Unpaired Image Captioning by Language Pivoting

Jiuxiang Gu[1], Shafiq Joty[2], Jianfei Cai[2], Gang Wang[3]

[1] ROSE Lab, Nanyang Technological University, Singapore
[2] SCSE, Nanyang Technological University, Singapore
[3] Alibaba AI Labs, Hangzhou, China
{jgu004, srjoty, asjfcai}@ntu.edu.sg, gangwang6@gmail.com

Abstract. Image captioning is a multimodal task involving computer vision and natural language processing, where the goal is to learn a mapping from the image to its natural language description. In general, the mapping function is learned from a training set of image-caption pairs. However, for some language, large scale image-caption paired corpus might not be available. We present an approach to this unpaired image captioning problem by language pivoting. Our method can effectively capture the characteristics of an image captioner from the pivot language (Chinese) and align it to the target language (English) using another pivot-target (Chinese-English) sentence parallel corpus. We evaluate our method on two image-to-English benchmark datasets: MSCOCO and Flickr30K. Quantitative comparisons against several baseline approaches demonstrate the effectiveness of our method.

Keywords: Image Captioning · Unpaired Learning

1 Introduction

Recent several years have witnessed unprecedented advancements in automatic image caption generation. This progress can be attributed *(i)* to the invention of novel deep learning framework that learns to generate natural language descriptions of images in an end-to-end fashion, and *(ii)* to the availability of large annotated corpora of images paired with captions such as MSCOCO [30] to train these models. The dominant methods are based on an encoder-decoder framework, which uses a deep convolutional neural network (CNN) to encode the image into a feature vector, and then use a recurrent neural network (RNN) to generate the caption from the encoded vector [29, 27, 44]. More recently, approaches of using attention mechanisms and reinforcement learning have dominated the MSCOCO captioning leaderboard [1, 39, 18].

Despite the impressive results achieved by the deep learning framework, one performance bottleneck is the availability of large paired datasets because neural image captioning models are generally *annotation-hungry* requiring a large amount of annotated image-caption pairs to achieve effective results [19]. However, in many applications and languages, such large-scale annotations are not readily available, and are expensive and slow to acquire. In these scenarios, unsupervised methods that can generate captions from unpaired data or semi-supervised methods that can exploit paired annotations from other domains or languages are highly desirable [5]. In this paper, we pursue the later research avenue, where we assume that we have access to image-caption paired instances in one language (Chinese), and our goal is to transfer this knowledge to a target language (English) for which we do not have such image-caption paired datasets. We also assume that we have access to a separate source-target (Chinese-English) parallel corpus to help us with the transformation. In other words, we wish to use the source language (Chinese) as a pivot language to bridge the gap between an input image and a caption in the target language (English).

The concept of using a pivot language as an intermediary language has been studied previously in machine translation (MT) to translate between a resource-rich language and a resource-scarce language [46, 42, 25, 6]. The translation task in this strategy is performed in two steps. A source-to-pivot MT system first translates a source sentence into the pivot language, which is in turn translated to the target language using a pivot-to-target MT system. Although related, image captioning with the help of a pivot language is fundamentally different from MT, since it involves putting together two different tasks – captioning and translation. In addition, the pivot-based *pipelined* approach to MT suffers from two major problems when it comes to image captioning. First, the conventional pivot-based MT methods assume that the datasets for source-to-pivot and pivot-to-target translations come from the same (or similar) domain(s) with similar styles and word distributions. However, as it comes to image captioning, captions in the pivot language (Chinese) and sentences in the (Chinese-English) parallel corpus are quite different in styles and word distributions. For instance, MSCOCO captioning dataset mostly consists of images of a large scene with object instances (nouns), whereas language parallel corpora are more generic. Second, the errors made in the source-to-pivot translation get propagated to the pivot-to-target translation module in the pipelined approach.

In this paper, we present an approach that can effectively capture the characteristics of an image captioner from the source language and align it to the target language using another source-target parallel corpus. More specifically, our pivot-based image captioning framework comprises an image captioner *image-to-pivot*, an encoder-decoder model that learns to describe images in the pivot language, and a *pivot-to-target* translation model, another encoder-decoder model that translates the sentence in pivot language to the target language, and these two models are trained on two separate datasets. We tackle the variations in writing styles and word distributions in the two datasets by adapting the language translation model to the captioning task. This is achieved by adapting both the encoder and the decoder of the pivot-to-target translation model. In particular, we regularize the word embeddings of the encoder (of pivot language) and the decoder (of target language) models to make them similar to image captions. We also introduce a joint training algorithm to connect the two models and enable them to interact with each other during training. We use AIC-ICC [47] and AIC-MT [47] as the training datasets and two datasets (MSCOCO and Flickr30K [37]) as the validation datasets. The results show that our approach yields substantial gains over the baseline methods on the validation datasets.

2 Background

Image Caption Generation. Image caption generation is a fundamental problem of automatically generating natural language descriptions of images. Motivated by the recent advances in deep neural networks [20, 13] and the release of large scale datasets [30, 37, 47], many studies [45, 15, 48, 22, 24, 52, 18, 17, 50] have used neural networks to generate image descriptions. Inspired by the success of encoder-decoder framework for neural machine translation (NMT) [9, 1], many researchers have proposed to use such a framework for image caption generation [44, 18]. One representative work in this direction is the method proposed by Vinyals *et al* [44]. They encode the image with a CNN and use a Long Short-Term Memory (LSTM) network as the decoder, and the decoder is trained to maximize the log-likelihood estimation of the target captions. After that, many approaches have been proposed to improve such encoder-decoder framework. One of the most commonly used approaches is the attention mechanism [49, 18]. Xu *et al* [49] use the attention mechanisms to incorporate the spatial attention on convolutional features of an image into decoder. Another improvement is to leverage the high-level visual attributes to enhance the sentence decoder [52, 51, 31]. Recently, Gu *et al* [19] propose a CNN-based image captioning model, which can explore both long-term and temporal information in word sequences for caption generation.

