

Unpaired Multimodal Neural Machine Translation via Reinforcement Learning

Anonymous Author (s)

Abstract. End-to-end neural machine translation (NMT) heavily relies on parallel corpora for training. However, high-quality parallel corpora are usually costly to collect. To tackle this problem, multimodal content, especially image, has been introduced to help build an NMT system without parallel corpora. In this paper, we propose a reinforcement learning (RL) method to build an NMT system by introducing a sequence-level supervision signal as a reward. Based on the fact that visual information can be a universal representation to ground different languages, we design two different rewards to guide the learning process, i.e., (1) the likelihood of generated sentence given source image and (2) the distance of attention weights given by image caption models. Experimental results on the Multi30K, IAPR-TC12, and IKEA datasets show that the proposed learning mechanism achieves better performance than existing methods.

1 Introduction

End-to-end neural machine translation (NMT) has shown its superiority on several resource-rich language pairs [2, 27, 9], which is mainly attributed to the quality and scale of available parallel corpora [6]. However, it is usually quite difficult to collect adequate high-quality parallel corpora, since preparing such corpora is very expensive and time-consuming.

To tackle the issue where no parallel corpora are available, pivot-based NMT methods indirectly learn the alignment of the source and target languages with the help of another language [8, 16, 33]. Although promising results have been obtained, this kind of methods still demands large scale parallel source-pivot and pivot-target corpora. On the other hand, nowadays, we can easily find abundant monolingual text documents with rich multimedia content as the side information, e.g., text with photos or videos posted to social networking sites and blogs^{1 2}. These visual medias are expected to be more or less correlated to the counterpart texts. How to utilize the multimodal content, especially image, to build NMT systems remains an open question.

To achieve modeling of source-to-target NMT using multimodal content only, the efforts in the literature can be divided into two classes. One is to learn a fixed-length modality-agnostic representation matching images and sentences in different languages in the same space [21, 10]. Although this approach enables translation without parallel corpora, the use of a fixed-length vector is a bottleneck in improving translation performance [2, 6]. The other class of work tries to generate pseudo parallel corpora by translating images into sentences in another language for given monolingual multimodal

¹ <http://blog.flickr.net/>

² <https://mobile.twitter.com/>

corpora using pre-trained image captioning model. Then, translation model could be learned with standard maximum likelihood estimation [6, 5]. Although this kind of approach is easy to implement, the mistakes made by the image captioning model would propagate to the translation model, thus hurt the performance.

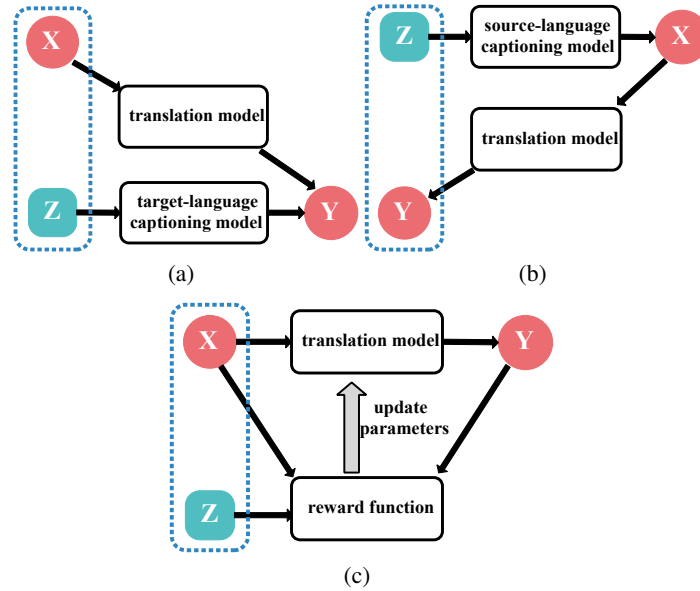


Fig. 1. (a) The teacher-student approach, (b) the 2-agent approach and (c) our reinforcement learning approach. X , Y and Z denote source, target sentences and image, respectively. We use a dashed-line box to denote that the image and sentence are paired.

In fact, to effectively leverage the visual information for NMT, an important fact should be noticed is that we can generally understand the content of images taken in other countries regardless of which language we speak. The major challenge here is how to leverage such a fact to guide the learning of translation model with unpaired multimodal content.

We address this challenge by casting the unpaired multimodal machine translation task as a reinforcement learning (RL) problem. Specifically, we introduce a sequence-level supervision signal by estimating the relevance between source and target sentences with the help of image, which aims to evaluate the quality of target sentence. Referring the translation model as *policy*, we formulate this sequence-level supervision signal as *reward* and directly optimize it. Intuitively, for a given source-language sentence and its corresponding image, a better translation in target language should have closer connection with the image. Based on this observation, we design two different reward functions to guide the learning process. Then, the policy is updated by REINFORCE algorithm [28]. Compared with previous methods [5, 6], our approach allows direct evaluation of source-target sentence pairs, without the need of translating images into sentences. Thus, this strategy avoids the problem of error propagation.

