

Zero-Resource Neural Machine Translation with Multi-Agent Communication Game

Yun Chen,[†] Yang Liu,[‡] Victor O.K. Li[†]

[†]Department of Electrical and Electronic Engineering, The University of Hong Kong

[‡]State Key Laboratory of Intelligent Technology and Systems

Tsinghua National Laboratory for Information Science and Technology

Department of Computer Science and Technology, Tsinghua University, Beijing, China

yun.chencreek@gmail.com; liuyang2011@tsinghua.edu.cn; vli@eee.hku.hk

Abstract

While end-to-end neural machine translation (NMT) has achieved notable success in the past years in translating a handful of resource-rich language pairs, it still suffers from the data scarcity problem for low-resource language pairs and domains. To tackle this problem, we propose an interactive multimodal framework for zero-resource neural machine translation. Instead of being passively exposed to large amounts of parallel corpora, our learners (implemented as encoder-decoder architecture) engage in cooperative image description games, and thus develop their own image captioning or neural machine translation model from the need to communicate in order to succeed at the game. Experimental results on the IAPR-TC12 and Multi30K datasets show that the proposed learning mechanism significantly improves over the state-of-the-art methods.

Introduction

Neural machine translation (NMT) (Kalchbrenner and Blunsom 2013; Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015), which directly models the translation process in an end-to-end way, has achieved state-of-the-art translation performance on resource-rich language pairs such as English-French and German-English (Johnson et al. 2016; Gehring et al. 2017; Vaswani et al. 2017). The success is mainly attributed to the quality and scale of available parallel corpora to train NMT systems. However, preparing such parallel corpora has remained a big problem in some specific domains or between resource-scarce language pairs. Zoph et al. (2016) indicate that NMT tends to obtain much worse translation quality than statistical machine translation (SMT) under small-data conditions.

As a result, developing methods to achieve neural machine translation without direct source-target parallel corpora has attracted increasing attention in the community recently. These methods utilize a third language (Firat et al. 2016; Johnson et al. 2016; Chen et al. 2017; Zheng, Cheng, and Liu 2017; Cheng et al. 2017) or modality (Nakayama and Nishida 2017) as a pivot to enable zero-resource source-to-target translation. Although promising results have been obtained, pivoting with a third language still demands large scale parallel source-pivot and pivot-target corpora. On the

other hand, large amounts of monolingual text documents with rich multimodal content are available on the web, e.g., text with photos or videos posted to social networking sites and blogs. How to utilize the monolingual multimodal content to build zero-resource NMT systems remains an open question.

Multimodal content, especially image, has been widely explored in the context of NMT recently. Most of the work focus on using image in addition to text query to reinforce the translation performance (Caglayan et al. 2016; Hitschler, Schamoni, and Riezler 2016; Calixto, Liu, and Campbell 2017). This task is called multimodal neural machine translation and has become a subtask in WMT16¹ and WMT17². In contrast, there exists limited work on bridging languages using multimodal content only. Gella et al. (2017) propose to learn multimodal multilingual representations of fixed length for matching images and sentences in different languages in the same space with image as a pivot. Nakayama and Nishida (2017) suggest putting a decoder on top of the fixed-length modality-agnostic representation to generate a translation in the target language. Although the approach enables zero-resource translation, the use of a fixed-length vector is a bottleneck in improving translation performance (Bahdanau, Cho, and Bengio 2015).

In this work, we introduce a multi-agent communication game within a multimodal environment (Lazaridou, Pham, and Baroni 2016; Havrylov and Titov 2017) to achieve direct modeling of zero-resource source-to-target NMT. We have two agents in the game: a captioner which describes an image in the source language and a translator which translates a source-language sentence to a target-language sentence. Apparently, the translator is our training target. The two agents collaborate with each other to accomplish the task of describing an image in the target language with message in the source language exchanged between the agents. Both agents get reward from the communication game and collectively learn to maximize the expected reward. Experiments on German-to-English and English-to-German translation tasks over the IAPR-TC12 and Multi30K datasets demonstrate that the proposed approach yields substantial gains over the baseline methods.

¹<http://www.statmt.org/wmt16/>

²<http://www.statmt.org/wmt17/>

Background

Given a source-language sentence \mathbf{x} and a target-language sentence \mathbf{y} , a NMT model aims to build a single neural network $P(\mathbf{y}|\mathbf{x}; \theta_{\mathbf{x} \rightarrow \mathbf{y}})$ that translates \mathbf{x} into \mathbf{y} , where $\theta_{\mathbf{x} \rightarrow \mathbf{y}}$ is a set of model parameters. For resource-rich language pairs, there exists a source-target parallel corpus $D_{\mathbf{x}, \mathbf{y}}$ to train the NMT model. The model parameters can be learned with standard maximum likelihood estimation on the parallel corpus:

$$\hat{\theta}_{\mathbf{x} \rightarrow \mathbf{y}} = \operatorname{argmax}_{\theta_{\mathbf{x} \rightarrow \mathbf{y}}} \left\{ \sum_{\langle \mathbf{x}, \mathbf{y} \rangle \in D_{\mathbf{x}, \mathbf{y}}} \log P(\mathbf{y}|\mathbf{x}; \theta_{\mathbf{x} \rightarrow \mathbf{y}}) \right\}. \quad (1)$$

Unfortunately, parallel corpora are usually not readily available for low-resource language pairs or domains. On the other hand, there exists monolingual multimodal content (images with text descriptions) in the source and target language. It is possible to bridge the source and target languages with the multimodal information (Nakayama and Nishida 2017) for an image is a universal representation across all languages.

