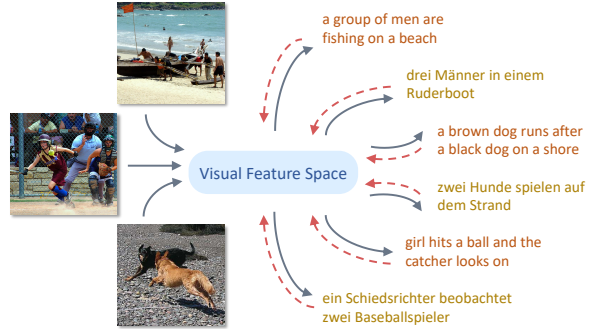


Figure 1: Our work shows how visual features capture multi-lingual information in image conditioned models (solid blue arrows) and how to pivot this information across languages during inference by incorporating feedback connections (dotted red arrows) from language back to visual feature space.

the common visual feature space used to generate text in the target languages also learns implicitly alignments between them and thus acts as its own form of “visual language”. Figure 1 shows some example images and textual descriptions in German and English, as well as a general outline of our approach. We demonstrate our findings by (1) showing that a textual description in a second language helps improving generated image description quality in a target language, and (2) showing how to use the visual feature space in an image encoder to translate sentences among target languages even in the absence of visual input. Stated otherwise, our claim is that *multi-lingual image captioning models can act as incidental machine translators*.

More broadly, our work explores the possibility of using visually grounded representation learning as a unifying medium across languages, where a single model is used for learning mappings across an exhaustive number of language pairs among target languages. We demonstrate our approach on two datasets of images annotated with German, English, and Japanese, English respectively.

## 2 Background

Our work is different from work in both general neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2015; Luong et al., 2015), and multimodal machine translation (MMT) (Elliott, 2018; Caglayan et al., 2019; Raunak et al., 2019) in that we do not use parallel corpora across languages. This distinction is important and perhaps confusing as we rely on the Multi30k dataset for which several versions and tasks exist (Elliott et al., 2016; Barrault et al., 2018). The first task, *task 1*, is perhaps the most popular, containing parallel text among languages (German, English, French and Czech) describing 30,000 images from the Flickr30k dataset (Young et al., 2014) with a single caption in each language. This task has often been used also as a pure machine translation benchmark by discarding the image information. The second task, *task 2*, is the one that concerns our work and is one of the tasks we leverage for training, which is the multilingual image description generation task, where each of the 30,000 images is annotated with 5 independent (unpaired) captions in German and English.

Using visual features as “pivoting” variables is related to using conditional latent variables to iteratively perform inference using backpropagation. A version of this idea was perhaps first mentioned in LeCun et al. (2006) as noted by Belanger and McCallum (2016). Besides work on Generative Adversarial Networks (Goodfellow et al., 2014), there are only a few works since then that have independently proposed to use iterative inference with backpropagation including Stoyanov et al. (2011); Domke (2013); Wang et al. (2018). We particularly adopt the single layer version of the most recently proposed feedback propagation approach of Wang et al. (2018) as it was more directly applied to convolutional neural networks for visual recognition. Unlike this previous work, we are the first to show that feedback propagation can leverage its latent space to use interactions among target variables even in the absence of any visual input at test time.

## 3 Method

As mentioned earlier, our base model consists of the image captioning model with “soft” attention proposed by Xu et al. (2015) but trained with independent textual decoders for each target language. In this model, the image encoder consists of a convolutional neural network and the textual decoders

consist of recurrent neural networks with Long Short Term Memory (LSTM) units. The output soft spatial attention vector computed from the input image is used as input for the decoders to generate captions in each target language. Let the input image be  $I$ , and let us consider the bilingual case of language  $a$  and language  $b$  where the targets are text sequences  $t_a$  and  $t_b$  respectively. The model can then be expressed as:

$$F(I) = [f_a(z), f_b(z)], \quad (1)$$

where  $z = g(I)$  is the output of a visual feature extractor  $g$ , and  $f_a$  and  $f_b$  are text decoders for each language that try to approximate  $t_a$  and  $t_b$  by producing a joint pseudo-distribution from where to sample text.