Exposure bias and loss-evaluation mismatch have been the major problems in sequence prediction tasks [38]. Exposure bias happens when a model is trained to predict a word given the previous ground-truth words but uses its own generated words during inference. The schedule sampling approach proposed in [2] can mitigate the exposure bias by selecting between the ground-truth words and the machine generated words according to the scheduled probability in training. Recently, the loss-evaluation mismatch problem has been well-addressed in sequence prediction tasks [38, 39, 32, 18]. Rennie *et al* [39] address both exposure bias and loss-evaluation problems with a self-critical learning, which utilizes the inference mode as the baseline in training. Gu *et al* [18] propose a coarse-to-fine learning approach which simultaneously solves the multi-stage training problem as well as the exposure bias issue. The most closely related to our approach is [22]. However, they construct a multilingual parallel dataset based on MSCOCO image corpus, while in our paper, we do not have such a multilingual corpus.

Neural Machine Translation. Neural machine translation is an approach that directly models the conditional probability of translating a sentence in source language into a sentence in target language. A natural choice to model such a decomposition is to use RNN-based models [26, 40, 1, 34, 23]. Recently, researchers have tried to improve the translation performance by introducing the attention mechanism [26, 41, 35, 10]. The attention-based translation model proposed by Kalchbrenner *et al* [26] is an early attempt to train the end-to-end NMT model. Luong *et al* [33] extend the basic encoder-decoder framework to multiple encoders and decoders. However, large-scale parallel corpora are usually not easy to obtain for some language pairs. This is unfortunate because NMT usually needs a large amount of data to train. As a result, improving NMT on resource-scarce language pairs has attracted much attention [55, 16].

Recently, many works have been done in the area of pivot strategies of NMT [11, 46, 42, 3, 53, 14, 25]. Pivot-based approach introduces a third language, named pivot language for which there exist source-pivot and pivot-target parallel corpora. The translation of pivot-based approaches can be divided into two steps: the sentence in the source language is first translated into a sentence in the pivot language, which is then translated to a sentence in the target language. However, such pivot-based approach has a major problem that the errors made in the source-to-pivot model will be forwarded to the pivot-to-target model. Recently, Cheng *et al* [7] introduce an autoencoder to reconstruct monolingual corpora. They further improve it in [8], in which they propose a joint training approach for pivot-based NMT.

3 Unpaired Image Captioning

Let $D_{i,x} = \{(i,x)^{(n_i)}\}_{n_i=0}^{N_i-1}$ denote the dataset with N_i image-caption pairs, and $D_{x,y} = \{(x,y)^{(n_x)}\}_{n_x=0}^{N_x-1}$ denote the translation dataset with N_x source-target sentence pairs. For notational simplicity, we use i to denote an image instance as well as the image modality. Similarly, we use x to represent a source sentence as well as a source/pivot language (Chinese), and y to represent a target sentence and a target language (English). Our ultimate goal is to learn a mapping function to describe an image i with a caption y. Formally,

$$y \sim \arg\max_{y} \left\{ P(y|i;\theta_{i\to y}) \right\} \tag{1}$$

where $\theta_{i\to y}$ are the model parameters to be learned in the absence of any paired data, $i^{(n_i)} \not\leftrightarrow y^{(n_y)}$. We use the pivot language x to learn the mapping: $i \xrightarrow{\theta_{i\to x}} x \xrightarrow{\theta_{x\to y}} y$. Note that image-to-pivot ($D_{i,x}$) and pivot-to-target ($D_{x,y}$) in our setting are two distinct datasets with possibly no common elements.

Fig. 1 illustrates our pivot-based image captioning approach. We have an image captioning model $P(x|i;\theta_{i\to x})$ to generate a caption in the pivot language from an image and a NMT model $P(y|x;\theta_{x\to y})$ to translate this caption into the target language. In addition, we have an autoencoder in the target language $P(\hat{y}|\hat{y};\theta_{\hat{y}\to\hat{y}})$ that guides the target language decoder to produce caption-like sentences. We train these components jointly so that they interact with each other. During inference, given an unseen image i to be described, we use the joint decoder:

$$y \sim \arg\max_{y} \left\{ P(y|i;\theta_{i\to x},\theta_{x\to y}) \right\} \tag{2}$$

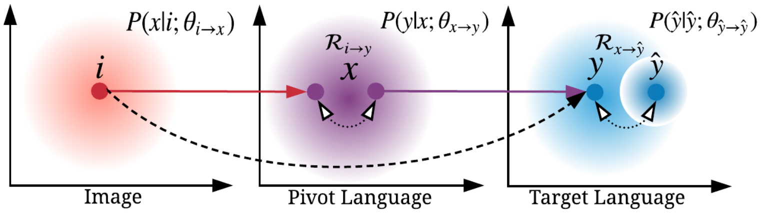


Fig. 1. Pictorial depiction of our pivot-based unpaired image captioning setting. Here, i, x, y, and $\hat{y}$ denote source image, pivot language sentence, target language sentence, and ground truth captions in target language, respectively. We use a dashed line to denote that there is no parallel corpus available for the pair. Solid lines with arrows represent decoding directions. Dashed lines inside a language (circle) denote stylistic and distributional differences between caption and translation data.

In the following, we first give an overview of neural methods for image captioning and machine translation using paired (parallel) data. Then, we present our approach that extends these standard models for unpaired image captioning with a pivot language.

3.1 Encoder-Decoder Models for Image Captioning and Machine Translation

Standard Image Captioning. For image captioning in the paired setting, the goal is to generate a caption $\tilde{x}$ from an image i such that $\tilde{x}$ is as similar to the ground truth caption x. We use $P_x(x|i;\theta_{i\to x})$ to denote a standard encoder-decoder based image captioning model with $\theta_{\mathrm{i\to x}}$ being the parameters. We first encode the given image to the image features v with a CNN-based image encoder: $v = \mathrm{CNN}(i)$. Then, we predict the image description x from the global image feature v. The training objective is to maximize the probability of the ground truth caption words given the image:

$$\tilde{\theta}_{i\to x} = \arg\max_{\theta_{i\to x}} \{\mathcal{L}_{i\to x}\} \tag{3}$$

$$= \arg\max_{\theta_{i\to x}} \{ \sum_{n_i=0}^{N_i-1} \sum_{t=0}^{M^{(n_i)}-1} \log P_x(x_t^{(n_i)}|x_{0:t-1}^{(n_i)}, i^{(n_i)};\theta_{i\to x})\} \tag{4}$$

where N_i is the number of image-caption pairs, $M^{(n_i)}$ is the length of the caption $x^{(n_i)}$, x_t denotes a word in the caption, and $P_x(x_t^{(n_i)}|x_{0:t-1}^{(n_i)}, i^{(n_i)})$ corresponds to the activation of the Softmax layer. The decoded word is drawn from:

$$x_t \sim \arg\max_{\mathcal{V}_{i\to x}^{x}} P(x_t|x_{0:t-1};i) \tag{5}$$

where $\mathcal{V}_{i\to x}^{x}$ is the vocabulary of words in the image-caption dataset $D_{i,x}$.