One way to ground a natural language to a visual image is through image captioning, which annotates a description for an input image with natural language through a CNN-RNN architecture (Xu et al. 2015; Karpathy and Fei-Fei 2015). Below, we call a pair of a text description and its counterpart image a ‘‘document’’ and use \mathbf{z} to denote an image. Given documents in the target language $D_{\mathbf{z}, \mathbf{y}} = \{\langle \mathbf{z}^n, \mathbf{y}^n \rangle\}_{n=1}^N$, an image caption model $P(\mathbf{y}|\mathbf{z}; \theta_{\mathbf{z} \rightarrow \mathbf{y}})$ can be built, which ‘‘translates’’ an image to a sentence in the target language. The model parameters $\theta_{\mathbf{z} \rightarrow \mathbf{y}}$ can be learned by maximizing the log-likelihood of the monolingual multimodal documents:

$$\hat{\theta}_{\mathbf{z} \rightarrow \mathbf{y}} = \operatorname{argmax}_{\theta_{\mathbf{z} \rightarrow \mathbf{y}}} \left\{ \sum_{\langle \mathbf{z}, \mathbf{y} \rangle \in D_{\mathbf{z}, \mathbf{y}}} \log P(\mathbf{y}|\mathbf{z}; \theta_{\mathbf{z} \rightarrow \mathbf{y}}) \right\}. \quad (2)$$

Inspired by the idea of pivot-based translation (Cheng et al. 2017; Zheng, Cheng, and Liu 2017), another way to achieve image-to-target translation is using a second language (the source language) as a pivot. As a result, image-to-target translation can be divided into two steps: the image is first translated to a source sentence using the image-to-source captioning model, which is then translated to a target sentence using the source-to-target translation model. We use two agents to represent the image-to-source captioning model and the source-to-target translation model. The image-to-target translation procedure can be simulated by a two-agent communication game, where agents cooperate with each other to play the game and collectively learn their model parameters based on the feedback. Below we formally define the game, which is a general learning framework for training zero-resource machine translation model with monolingual multimodal documents only.

Two-agent Communication Game

Problem Formulation

Given monolingual documents in the source language $D_{\mathbf{z}, \mathbf{x}} = \{\langle \mathbf{z}^{(m)}, \mathbf{x}^{(m)} \rangle\}_{m=1}^M$ and in the target language

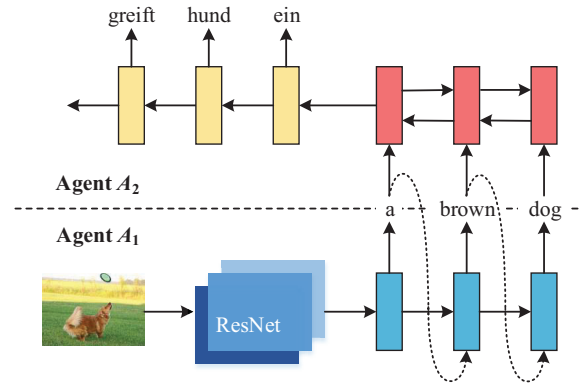


Figure 1: Zero-resource neural machine translation through a two-agent communication game within a multimodal environment. Agent A_1 is an image captioning model implemented with CNN-RNN architecture (Xu et al. 2015); Agent A_2 is a neural machine translation model implemented with RNNSearch (Bahdanau, Cho, and Bengio 2015). Taking an image as input, agent A_1 sends a message in the source language to agent A_2 , which is translated by agent A_2 to a target-language sentence to win a reward.

$D_{\mathbf{z}, \mathbf{y}} = \{\langle \mathbf{z}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^N$, our aim is to learn a model which translates source sentence \mathbf{x} to target sentence \mathbf{y} . Importantly, $D_{\mathbf{z}, \mathbf{x}}$ and $D_{\mathbf{z}, \mathbf{y}}$ do not overlap; they do not share the same images at all. Our model consists of a *captioner* $P(\mathbf{x}|\mathbf{z}; \theta_{\mathbf{z} \rightarrow \mathbf{x}})$ which translates image to source sentence and a *translator* $P(\mathbf{y}|\mathbf{x}; \theta_{\mathbf{x} \rightarrow \mathbf{y}})$ which translates source sentence to target sentence, where $\theta_{\mathbf{z} \rightarrow \mathbf{x}}$ and $\theta_{\mathbf{x} \rightarrow \mathbf{y}}$ are model parameters. To make the task a real zero-resource scenario, we assume that no parallel corpora are available even in the validation set. That is to say, we only have monolingual documents $D_{\mathbf{z}, \mathbf{x}}^{val} = \{\langle \mathbf{z}^m, \mathbf{x}^m \rangle\}_{m=1}^{M^{val}}$ and $D_{\mathbf{z}, \mathbf{y}}^{val} = \{\langle \mathbf{z}^n, \mathbf{y}^n \rangle\}_{n=1}^{N^{val}}$ for validation.

In the following we describe two models: (i) the PRE. (pre-training) model that only trains the translator in the two-agent game as the captioner can be pre-trained with $D_{\mathbf{z}, \mathbf{x}}$; (ii) the JOINT model that jointly optimize the captioner and translator through reinforcement learning in the communication game.