Our main contributions are summarized as follows:

- (1) To effectively leverage the visual information for NMT, we proposed to cast the unpaired multimodal NMT task as a RL problem.
- (2) We introduce sequence-level reward by estimating relevance between source and target sentences with the help of image. Specifically, we propose two kinds of reward to guide the training of NMT model.
- (3) Experiments on three translation tasks over three datasets show that the proposed rewards can provide good supervision on unlabeled multimodal corpora, and achieve better performance than existing methods.

2 Background

Neural Machine Translation Given the source language space \mathcal{X} and target language space \mathcal{Y} , an NMT model takes a sample from \mathcal{X} as input and maps to space \mathcal{Y} . In common practice, the NMT model is represented by a conditional distribution $P_\theta(Y|X)$ parameterized by θ , where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. In standard supervised learning, given a parallel corpus $\mathcal{D}_{X,Y}$, the translation model is learned by maximizing the likelihood of the training data:

$$\mathcal{L}(\theta) = \sum_{\langle X,Y \rangle \in \mathcal{D}_{X,Y}} \log P_\theta(Y|X). \quad (1)$$

Grounding Language to Visual Image There exists monolingual multimodal content (images with text descriptions) on the Web. It is possible to ground natural language to a visual image through image captioning, which annotates a description for an input image with natural language [31, 17]. Given a multimodal corpus in $\mathcal{D}_{Z,Y}$ where Z is an image and Y is a sentence describing Z in target language, an image captioning model $P_{\phi_{Z \rightarrow Y}}(Y|Z)$ can be built, which “translates” an image to a sentence. The model parameters $\phi_{Z \rightarrow Y}$ can be learned by maximizing the log-likelihood of $\mathcal{D}_{Z,Y}$:

$$\mathcal{L}(\phi_{Z \rightarrow Y}) = \sum_{\langle Z,Y \rangle \in \mathcal{D}_{Z,Y}} \log P_{\phi_{Z \rightarrow Y}}(Y|Z) \quad (2)$$

Unpaired Multimodal NMT In fact, parallel corpora are usually not readily available for low-resource language pairs or domains. Moreover, if these corpora are directly used for training, the linguistic dissimilarity and language mismatch between source language and target language will seriously decrease the translation performance [11]. Fortunately, it is possible to bridge the source and target languages with the multimodal information. The problem can also be called unsupervised multimodal neural machine translation [26]. Assuming that there are a source-language multimodal corpus $\mathcal{D}_{Z,X} = \{\langle Z^{(m)}, X^{(m)} \rangle\}_{m=1}^M$ and a target-language multimodal corpus $\mathcal{D}_{Z,Y} = \{\langle Z^{(n)}, Y^{(n)} \rangle\}_{n=1}^N$ where $\mathcal{D}_{Z,X}$ and $\mathcal{D}_{Z,Y}$ don’t have to overlap, to achieve modeling of source-target NMT, intuitively there are two ways. First, as shown in Figure 1(a), given $\mathcal{D}_{Z,Y}$, we can build a model $P_{\phi_{Z \rightarrow Y}}(Y|Z)$ which can translate an image to a target-language sentence. Thus, for source-language corpus $\mathcal{D}_{Z,X}$, we can translate the images to target-language sentences using $P_{\phi_{Z \rightarrow Y}}(Y|Z)$, forming pseudo source-target

sentence pairs. Thus, source-target translation model can be build with Maximum Likelihood Estimation (MLE) training. This procedure is the same as the *teacher-student* approach in [5], except for replacing the pivot language with image. Similarly, for target-language corpus $\mathcal{D}_{Z,Y}$, we could form pseudo source-target corpus by translating images to source-language sentences with pre-trained image captioning model $P_{\phi_{Z \rightarrow X}}(X|Z)$, then achieve modeling of source-target translation via the *2-agent* communication game [6]. This procedure is shown in Figure 1(b). In these approaches, the mistakes made by the image captioning model would be propagated to the translation model, thus hurt the translation performance. Different from these methods, in this paper, we propose a reinforcement learning approach to learn translation model. Below we formally define the RL training procedure, which is a general learning framework for training NMT model with unpaired multimodal documents only.

3 Methodology

3.1 Problem Definition

In this section, we define the problem of unpaired neural machine translation. On both the source and the target sides, the data comes in the paired form of $(x, z) \in \mathcal{X} \times \mathcal{Z}$ and $(y, z) \in \mathcal{Y} \times \mathcal{Z}$. Here we define two kinds of tasks. **Zero-resource translation:** in this setting, the image that corresponds to x and the image that corresponds to y don't overlap, i.e., not the same image. **Translation with comparable sentences:** in this settings, the source language x and target language y describe the same images. The main purpose is to learn a multi-modal translation model $X \rightarrow Y$ with the help of image Z . Note there is no explicit paired information cross two languages, making it hard to straightforwardly optimize the supervised likelihood. Our method can achieve excellent performance on both of the tasks.

3.2 Overview

In this section, we introduce a more straightforward approach to build NMT system leveraging the property that visual information can be a universal representation to ground different languages. The basic idea is that based on the property of visual information, we can estimate the relevance between source and target sentence by exploiting the relation between sentences and images, and explicitly optimize the estimated relevance. Formally, we formulate the translation task as an RL problem as follows.