Neural Machine Translation. Given a pair of source and target sentences (x, y), the NMT model $P_y(y|x; \theta_{x\to y})$ computes the conditional probability:

$$P_y(y|x) = \prod_{t=0}^{N-1} P(y_t|y_{0:t-1}; x_{0:M-1}) \tag{6}$$

where M and N are the lengths of the source and target sentences, respectively. The maximum-likelihood training objective of the model can be expressed as:

$$\begin{aligned} \tilde{\theta}_{x\to y} &= \arg\max_{\theta_{x\to y}} \big\{\mathcal{L}_{x\to y}\big\} && (7) \\ &= \arg\max_{\theta_{x\to y}} \Big\{ \sum_{n_x=0}^{N_x-1} \sum_{t=0}^{N^{(n_x)}-1} \log P_y(y_t^{(n_x)}|y_{0:t-1}^{(n_x)}; x^{(n_x)}; \theta_{x\to y})\Big\} && (8) \end{aligned}$$

During inference we calculate the probability of the next symbol given the source sentence encoding and the decoded target sequence so far, and draw the word from the dictionary according to the maximum probability:

$$y_t \sim \arg\max_{\mathcal{V}^y_{x\to y}} P(y_t|y_{0:t-1}; x_{0:M-1}) \tag{9}$$

where $\mathcal{V}^y_{x\to y}$ is the vocabulary of the target language in the translation dataset $D_{x,y}$.

Unpaired Image Captioning by Language Pivoting. In the unpaired setting, our goal is to generate a description y in the target language for an image i without any pair information. We assume, there is a second language x called "pivot" for which we have (separate) image-pivot and pivot-target paired datasets. The image-to-target model in the pivot-based setting can be decomposed into two sub-models by treating the pivot sentence as a latent variable:

$$P(y|i; \theta_{i\to x}, \theta_{x\to y}) = \sum_x P_x(x|i; \theta_{i\to x}) P_y(y|x; \theta_{x\to y}) \tag{10}$$

where $P_x(x|i; \theta_{i\to x})$ and $P_y(y|x; \theta_{x\to y})$ are the image captioning and NMT models, respectively. Due to the exponential search space in the pivot language, we approximate the captioning process with two steps. The first step translates the image i into a pivot language sentence $\tilde{x}$. Then, the pivot language sentence is translated to a target language sentence $\tilde{y}$. To learn such a pivot-based model, a simple approach is to combine the two loss functions in Equations (4) and (8) as follows:

$$\mathcal{J}_{i\to x, x\to y} = \mathcal{L}_{i\to x} + \mathcal{L}_{x\to y} \tag{11}$$

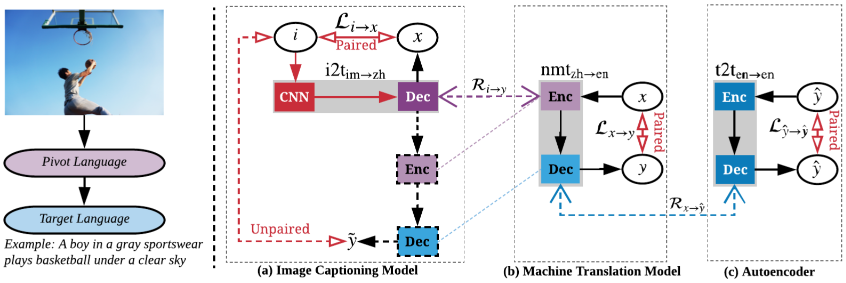
During inference, the decoding decision is given by:

$$\begin{aligned} \tilde{x} &= \arg\max_x \big\{P_x(x|i; \tilde{\theta}_{i\to x})\big\} && (12) \\ \tilde{y} &= \arg\max_y \big\{P_y(y|\tilde{x}; \tilde{\theta}_{x\to y})\big\} && (13) \end{aligned}$$

where $\tilde{x}$ is the image description generated from i in the pivot language, $\tilde{y}$ is the translation of $\tilde{x}$, and $\tilde{\theta}_{i\to x}$ and $\tilde{\theta}_{x\to y}$ are the learned model parameters.

However, this *pipelined* approach to image caption generation in the target language suffers from couple of key limitations. First, image captioning and machine translation are two different tasks. The image-to-pivot and pivot-to-target models are quite different in terms of vocabulary and parameter space because they are trained on two possibly unrelated datasets. Image captions contain description of objects in a given scene, whereas machine translation data is more generic, in our case containing news event descriptions, movie subtitles, and conversational texts. They are two different domains with differences in writing styles and word distributions. As a result, the captions generated by the pipeline approach may not be similar to human-authored captions. Fig. 1 distinguishes between the two domains of pivot and target sentences: caption domain and translation domain (see second and third circles). The second limitation is that the errors made by the image-to-pivot captioning model get propagated to the pivot-to-target translation model.

To overcome the limitations of the pivot-based caption generation, we propose to reduce the discrepancy between the image-to-pivot and pivot-to-target models, and to train them jointly so that they learn better models by interacting with each other during training. Fig. 2 illustrates our approach. The two models share some common aspects that we can exploit to connect them as we describe below.



(a) Image Captioning Model (b) Machine Translation Model (c) Autoencoder

Fig. 2. Illustration of our image captioning model with pivot language. The image captioning model first transforms an image into latent pivot sentences, from which our machine translation model generates the target caption.