The Game

As illustrated in Fig. 1, we propose a simple communication game with two agents, the captioner A_1 and the translator A_2 . Sampling a monolingual document $\langle \mathbf{z}, \mathbf{y} \rangle$ from $D_{\mathbf{z}, \mathbf{y}}$, the game is defined as follows:

- 1 A_1 is shown the image and is told to describe the image with a source-language sentence \mathbf{x}_{mid} .
- 2 A_2 is shown the middle sentence \mathbf{x}_{mid} generated by A_1 without the image information. It is told to translate \mathbf{x}_{mid} to a target-language sentence.
- 3 The environment evaluates the consistency of the translated target sentence and the gold-standard target-

language sentence y and then both agents receive a reward.

The captioner A_1 and the translator A_2 must work together to achieve a good reward. A_1 should learn how to provide accurate image description in the source language and A_2 should be good at translating a source sentence to a target sentence. This game can be played for an arbitrary number of rounds, and the captioner A_1 and the translator A_2 will get trained through this reinforcement procedure (e.g., by means of the policy gradient methods). In this way, we develop a general learning framework for training zero-resource machine translation model (the translator A_2) with monolingual multimodal documents only through a multi-agent communication game. As we do not assume the specific architectures of the captioner and the translator, our proposed learning framework is transparent to architectures and can be applied to any end-to-end image captioning and NMT systems.

Implementation

For a game beginning with a monolingual document $\langle \mathbf{z}, \mathbf{y} \rangle \in D_{z,y}$, we use \mathbf{x}_{mid} to denote the exchanged source sentence between agents. The goal of training is to find the parameters of the agents that maximize the expected reward:

$$\mathcal{E}(\theta_{z \rightarrow x}, \theta_{x \rightarrow y}) = \mathbb{E}_{P(\mathbf{x}_{mid}|\mathbf{z};\theta_{z \rightarrow x})}[r(\mathbf{y}, \mathbf{x}_{mid}, \theta_{x \rightarrow y})]. \quad (3)$$

We follow He et al. (2016a) and define the reward as the log probability of agent A_2 generates \mathbf{y} from \mathbf{x}_{mid} :

$$r(\mathbf{y}, \mathbf{x}_{mid}, \theta_{x \rightarrow y}) = \log P(\mathbf{y}|\mathbf{x}_{mid}; \theta_{x \rightarrow y}). \quad (4)$$

As a result, the expected reward in the multi-agent communication game can be re-written as:

$$\mathcal{E}(\theta_{z \rightarrow x}, \theta_{x \rightarrow y}) = \mathbb{E}_{P(\mathbf{x}_{mid}|\mathbf{z};\theta_{z \rightarrow x})}[\log P(\mathbf{y}|\mathbf{x}_{mid}; \theta_{x \rightarrow y})]. \quad (5)$$

In training, we optimize the parameters of the captioner and translator through policy gradient methods for expected reward maximization:

$$\begin{aligned} & \hat{\theta}_{z \rightarrow x}, \hat{\theta}_{x \rightarrow y} \\ &= \operatorname{argmax}_{\theta_{z \rightarrow x}, \theta_{x \rightarrow y}} \left\{ \sum_{\langle \mathbf{z}, \mathbf{y} \rangle \in D_{z,y}} \mathcal{E}(\theta_{z \rightarrow x}, \theta_{x \rightarrow y}) \right\}. \end{aligned} \quad (6)$$

We compute the gradient of $\mathcal{E}(\theta_{z \rightarrow x}, \theta_{x \rightarrow y})$ with respect to parameters $\theta_{z \rightarrow x}$ and $\theta_{x \rightarrow y}$. According to the policy gradient theorem (Sutton et al. 1999), it is easy to verify that:

$$\begin{aligned} & \nabla_{\theta_{z \rightarrow x}} \mathcal{E}(\theta_{z \rightarrow x}, \theta_{x \rightarrow y}) \\ &= \mathbb{E}[r \nabla_{\theta_{z \rightarrow x}} \log P(\mathbf{x}_{mid}|\mathbf{z}; \theta_{z \rightarrow x})], \end{aligned} \quad (7)$$

$$\begin{aligned} & \nabla_{\theta_{x \rightarrow y}} \mathcal{E}(\theta_{z \rightarrow x}, \theta_{x \rightarrow y}) \\ &= \mathbb{E}[\nabla_{\theta_{x \rightarrow y}} \log P(\mathbf{y}|\mathbf{x}_{mid}; \theta_{x \rightarrow y})], \end{aligned} \quad (8)$$

in which the expectation is taken over \mathbf{x}_{mid} and $r = \log[P(\mathbf{y}|\mathbf{x}_{mid}; \theta_{x \rightarrow y})]$.