Specifically, the NMT model can be viewed as an *agent*, which interacts with the *environment*. The parameters of this agent defines a *policy*, whose execution results in the agent picking an *action* a . In this case, an action refers to generating the next token at each time step. After taking each action, the agent updates its state. A terminal *reward* is received once the agent finished generating a complete sequence \hat{Y} , denoted as $R(\hat{Y})$. Note that the reward $R(\hat{Y})$ is a sentence-level reward, i.e., a scalar for each complete sentence \hat{Y} . Then, the goal of the training is to maximize the expected total reward. As show in Figure 1(c), our approach is different from the teacher-student and 2-agent method since during training we don't need to use image captioning model to translate



Source: A brown dog and a black dog are running
Generated: Ein brauner Hund und ein schwarzer Hund am Strand

Fig. 2. An example of En-De translation from the Multi30K dataset. The figure shows a picture of two dogs. The generated language of translation model matches the image well, but has a serious mismatch with the source language.

images into sentences. Instead, we leverage the alignment information between images and sentences to estimate the relevance between sentences. Thus, the problem of errors produced by the image captioning model propagating and hurting the translation model can be relieved.

3.3 Reward Computation

It is critical to set up appropriate rewards $R(\hat{Y})$ for RL training. In this section, we propose two methods to obtain the reward based on two observations respectively. To be first, the target language is a description of the image, so the generated language and the corresponding image need to be well matched.

(1) The likelihood of the generated sample given the source image. Our first observation is that for an image-sentence pair $\langle Z, X \rangle \in \mathcal{D}_{Z,X}$ and a generated target-language sentence \hat{Y} , if \hat{Y} is a good translations for X , \hat{Y} should have close connection with Z . Based on this observation, if \hat{Y} has a closer connection with Z , it should be a better translation for X . Since an image captioning model could tell the probability of a sentence given an image, we train a target-language image captioning model $P_{\phi_{Z \rightarrow Y}}(Y|Z)$ with corpus $\mathcal{D}_{Z,Y}$. Then, the reward is set as:

$$R1(\hat{Y}) = \log P_{\phi_{Z \rightarrow Y}}(\hat{Y}|Z), \quad (3)$$

The above proposed method can improve the consistency between the generated language and corresponding images, but it also has a serious problem. As shown in Figure 2, the generated translation in German, which means the brown dog and the black dog are running on the beach. It matches the image well, but it has a serious mismatch between the source language. The source language focuses on the dog’s movements, but

the generated language focuses on the location beach, resulting in inconsistent translations. So the attention differences paid to the image areas cause the inconsistency in the translation. To tackle this problem, we propose to use the distance of attention weights of the source and generated language as the reward.

(2) The distance of attention weights from image captioning models. We have observed that for an image-sentence pair $\langle Z, X \rangle \in \mathcal{D}_{Z,X}$ and a generated target-language sentence \hat{Y} , if X and \hat{Y} are translations for each other, \hat{Y} and X should have similar alignment information with Z . Since an image captioning model with attention mechanism can tell the alignment information between images and sentences with attention weights, we compute $R(\hat{Y})$ as the distance of the attention weights obtained from pre-trained image caption models for both languages. Specifically, we represent the source sentence and generated target sentence as $X = (x_1, x_2, \dots, x_S)$ and $\hat{Y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_T)$, where S and T denote the length of source and target sentences respectively. We use pre-trained CNNs [14] for image feature extraction and then denote the image as a matrix $Z = (z_1, z_2, \dots, z_L)$, where each of the L rows consists of a feature vector and the feature vector represents one grid in the image [4]. Then, for X and \hat{Y} , using soft attention computed by image captioning models $P_{\phi_{Z \rightarrow X}}(X|Z)$ and $P_{\phi_{Z \rightarrow Y}}(Y|Z)$, we could obtain the normalized alignment matrix between all the image patches and the target word to be emitted at time step, i.e., the attention weights $A_X^Z = (a_{x,1}^Z, a_{x,2}^Z, \dots, a_{x,S}^Z)$ and $A_{\hat{Y}}^Z = (a_{\hat{y},1}^Z, a_{\hat{y},2}^Z, \dots, a_{\hat{y},T}^Z)$ respectively, where each column $a_{x,s}^Z$ or $a_{\hat{y},t}^Z$ is a L -dimension vector representing the attention vector of the current word. Then, we compute the sum of the attention vector as:

$$\begin{aligned}\alpha_X^Z &= \sum_{s=1}^S a_{x,s}^Z, \\ \alpha_{\hat{Y}}^Z &= \sum_{t=1}^T a_{\hat{y},t}^Z.\end{aligned}\tag{4}$$

α_X^Z and $\alpha_{\hat{Y}}^Z$ are both L -dimension vectors. Since S and T have no guarantee to be equal, to make α_X^Z and $\alpha_{\hat{Y}}^Z$ comparable, we normalize them as $\hat{\alpha}_X^Z$ and $\hat{\alpha}_{\hat{Y}}^Z$ respectively. Then, the reward is computed as:

$$R2(\hat{Y}) = 1 - \text{distance}(\hat{\alpha}_X^Z, \hat{\alpha}_{\hat{Y}}^Z).\tag{5}$$

where *distance* is computed with cosine similarity in this work. Note that here we use a simple and effective method to calculate the attention weights distance. There are many other potential designs. We leave these for future work.