Connecting Image-to-Pivot and Pivot-to-Target. One way to connect the two models is to share the corresponding embedding matrices by defining a common embedding matrix for the decoder of image-to-pivot and the encoder of pivot-to-target. However, since the caption and translation domains are different, their word embeddings should also be different. Therefore, rather than having a common embedding matrix, we add a regularizer $\mathcal{R}_{i\to y}$ that attempts to bring the input embeddings of the NMT model close to the output embeddings of image captioning model by minimizing their l_2 distance. Formally,

$$\mathcal{R}_{i\to y}(\theta_{i\to x}^{w_x}, \theta_{x\to y}^{w_x}) = -\sum_{w_x \in \mathcal{V}_{i\to x}^{x} \cap \mathcal{V}_{x\to y}^{x}} ||\theta_{i\to x}^{w_x} - \theta_{x\to y}^{w_x}||_2 \tag{14}$$

where w_x is a word in the pivot language that is shared by the two embedding matrices, and $\theta_{i\to x}^{w_x} \in \mathbb{R}^d$ denotes the vector representation of w_x in the source-to-pivot model, and $\theta_{x\to y}^{w_x} \in \mathbb{R}^d$ denotes the vector representation of w_x in the pivot-to-target model. Note that, here we adapt $\theta_{x\to y}^{w_x}$ towards $\theta_{i\to x}^{w_x}$, that is, $\theta_{i\to x}^{w_x}$ is already a learned model and kept fixed during adaptation.

Adapting the encoder embeddings of the NMT model does not guarantee that the decoder of the model will produce caption-like sentences. For this, we need to also adapt the decoder embeddings of the NMT model to the caption data. We first use the target-target parallel corpus $D_{\hat{y},\hat{y}} = \{(\hat{y}^{(n_{\hat{y}})}, \hat{y}^{(n_{\hat{y}})})\}_{n_{\hat{y}}=0}^{N_{\hat{y}}-1}$ to train an autoencoder $P(\hat{y}|\hat{y};\theta_{\hat{y}\to\hat{y}})$, where $\theta_{\hat{y}\to\hat{y}}$ are the parameters of the autoencoder. The maximum-likelihood training objective of autoencoder can be expressed as:

$$\tilde{\theta}_{\hat{y}\to\hat{y}} = \arg\max_{\theta_{\hat{y}\to\hat{y}}} \{\mathcal{L}_{\hat{y}\to\hat{y}}\} \tag{15}$$

where $\mathcal{L}_{\hat{y}\to\hat{y}}$ is the cross-entropy (XE) loss. The autoencoder then "teaches" the decoder of the translation model $P(y|x;\theta_{x\to y})$ to learn similar word representations. This is again achieved by minimizing the l_2 distance between two vectors:

$$\mathcal{R}_{x\to\hat{y}}(\theta_{x\to y}^{w_y}, \theta_{\hat{y}\to\hat{y}}^{w_y}) = -\sum_{w_y \in \mathcal{V}_{x\to y}^{y} \cap \mathcal{V}_{\hat{y}\to\hat{y}}^{y}} ||\theta_{x\to y}^{w_y} - \theta_{\hat{y}\to\hat{y}}^{w_y}||_2 \tag{16}$$

where $\mathcal{V}_{\hat{y}\to\hat{y}}^{y}$ is the vocabulary of y in $D_{\hat{y},\hat{y}}$, and w_y is a word in the target language that is shared by the two embedding matrices. By optimizing Equation (16), we try to make the learned caption share a similar style as the target captions.

Joint Training. In training, our goal is to find a set of source-to-target model parameters that maximizes the training objective:

$$\mathcal{J}_{i\to x,x\to y,y\to\hat{y}} = \mathcal{L}_{i\to x} + \mathcal{L}_{x\to y} + \mathcal{L}_{\hat{y}\to\hat{y}} + \lambda\mathcal{R}_{i\to x,x\to y,y\to\hat{y}} \tag{17}$$

$$\mathcal{R}_{i\to x,x\to y,y\to\hat{y}} = \mathcal{R}_{i\to y}(\theta_{i\to x}^{w_x}, \theta_{x\to y}^{w_x}) + \mathcal{R}_{x\to\hat{y}}(\theta_{x\to y}^{w_y}, \theta_{\hat{y}\to\hat{y}}^{w_y}) \tag{18}$$

where λ is the hyper-parameter used to balance the preference between the loss terms and the connection terms. Since both the captioner $P_x(x|i;\theta_{i\to x})$ and the translator $P_y(y|x;\theta_{x\to y})$ have large vocabulary sizes (see Table 1), it is hard to train the joint model with an initial random policy. Thus, in practice, we pre-train the captioner, translator and autoencoder first, and then jointly optimize them with Equation (17).

4 Experiments

Datasets. In our experiments, we choose the two independent datasets used from AI Challenger (AIC) [47]: AIC Image Chinese Captioning (AIC-ICC) and

AIC Chinese-English Machine Translation (AIC-MT), as the training datasets, while using MSCOCO and Flickr30K English captioning datasets as the test datasets. Table 1 shows the statistics of the datasets used in our experiments.

Table 1. Statistics of the datasets used in our experiments, where "im" denotes the image, "zh" denotes Chinese, and "en" denotes English.

	Dataset	Lang.	Source		Target	
			# Image/Sent.	Vocab. Size	# Sent.	Vocab. Size
Training	AIC-ICC	im → zh	240K	–	1,200K	4,461
	AIC-MT	zh → en	10,000K	50,004	10,000K	50,004
Testing	MSCOCO	im → en	123K	–	615K	9,487
	Flickr30K	im → en	30K	–	150K	7,000

Training Datasets. For image-to-Chinese captioning training, we follow the settings in AIC-ICC [47] and take 210,000 images for training and 30,000 images for model validation. Each image contains five reference Chinese captions, and each of the captions contains most of the common daily scenes in which a person usually appears. We use the "*Jieba*"[4], a Chinese text segmentation module, for word segmentation. We truncate all the captions longer than 16 tokens and prune the vocabulary by dropping words with a frequency less than 5, resulting in a vocabulary size of 4,461 words.

We take Chinese as the pivot language and learn a Chinese-to-English translation model on AIC-MT. AIC-MT consists of 10,000K Chinese-English parallel sentences, and the English sentences are extracted from English learning websites and movie subtitles. We reserve 4K sentence pairs for validation and 4K sentence pairs for testing. During preprocessing, we remove the empty lines and retain sentence pairs with no more than 50 words. We prune the vocabulary and end up with a vocabulary of size 50,004 words including special Begin-of-Sentence (BOS) and End-of-Sentence (EOS) tokens. To guide the target decoder to generate caption-like sentences, an autoencoder is also trained with the target image descriptions extracted from MSCOCO. In our training, we extract 60,000 image descriptions from the MSCOCO training split, and randomly sort these samples.