Unfortunately, Eqn. 7 and 8 are intractable to calculate due to the exponential search space of $\mathcal{X}(\mathbf{z})$. Following (He et al. 2016a), we adopt beam search for gradient estimation. Compared with random sampling, beam search can

Algorithm 1 The learning algorithm in the two-agent communication game

- 1: **Input:** Monolingual multimodal documents $D_{z,y} = \{\langle \mathbf{z}^n, \mathbf{y}^n \rangle\}_{n=1}^N$, initial image-to-source caption model $\theta_{z \rightarrow x}$, initial source-to-target translation model $\theta_{x \rightarrow y}$, beam search size K , learning rates $\gamma_{1,t}, \gamma_{2,t}$.
- 2: **repeat**
- 3: $t = t + 1$.
- 4: Sample a document $\langle \mathbf{z}^n, \mathbf{y}^n \rangle$ from $D_{z,y}$.
- 5: Set $\mathbf{z} = \mathbf{z}^n, \mathbf{y} = \mathbf{y}^n$.
- 6: Generate K sentences $\mathbf{x}_{mid,1}, \dots, \mathbf{x}_{mid,K}$ using beam search according to the image captioning model $P(\mathbf{x}|\mathbf{z}; \theta_{z \rightarrow x})$.
- 7: **for** $k = 1, \dots, K$ **do**
- 8: Set the reward for the k th sampled sentence as $r_k = \log P(\mathbf{y}|\mathbf{x}_{mid,k}; \theta_{x \rightarrow y})$.
- 9: **end for**
- 10: Compute the stochastic gradient of $\theta_{z \rightarrow x}$:

$$\nabla_{\theta_{z \rightarrow x}} \hat{\mathbb{E}}[r] = \frac{1}{K} \sum_{k=1}^K [r_k \nabla_{\theta_{z \rightarrow x}} \log P(\mathbf{x}_{mid,k}|\mathbf{z}; \theta_{z \rightarrow x})].$$

- 11: Compute the stochastic gradient of $\theta_{x \rightarrow y}$:

$$\nabla_{\theta_{x \rightarrow y}} \hat{\mathbb{E}}[r] = \frac{1}{K} \sum_{k=1}^K [\nabla_{\theta_{x \rightarrow y}} \log P(\mathbf{y}|\mathbf{x}_{mid,k}; \theta_{x \rightarrow y})].$$

- 12: Model updates:

$$\begin{aligned} \theta_{z \rightarrow x} &\leftarrow \theta_{z \rightarrow x} + \gamma_{1,t} \nabla_{\theta_{z \rightarrow x}} \hat{\mathbb{E}}[r], \\ \theta_{x \rightarrow y} &\leftarrow \theta_{x \rightarrow y} + \gamma_{2,t} \nabla_{\theta_{x \rightarrow y}} \hat{\mathbb{E}}[r]. \end{aligned}$$

- 13: **until** convergence
-

help to avoid the very large variance and sometimes unreasonable results brought by image captioning (Ranzato et al. 2015). Specifically, we run beam search with the captioner A_1 to generate top- K high-probability middle outputs in the source language, and use the averaged value on the middle outputs to approximate the true gradient. Algorithm 1 shows the detailed learning algorithm.

Training

Since the captioner A_1 has a very large action space to generate source description \mathbf{x}_{mid} , it is extremely difficult to learn with an initial random policy. Specifically, the search space for A_1 is of size $\mathcal{O}(\mathcal{V}^T)$, where \mathcal{V} is the number of words in the source vocabulary (more than 1000 in our experiments) and T is the length of the sentence (around 10 to 20 in our experiments). Thus, we pre-train the captioner A_1 with maximum likelihood estimation leveraging monolingual dataset $D_{z,x}$:

$$\hat{\theta}_{z \rightarrow x}^{\text{pre}} = \operatorname{argmax}_{\theta_{z \rightarrow x}} \left\{ \sum_{\langle \mathbf{z}, \mathbf{x} \rangle \in D_{z,x}} \log P(\mathbf{x}|\mathbf{z}; \theta_{z \rightarrow x}) \right\}. \quad (9)$$

We initialize the captioner A_1 with the pre-trained image caption model and randomly initialize the translator A_2 . To avoid the randomly initialized agent A_2 doing harm to the pre-trained captioner A_1 , we fix A_1 for the initial few epochs

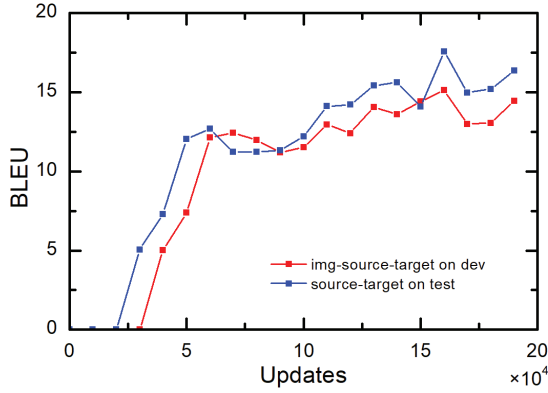


Figure 2: Validation of the PRE. model on monolingual documents $D_{z,y}^{val}$ for the IAPR-TC12 German-to-English translation task. BLEU score of the translator A_2 on the test set correlates very well with BLEU score on $D_{z,y}^{val}$ using our proposed validation criterion.

and only optimize agent A_2 . Then we adopt two different training approaches:

1. The PRE. (pre-training) model: We keep the captioner fixed and only train the translator in the two-agent game. Thus, the training objective is:

$$\begin{aligned} & \mathcal{J}_{\text{PRE.}}(\theta_{x \rightarrow y}) \\ &= \sum_{\langle \mathbf{z}, \mathbf{y} \rangle \in D_{z,y}} \mathbb{E}_{P(\mathbf{x}_{mid} | \mathbf{z}; \theta_{z \rightarrow x}^{\text{pre}})} [\log P(\mathbf{y} | \mathbf{x}_{mid}; \theta_{x \rightarrow y})]. \end{aligned} \quad (10)$$