While the trained model amounts to a traditional image captioning model under a multi-lingual objective, at test time we experiment with the following settings: (1) Predicting image descriptions in multiple languages conditioned on the visual input, (2) predicting text in one language conditioned on the visual input and text in a second language (or languages), and (3) predicting text in one language conditioned on the other language (or languages) but with no visual input. The first case can be performed directly by standard decoding techniques on the outputs  $f_a(z)$  and  $f_b(z)$  such as beam search. So we explain here in detail the latter two cases:

**Visual Input + Second Language** In order to use the latent feature space to predict  $t_a$  conditioned on  $t_b$  and  $I$ , we estimate a pivoting variable  $\hat{z}$  by iteratively minimizing using backpropagation the following:

$$\hat{z} = \arg \min_z E(t_b, f_b(z)), \quad (2)$$

where  $E$  is an energy function that measures the compatibility between  $t_b$  and  $f_b(z)$ . In other words we try to synthesize a feature  $\hat{z}$  that can plausibly generate the target text in language  $b$ . Pivoting variable  $z$  in the first iteration is computed from input image  $I$  as  $z = g(I)$ . In practice we used the same loss function used to approximate our text decoders for our energy function during inference (cross entropy loss). This general approach referred as energy-based inference in LeCun et al. (2006) is referred as feedback-based inference in Wang et al. (2018) and  $z$  as a pivoting variable, we adopt this later terminology. The target text description in language  $a$  can be obtained by standard decoding

<i>Input</i>	<i>Target</i>	<i>BLEU-4</i>	<i>ROUGE-L</i>	<i>CIDEr</i>
Image	DE	16.29	40.85	44.88
Image	EN	24.89	47.22	51.60
Image + EN	DE	21.36	46.51	58.57
Image + DE	EN	27.22	50.34	61.61
EN	DE	15.23	41.79	40.45
DE	EN	18.37	44.43	40.15

Table 1: Results on Multi30k dataset with German (DE) and English (EN) unpaired textual captions.

techniques such as beam search from the pseudo-distribution  $f_a(\hat{z})$ .

**No Visual Input** In our third type of inference we use the latent feature space to predict  $t_a$  conditioned exclusively on  $t_b$  but without access to any image input. We optimize the same expression as in Equation 2 but initialize  $z$  as  $z = g(\xi)$  instead, where  $\xi$  is a trivial input image with pixel values sampled from a gaussian distribution  $\mathcal{N}(\mu, \sigma^2)$  with a mean and standard deviation estimated from pixel values in the training data. In this case the final value of the visual feature  $\hat{z}$  is iteratively synthesized only from the textual information in  $t_b$ . As in the previous case, the target textual description in language  $a$  can be obtained by standard decoding techniques from the pseudo-distribution  $f_a(\hat{z})$ .

The approach outlined in this section is general and can be extended for arbitrary languages  $a$  and  $b$  and to an arbitrary number of languages by adding more textual decoders such that  $F(I) = [f_1(z), f_2(z), \dots, f_n(z)]$ , and for arbitrary conditioning during inference such that Equation 2 becomes:

$$\hat{z} = \arg \min_z \sum_{k \in K} E_k(t_k, f_k(z)),$$

where  $K \subset V$  is the support subset of languages used as feedback during inference, and  $V$  is the set of all target languages.

In addition, the presented approach is also agnostic to the neural network architecture of the underlying language grounding model as long as the model is end-to-end differentiable.

## 4 Experiments

**Data** We use *task 2* in Multi30k (Elliott et al., 2016), which has 29,000, 1,014, and 1,000 images for training, validation, and testing respectively. Each image has 5 English and 5 German unpaired textual descriptions. Therefore, there are

<i>Input</i>	<i>Target</i>	<i>BLEU-4</i>	<i>ROUGE-L</i>	<i>CIDEr</i>
Image	JP	40.36	57.42	101.03
Image	EN	32.68	51.99	99.79
Image + EN	JP	42.33	58.92	110.82
Image + JP	EN	34.29	53.22	108.53
EN	JP	31.92	52.35	84.64
JP	EN	24.75	46.81	81.38

Table 2: Results COCO+STAIR with Japanese (JP) and English (EN) unpaired textual captions.