Validation Datasets. We validate the effectiveness of our method on MSCOCO and Flickr30K datasets. The images in MSCOCO typically contain multiple objects with significant contextual information. Likewise, each image in MSCOCO also has five reference description, and most of these descriptions are depicting humans participating in various activities. We use the same test splits as that in [27]. For MSCOCO, we use 5,000 images for validation and 5,000 images for testing, and for Flickr30K, we use 1,000 images for testing.

[4] https://github.com/fxsjy/jieba

4.1 Implementation Details

Architecture. As can be seen in Fig. 2, we have three models used in our image captioner. The first model $\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$ learns to generate the Chinese caption x from a given image i. It is a standard CNN-RNN architecture [44], where word outputted from the previous time step is taken as the input for the current time step. For each image, we encoder it with ResNet-101 [21], and then apply average pooling to get a vector of dimensions 2,048. After that, we map the image features through a linear projection and get a vector of dimensions 512. The decoder is implemented based on an LSTM network. The dimensions of the LSTM hidden states and word embedding are fixed to 512 for all of the models discussed in this paper. Each sentence starts with a special BOS token, and ends with an EOS token.

The second model $\mathrm{nmt}_{\mathrm{zh}\to\mathrm{en}}$ learns to translate the Chinese sentence x to the English sentence y. It has three components: a sentence encoder, a sentence decoder, and an attention module. The words in the pivot language are first mapped to word vectors and then fed into a bidirectional LSTM network. The decoder predicts the target language words based on the encoded vector of the source sentence as well as its previous outputs. The encoder and the decoder are connected through an attention module which allows the decoder to focus on different regions of the source sentence during decoding.

The third model $\mathrm{t2t}_{\mathrm{en}\to\mathrm{en}}$ learns to produce the caption-style English sentence $\hat{y}$. It is essentially an autoencoder trained on a set of image descriptions extracted from MSCOCO, where the encoder and the decoder are based on one-layer LSTM network. The encoder reads the whole sentence as input and the decoder is to reconstruct the input sentence.

Training Setting. All the modules are randomly initialized before training except the image CNN, for which we use a pre-trained model on ImageNet. We first independently train the image Chinese captioner, the Chinese-to-English translator, and the autoencoder with the cross-entropy loss on AIC-ICC, AIC-MT, and MSCOCO corpus, respectively. During this stage, we use Adam [28] algorithm to do model updating with a mini-batch size of 100. The initial learning rate is $4e^{-4}$, and the momentum is 0.9. The best models are selected according to the validation scores, which are then used for the subsequent joint training. Specifically, we combine the just trained models with the connection terms, and conduct a joint training with Equation (17). We set the hyper-parameter λ to 1.0, and train the joint model using Adam optimizer with a mini-batch size of 64 and an initial learning rate of $2e^{-4}$. Weight decay and dropout are applied in this training phase to prevent over-fitting.

Testing setting. During testing, the output image description is first formed by drawing words in pivot language from $\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$ until an EOS token is reached, and then translated with $\mathrm{nmt}_{\mathrm{zh}\to\mathrm{en}}$ to the target language. Here we use beam search for the two inference procedures. Beam search is an efficient decoding method for RNN-based models, which keeps the top-k hypotheses at each time step, and considers them as the candidates to generate a new top-k hypotheses at the next time step. We set a fixed beam search size of $k = 5$ for $\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$

and $k = 10$ for $\mathrm{nmt}_{\mathrm{zh}\to\mathrm{en}}$. We evaluate the quality of the generated image descriptions with the standard evaluation metrics: BLEU [36], METEOR [12], and CIDEr [43]. Since BLEU aims to assess how similar two sentences are, we also evaluate the diversity of the generated sentence with Self-BLEU [54], which takes one sentence as the hypothesis and the others as the reference, and then calculates BLEU score for every generated sentence. The final Self-BLEU score is defined as the average BLEU scores of the sentences.

4.2 Quantitative Analysis

Results of Image Chinese Captioning. Table 2 shows the comparison results on the AIC-ICC validation set, where B@n is short for BLEU-n. We compare our $\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$ model with the baseline [47] (named AIC-I2T). Both AIC-I2T and our image caption model ($\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$) are trained with cross-entropy loss. We can see that our model outperforms the baseline in all the metrics. This might be due to different implementation details, *e.g.*, AIC-I2T utilizes Inception-v3 for the image CNN while we use ResNet-101.

Table 2. Performance comparisons on AIC-ICC. The results of $\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$ are achieved via beam search.

Image Chinese Captioning					
Approach	B@1	B@2	B@3	B@4	CIDEr
$\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$	**77.8**	**65.9**	**55.5**	**46.6**	**144.2**
AIC-I2T [47]	76.5	64.8	54.7	46.1	142.5

Table 3. Performance comparisons on AIC-MT test dataset. Note that our $\mathrm{nmt}_{\mathrm{zh}\to\mathrm{en}}$ model uses beam search.

Chinese-to-English Translation		
Approach	Accuracy	Perplexity
$\mathrm{nmt}_{\mathrm{zh}\to\mathrm{en}}$	55.0	8.9
Google Translation	**57.8**	–

Results of Chinese-to-English Translation. Table 3 provides the comparison between our attention-based machine translator with the online Google translator on AIC-MT test split. We use the "googletrans"[5], a free Python tool that provides Google translator API. The perplexity value in the second column is the geometric mean of the inverse probability for each predicted word. Our attention-based NMT model ($\mathrm{nmt}_{\mathrm{zh}\to\mathrm{en}}$) is trained on AIC-MT training set. We can see that our model is slightly worse than online Google translation in accuracy. This not surprising considering that Google's translator is trained on much larger datasets with more vocabulary coverage, and it is a more complex system that ensembles multiple NMT models.

Results of Unpaired Image English Captioning. Table 4 shows the comparisons among different variants of our method on MSCOCO dataset. Our upper bound is achieved by an image captioning model $\mathrm{i2t}_{\mathrm{im}\to\mathrm{en}}$ that is trained with paired English captions. $\mathrm{i2t}_{\mathrm{im}\to\mathrm{en}}$ shares the same architecture as $\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$, except that they have different vocabulary sizes. The lower bound is achieved by pipelinining $\mathrm{i2t}_{\mathrm{im}\to\mathrm{zh}}$ and $\mathrm{nmt}_{\mathrm{zh}\to\mathrm{en}}$. In the pipeline setting, these two models

[5] https://pypi.python.org/pypi/googletrans

are trained on AIC-ICC and AIC-MT, respectively. We also report the results of our implementation of FC-2K [39], which adopts a similar architecture.