2. The JOINT model: We jointly optimize the captioner and the translator through reinforcement learning in the communication game. To encourage the captioner A_1 to correctly describe the image in the source language, we use half documents from monolingual dataset $D_{z,x} = \{ \langle \mathbf{z}^{(m)}, \mathbf{x}^{(m)} \rangle \}_{m=1}^M$ in each mini batch to constrain the parameters of A_1 . We train to maximize the weighted sum of the reward based on documents from $D_{z,y}$ and the log-likelihood of image-to-source captioning model $\theta_{z \rightarrow x}$ on documents from $D_{z,x}$. The objective becomes:

$$\begin{aligned} & \mathcal{J}_{\text{JOINT}}(\theta_{z \rightarrow x}, \theta_{x \rightarrow y}) \\ &= \sum_{\langle \mathbf{z}, \mathbf{y} \rangle \in D_{z,y}} \mathbb{E}_{P(\mathbf{x}_{mid} | \mathbf{z}; \theta_{z \rightarrow x})} [\log P(\mathbf{y} | \mathbf{x}_{mid}; \theta_{x \rightarrow y})] \\ & \quad + \lambda \sum_{\langle \mathbf{z}, \mathbf{x} \rangle \in D_{z,x}} \log P(\mathbf{x} | \mathbf{z}; \theta_{z \rightarrow x}). \end{aligned} \quad (11)$$

Validation

Since we do not have access to parallel sentences even at validation time, we need to have a criterion for model and hyper-parameters selection for the PRE. and JOINT methods. We validate the model on $D_{z,y}^{val}$.

Given model parameters $\theta_{z \rightarrow x}$ and $\theta_{x \rightarrow y}$ at some iteration and a monolingual document $\langle \mathbf{z}, \mathbf{y} \rangle$ from $D_{z,y}^{val}$, we output the image’s description in the target language with the

Table 1: Dataset statistics.

		IAPR-TC12	Multi30K
Num. of images		20,000	31,014
Vocabulary	En	1207	5877
	De	1511	7225
Avg. length of descriptions	En	18.8	14.1
	De	15.9	11.7

following decision rule using beam search:

$$\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x}} \left\{ P(\mathbf{x} | \mathbf{z}; \theta_{z \rightarrow x}) \right\} \quad (12)$$

$$\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y}} \left\{ P(\mathbf{y} | \hat{\mathbf{x}}; \theta_{x \rightarrow y}) \right\} \quad (13)$$

With the generated $\hat{\mathbf{y}}$ as the hypothesis and the gold-standard description \mathbf{y} as the reference, we use the BLEU score as the validation criterion, namely, we choose the translation model with the highest BLEU score. Figure 2 shows a typical example of the correlation between this measure and the translation performance of the translator A_2 on test set.

Experiments

Data Set

We evaluate our model on two publicly available multilingual image-description datasets as in (Nakayama and Nishida 2017). The IAPR-TC12 dataset (Grubinger et al. 2006), which consists of English image descriptions and the corresponding German translations, has a total of 20K images. Each image contains multiple descriptions and each description corresponds to a different aspect of the image. Since the first sentence is likely to describe the most salient objects (Grubinger et al. 2006), we use only the first description of each image. We randomly split the dataset into training, validation and test sets with 18K, 1K and 1K images respectively. The recently published Multi30K dataset (Elliott et al. 2016), which is a multilingual extension of Flickr30k corpus (Young et al. 2014), has 29K, 1K and 1K images in the training, validation and test splits respectively with English and German image descriptions (Elliott et al. 2016). There are two types of multilingual annotations in the dataset: (i) a corpus of one English description per image and its German translation; and (ii) a corpus of 5 independently collected English and German descriptions per image. Since the corpus of independently collected English and German descriptions better fit the noisy multimodal content on the web, we adopt this corpus in our experiments. Note that although these descriptions describe the same image, they are not translations of each other.

For preprocessing, we use the scripts in the Moses SMT Toolkit (Koehn et al. 2007) to normalise and tokenize English and German descriptions. For the IAPR-TC12 dataset, we construct the vocabulary with words appearing more than 5 times in the training splits and replace those appearing less than 5 times with UNK symbol. For the Multi30K dataset, we adopt a joint byte pair encoding (BPE) (Sennrich, Had-

Table 2: Splits for experiments. Each image is annotated with one English or German sentence for the IAPR-TC12 dataset, while each image is described by 5 English or German sentences for the Multi30K dataset.

Split	Pair	IAPR-TC12			Multi30K		
		img	En	De	img	En	De
Train	img-En	9,000	9,000	-	14,500	72,500	-
	img-De	9,000	-	9,000	14,500	-	72,500
Validation	img-En	500	500	-	507	2,535	-
	img-De	500	-	500	507	-	2,535
Test	En-De	-	1,000	1,000	-	5,000	5,000

dow, and Birch 2016) with 10K merge operations on English and German descriptions to reduce vocabulary size. To comply with the zero-resource setting, we randomly split the images in the training and validation datasets into two parts with equal size. One part constructs the image-English split and the other part the image-German split. Unnecessary modalities for each split (e.g., German descriptions for image-English split) are ignored. Note that the two splits have no overlapping images, and we have no direct English-German parallel corpus. Table 1 and 2 summarizes data statistics.

