Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage

Ashish V. Thapliyal
Google Research
asht@google.com

Radu Soricut
Google Research
rsoricut@google.com

Abstract

Cross-modal language generation tasks such as image captioning are directly hurt in their ability to support non-English languages by the trend of data-hungry models combined with the lack of non-English annotations. We investigate potential solutions for combining existing language-generation annotations in English with translation capabilities in order to create solutions at web-scale in both domain and language coverage. We describe an approach called Pivot-Language Generation Stabilization (PLuGS), which leverages directly at training time both existing English annotations (gold data) as well as their machine-translated versions (silver data); at run-time, it generates first an English caption and then a corresponding target-language caption. We show that PLuGS models outperform other candidate solutions in evaluations performed over 5 different target languages, under a large-domain testset using images from the Open Images dataset. Furthermore, we find an interesting effect where the English captions generated by the PLuGS models are better than the captions generated by the original, monolingual English model.

1 Introduction

Data hungry state-of-the-art neural models for language generation have the undesired potential to widen the quality gap between English and non-English languages, given the scarcity of non-English labeled data. One notable exception is machine translation, which benefits from large amounts of bilingually or multilingually annotated data. But cross-modal language generation tasks, such as automatic image captioning, tend to be directly hurt by this trend: existing datasets such as Flickr (Young et al., 2014a), MSCOCO (Lin et al., 2014), and Conceptual Captions (Sharma et al., 2018) have extensive labeled data for English, but labeled data is extremely scarce in other languages (Elliott et al., 2016) (at 2 orders of magnitude less for a couple of languages, and none for the rest).

In this paper, we conduct a study aimed at answering the following question: given a large annotated web-scale dataset such as Conceptual Captions (Sharma et al., 2018) in one language, and a baseline machine translation system, what is the optimal way to scale a cross-modality language generation system to new languages at web-scale?

We focus our study on the task of automatic image captioning, as a representative for cross-modal language generation where back-and-forth consistency cannot be leveraged in a straightforward manner [1]. In this framework, we proceed to test several possible solutions, as follows: (a) leverage existing English (En) image captioning datasets to train a model that generates En captions, which are then translated into a target language X; we call this approach Train-Generate-Translate (TGT); (b) leverage existing En captioning datasets and translation capabilities to first translate the data into the target language X, and then train a model that generates X -language captions; we call this approach Translate-Train-Generate (TTG); (c) stabilize the TTG approach by directly using the En gold data along with the translated training data in the X language (silver data) to train a model that first generates En captions (conditioned on the image), and then generates X -language captions (conditioned on the image and the generated En caption); this approach has En acting as a pivot language between the input modality and the X - language output text, stabilizing against and reduc-

[1] We chose to focus on the cross-modality version of this problem because for the text-only modality the problem is less severe (due to existing parallel data) and also more studied (Artetxe et al., 2018), as it is amenable to exploiting back-and-forth consistency as a powerful learning signal.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 160–170
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

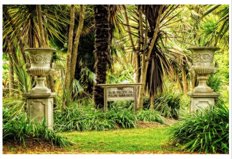
Image	**TGT** **Train Generate Translate**	**TTG** **Translate Train Generate**	**PLuGS** **Pivot Language Generation Stabilization**
	Das Logo ist auf dem Computer zu sehen. (*the logo can be seen on the computer.*)	**Bild mit dem Titel Live mit einem Schritt** (*Image titled Live with a step*)	the iphone is seen in this undated image . <de> **Das iPhone ist in diesem undatierten Bild zu sehen .**
	Autoverkehr an einem regnerischen Tag (*car traffic on a rainy day*)	**Polizeiauto auf der Straße** (*police car on the street*)	a car in the city <de> **ein auto in der stadt**
	**Bronzestatue im Garten** (*bronze statue in the garden*)	**eine Stadt im Garten** (*a city in the garden*)	the entrance to the gardens <de> **der Eingang zu den Gärten**

Figure 1: Examples of captions produced in German by Train-Generate-Translate (TGT), Translate-Train-Generate (TTG), and Pivot Language Generation Stabilization (PLuGS) approaches. Captions are shown in bold font. For TGT and TTG outputs, we show the English translation in parenthesis beside the caption. For the PLuGS outputs we mark the Stabilizer in the output using a light gray background. We do not explicitly show a translation for PLuGS outputs since the Stabilizer is already a translation.

ing potential translation noise. We call the latter the Pivot-Language Generation Stabilization (PLuGS) approach. Examples of outputs produced by these three solutions are shown in Fig. 1.