3.4 Objective Function

Given the source-language multimodal corpus $\mathcal{D}_{Z,X} = \{\langle Z^{(m)}, X^{(m)} \rangle\}_{m=1}^M$, the goal of RL training is to maximize the expected reward:

$$\begin{aligned}\mathcal{O}_{RL} &= \sum_{m=1}^M \mathbb{E}_{\hat{Y} \sim P_{\theta}(\hat{Y}|X^{(m)})} R(\hat{Y}) \\ &= \sum_{m=1}^M \sum_{\hat{Y} \in \mathcal{Y}} P_{\theta}(\hat{Y}|X^{(m)}) R(\hat{Y})\end{aligned}\tag{6}$$

where \mathcal{Y} is the space of all candidate translation sentences, which is exponentially large due to the large vocabulary size, making it impossible to exactly maximize \mathcal{O}_{RL} . In practice, REINFORCE [28] is usually leveraged to approximate the above expectation via sampling \hat{Y} from the policy P_θ , leading to the gradient of θ as :

$$\nabla_\theta \mathcal{O}_{RL} = \sum_{m=1}^M R(\hat{Y}) \nabla_\theta \log P_\theta(\hat{Y} | X^{(m)}) \quad (7)$$

Since REINFORCE algorithm suffers from high variance in gradient estimation caused by using single sample \hat{Y} to estimate the expectation, to reduce the variance, we subtract an average reward from the returned reward as in [23].

Algorithm 1 REINFORCE algorithm for multimodal NMT

Require: Initial policy $P_\theta(Y|X)$, source image captioning model $P_{\phi_{Z \rightarrow X}}$ and target image captioning model $P_{\phi_{Z \rightarrow Y}}$ with random weights θ , $\phi_{Z \rightarrow X}$ and $\phi_{Z \rightarrow Y}$ respectively; a reward function $R(\hat{Y})$; monolingual multimodal corpora $\mathcal{D}_{Z,X} = \{\langle Z^{(m)}, X^{(m)} \rangle\}_{m=1}^M$ and $\mathcal{D}_{Z,Y} = \{\langle Z^{(n)}, Y^{(n)} \rangle\}_{n=1}^N$

- 1: Pre-train $P_{\phi_{Z \rightarrow X}}$, $P_{\phi_{Z \rightarrow Y}}$ and $P_\theta(Y|X)$
- 2: Initial delayed policy $P'_{\theta'}$ with the same weight: $\theta' = \theta$
- 3: **repeat**
- 4: Randomly receive an instance $\langle Z, X \rangle \in \mathcal{D}_{Z,X}$
- 5: Generate a sequence of actions \hat{Y} from P'
- 6: Set the reward of the generated sequence as $r = R(\hat{Y})$
- 7: Update policy weight θ using the training objective in Eqn. (9)
- 8: Update delayed policy with a constant γ : $\theta' = \gamma\theta + (1 - \gamma)\theta'$
- 9: **until** model converged

3.5 Training Details

The entire training process is described in Algorithm 1, which consists of two steps. We first pre-train the image captioning and translation models, and then train the translation model with image captioning model fixed via RL. It is important to note that our translation model only pretrains on the same dataset in order to provide an initialization that alleviates the instability of reinforcement learning [29]. We do not leverage additional data.

Specifically, since image captioning models are required when computing reward, we pre-train the captioning models with maximum likelihood estimation leveraging monolingual datasets $\mathcal{D}_{Z,X}$ and $\mathcal{D}_{Z,Y}$. For the translation model, as discussed in [23], the large action space (since the vocabulary is large) makes it extremely difficult to learn with an initial random policy. Thus, we pre-train the policy with some warm-start schemes. Since $\mathcal{D}_{Z,X}$ and $\mathcal{D}_{Z,Y}$ are not guaranteed to overlap, there may be two real-world situations: (1) $\mathcal{D}_{Z,X}$ and $\mathcal{D}_{Z,Y}$ don't overlap. In this scenario, we can pre-train the translation model with the 2-agent approach [6] and teacher-student (shorted as TS) approach [5]; (2) It is also possible in the real-world that each image is annotated with some source-language descriptions and some target-language descriptions, i.e., $\mathcal{D}_{Z,X}$

and $\mathcal{D}_{Z,Y}$ overlap with the same Z . It is worth mentioning that since each sentence is usually generated independently by different people, any source-target pair of descriptions for a given image could be considered a comparable translation pair but not translations of each other. Therefore, it is possible to use this kind of corpora either by considering the cross product of each source and target descriptions [3]. We adopt the cross product of each source and target descriptions as training corpus and use MLE to pre-train the translation model on the same datasets. The MLE objective is as follows:

$$\mathcal{O}_{MLE} = \sum_{i=1}^M \log P(Y^{(m)} | X^{(m)}) \quad (8)$$

After pre-training of image captioning and translation models, we start the RL training process. We apply deep RL techniques [1] by adopting delayed policy with the purpose to prevent divergence. In order to further stabilize the training process, we linearly combine the MLE training objective and RL objective [30, 20] as follows:

$$\mathcal{O}_{COM} = (1 - \alpha) * \mathcal{O}_{MLE} + \alpha * \mathcal{O}_{RL}. \quad (9)$$

Especially, the MLE training objective is the same as the pre-training procedure.