145,000, 5,070, and 5,000 captions for training, validation and testing for each language. We jointly train the image captioning model to generate captions for both languages. Multi30k provides pre-processed lowercase tokens for all the sentences. We also use STAIR Captions (Yoshikawa et al., 2017), which contains Japanese captions for all images in the MS COCO dataset (Lin et al., 2014). The Japanese captions are also collected independently from the English captions in MS COCO, thus not being paired.

**Model** We use Resnet-50 (He et al., 2016) in the image encoder and keep the same settings as in (Xu et al., 2015). The attention, embedding and decoder dimensions are all set to 512. During training, we use teacher-forcing for several epochs and finetune the whole model including the image encoder using cross entropy losses over the vocabulary of words for each language. The learning rate for text decoders is 4e-4 and 1e-4 for the image encoder. During feedback propagation, we choose the intermediate representation after the Conv-40 layer in Resnet-50 as pivot variable (We chose this layer over Conv-22 and Conv-49 using a held out set) and we empirically determine the number of steps and update rate in the iterative optimization empirically<sup>1</sup>. For the text decoders, the vocabulary size for all the languages is 10,000. All captions are sampled using beam search with a beam size of 5.

**Results** Table 1 and Table 2 shows our results on Multi30k and COCO+STAIR respectively under six possible different scenarios depending on inputs and outputs and reporting BLEU-4, ROUGE-L and CIDEr evaluation metrics. Our results are remarkably consistent across languages and datasets in that (1) —*a caption from a second language always improves image caption quality in*

<sup>1</sup>code is available at <https://github.com/uvavision/visual-pivoting>

the first language, this is true for all pairs and directions English-German, German-English, English-Japanese, Japanese-English (2) In both datasets, but especially in the Japanese-English, English-Japanese case, *models show a remarkable ability to learn alignments between languages even in the absence of visual input*. This difference in gains might be due to COCO+STAIR having a larger training data. Qualitative results are shown in Figure 2 for both image + second language caption generation, and caption to caption translation. For instance in the top example, the gender of the subject is identified from the visual input but the location is clearly leveraging the input German caption.

Since the sentences are only paired with the underlying image, we might have an input caption as “The young boy is playing with a red ball”, and five reference captions such as “Ein Junge spielt mit dem Sand” (a young boy plays in the sand). How well would a machine translation system perform on this task? We used Google Translate for this purpose and found that it obtains BLEU: 16.75, ROUGE: 42.54 and CIDEr: 50.09 on English to German in the Multi30k dataset. These numbers are contrasted with our results in the last row of Table 1 where our method obtains comparable results with BLEU: 18.37, ROUGE: 44.43 and CIDEr: 40.15. Google Translate which is a system not tuned specifically for this data, only performs significantly better in terms of CIDEr scores which is a metric that rewards matches in infrequent n-grams.

## 5 Related Work

Our work is closely related to the problem of lexicon induction from images which has been used to address the issue when paired texts are not available for machine translation. Works that have leveraged visual features to build such lexicon include Bergsma and Van Durme (2011); Kiela et al. (2015); Hewitt et al. (2018). Other works with similar goals include Hitschler et al. (2016) where visual features are used to assemble a weakly supervised set of text pairs, Gu et al. (2018) where the objective is to leverage both image-caption pairs and multilingual parallel corpora, and Gella et al. (2017) where the images are used as pivot between languages to learn multimodal multilingual common representations. Our work leverages only unpaired data and does not aim to train a machine translation model or obtain multimodal representations explicitly. Related to our goals is also work




INPUTS	OUTPUTS
 ein Mann fängt das Ball am Strand.	<b>image:</b> A man in a white shirt is jumping in the air.  <b>image + de:</b> A man is playing with a red ball on the beach.
 eine Frau in gestreiftem Shirt klettert an einer Felswand	<b>image:</b> a man is standing on a rock overlooking a valley.  <b>image + de:</b> a woman in a striped shirt is standing on a rock .
 新聞紙の上に無数のはさみがおいてある	<b>image:</b> A group of blue and white cake on a table.  <b>image + jp:</b> A table topped with lots of blue and white scissors.
INPUTS	OUTPUTS
ein Kleinkind spielt mit einer gelben Plastikschiippe.	a baby is playing with a yellow ball in the grass.
der Bub spielt mit dem Sand.	a child is playing in the sand.
ein Junge spielt mit einer Spielzeugschaufel auf steinigem Boden.	a young boy playing with a toy in a patch of grass.
デスクにパソコンが置いてある	a laptop computer sitting on top of a desk.
木製のテーブルと棚にパソコンとプリンターが置いてある	a room with a wooden door and a door.
デスク上のパソコンの横に水が入ったペットボトルが置かれている	a black cat sitting on top of a computer desk.
デスクの上にパソコンやライト、本が置かれている	a desk with a laptop and a book.