We perform extensive evaluations across five different languages (French, Italian, German, Spanish, Hindi) to compare these three approaches. The results indicate that the bilingual PLuGS models consistently perform the best in terms of captioning accuracy. Since there is very little support in the literature regarding the ability of standard evaluation metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016) to accurately measure captioning accuracy for non-English languages, our evaluations are done using fine-grained, side-by-side human evaluations using paid raters; we explain the evaluation protocol in detail in Sec. 5.

Besides the evaluations on bilingual PLuGS models, we also train and evaluate a multilingual PLuGS model, in which all five non-English languages considered are supported through a single model capable of generating outputs in all 5 languages. The results indicate that similar languages are reinforcing each other in the common representation space, showing quantitative gains for the Romance languages involved in our experiments. A related but perhaps less expected result is that the English captions generated by PLuGS models (what we call the Stablizer outputs) are better, as measured using side-by-side human evaluations, than captions generated by the original, monolingual English model.

There is a final additional advantage to having PLuGS models as a solution: in real-world applications of image captioning, quality estimation of the resulting captions is an important component that has recently received attention (Levinboim et al., 2019). Again, labeled data for quality-estimation (QE) is only available for English[2], and generating it separately for other languages of interest is expensive, time-consuming, and scales poorly. The TGT approach could directly apply a QE model at run-time on the En caption, but the subsequent translation step would need to be perfect in order not to ruin the predicted quality score. The TTG ap-

[2] https://github.com/google-research-datasets/Image-Caption-Quality-Dataset

proach cannot make use at run-time of an En QE model without translating the caption back to English and thus again requiring perfect translation in order not to ruin the predicted quality score. In contrast, the PLuGS approach appears to be best suited for leveraging an existing En QE model, due to the availability of the generated bilingual output that tends to maintain consistency between the generated EN- & X-language outputs, with respect to accuracy; therefore, directly applying an English QE model appears to be the most appropriate scalable solution.

2 Related Work

There is a large body of work in automatic image captioning for English, starting with early work (Hodosh et al., 2013; Donahue et al., 2014; Karpathy and Fei-Fei, 2015; Kiros et al., 2015; Xu et al., 2015) based on data offered by manually annotated datasets such as Flickr30K (Young et al., 2014b) and MS-COCO (Lin et al., 2014), and more recently with work using Transformer-based models (Sharma et al., 2018; Zhao et al., 2019; Changpinyo et al., 2019) based on the web-scale Conceptual Captions dataset (Sharma et al., 2018).

Generating image captions in languages other than English has been explored in the context of the WMT 2017-2018 multimodal translation sub-task on multilingual caption generation (Elliott et al., 2017). The goal of the task is to generate image captions in German and French, using a small training corpus with images and captions available in English, German and French (based on Flickr30K). In the context of that work, we use the results reported in (Caglayan et al., 2019) to quantitatively compare it against our approach.

Another relevant connection is with the work in (Jaffe, 2017), which explores several LSTM-based encoder-decoder models that generate captions in different languages. The model most similar to our work is their Dual Attention model, which first generates an English caption, then an LSTM with attention over the image and the generated English caption produces a German caption. Their quantitative evaluations do not find any additional benefits for this approach.

Our work is related to this idea, but there are key technical differences. In the PLuGS approach, we train an end-to-end model based on a Transformer (Vaswani et al., 2017) decoder that exploits the generated English-prefix via the self-attention mechanism to learn to predict the non-English target caption, conditioned on the English tokens at multiple levels through the decoder stack. Moreover, we approach this study as the search for a solution for web-scale multi-language image captioning: we employ the web-sized Conceptual Captions dataset for training, and consider the effects of using captions across multiple languages, as well as multi-language/single-model setups.

3 Model Architecture

We model the output caption using a sequence-generation approach based on Transformer Networks (Vaswani et al., 2017). The output is the sequence of sub-tokens comprising the target caption. As shown in Fig. 2, the input sequence is obtained by concatenating the following features.

Global Image Embedding: We use a global image representation using the Graph-RISE model (Juan et al., 2019), a ResNet-101 model (He et al., 2016) trained for image classification at ultra-fine granularity levels. This model produces a compact image embedding i of dimension $D_i = 64$. This embedding is projected to match Transformer dimensions (set to 512 in most of our experiments) by a 2 layer DNN with linear activation and fed as the first element in the sequence of inputs to the encoder.