Table 4. Results of unpaired image-to-English captioning on MSCOCO 5K and Flickr30K 1K test splits, where M is short for METEOR.

Approach	Lang.	B@1	B@2	B@3	B@4	M	CIDEr
MSCOCO							
i2t$_{\text{im}\rightarrow\text{en}}$ (Upper bound, XE Loss)	en	73.2	56.3	42.0	31.2	25.3	95.1
FC-2K [39] (ResNet101, XE Loss)	en	–	–	–	29.6	25.2	94.0
i2t$_{\text{im}\rightarrow\text{zh}}$ + nmt$_{\text{zh}\rightarrow\text{en}}$ ($\mathcal{R}_{i\rightarrow x,x\rightarrow y,y\rightarrow\hat{y}}$)	en	**46.2**	**24.0**	**11.2**	**5.4**	13.2	**17.7**
i2t$_{\text{im}\rightarrow\text{zh}}$ + nmt$_{\text{zh}\rightarrow\text{en}}$ ($\mathcal{R}_{i\rightarrow x,x\rightarrow y}$)	en	45.5	23.6	11.0	5.3	13.1	17.3
i2t$_{\text{im}\rightarrow\text{zh}}$ + nmt$_{\text{zh}\rightarrow\text{en}}$ (Lower bound)	en	42.0	20.6	9.5	3.9	12.0	12.3
i2t$_{\text{im}\rightarrow\text{zh}}$ + Online Google Translation	en	42.2	21.8	10.7	5.3	**14.5**	17.0
Flickr30K							
i2t$_{\text{im}\rightarrow\text{en}}$ (Upper bound, XE Loss)	en	63.1	43.8	30.2	20.7	17.7	40.1
i2t$_{\text{im}\rightarrow\text{zh}}$ + nmt$_{\text{zh}\rightarrow\text{en}}$ ($\mathcal{R}_{i\rightarrow x,x\rightarrow y,y\rightarrow\hat{y}}$)	en	**49.7**	**27.8**	**14.8**	**7.9**	13.6	**16.2**
i2t$_{\text{im}\rightarrow\text{zh}}$ + nmt$_{\text{zh}\rightarrow\text{en}}$ ($\mathcal{R}_{i\rightarrow x,x\rightarrow y}$)	en	48.7	26.1	12.8	6.4	13.0	14.9
i2t$_{\text{im}\rightarrow\text{zh}}$ + nmt$_{\text{zh}\rightarrow\text{en}}$ (Lower bound)	en	45.9	25.2	13.1	6.9	12.5	13.9
i2t$_{\text{im}\rightarrow\text{zh}}$ + Online Google Translation	en	46.2	25.4	13.9	7.7	**14.4**	15.8

For unpaired image-to-English captioning, our method with the connection term on pivot language ($\mathcal{R}^{w_x}_{i\rightarrow x,x\rightarrow y}$) outperforms the method of combining i2t$_{\text{im}\rightarrow\text{zh}}$ with online Google translation in terms of B@n and CIDEr metrics, while obtaining significant improvements over the lower bound. This demonstrates the effectiveness of the connection term on the pivot language. Moreover, by adding the connection term on the target language, our model with the two connection terms ($\mathcal{R}_{i\rightarrow x,x\rightarrow y,y\rightarrow\hat{y}}$) further improves the performance. This suggests that a small corpus in the target domain is able to make the decoder to generate image descriptions that are more like captions. The connection terms help to bridge the word representations of the two different domains. The captions generated by Google translator have higher METEOR. We speculate the following reasons. First, Google Translator generates longer captions than ours. Since METEOR computes the score not only on the basis of n-gram precision but also of uni-gram recall, its default parameters favor longer translations than other metrics [4]. Second, in addition to exact word matching, METEOR considers matching of word stems and synonyms. Since Google translator is trained on a much larger corpus than ours, it generates more synonymous words. Table 4 also shows the results of unpaired image English captioning on Flickr30K, where we can draw similar conclusions.

We further evaluate the diversity of the generated image descriptions using Self-BLEU metric. Table 5 shows the detailed Self-BLEU scores. It can be seen that our method generates image descriptions with the highest diversity, compared with the upper and lower bounds. For better comparison, we also calculate the Self-BLEU scores calculated on ground truth captions.

Table 5. Self-BLEU scores on MSCOCO 5K test split. Note that lower Self-BLEU scores imply higher diversity of the image descriptions.

Approach	Lang.	Self-B@2	Self-B@3	Self-B@4	Self-B@5
$\text{i2t}_{\text{im}\rightarrow\text{en}}$ (GT Captions)	en	85.0	67.8	49.1	34.4
$\text{i2t}_{\text{im}\rightarrow\text{en}}$ (Upper bound)	en	99.0	97.5	94.6	90.7
$\text{i2t}_{\text{im}\rightarrow\text{zh}} + \text{nmt}_{\text{zh}\rightarrow\text{en}}$ ($\mathcal{R}_{i\rightarrow x, x\rightarrow y, y\rightarrow \hat{y}}$)	en	**95.6**	**91.7**	**86.5**	**80.2**
$\text{i2t}_{\text{im}\rightarrow\text{zh}} + \text{nmt}_{\text{zh}\rightarrow\text{en}}$ (Lower bound)	en	98.1	95.9	92.3	87.6