4 Experiments

4.1 Datasets

Our method is evaluated on three publicly available multilingual multimodal datasets, i.e., Multi30K and IAPR-TC12 as in [21, 6], and IKEA dataset [34]. Specifically, Multi30K [7] is a multilingual extension of Flickr30k corpus [32]. It has 29K, 1K and 1K images in the training, validation and test splits respectively with English and German descriptions. We adopt Multi30K task2 corpus in our experiments, which consists of 5 independently collected English and German descriptions per image, i.e., these descriptions are not translations of each other. For this corpus, we evaluate our method on the German-English (De-En) translation task. The IAPR-TC12 dataset [12] has a total of 20K images as well as each image’s multiple English descriptions and the corresponding German translations. Following [6], we use only the first description of each image and split the dataset into training, validation and test sets with 18K, 1K and 1K images respectively. For IAPR-TC12 dataset, we evaluate our approach on both German-English (De-En) and English-German (En-De) tasks. To our knowledge, task2 of Multi30K dataset and IAPR-TC12 dataset only have on language pair of English and German. To better understand the proposed method, we further evaluate our method on the English-German (En-De) and English-French (En-Fr) tasks of IKEA dataset [34]. The data splits are the same as [34].

To fit the situation where source and target multimodal corpora don’t overlap, following [6], we randomly split the images in the training and validation datasets into two parts with equal size. Thus, the two splits have no overlapping images, and we have no direct English- German parallel corpus. The sentences in the datasets are normalized and tokenized with the Moses Toolkit [19]. For Multi30K and IKEA dataset, we adopt

Table 1. BLEU scores on Multi30K German-English translation test set with testing methods Test-1 and Test-2 in zero-resource scenario.

Training strategy	Test-1	Test-2
3-way model	15.9	14.2
UMNMT	19.9	18.2
2-agent-PRE	19.5	18.0
2-agent-JOINT	20.1(+0.6)	18.2(+0.2)
TS-PRE	19.8	19.2
TS-JOINT	20.3(+0.5)	19.5(+0.3)
2-agent-PRE+RL-R1	21.2(+1.7)	19.4(+1.4)
2-agent-PRE+RL-R2	21.5(+2.0)	19.9(+1.9)
TS-PRE+RL-R1	20.9(+1.1)	20.1(+0.9)
TS-PRE+RL-R2	21.1(+1.3)	20.3(+1.1)

joint byte pair encoding [24] with 10K merge operations to reduce vocabulary size. For IAPR-TC12 dataset, we construct the vocabulary with words appearing more than 5 times in the training set and replace the remaining words with UNK.

4.2 Baseline Methods

To demonstrate the effectiveness of our method, we compare our implementations with state-of-the-art baselines as follows.

- (1) *3-way model* [21]. This method adopts end-to-end training strategy and trains the decoder with image and description.
- (2) *UMNMT* [26]. This method is the state-of-the-art zero-resource multimodal neural machine translation method. It adopts Transformer model with controllable attention mechanism that encode both image and language and leverages cycle-consistency loss. For fair comparison, we train the model from the used datasets and do not use the model that was not pre-trained on the tens of millions scale data. We also use the same Transformer architecture with same number of layers and feature dimensions.
- (3) *2-agent-PRE*, *2-agent-JOINT* [6]. In this method, 2-agent-PRE keeps the captioner fixed and only trains the translator until model converges, then 2-agent-JOINT jointly trains the captioner and translator based on 2-agent-PRE.
- (4) *TS-PRE*, *TS-JOINT* [5]. Similarly, TS-PRE keeps the captioner fixed and only trains the translator until model converges, then TS-JOINT jointly trains the captioner and translator based on TS-PRE.

4.3 Implementation Details

To extract image features, we follow the suggestion of [3] and adopt ResNet-50 network [14] pre-trained on ImageNet without fine-tuning. We use the (14,14,1024) feature map of the res4fx (end of Block-4) layer after ReLU. Then, we vectorise this 3-tensor into a 196×1024 matrix.

Table 2. BLEU scores on IAPR-TC12 English-German and German-English translation test sets in zero-resource scenario.

Training strategy	De-En	En-De
3-way model	13.9	8.6
UMNMT	19.3	14.5
2-agent-PRE	18.7	14.4
2-agent-JOINT	19.2(+0.5)	14.6(+0.2)
TS-PRE	17.1	13.9
TS-JOINT	17.5(+0.4)	14.1(+0.2)
2-agent-PRE+RL-R1	20.1(+1.4)	15.6(+1.2)
2-agent-PRE+RL-R2	20.5(+1.8)	15.7(+1.3)
TS-PRE+RL-R1	18.9(+1.8)	15.0(+1.1)
TS-PRE+RL-R2	19.3(+2.2)	15.5(+1.6)

We adopt the Transformer³ model with *base* setting as defined in [27] for all the translation tasks. For pre-training, we consider two situations: (1) when source and target multimodal corpora don't overlap, we pre-train the translation model with 2-agent-PRE and TS-PRE on both datasets in this *zero-resource* scenario. In this scenario, our proposed reward is calculated based on the source image that corresponds to the source sentence; (2) when source and target multimodal corpora overlap with the same images, we pre-train the translation model with MLE on Multi30K dataset with these *comparable sentences*. The optimizer used for MLE is Adam [18], and we follow the same learning rate schedule in [27]. During training, roughly 4, 096 source tokens and 4, 096 target tokens are paired in one mini batch. Each model is trained using a single Tesla K80 GPU. For RL training, the model is initialized with parameters of the pre-trained model, and we continue training it with learning rate 0.0001. Hyper-parameter α is 0.5 and 0.7 for R1 and R2 respectively, and the delay constant γ is 0.1.