Figure 2: Here we showcase interesting examples of the types of translations obtained with our approach. Casing and color coding were added manually.

aiming to translate neural network internal representations into natural language e.g. (Andreas et al., 2017; Evtimova et al., 2018). Moreover, general work in multimodal machine translation under supervised/unsupervised learning is also related to our work. Elliott and Kádár (2017) and Helcl et al. (2018) investigate visually grounded representations to improve supervised multimodal machine translation, and ignore input images at test time. Using reinforcement learning, Chen et al. (2018) jointly optimizes a captioner and a neural machine translator to achieve unsupervised multimodal machine translation, while Su et al. (2019) and Huang et al. (2020) explore transformers (Vaswani et al.,



2017) to construct a text encoder-decoder for the same goal. Our work is different from referred multimodal machine translation works since our work starts from multilingual image captioning and is applied to machine translation, while some of the other methods start from a multimodal machine translation and are applied to machine translation, however building models that take advantage from these two tasks is a possible avenue for future work. Many of previous methods rely on pre-training on external data for either captioning or machine translation and finetune models using *task 1* data from Multi30k, while we rely on only the provided *task 2* data from Multi30k. For example, Su et al. (2019) and Huang et al. (2020) both utilize WMT News Crawl datasets to pre-train machine translation models.

## 6 Conclusions

We show that visual feature space can be used as a pivot for transferring information across languages. We demonstrated this by showing how having access to captions in a second language can improve the generated caption quality in a target language. Moreover, we present the key result that we can perform arbitrary mappings among target languages in an image conditioned model, even when removing the requirement of visual input, essentially demonstrating the model learns mappings across languages similar to machine translation models.

**Acknowledgments** This work was partially funded by gifts from Adobe Research. We are also thankful for positive comments and suggestions from anonymous reviewers.

## References

- Jacob Andreas, Anca Dragan, and Dan Klein. 2017. Translating neuralese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 232–242.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chirag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT ’18)*.
- David Belanger and Andrew McCallum. 2016. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.
- Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *Twenty-Second International Joint Conference on Artificial Intelligence*.
- Ozan Caglayan, Pranava Swaroop Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4159–4170.
- Iacer Calixto, Teófilo de Campos, and Lucia Specia. 2012. Images as context in statistical machine translation. In *Proceedings of the workshop on vision and language, VL*.
- Yun Chen, Yang Liu, and Victor OK Li. 2018. Zero-resource neural machine translation with multi-agent communication game. *arXiv preprint arXiv:1802.03116*.
- Justin Domke. 2013. Learning graphical model parameters with approximate marginal inference. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2454–2467.
- Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.
- Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR, abs/1510.04709*, 3.
- Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*.
- Desmond Elliott and Akos Kádár. 2017. Imagination improves multimodal translation. *arXiv preprint arXiv:1705.04350*.
- Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. 2018. Emergent communication in a multi-modal, multi-step referential game. *International Conference on Learning Representations*.
- Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. *arXiv preprint arXiv:1707.07601*.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

- Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired image captioning by language pivoting. In *The European Conference on Computer Vision (ECCV)*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. Cuni system for the wmt18 multimodal translation task. *arXiv preprint arXiv:1811.04697*.
- John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.
- Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2399–2409.
- Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. *arXiv preprint arXiv:2005.03119*.
- Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.
- Douwe Kiela, Ivan Vulic, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. ACL; East Stroudsburg, PA.
- Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.
- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.
- Vikas Raunak, Sang Keun Choe, Quanyang Lu, Yi Xu, and Florian Metze. 2019. On leveraging the visual modality for neural machine translation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 147–151.
- Veselin Stoyanov, Alexander Ropson, and Jason Eisner. 2011. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 725–733.
- Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. 2019. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10482–10491.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Tianlu Wang, Kota Yamaguchi, and Vicente Ordonez. 2018. Feedback-prop: Convolutional neural network inference under partial evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 898–907.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.
- Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.