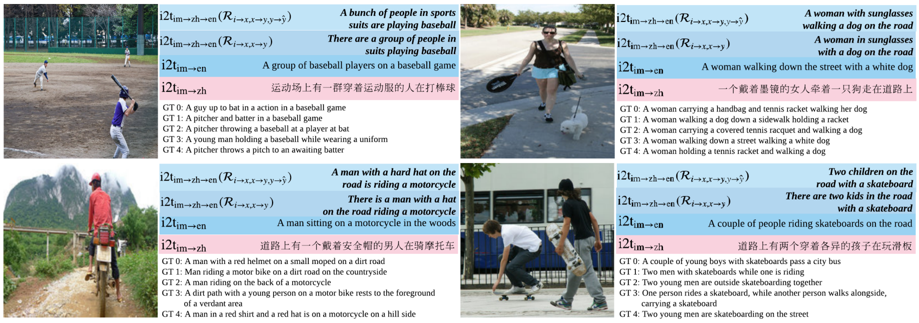
We also conduct a human evaluation of the generated captions for different models as well as the ground truth captions. A total number of 12 evaluators of different educational background were invited, and a total of 1.2K samples were randomly selected from the test split for the user study. Particularly, we measure the caption quality from two aspects: *relevant* and *resemble*. The *relevant* metric indicates whether the caption is correct according to the image content. The *resemble* metric assesses to what extent the systems produce captions that resemble human-authored captions. The evaluators assessed the quality in 5 grades: 1-very poor, 2-poor, 3-barely acceptable, 4-good, 5-very good. Each evaluator assessed randomly chosen 100 images. The results presented in Table 6 demonstrate that our approach can generate relevant and human understandable image captions as the paired (Upper bound) approach.

Table 6. Evaluation results of user assessment on MSCOCO 1.2K test split.

Approach	$\text{i2t}_{\text{im}\rightarrow\text{zh}} + \text{nmt}_{\text{zh}\rightarrow\text{en}}$ ($\mathcal{R}_{i\rightarrow x, x\rightarrow y, y\rightarrow \hat{y}}$)	Upper Bound	Ground-Truth
Relevant	3.81	3.99	4.68
Resemble	3.78	4.05	4.48

4.3 Qualitative Results

We provide some captioning examples in Fig. 3 for a better understanding of our model. We show in different color the generated captions for several images by the three models along with the ground truth (GT) captions. From these exemplary results, we can see that, compared with the paired model $\text{i2t}_{\text{im}\rightarrow\text{en}}$, our pivot-based unpaired model $\text{i2t}_{\text{im}\rightarrow\text{zh}\rightarrow\text{en}}(\mathcal{R}_{i\rightarrow x, x\rightarrow y, y\rightarrow \hat{y}})$ often generates more diverse captions; thanks to the additional translation data. At the same time, our model can generate caption-like sentences by bridging the gap between the datasets and by joint training of the model components. For example, with the detected people in the first image, our model generates the sentence with "*a bunch of people in sports suits*", which is more diverse than the sentence with "*a group of baseball players*" generated by the paired model.



$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}(\mathcal{R}_{i\rightarrow x,x\rightarrow y,y\rightarrow\hat{y}})$ ***A bunch of people in sports suits are playing baseball***

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}(\mathcal{R}_{i\rightarrow x,x\rightarrow y})$ ***There are a group of people in suits playing baseball***

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{en}}$ A group of baseball players on a baseball game

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}}$ 运动场上有一群穿着运动服的人在打棒球

GT 0: A guy up to bat in a action in a baseball game
GT 1: A pitcher and batter in a baseball game
GT 2: A pitcher throwing a baseball at a player at bat
GT 3: A young man holding a baseball while wearing a uniform
GT 4: A pitcher throws a pitch to an awaiting batter

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}(\mathcal{R}_{i\rightarrow x,x\rightarrow y,y\rightarrow\hat{y}})$ ***A woman with sunglasses walking a dog on the road***

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}(\mathcal{R}_{i\rightarrow x,x\rightarrow y})$ ***A woman in sunglasses with a dog on the road***

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{en}}$ A woman walking down the street with a white dog

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}}$ 一个戴着墨镜的女人牵着一只狗走在道路上

GT 0: A woman carrying a handbag and tennis racket walking her dog
GT 1: A woman walking a dog down a sidewalk holding a racket
GT 2: A woman carrying a covered tennis racquet and walking a dog
GT 3: A woman walking down a street walking a white dog
GT 4: A woman holding a tennis racket and walking a dog

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}(\mathcal{R}_{i\rightarrow x,x\rightarrow y,y\rightarrow\hat{y}})$ ***A man with a hard hat on the road is riding a motorcycle***

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}(\mathcal{R}_{i\rightarrow x,x\rightarrow y})$ ***There is a man with a hat on the road riding a motorcycle***

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{en}}$ A man sitting on a motorcycle in the woods

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}}$ 道路上有一个戴着安全帽的男人在骑摩托车

GT 0: A man with a red helmet on a small moped on a dirt road
GT 1: Man riding a motor bike on a dirt road on the countryside
GT 2: A man riding on the back of a motorcycle
GT 3: A dirt path with a young person on a motor bike rests to the foreground of a verdant area
GT 4: A man in a red shirt and a red hat is on a motorcycle on a hill side

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}(\mathcal{R}_{i\rightarrow x,x\rightarrow y,y\rightarrow\hat{y}})$ ***Two children on the road with a skateboard***

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}(\mathcal{R}_{i\rightarrow x,x\rightarrow y})$ ***There are two kids in the road with a skateboard***

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{en}}$ A couple of people riding skateboards on the road

$\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}}$ 道路上有两个穿着各异的孩子在玩滑板

GT 0: A couple of young boys with skateboards pass a city bus
GT 1: Two men with skateboards while one is riding
GT 2: Two young men are outside skateboarding together
GT 3: One person rides a skateboard, while another person walks alongside, carrying a skateboard
GT 4: Two young men are skateboarding on the street

Fig. 3. Examples of the generated sentences on MSCOCO test images, where $\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}}$ is the image captioner trained on AIC-ICC, $\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{en}}$ is the image captioner trained on MSCOCO, $\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}$ $(\mathcal{R}_{i\rightarrow x,x\rightarrow y,y\rightarrow\hat{y}})$ and $\mathrm{i2t}_{\mathrm{im}\rightarrow\mathrm{zh}\rightarrow\mathrm{en}}$ $(\mathcal{R}_{i\rightarrow x,x\rightarrow y})$ are our proposed models for unpaired image captioning, and GT stands for ground truth caption.

5 Conclusion

In this paper, we have proposed an approach to unpaired image captioning with the help of a pivot language. Our method couples an image-to-pivot captioning model with a pivot-to-target NMT model in a joint learning framework. The coupling is done by adapting the word representations in the encoder and the decoder of the NMT model to produce caption-like sentences. Empirical evaluation demonstrates that our method consistently outperforms the baseline methods on MSCOCO and Flickr30K image captioning datasets. In our future work, we plan to explore the idea of 'back-translation' to create pseudo Chinese-English translation data for English captions, and adapt our decoder language model by training on this pseudo dataset.