Object Labels Embeddings: Detecting the presence of certain objects in the image (e.g. “woman”, “flag”, “laptop”) can help generate more accurate captions, since a good caption should mention the more salient objects. The object labels are generated by an object detection model which is run over the entire image. The output labels are then converted to vectors using word embeddings to obtain what we call object-label embeddings.

More precisely, we detect object labels over the entire image using a ResNet-101 object-detection classifier trained on the JFT dataset (Hinton et al., 2015). The classifier produces a list of detected object-label identifiers, sorted in decreasing order by the classifier’s confidence score; we use the first sixteen of these identifiers. The identifiers are then mapped to embeddings o_j using an object-label embedding layer which is pre-trained to predict label co-occurrences in web documents, using a word2vec approach (Mikolov et al., 2013). The resulting sequence of embeddings is denoted $O = (o_1, \dots, o_{|O|})$, where each o_j has dimension $D_o =$

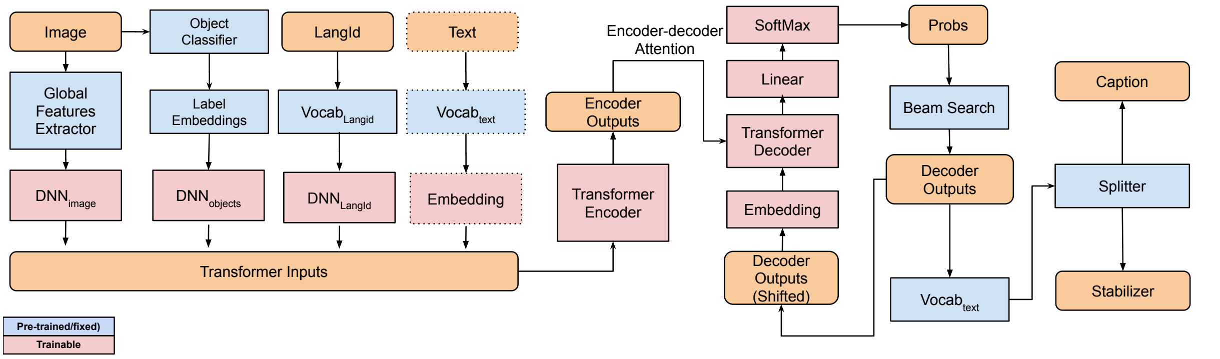


Figure 2: The Transformer based PLuGS model. The text on the input side is used for the translation and multi-modal translation experiments with the Multi30K dataset. For image captioning, no text input is provided.

256. Each member of this sequence of embeddings is projected to match Transformer dimensions by a 2 layer DNN with linear activation. This sequence of projected object-label embeddings is fed to the encoder together with the global image embedding.

LangId Embeddings: When training language-aware models, we add as input the language of the target sequence. We specify the language using a language identifier string such as *en* for English, *de* for German, etc. We call this the LangId of the target sequence or target LangId in short. Given the target LangId, we encode it using a LangId vocabulary, project it to match Transformer dimensions with a 2 layer DNN, then append it to the encoder input sequence.

Text Embeddings: All text (input or output) is encoded using byte-pair encoding (Sennrich et al., 2016) with a shared source-target vocabulary of about 4000 tokens, then embedded as described in (Vaswani et al., 2017), resulting in a sequence of text embeddings. The embeddings dimensions are chosen to match the Transformer dimensions. When performing the translation (MT) and multimodal translation (MMT) experiments in Sec. 6.1, the sequence of source text embeddings are fed to the encoder after the LangId embedding. Additionally, we reserve a token-id in the text vocabulary for each language (e.g. $\langle de \rangle$ for German) for use as a separator in the PLuGS model output and also have a separate start-of-sequence token for each language.

Decoding: We decode with beam search with beam width 5.

PLuGS: For PLuGS models, in addition to the target caption we require the model to generate a pivot-language (En) caption which we call the Stabilizer. Specifically, we train the model over target sequences of the form Stabilizer + $\langle$separator$\rangle$ + Caption.