Experimental Setup

To extract image features, we follow the suggestion of (Caglayan et al. 2016) and adopt ResNet-50 network (He et al. 2016b) pre-trained on ImageNet without finetuning. We use the (14,14,1024) feature map of the res4fx (end of Block-4) layer after ReLU. For some baseline methods that do not support attention mechanism, we extract 2048-dimension feature after the pool5 layer. We follow the architecture in (Xu et al. 2015) for image captioning and standard RNNSearch architecture (Bahdanau, Cho, and Bengio 2015) for translation. We leverage *dl4mt*³ and *arctic-captions*⁴ for all our experiments. The beam search size is 2 in the middle image-to-source caption generation. During validation and testing, we set the beam search size to be 5 for both the captioner and the translator. All models are quantitatively evaluated with BLEU (Papineni et al. 2002). For the Multi30K dataset, each image is paired with 5 English descriptions and 5 German descriptions in the test set. We follow the setting in (Caglayan et al. 2016) and let the NMT generate a target description for each of the 5 source sentences and pick the one with the highest probability as the translation result. The evaluation is performed against the corresponding five target descriptions in the testing phase.

We compare our approach with four baseline methods:

- 1 Random: For a sentence in source language, we randomly select a document in $D_{z,y}$ whose caption would be output as the translation result.
- 2 TFIDF: For a source sentence, we first search the nearest document in $D_{z,x}$ in terms of cosine similarity of TFIDF text features. Then, for the coupled image of that document, we retrieve the most similar document in $D_{z,y}$ in

³*dl4mt-tutorial*: <https://github.com/nyu-dl>

⁴*arctic-captions*: <https://github.com/kelvinxu/arctic-captions>

Table 3: Comparison with previous work on German-to-English and English-to-German translation with zero resource over the IAPR-TC12 and Multi30K datasets.

	IAPR-TC12		Multi30K	
	De-En	En-De	De-En	En-De
Random	2.5	1.2	2.0	0.8
TFIDF	10.2	7.5	2.4	1.0
3-way model	13.9	8.6	15.9	10.1
TS model	14.1	9.2	16.4	10.3
PRE.	17.6	11.9	16.0	12.1
JOINT	18.6	14.2	19.6	16.6

terms of cosine similarity of pool5 feature vectors. We output the coupled target sentence of the retrieved document as our translation.

- 3 3-way model (Nakayama and Nishida 2017): We leverage the best 3-way model proposed in (Nakayama and Nishida 2017) as our baseline, which adopts end to end training strategy and trains the decoder with image and description.
- 4 TS model (Chen et al. 2017): This method is originally designed for leveraging a third language as a pivot to enable zero-resource NMT. We follow the teacher-student framework and replace the pivot language with image. The teacher model is an image-to-target captioning model and the student model is the zero-resource source-to-target translation model.

Comparison with Baselines

Table 3 shows the translation performance of our proposed zero-resource NMT on the IAPR-TC12 and Multi30K datasets in comparison with other baselines. Note that the TS model, PRE. model and JOINT model all utilize attention mechanism, while the 3-way model does not support attention. It is clear from the table that in all the cases, our proposed JOINT model outperforms all the other baselines. Even with fixed captioner, the PRE. model can outperform the baselines in most cases.

Specifically, the JOINT model outperforms the 3-way model by +4.7 BLEU score on German-to-English translation and +5.6 BLEU score on English-to-German translation over the IAPR-TC12 dataset; +3.7 BLEU score on German-to-English translation and +6.5 BLEU score on English-to-



Source (German)	Translation (English)	Reference image
1. ein mädchen in einem kleid springt auf einem grünen rasen . 2. ein kleines mädchen beim springen . 3. ein kleines mädchen springt auf einer wiese herum . 4. ein mädchen springt auf einer wise . 5. ein kleines mädchen in einem rosa kleid hüpft im gras	PRE.: a girl in a pink dress is jumping in the air . JOINT: a little girl in a pink dress is jumping in the air .	
die grauen mauern und grünen terrassen einer ruine auf einem berg , mit einem sehr markanten berg dahinter und einer bergkette im hintergrund .	PRE.: a ruin with grey walls and green terraces in the foreground . JOINT: the grey walls and green terraces of ruins on top of a mountain , with a very distinctive mountain behind them and a wooded mountain range in the background .	

Figure 3: Examples of target translations from the test set using zero-resource NMT trained by our proposed PRE. and JOINT models. The first example is from the Multi30K dataset, while the second is from the IAPR-TC12 dataset.

German translation for the Multi30K dataset. The performance gap can be explained since the 3-way model attempts to encode a whole input sentence or image into a single fixed-length vector, rendering it difficult for the neural network to cope with long sentences or images with a lot of clutter. In contrast, the JOINT model adopts the attention mechanism for the captioner and the translator. Rather than compressing an entire image or source sentence into a static representation, attention allows for a model to automatically attend to parts of a source sentence or image that are relevant to predicting a target-side word. This explanation is in line with the observation that all three models with attention mechanism outperform the 3-way model.