Acknowledgments

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, and the Infocomm Media Development Authority, Singapore. We gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC) for our research at NTU ROSE Lab, Singapore.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
2. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: NIPS. pp. 1171–1179 (2015)
3. Bertoldi, N., Barbaiani, M., Federico, M., Cattoni, R.: Phrase-based statistical machine translation with pivot languages. In: IWSLT. pp. 143–149 (2008)
4. Cer, D., Manning, C.D., Jurafsky, D.: The best lexical metric for phrase-based statistical mt system optimization. In: NAACL. pp. 555–563 (2010)
5. Chen, T.H., Liao, Y.H., Chuang, C.Y., Hsu, W.T., Fu, J., Sun, M.: Show, adapt and tell: Adversarial training of cross-domain image captioner. In: ICCV. pp. 521–530 (2017)
6. Chen, Y., Liu, Y., Li, V.O.: Zero-resource neural machine translation with multi-agent communication game. In: AAAI. pp. 5086–5093 (2018)
7. Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., Liu, Y.: Semi-supervised learning for neural machine translation. In: ACL. pp. 1965–1974 (2016)
8. Cheng, Y., Yang, Q., Liu, Y., Sun, M., Xu, W.: Joint training for pivot-based neural machine translation. In: IJCAI. pp. 3974–3980 (2017)
9. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. pp. 1724–1734 (2014)
10. Cohn, T., Hoang, C.D.V., Vymolova, E., Yao, K., Dyer, C., Haffari, G.: Incorporating structural alignment biases into an attentional neural translation model. In: ACL. pp. 876–885 (2016)
11. Cohn, T., Lapata, M.: Machine translation by triangulation: Making effective use of multi-parallel corpora. In: ACL. pp. 728–735 (2007)
12. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: ACL. pp. 376–380 (2014)
13. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: CVPR. pp. 2393–2402 (2018)
14. El Kholy, A., Habash, N., Leusch, G., Matusov, E., Sawaf, H.: Language independent connectivity strength features for phrase pivot statistical machine translation. In: ACL. pp. 412–418 (2013)
15. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: CVPR. pp. 1473–1482 (2015)
16. Firat, O., Sankaran, B., Al-Onaizan, Y., Vural, F.T.Y., Cho, K.: Zero-resource translation with multi-lingual neural machine translation. In: EMNLP. pp. 268–277 (2016)
17. Gu, J., Cai, J., Joty, S., Niu, L., Wang, G.: Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: CVPR. pp. 7181–7189 (2018)
18. Gu, J., Cai, J., Wang, G., Chen, T.: Stack-captioning: Coarse-to-fine learning for image captioning. In: AAAI. pp. 6837–6844 (2018)
19. Gu, J., Wang, G., Cai, J., Chen, T.: An empirical study of language cnn for image captioning. In: ICCV. pp. 1222–1231 (2017)
20. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks. Pattern Recognition pp. 354–377 (2017)

21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
22. Hitschler, J., Schamoni, S., Riezler, S.: Multimodal pivots for image caption translation. In: ACL. pp. 2399–2409 (2016)
23. Jean, S., Cho, K., Memisevic, R., Bengio, Y.: On using very large target vocabulary for neural machine translation. In: ACL. pp. 1–10 (2015)
24. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding long-short term memory for image caption generation. In: ICCV. pp. 2407–2415 (2015)
25. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al.: Google's multilingual neural machine translation system: enabling zero-shot translation. TACL pp. 339–352 (2016)
26. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: EMNLP. pp. 1700–1709 (2013)
27. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. pp. 3128–3137 (2015)
28. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
29. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: Understanding and generating image descriptions. In: CVPR. pp. 1601–1608 (2011)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
31. Liu, C., Sun, F., Wang, C., Wang, F., Yuille, A.: Mat: A multimodal attentive translator for image captioning. In: IJCAI. pp. 4033–4039 (2017)
32. Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K.: Improved image captioning via policy gradient optimization of spider. In: ICCV. pp. 873–881 (2017)
33. Luong, M.T., Le, Q.V., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task sequence to sequence learning. In: ICLR (2016)
34. Luong, M.T., Sutskever, I., Le, Q.V., Vinyals, O., Zaremba, W.: Addressing the rare word problem in neural machine translation. In: ACL. pp. 11–19 (2015)
35. Mi, H., Sankaran, B., Wang, Z., Ittycheriah, A.: Coverage embedding models for neural machine translation. In: EMNLP. pp. 955–960 (2016)
36. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
37. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV. pp. 2641–2649 (2015)
38. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. In: ICLR (2016)
39. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR. pp. 7008–7024 (2017)
40. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS. pp. 3104–3112 (2014)
41. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Modeling coverage for neural machine translation. In: ACL. pp. 76–85 (2016)
42. Utiyama, M., Isahara, H.: A comparison of pivot methods for phrase-based statistical machine translation. In: NAACL. pp. 484–491 (2007)
43. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR. pp. 4566–4575 (2015)
44. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. pp. 3156–3164 (2015)

45. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. PAMI pp. 652–663 (2017)
46. Wu, H., Wang, H.: Pivot language approach for phrase-based statistical machine translation. Machine Translation pp. 165–181 (2007)
47. Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al.: Ai challenger: A large-scale dataset for going deeper in image understanding. arXiv preprint arXiv:1711.06475 (2017)
48. Wu, Q., Shen, C., Liu, L., Dick, A., Hengel, A.v.d.: What value do explicit high level concepts have in vision to language problems? In: CVPR. pp. 203–212 (2016)
49. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057 (2015)
50. Yang, X., Zhang, H., Cai, J.: Shuffle-then-assemble: Learning object-agnostic visual relationship features. In: ECCV (2018)
51. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: ICCV. pp. 22–29 (2017)
52. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR. pp. 4651–4659 (2016)
53. Zahabi, S.T., Bakhshaei, S., Khadivi, S.: Using context vectors in improving a machine translation system with bridge language. In: ACL. pp. 318–322 (2013)
54. Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., Yu, Y.: Texygen: A benchmarking platform for text generation models. In: SIGIR. pp. 1097–1100 (2018)
55. Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. In: EMNLP. pp. 1568–1575 (2016)