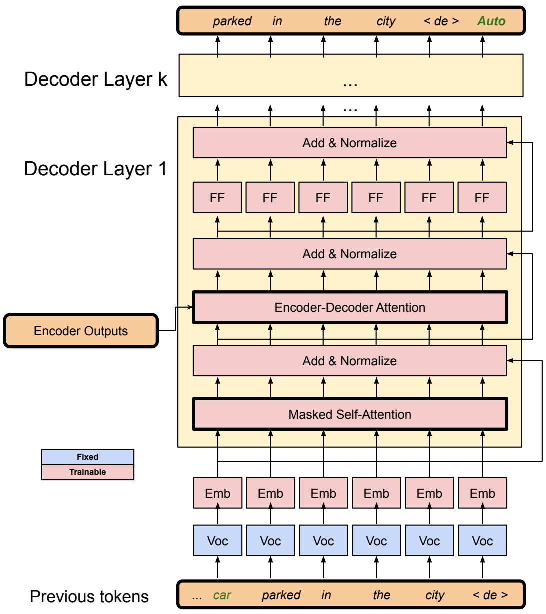


Figure 3: Caption's dependence on the Stabilizer. The target-language caption is conditioned on the Stabilizer through the Masked Self-Attention in the decoder, and on the input image through the Encoder-Decoder attention that attends to the outputs of the last encoder layer. Note that in this figure, FF stands for the feed forward network, Voc stands for the (fixed) text vocab, and Emb stands for the (trainable) text embeddings.

We use $\langle$\$LangId$\rangle$ as the separator (i.e., for German captions we use $\langle de \rangle$ as the separator). This approach has the advantage that it can be applied to multilingual models as well. We subsequently split the model output based on the separator to obtain two strings: the Stabilizer and the Caption.

Note an important technical advantage here: as shown in Fig. 3, after initially generating the Stabilizer output, the Transformer decoder is capable of exploiting it directly via the self-attention mechanism, and learn to predict the non-English Caption tokens conditioned (via teacher-forcing) on the gold-data English tokens at multiple levels through the decoder stack, in addition to the cross-attention mechanism attending to the inputs. As our results indicate, the models are capable of maintaining this advantage at run-time as well, when auto-regressive decoding is performed.

4 Datasets

We perform our experiments using two different benchmarks. We use the Multi30K (Elliott et al., 2016) dataset in order to compare the effect of the PLuGS model using a resource that has been widely used in the community. We focus on Task 1 for French from (Caglayan et al., 2019), generating a translation in French based on an image and an English caption as input. The training set consists of images from the Flickr30K train and validation splits, along with the corresponding French captions. The validation split consists of test2016 images and captions, and the test split consists of the test2017 images and captions.

For the core results in this paper, we use the Conceptual Captions dataset (Sharma et al., 2018) as our English-annotated generation labels, in order to capture web-scale phenomena related to image captioning. In addition, we use Google Translate as the translation engine (both for the run-time translations needed for the TGT approach and the training-time translations needed for the TTG and PLuGS approaches), targeting French, Italian, German, Spanish, and Hindi as target languages. We use the standard training and validation splits from Conceptual Captions for developing our models. We report the results using a set of 1,000 randomly samples images from the Open Images Dataset (Kuznetsova et al., 2018). We refer to this test set as OID1k when reporting our results.

5 Evaluation

In the experiments done using the Multi30K dataset, we are reporting results using the METEOR (Banerjee and Lavie, 2005) metric, in line with previous work. For the experiments performed using the Conceptual Captions dataset, we have found that automated evaluation metrics for image captioning such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016) cannot accurately measure captioning accuracy for non-English languages. However, we are reporting CIDEr numbers as a point of comparison, and contrast these numbers with human evaluation results. We describe the human evaluation framework we use next.