Although the TS model also adopts the attention mechanism, its performance is dominated by the performance of the teacher model, namely the image captioning model. During the learning process of the student model, the teacher model is kept fixed all the time and the student tries to mimic the decoding behavior of the teacher in the teacher-student framework. The description of the same image can vary a lot in different languages as indicated in (van Miltenburg, Elliott, and Vossen 2017). Thus, the teacher model’s captioning result in target language is not necessarily the translation of the image’s coupled source-language description. On the contrary, the JOINT model jointly optimize the captioner and the translator to win a two-agent communication game. The two agents are complementary to each other and are trained jointly to maximize expected reward. This mechanism helps to solve the problem of cross-linguistic differences in image description (van Miltenburg, Elliott, and Vossen 2017), resulting in better zero-resource NMT model.

Figure 3 shows translation examples of the zero-resource NMT trained by the PRE. and JOINT methods. We also list the corresponding image for reference. The first example is from the Multi30K dataset, while the second is from the IAPR-TC12 dataset. Apparently, the JOINT model has successfully learned how to translate even with zero-resource.

Effect of Joint Training

We also compare the performance of the captioner with (JOINT) and without (PRE.) the joint training. Table 4 shows the result. Cheng et al. (2017) demonstrate that in

Table 4: Comparison the captioner’s performance of our proposed PRE. and JOINT model.

Model	IAPR-TC12		Multi30K	
	img-de	img-en	img-de	img-en
PRE.	10.7	18.0	10.9	22.7
JOINT	10.5	17.3	12.3	21.3

pivot-based NMT, where the pivot is a third language, the source-pivot and pivot-target translation models can be improved simultaneously through joint training. In our setting, image-to-source captioning does not necessarily improve with joint training. It gets slightly worse on three out of four translation tasks. Using a third language as a pivot, the source-to-pivot and pivot-to-target models are symmetrical as they are both NMT models. In contrast, in our setting the captioner and the translator are not symmetrical anymore since NMT is much easier to improve than image caption. We suspect that the translator’s performance dominates the multi-agent game.

Comparison with Oracle

Figure 4 compares the JOINT model with ORACLE that uses direct source-target parallel or comparable corpora on the IAPR-TC12 and Multi30K dataset.

For the Multi30K dataset, each image is annotated with 5 German sentences and 5 English sentences. The training dataset for ORACLE can be constructed by the cross product of 5 source and 5 target descriptions which results in a total of 25 description pairs for each image or by only taking the 5 pairwise descriptions. We follow (Caglayan et al. 2016) and use the pairwise way.

For the IAPR-TC12 dataset, the JOINT model is roughly comparable with ORACLE when the number of parallel sentences are limited to about 15% that of our monolingual ones. For the Multi30K dataset, the JOINT model obtains comparable results with ORACLE trained by number of comparable sentence pairs about 40% that of our monolingual corpus. It obtains 75% of the BLEU score on German-to-English translation and 65% of the BLEU score on English-to-German translation for NMT trained with full

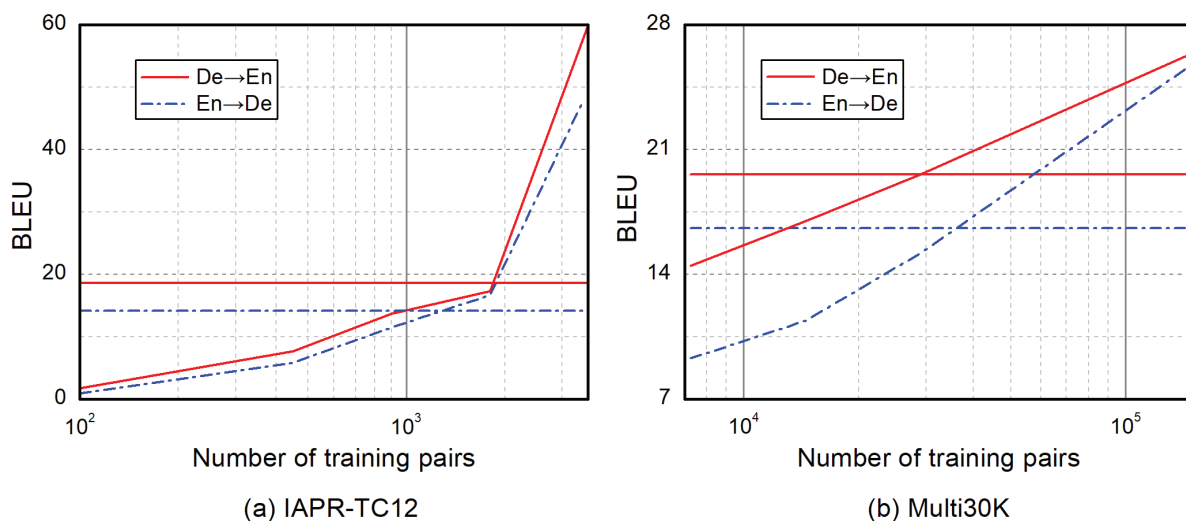


Figure 4: Comparison with ORACLE that uses direct source-target parallel or comparable corpora. The JOINT model corresponds to horizontal lines.

corpora. Although there is still a significant gap in performance as compared to ORACLE trained with parallel corpus, this is encouraging since our approach only uses monolingual multimodal documents.