For evaluation, all models are quantitatively evaluated with BLEU [22]. Especially, for the Multi30K dataset, since each image is paired with 5 English and 5 German descriptions in the test set, we adopt two methods to evaluate the translation models: (1) We follow the setting in [6], generating a target description for each source sentences and picking the one with the highest probability. The evaluation is performed against the corresponding 5 target descriptions. This method is denoted as *Test-1*; (2) We generate a target description for each 5 source sentences and calculate BLEU score for each generated description against the corresponding 5 target descriptions. Then, we use the average of the calculated BLEU scores as the final result. This method is denoted as *Test-2*. During validation and testing, we set the beam search size to be 5 for the translation model.

4.4 Main Results

We first evaluate our proposed different strategies in comparison with baselines on Multi30K, IAPR-TC12 and IKEA datasets .

³ <https://github.com/tensorflow/tensor2tensor>.

Table 3. BLEU scores on IKEA English-German and English-French translation test sets in zero-resource scenario.

Training strategy	En-De	En-Fr
3-way model	22.1	23.3
UMNMT	33.5	34.7
2-agent-PRE	33.2	34.4
2-agent-JOINT	33.6(+0.4)	34.6(+0.2)
TS-PRE	32.8	34.3
TS-JOINT	33.1(+0.3)	34.4(+0.1)
2-agent-PRE+RL-R1	34.4(+1.2)	35.6(+1.2)
2-agent-PRE+RL-R2	34.7(+1.5)	36.0(+1.6)
TS-PRE+RL-R1	34.2(+1.4)	35.7(+1.4)
TS-PRE+RL-R2	34.6(+1.8)	35.9(+1.6)

Zero-resource translation. We first show the results when the two monolingual multimodal corpora don’t overlap. We first evaluate our method on Multi30K De-En translation task with the testing methods Test-1 and Test-2. The results are shown in Table 1. In this scenario, TS-PRE and 2-agent-PRE are adopted for pre-training translation models before RL procedure. RL-R1 and RL-R2 represent our reinforcement learning method with rewards R1 and R2 respectively. From Table 1, we can see that both the reinforcement learning method with reward R1 and R2 outperform the 2-agent and teacher-student methods across testing methods. Our best performed methods are RL-R2 pre-trained with 2-agent-PRE for Test-1, and RL-R2 pre-trained with TS-PRE for Test-2. These two methods have an improvement of 1.4/0.8 BLEU points over the best baselines.

We also evaluate our method on IAPR-TC12 En-De and De-En translation tasks. The results are shown in Table 2. We can see that our proposed method also outperforms all the baseline approaches on both translation tasks. Specifically, our best methods have an improvement of 1.3/1.1 BLEU points on De-En/En-De translation over the best baselines.

Similarly, for the results on IKEA En-De and En-Fr translation tasks shown in Table 3, our proposed method achieves superior performance compared with baselines. Our method can obtain obvious improvement on both language pairs of En-De and En-Fr.

From the results above, we can have the following findings. First, our approach achieves the best results on all three data sets, substantially exceeding the baselines. It demonstrates the effectiveness of our method. Then we can find that our proposed R2 works better than R1 in all cases. It verifies the importance of modeling the distance of attention weights of source and generated sentences into translation models. Last, the state-of-the-art UMNMT method has achieved only a small improvement. On the one hand, it introduces too many parameters in the training process, which makes the model difficult to train. Moreover, it shows the superiority of using reinforcement learning to reinforce the consistency between images and sentences.

Translation with comparable sentences. We also evaluate our method when the two monolingual multimodal corpora overlap with the same images on Multi30K dataset.

Table 4. Comparison with previous work with comparable sentences over the Multi30K test set with Test-1 and Test-2.

Training strategy	Test-1	Test-2
PRE	30.4	27.5
PRE + 2-agent-JOINT	30.2	27.1
PRE + TS-JOINT	30.1	26.9
PRE + RL-R1	32.1	29.1
PRE + RL-R2	32.8	29.5