5.1 Human Side-by-Side Evaluation

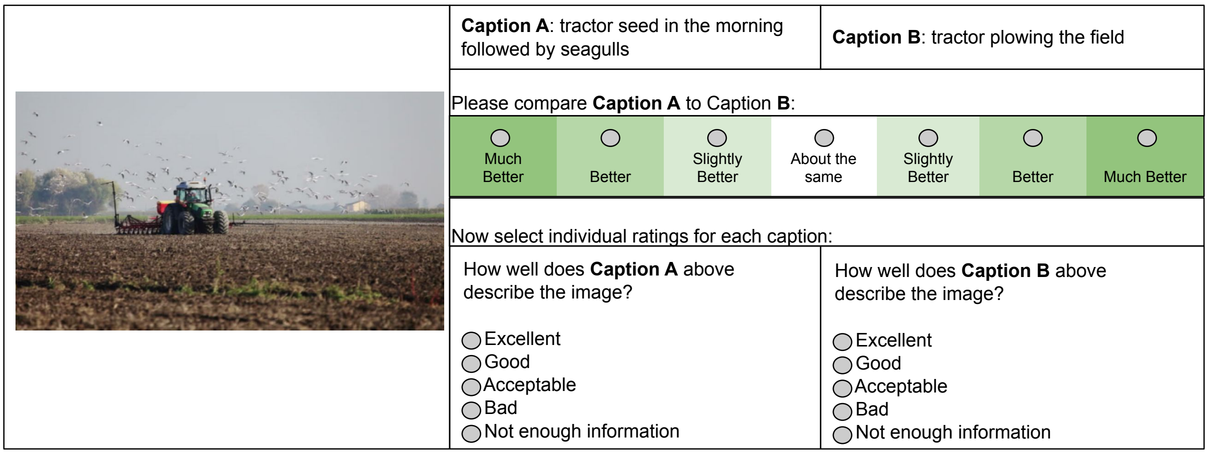
We perform side-by-side human evaluation for comparing model outputs. To compare two image captioning models A (baseline) vs B, we generate captions for these images with each model and ask human raters to compare them. As illustrated in Fig. 4, the raters are shown the image with the two captions randomly placed to the left vs. right, and are asked to compare the captions on a side-by-side rating scale. In addition, they are asked to also provide an absolute rating for each caption. The absolute rating provides a cross-check on the comparison. Each image and associated captions are rated by three raters in our experiments.

We calculate the following statistics using the resulting side-by-side rating comparisons:
$Wins$: Percent of images where majority of raters (i.e. 2 out of 3) marked Caption B as better (after derandomization).
$Losses$: Percent of images where majority of raters marked Caption A as better.
$Gain_{sxs} = Wins - Losses$

We also calculate the following statistics using the resulting absolute ratings:
A_{Accept} = Percent of images where majority of raters mark caption A as Acceptable, Good, or Excellent.
B_{Accept} = Percent of images where majority of raters mark caption B as Acceptable, Good, or Excellent.
$Gain_{Accept} = B_{Accept} - A_{Accept}$

The advantages of the $Gain_{sxs}$ and $Gain_{Accept}$ metrics is that they are intuitive, i.e., they measure the absolute increase in accuracy between the two experimental conditions[3]

[3] Inter-rater agreement analysis shows that for each evaluation comparing two models, two of the three raters agree on $Win/Loss/Same$ for 90% to 95% of the items. Further, for more than 98% of the items using the difference between the absolute ratings gives the same $Win/Loss/Same$ values as obtained from the side-by-side ratings. Also, for 80% to 85% of the absolute ratings, two of the three raters agree on the rating.



Caption A: tractor seed in the morning followed by seagulls

Caption B: tractor plowing the field

Please compare **Caption A** to Caption **B**:

○	○	○	○	○	○	○
Much Better	Better	Slightly Better	About the same	Slightly Better	Better	Much Better

Now select individual ratings for each caption:

How well does **Caption A** above describe the image?

○Excellent
○Good
○Acceptable
○Bad
○Not enough information

How well does **Caption B** above describe the image?

○Excellent
○Good
○Acceptable
○Bad
○Not enough information

Figure 4: Side-by-side human evaluation of two image captions. The same template is used for evaluating English as well as the 5 languages targeted.

5.2 Training Details

Multi30K: For the experiments using this dataset, we use a Transformer Network (Vaswani et al., 2017) with 3 encoder and 3 decoder layers, 8 heads, and model dimension 512. We use the Adam optimizer (Kingma and Ba, 2015), and do a hyperparameter search over learning rates $\{3e^{-4}, e^{-4}, 3e^{-5}, e^{-5}\}$ with linear warmup over 16000 steps followed by exponential decay over $\{50k, 100k\}$ steps. We use $5e^{-6}$ as the weight for L_2 regularization. We train with a batch size of 1024, using a dropout of 0.3, on 8 TPU (You et al., 2019) cores.

Conceptual Captions: For all except large multilingual models, we use a vanilla Transformer with 6 encoder and decoder layers, 8 heads, and model dimension 512. We use the SGD optimizer, and do a hyperparameter search over learning rates $\{0.12, 0.15, 0.18, 0.21, 0.24\}$ with linear warmup over 16000 steps followed by exponential decay over $\{350k, 450k\}$ steps. For multilingual models, we also use linear warmup over 80000 steps. We use $1e^{-5}$ as the weight for L_2 regularization. We train with a batch size of 4096, using a dropout of 0.3 on 32 TPU (You et al., 2019) cores.

For large multilingual models, we use a Transformer with 10 encoder and decoder layers, 12 heads, and model dimension 768[4] We also use a smaller learning rate of 0.09.