Related Work

Training NMT models without source-target parallel corpora by leveraging a third language or image modality has attracted intensive attention in recent years. Utilizing a third language as a pivot has already achieved promising translation quality for zero-resource NMT. Firat et al. (2016) pre-train multi-way multilingual model and then fine-tune the attention mechanism with pseudo parallel data generated by the model to improve zero-resource translation. Johnson et al. (Johnson et al. 2016) adopt a universal encoder-decoder network in multilingual scenarios to naturally enable zero-resource translation. In addition to the above multilingual methods, several authors propose to train the zero-resource source-to-target translation model directly. Chen et al. (2017) propose a teacher-student framework under the assumption that parallel sentences have close probabilities of generating a sentence in a third language. Zheng et al. (2017) maximize the expected likelihood to train the intended source-to-target model. However, all these methods assume that source-pivot and pivot-target parallel corpora are available. Another line is to bridge zero-resource language pairs via images. Nakayama and Nishida (2017) train multimodal encoders to learn modality-agnostic multilingual representation of fix length using image as a pivot. On top of the fix-length representation, they build a decoder to output a translation in the target language. Although the performance is limited by the fix-length representation, their work shows that zero-resource neural machine translation with an image pivot is possible.

Multimodal neural machine translation, which introduce image modality into NMT as an additional information

source to reinforce the translation performance, has received much attention in the community recently. A lot of work have shown that image modality can benefit neural machine translation, hopefully by relaxing ambiguity in alignment that cannot be solved by texts only (Hitschler, Schamoni, and Riezler 2016; Calixto, Liu, and Campbell 2017). Note that their setting is much easier than ours because in their setting multilingual descriptions for the same images are available in the training dataset and an image is part of the query in both training and testing phases.

Conclusion

In this work, we propose a multi-agent communication game to tackle the challenging task of training a zero-resource NMT system from just monolingual multimodal data. In contrast with previous studies that learn a modality-agnostic multilingual representation, we successfully deploy the attention mechanism to the target zero-resource NMT model by encouraging the agents to cooperate with each other to win a image-to-target translation game. Experiments on German-to-English and English-to-German translation over the IAPR-TC12 and Multi30K datasets show that our proposed multi-agent learning mechanism can significantly outperform the state-of-the-art methods.

In the future, we plan to explore whether machine translation can perform satisfactorily with automatically crawled noisy multimodal data from the web. Since our current method is intrinsically limited to the domain where texts can be grounded to visual content, it is also interesting to explore how to further extend the learned translation model to handle generic documents.

Acknowledgments

This work is partially supported by the National Key R&D Program of China (No. 2017YFB0202204), National Natural Science Foundation of China (No. 61522204, No.

61331013) and the HKU Artificial Intelligence to Advance Well-being and Society (AI-WiSe) Lab.

References

- Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.
- Caglayan, O.; Aransa, W.; Wang, Y.; Masana, M.; García-Martínez, M.; Bougares, F.; and Loïc Barrault and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? In *WMT*.
- Calixto, I.; Liu, Q.; and Campbell, N. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*.
- Chen, Y.; Liu, Y.; Cheng, Y.; and Li, V. O. K. 2017. A teacher-student framework for zero-resource neural machine translation. In *ACL*.
- Cheng, Y.; Yang, Q.; Liu, Y.; Sun, M.; and Xu, W. 2017. Joint training for pivot-based neural machine translation. In *IJCAI*.
- Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, 70–74.
- Firat, O.; Sankaran, B.; Al-Onaizan, Y.; Yarman-Vural, F. T.; and Cho, K. 2016. Zero-resource translation with multilingual neural machine translation. In *EMNLP*.
- Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. 2017. Convolutional sequence to sequence learning. In *ICML*.
- Gella, S.; Sennrich, R.; Keller, F.; and Lapata, M. 2017. Image pivoting for learning multilingual multimodal representations. *CoRR* abs/1707.07601.
- Grubinger, M.; Clough, P.; Müller, H.; and Deselaers, T. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *LREC*, 13–23.
- Havrylov, S., and Titov, I. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *CoRR* abs/1705.11192.
- He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.-Y.; and Ma, W.-Y. 2016a. Dual learning for machine translation. In *NIPS*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778.
- Hitschler, J.; Schamoni, S.; and Riezler, S. 2016. Multimodal pivots for image caption translation. In *ACL*.
- Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F. B.; Wattenberg, M.; Corrado, G. S.; Hughes, M.; and Dean, J. 2016. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *CoRR* abs/1611.04558.
- Kalchbrenner, N., and Blunsom, P. 2013. Recurrent continuous translation models. In *EMNLP*.
- Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.
- Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.
- Lazaridou, A.; Pham, N. T.; and Baroni, M. 2016. Towards multi-agent communication-based language learning. *CoRR* abs/1605.07133.
- Nakayama, H., and Nishida, N. 2017. Zero-resource machine translation by multimodal encoderdecoder network with multimedia pivot. *Machine Translation* 31:49–64.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.
- Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence level training with recurrent neural networks. *CoRR* abs/1511.06732.
- Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units.
- Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks.
- Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*.
- van Miltenburg, E.; Elliott, D.; and Vossen, P. T. J. M. 2017. Cross-linguistic differences and similarities in image descriptions. *CoRR* abs/1707.01736.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *CoRR* abs/1706.03762.
- Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.
- Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.
- Zheng, H.; Cheng, Y.; and Liu, Y. 2017. Maximum expected likelihood estimation for zero-resource neural machine translation. In *IJCAI*.
- Zoph, B.; Yuret, D.; May, J.; and Knight, K. 2016. Transfer learning for low-resource neural machine translation. In *EMNLP*.