Reference image	Source (German)	Target (English)
	<ol style="list-style-type: none"> 1. ein junges mädchen beim laufen . 2. ein mädchen rennt in einem abgesperrten bereich entlang . 3. eine läuferin läuft . 4. das mädchen rennt in einem wettspiel . 5. spielende kinder auf einer wiese neben einer sandbahn , ein mädchen läuft entlang der bahn 	<p>Ref 1: a young girl in dark shorts and a blue tank top runs on the grass near an orange cone and tape .</p> <p>Ref 2: a young lady wearing blue and black is running past an orange cone .</p> <p>Ref 3: a young woman running by an orange ribbon .</p> <p>Ref 4: a young girl running by herself in a park .</p> <p>Ref 5: a girl in a blue tank top winning a race .</p> <p>2-agent-PRE: a woman running in a race .</p> <p>RL-R1: a woman in black is running in a race .</p> <p>RL-R2: a woman in a blue shirt is running in a race .</p>
	<ol style="list-style-type: none"> 1. ein junger mann mit schutzhelm klettert eine felswand herauf . 2. bergsteiger in weißen helm hält sich an einer lila seil . 3. ein mann beim klettern mit seil gesichert . 4. der mann mit dem weißen helm im klettergeschirr klettert an dem fels . 5. ein bergkletterer übt an einer steinwand in geringer höhe das klettern . 	<p>Ref 1: a man in a white shirt and helmet is using climbing equipment .</p> <p>Ref 2: a man with a white shirt and helmet is climbing a mountain .</p> <p>Ref 3: a young man wearing a white helmet climbing up a rock wall</p> <p>Ref 4: a man in a harness climbing a rock wall</p> <p>Ref 5: man rock climbing looking up the rock</p> <p>2-agent-PRE</p> <p>Hyp 1: a man is walking up a mountain .</p> <p>Hyp 2: a man in a red shirt is rock climbing .</p> <p>Hyp 3: a man is hanging from a pole .</p> <p>Hyp 4: a man in a white hard hat is climbing a mountain .</p> <p>Hyp 5: a man in a red shirt is rock climbing .</p> <p>RL-R2</p> <p>Hyp 1: a man in a white hard hat is standing on a mountain .</p> <p>Hyp 2: a man in a white shirt is rock climbing .</p> <p>Hyp 3: a man in a white shirt is hanging from a pole .</p> <p>Hyp 4: a man in a white hard hat is standing on a rock .</p> <p>Hyp 5: a man in a white shirt is rock climbing .</p>

Fig. 3. Examples of translations from the Multi30K test set. The first example is translated using Test-1 as [6], while the second example is translated using Test-2. For the first example, we pre-train the translation model with 2-agent-PRE, then continue training with RL-R1 and RL-R2. For the second example, we only show the translation results of the 2-agent-PRE and the better performed RL-R2. We highlight the words that distinguish the systems’ results in **blue**, **red** and **green**. **Red** words are marked for correct translations in hypotheses compared with **blue** words in references, and **green** words are marked for incorrect translations in hypotheses compared to **blue** words.

The translation model is pre-trained with MLE on the cross product of each source and target sentences, which is denoted as PRE. The BLEU scores of different training strategies are shown in Table 4. We can see that our method with both rewards achieves obvious improvement over MLE pre-training, while the 2-agent and teacher-student methods can’t gain any improvements in this scenario. The possible reason is that the generated sentences by the image captioning model in the baseline methods are low-quality compared two the cross product of each source and target sentences, thus can’t help the translation model to get better performance.

Table 5. BLEU scores for different α on De-En translation Multi30K test set for RL-R1.

α	0.1	0.3	0.5	0.7	0.9
test1	30.9	31.7	32.1	31.2	30.7
test2	27.8	28.5	29.1	28.1	27.9

Table 6. BLEU scores for different α on De-En translation Multi30K test set for RL-R2.

α	0.1	0.3	0.5	0.7	0.9
test1	30.4	30.7	32.1	32.8	30.9
test2	27.7	27.8	28.7	29.5	28.0

4.5 Impact of Hyper-parameter

As shown in Eqn. (9), the hyper-parameter α controls the trade-off between MLE and RL objectives. To show the impact of this hyper-parameter, we evaluate the model performance on De-En Multi30K test set with different α in the scenario of corpora overlapping. Specifically, for RL-R1 and RL-R2, we set α both to be [0.1, 0.3, 0.5, 0.7, 0.9] in our experiments. The results are presented in Table 5 and Table 6. We find that when α is set to be 0.5 and 0.7 for RL-R1 and RL-R2 respectively, our method achieves the best performance.

4.6 Case Study

In Figure 3, we provide some qualitative comparisons between the translations from the pre-training method 2-agent-PRE and our RL method. In the first example, our RL-R1 properly translates the words “in back” and RL-R2 properly translates the words “in a blue shirt”, while PRE didn’t tell anything about the wearing. In the second example, each translation of our RL-R2 correctly translates “white shirt” or “white hard hat”, while the translations of PRE tends to include the wrong words “red shirt” or include nothing about the wearing. From these examples, we can see that our RL method can guide the translation model to focus more on the image, thus correctly translate the information in the image and improve translation quality.

5 Related Work

The related research topics to our concerns can be classified into the following three categories: (1) multimodal neural machine translation, (2) pivot-based neural machine translation and (3) reinforcement learning for sequence prediction.

Multimodal Neural Machine Translation. The aim of this task is to use images as well as parallel corpora to improve the translation performance. It has been shown that image modality can benefit NMT by relaxing ambiguity in alignment that cannot be solved by texts only [4, 15, 34]. This task is much easier than ours because in its 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 [6]. The unsupervised multimodal NMT is proposed in [26]. However, they did not consider the consistency of

the image and language at the sentence level. Their introduction of too many parameters also hurts training.