6 Experiments and Results

6.1 Multi30K

In order to compare our work to related work we train our models on the Multi30K dataset and compared our results to the results in (Caglayan et al., 2019). We focus on Task 1: generate a French translation based on an image and English caption as input. Table 1 shows the results on the Multi30K dataset for Multimodal Translation. Note that since (Caglayan et al., 2019) does not show numbers for the pure (no caption input) image captioning task, we show numbers for the D_4 condition, where only the first 4 tokens of the English caption are provided as input to the image captioning model.

We see that the PLuGS model is able to produce numbers for MT and MMT that are close to the baseline, even thought it is just an image captioning model augmented to handle these tasks. For the D_4 task, which is the closest to image captioning, the PLuGS model shows improvement over the baseline. Furthermore, the results contain preliminary indications that the PLuGS approach produces better results compared to the non-PLuGS approach

Task	Baseline	non-PLuGS	PLuGS
MT	70.6	66.6	67.7
MMT	70.9	64.7	65.6
IC-D_4	32.3	30.6	32.8

Table 1: Multi30K test set METEOR scores for Translation (MT), Multi Modal Translation (MMT), and Image Captioning (IC-D_4). The baseline is from task 1 of (Caglayan et al., 2019).

[4] Dimension chosen so that we maintain 64 dimensions per head.

Lang	$Wins$	$Losses$	$Gain_{sxs}$	PLuGS$_{Accept}$	TGT$_{Accept}$	$Gain_{Accept}$
Fr	22.8	19.4	3.4	68.7	66.5	2.2
It	22.5	18.3	4.2	52.1	49.9	2.2
De	22.6	19.1	3.5	69.2	67.7	1.5
Es	27.0	22.1	4.9	58.8	56.9	1.9
Hi	26.8	23.8	3.0	78.6	75.9	2.7
	$Wins$	$Losses$	$Gain_{sxs}$	PLuGS$_{Accept}$	TTG$_{Accept}$	$Gain_{Accept}$
Fr	18.2	17.3	0.9	66.2	64.2	2.0
It	23.7	20.8	2.9	55.1	52.2	2.9
De	21.9	19.6	2.3	64.3	63.0	1.3
Es	24.9	23.8	1.1	57.7	56.8	0.9
Hi	27.4	25.5	1.9	71.3	69.6	1.7

Table 2: SxS performance of PLuGS vs. TGT models (upper half) and PLuGS vs. TTG models (lower half), across five target languages on OID1k. The PLuGS models perform better on both $Gain_{SxS}$ and $Gain_{Accept}$ metrics, for all five languages.

Lang	TGT	TTG	PLuGS	PLuGS-TGT	PLuGS-TTG
Fr	0.7890	0.7932	0.7820	-0.0070	-0.0112
It	0.7729	0.7760	0.7813	0.0084	0.0053
De	0.6220	0.6079	0.6170	0.0050	0.0091
Es	0.8042	0.7907	0.7854	-0.0188	-0.0053
Hi	0.7026	0.7149	0.7155	0.0129	0.0006

Table 3: CIDEr scores on CC-1.1 validation set for PLuGS, TGT, and TTG models for five languages.

(+2.2 METEOR).

6.2 Conceptual Captions

In this section, we evaluate the performance of models trained using Conceptual Captions, as detailed in Sec. 4. Table 2 presents the results on the OID1k testset for the SxS human evaluations between the TGT and PLuGS models (upper half), and between the TTG and PLuGS models (lower half). The results show that, for all five languages, the PLuGS model captions are consistently superior to the TGT captions on both $Gain_{SxS}$ and $Gain_{Accept}$ metrics. The $Gain_{SxS}$ are between 3% and 5% absolute percentages between TGT and PLuGS models, and 1% and 3% absolute percentages between TTG and PLuGS models, with similar trends for the $Gain_{Accept}$ metric.

Table 3 presents the CIDEr scores on the validation set of the Conceptual Captions v1.1 (CC-1.1). The CIDEr metric fails to capture any meaningful correlation between its scores and the results of the SxS human evaluations.

6.3 Multilingual Models

We further explore the hypothesis that adding more languages inside one single model may perform even better, as a result of both translation noise canceling out and the languages reinforcing each other in a common representation space. In this vein, we rename the bilingual version as PLuGS-2L, and train several additional models: a TTG-5L model, which uses a LangId token as input and uses for training all translated captions for all five languages and English; a TTG$_{\text{large}}$-5L model, for which we simply increased the capacity of the Transformer network (see Sec. 5.2); and a PLuGS-5L model, which is trained using groundtruth labels that are concatenations (using the LangId token as separator) between golden groundtruth En labels and their translated versions, for all five target languages.

Results using CIDEr are shown in Table 4. Across all languages, the TTG-5L models show a large gap in the CIDEr scores as compared to the TTG monolingual models. Using more capacity in the TTG$_{\text{large}}$-5L model closes the gap only slightly. However, the effect of using pivot-language stabilizers tends to be consistently larger, in terms of CIDEr improvements, than the ones obtained by increasing the model capacity.

To accurately evaluate the impact of multilinguality, we also perform SxS evaluations between the PLuGS-2L (as the base condition) vs.

Lang	TTG	PLuGS-2L	TTG-5L	TTG$_{\text{large}}$-5L	PLuGS-5L
Fr	0.7932	0.7820	0.6834	0.7064	0.7264
It	0.7760	0.7813	0.6538	0.6885	0.6978
De	0.6079	0.6170	0.4992	0.5367	0.5503
Es	0.7907	0.7854	0.7093	0.7203	0.7284
Hi	0.7149	0.7155	0.5891	0.6201	0.6641

Table 4: CIDEr scores on CC-1.1 validation set for bilingual and multilingual models.

Lang	$Wins$	$Losses$	$Gain_{sxs}$	B_{Accept}	A_{Accept}	$Gain_{Accept}$
Fr	21.3	18.3	3.0	69.8	68.7	1.1
It	22.2	18.2	4.0	56.4	55.5	0.9
Hi	26.8	27.0	-0.2	75.6	79.5	-3.9

Table 5: SxS performance of PLuGS-5L vs. PLuGS-2L models for three languages.

PLuGS-5L (as the test condition) models, over three languages (French, German, and Hindi). As shown in Table 5, the PLuGS-5L model performs better on French and Italian (3% and 4% better on $Gain_{sxs}$), while performing worse on Hindi compared to the bilingual PLuGS Hindi model (-0.2% on $Gain_{sxs}$, -3.9% on $Gain_{Accept}$). The results are encouraging, and indeed support the hypothesis that similar languages are reinforcing each other in the common representation space, explaining the gain observed for the Romance languages and the detrimental impact on Hindi.

We also note here that the human evaluation results, except for Hindi, come in direct contradiction to the CIDEr metric results, which indicate a large performance hit for PLuGS-5L vs. PLuGS-2L, across all languages. This reflects again the extreme care needed when judging the outcome of such experiments based on the existing automatic metrics.

6.4 Stabilizers Used as English Captions

As already mentioned, the PLuGS models generate outputs of the form Stabilizer + ⟨LangId⟩ + Caption. We therefore ask the following question: how does the quality of the Stabilizer output compare to the quality of captions produced by the baseline English model (that is, the same model whose captions are translated to the target languages in the TGT approach)?

We perform SxS human evaluations over Stabilizer captions (English) for three different PLuGS-2L models (trained for French, German, and Spanish). As shown in Table 6, the somewhat unexpected answer is that these Stabilizer outputs are consistently better, as English captions, compared to the ones produced by the original monolingual English captioning model. The $Gain_{sxs}$ are between 5% and 6% absolute percentage improvements, while $Gain_{Accept}$ also improves up to 3.4% absolute for the PLuGS-Fr model.

We again note that the CIDEr metric is not able to correctly capture this trend, as shown by the results in Table 7, which indicate a flat/reverse trend.

6.5 Caption is Translation of Stabilizer

So far, we have verified that both the target-language Caption and the Stabilizer English outputs for the PLuGS-2L models are better compared to the alternative ways of producing them. Additionally, we want to check whether the Stabilizer and the target-language Caption are actually translations of each other, and not just independently good captions associated with the input image. In Table 9, we show the BLEU-4 score of the translation of the Stabilizer output for the PLuGS-2L models, compared to the corresponding PLuGS-2L Caption treated as a reference, using the images in the OID1k test set. The high BLEU scores are indeed confirming that the Caption outputs are close translations of the Stabilizer English outputs. This allows us to conclude that PLuGS models are indeed performing the double-duty of captioning and translation.