Pivot-based Neural Machine Translation. Another line of work has been to train NMT system from non-parallel data with the help of another modality, which is called the pivot-based machine translation. Specifically, researchers have tried to build multilingual NMT systems trained by other language pairs to enable translation [8, 16] with non-parallel data for the intended translation pair. In addition to the multilingual methods, several authors proposed to train the translation model in more direct ways. For example, [5] proposed a teacher-student framework under the assumption that parallel sentences have close probabilities of generating a sentence in a third language. [33] maximized the expected likelihood to train the intended source-to-target model. Nonetheless, all these methods still require that source-pivot and pivot-target parallel corpora are available. Besides languages, images are also used as the pivot to build NMT systems. Zero-resource NMT by utilizing image as a pivot was first achieved by training multimodal encoders to share common semantic representation [21]. To overcome the bottleneck of the fixed-length vector in this method, [6] proposed a 2-agent approach which jointly trains the translation and image captioning model.

Reinforcement Learning for Sequence Prediction. In sequence prediction task, reinforcement learning is always used to learn and refine model parameters according to task-specific reward signals [29, 13]. In [23], the authors proposed to train a neural translation model with the objective of optimizing the sentence-level BLEU score. [25] proposed to adopt minimum risk training to minimize the task specific loss on NMT training data. Instead of the REINFORCE algorithm used in the above two works, [1] further optimize the policy by actor-critic algorithm.

6 Conclusion and Future Work

In this work, to tackle the challenging task of training an NMT system from just unpaired multimodal data, we successfully deploy a reinforcement learning (RL) method to build NMT system by introducing a sequence-level supervision signal as reward. Experiments on German-English, English-German and English-French translation over the IAPR-TC12, Multi30K and IKEA datasets show that our proposed reinforcement learning mechanism can significantly outperform the existing methods.

In the future, we will continue trying to ground the visual context into translation model, such as using a shared encoder over source, target sentences and image, and then constraining them to be similar in order to constrain both the source and target representations to be faithful to the image. Moreover, we also would like to better understand the proposed method on larger training corpora and alternative language pairs.

References

1. Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., Bengio, Y.: An actor-critic algorithm for sequence prediction (2017)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Third International Conference on Learning Representations (2015)

3. Caglayan, O., Aransa, W., Wang, Y., Masana, M., Garcia-Martinez, M., Bougares, F., Barrault, L., Van de Weijer, J.: Does multimodality help human and machine for translation and image captioning? In: First Conference on Machine Translation. vol. 2, pp. 627–633 (2016)
4. Calixto, I., Liu, Q., Campbell, N.: Doubly-attentive decoder for multi-modal neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1913–1924 (2017)
5. Chen, Y., Liu, Y., Cheng, Y., Li, V.O.: A teacher-student framework for zero-resource neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1925–1935 (2017)
6. Chen, Y., Liu, Y., Li, V.O.: Zero-resource neural machine translation with multi-agent communication game. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
7. Elliott, D., Frank, S., Sima'an, K., Specia, L.: Multi30k: Multilingual english-german image descriptions. In: Proceedings of the 5th Workshop on Vision and Language. pp. 70–74 (2016)
8. Firat, O., Sankaran, B., Al-Onaizan, Y., Vural, F.T.Y., Cho, K.: Zero-resource translation with multi-lingual neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 268–277 (2016)
9. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning—Volume 70. pp. 1243–1252. JMLR. org (2017)
10. Gella, S., Sennrich, R., Keller, F., Lapata, M.: Image pivoting for learning multilingual multimodal representations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2839–2845 (2017)
11. Graça, Y.K.M., Ney, H.: When and why is unsupervised neural machine translation useless? In: 22nd Annual Conference of the European Association for Machine Translation. p. 35 (2020)
12. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In: International workshop ontoImage. vol. 2 (2006)
13. Hashimoto, K., Tsuruoka, Y.: Accelerated reinforcement learning for sentence generation by vocabulary prediction. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3115–3125 (2019)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hitschler, J., Schamoni, S., Riezler, S.: Multimodal pivots for image caption translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 2399–2409 (2016)
16. 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. Transactions of the Association for Computational Linguistics **5**, 339–351 (2017)
17. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Third International Conference on Learning Representations (2015)
19. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational

- linguistics companion volume proceedings of the demo and poster sessions. pp. 177–180 (2007)
20. Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., Jurafsky, D.: Adversarial learning for neural dialogue generation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2157–2169 (2017)
 21. Nakayama, H., Nishida, N.: Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation* **31**(1-2), 49–64 (2017)
 22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
 23. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. In: Fourth International Conference on Learning Representations (2016)
 24. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1715–1725 (2016)
 25. Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., Liu, Y.: Minimum risk training for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1683–1692 (2016)
 26. Su, Y., Fan, K., Bach, N., Kuo, C.C.J., Huang, F.: Unsupervised multi-modal neural machine translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10482–10491 (2019)
 27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
 28. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* **8**(3-4), 229–256 (1992)
 29. Wu, L., Tian, F., Qin, T., Lai, J., Liu, T.Y.: A study of reinforcement learning for neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3612–3621 (2018)
 30. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
 31. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)
 32. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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 (2014)
 33. Zheng, H., Cheng, Y., Liu, Y.: Maximum expected likelihood estimation for zero-resource neural machine translation. In: IJCAI. pp. 4251–4257 (2017)
 34. Zhou, M., Cheng, R., Lee, Y.J., Yu, Z.: A visual attention grounding neural model for multimodal machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3643–3653 (2018)