6.6 Stabilizers Used for Quality Estimation

Finally, we perform an experiment to understand the extent to which the quality of the Stabilizer outputs is correlated with the quality of the target-language Captions, so that a QE model (Levinboim et al., 2019) trained for English can be applied directly on PLuGS model outputs (more specifically,

Model	$Wins$	$Losses$	$Gain_{sxs}$	B_{Accept}	A_{Accept}	$Gain_{Accept}$
PLuGS-Fr	26.9	21.8	5.1	70.4	67.0	3.4
PLuGS-De	26.6	21.3	5.3	70.4	69.7	0.7
PLuGS-Es	28.0	21.8	6.2	69.7	67.8	1.9

Table 6: Performance of Stabilizers used as captions from PLuGS models for three languages vs the captions produced by the baseline English model. The PLuGS Stabilizer outputs are better captions across all three languages.

Model	PLuGS	Baseline	Diff
PLuGS-Fr	0.8663	0.8772	-0.0139
PLuGS-De	0.8680	0.8772	-0.0092
PLuGS-Es	0.8590	0.8772	-0.0182

Table 7: CIDEr scores on CC-1.1 validation set for Baseline and PLuGS-Stabilizer outputs (English captions).

Model	Spearman ρ		
	TGT	TTG	PLuGS
PLuGS-Fr	0.3017	0.3318	**0.5982**
PLuGS-De	0.3246	0.2900	**0.5862**
PLuGS-Es	0.2928	0.3201	**0.5566**

Table 8: Spearman correlation of Stabilizer vs TGT, TTG and PLuGS Captions across three languages.

on the Stabilizer outputs). To that end, we perform human evaluations of stand-alone captions.

In this type of evaluation, the raters are shown an image along with a single caption, and are asked to provide an absolute rating for the caption on a 4-point scale. As before, we define the metric *Accept* = Percent of images where majority of raters (2 of 3) marked Caption as Acceptable, Good or Excellent. Since these ratings are obtained individually for captions, we can use them to measure cross-lingual quality correlations.

6.6.1 Quality Correlation between Stabilizer and Caption

We use the stand-alone caption evaluation results to compute quality correlations. Table 8 shows the correlation between the median human rating for the Stabilizer (English caption) vs Caption (target-language caption) for the PLuGS models considered. We see that the correlation is much higher compared to the baselines, calculated by computing the correlation of the median rating for the Stabilizer vs Caption (target-language) generated by the TGT and TTG approaches.

These results confirm that the PLuGS approach appears to be best suited for leveraging an existing En QE model, due to the availability of the generated Stabilizer output that tends to maintain consistency between the English and the target-language caption, with respect to content accuracy.

	Fr	It	De	Es	Hi
BLEU	93.3	92.9	88.2	93.9	88.2

Table 9: The BLEU-4 score of the translation of the stabilizer against the caption treated as the reference.

7 Conclusions

We present a cross-modal language generation approach called PLuGS, which successfully combines the availability of an existing gold annotation (usually in English) with the availability of translation engines that automatically produce silver-data annotations. The result is a multilingual engine capable of generating high-quality outputs in the target languages, with no gold annotations needed for these languages.

We show that, for image captioning, the PLuGS approach out-performs other alternatives, while also providing the ability to pack multiple languages in a single model for increased performance. Surprisingly, by considering the generated outputs in the original language of the annotation (Stabilizer outputs), we find that the quality of the Stabilizers is higher compared to the outputs of a model trained on the original annotated data.

Overall, our results can be understood as a successful instance of transfer learning from a unimodal task (text-to-text translation) to a cross-modal task (image-to-text generation), which allows us to indirectly leverage the abundance of text-only parallel data annotations across many languages to improve the quality of an annotation-poor cross-modal setup.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *ECCV*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Ozan Caglayan, Pranava 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4159–4170.

Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In *EMNLP-IJCNLP*.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2014. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 215–233.

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, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*.

Alan Jaffe. 2017. Generating image descriptions using multilingual data. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 458–464.

Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-rise: Graph-regularized image semantic embedding. *CoRR*, abs/1902.10814.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

T. Levinboim, A. Thapliyal, P. Sharma, and R. Soricut. 2019. Quality estimation for image captions based on large-scale human evaluations. *arXiv preprint arXiv:1909.03396*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proceedings of ECCV*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the ACL*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of CVPR*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML*.

Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, and Kurt Keutzer. 2019. Fast deep neural network training on distributed systems and cloud tpus. *IEEE Trans. Parallel Distrib. Syst.*, 30(11):2449–2462.

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

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

Sanqiang Zhao, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2019. Informative image captioning with external sources of information. In *ACL*.