From Words to Sentences: A Progressive Learning Approach for Zero-resource Machine Translation with Visual Pivots

Shizhe Chen[1*], **Qin Jin**[1†] and **Jianlong Fu**[2]
[1]Renmin University of China, Beijing, P.R. China
[2]Microsoft Research Asia, Beijing, P.R. China
{cszhe1, qjin}@ruc.edu.cn, jianf@microsoft.com

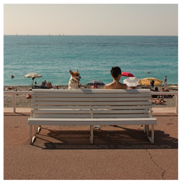
Abstract

The neural machine translation model has suffered from the lack of large-scale parallel corpora. In contrast, we humans can learn multi-lingual translations even without parallel texts by referring our languages to the external world. To mimic such human learning behavior, we employ images as pivots to enable zero-resource translation learning. However, a picture tells a thousand words, which makes multi-lingual sentences pivoted by the same image noisy as mutual translations and thus hinders the translation model learning. In this work, we propose a progressive learning approach for image-pivoted zero-resource machine translation. Since words are less diverse when grounded in the image, we first learn word-level translation with image pivots, and then progress to learn the sentence-level translation by utilizing the learned word translation to suppress noises in image-pivoted multi-lingual sentences. Experimental results on two widely used image-pivot translation datasets, IAPR-TC12 and Multi30k, show that the proposed approach significantly outperforms other state-of-the-art methods.

1 Introduction

The recent success of neural machine translation (NMT) [Bahdanau *et al.*, 2015] has greatly benefited from large-scale high-quality parallel corpora. However, such NMT models are data-hungry and perform poorly without sufficient parallel data [Zoph *et al.*, 2016]. Due to the high expense of collecting parallel texts, more and more researchers are paying attention to develop NMT models under the zero-resource condition where no parallel source-target texts are available.

Inspired by how we humans learn a novel language when no parallel texts are available, for example connecting sentences in two languages that describe the same image, researchers have proposed to employ images as pivots for zero-resource machine translation, which can benefit from abundant images with mono-lingual descriptions on the Internet [Sharma *et al.*, 2018]. The shared principle in previous image-pivoted works [Nakayama and Nishida, 2017; Lee *et al.*, 2018; Chen *et al.*, 2018] assumes that the source sentence and the target sentence are semantically equivalent as they are describing the same image.



Groundtruth English Caption:
a woman in a swimsuit sitting on a bench with a dog.

Groundtruth German Caption:
die frau und der hund sitzen auf der bank vor dem meer.
(the woman and the dog sitting on the bench in front of the sea.)

Generated German Caption:
ein mann und eine frau sitzen auf einer bank am strand.
(a man and a woman are sitting on a bench at the beach.)

Figure 1: A picture tells a thousand words - illustration of the challenge in image-pivoted zero-resource machine translation. Only few words in red are correct translations even in groundtruth captions of the same image. Captioning mistakes in blue can further make such multi-lingual captions noisy as mutual translation. We translate German captions in brackets for non-German readers.

However, due to the one-to-many relationship between images and captions (as known as *"a picture tells a thousand words"*), multi-lingual sentences describing the same image are not necessarily good mutual translations. As shown in Figure 1, although the captions in English and German both accurately describe the image, they capture different aspects in the image and only few words in the captions are correct mutual translations. Moreover, since images with multi-lingual captions are hard to obtain in realistic settings, an image caption model is usually adopted to generate multi-lingual caption sentences for the image. However, due to the imperfection of caption models, the semantic discrepancy of generated multi-lingual captions could be even larger.

In this work, we propose a progressive learning approach to overcome above challenges in the image-pivoted zero-resource machine translation. We propose to learn the translation in an easy-to-advanced progressive way, by firstly grasping the word-level translation with the help of image pivots and then progressing to learn more challenging sentence-level translation with the assistance of word translation and image pivots. To be specific, since words grounded in certain regions of the image are less diverse than sentences, word-level translation can be more effectively learned based on image pivots. The multi-lingual word representations learned

[*] This work was performed when Shizhe Chen was visiting Microsoft Research Asia as a research intern.

[†] Qin Jin is the corresponding author.

from the word translation and the image pivots altogether are used to enhance the sentence translation from two aspects: i) suppressing noises in image-pivoted multi-lingual sentences via re-weighting sentences at fine-grained token-level; ii) supporting the learning of language-agnostic sentence representation for cross language decoding via auto-encoding. The two aspects are complementary to train the NMT model. We carry out extensive experiments on two benchmark image-pivot machine translation datasets: IAPR-TC12 and Multi30k. Our proposed approach significantly outperforms other state-of-the-art image-pivot methods.

2 Related Work

The encoder-decoder based neural machine translation (NMT) model [Cho *et al.*, 2014; Bahdanau *et al.*, 2015] has achieved great success in recent years. However, it requires large-scale parallel texts for training, which performs poorly without sufficient data [Zoph *et al.*, 2016; Castilho *et al.*, 2017]. There are mainly three types of methods to avoid the reliance on source-target parallel data, namely third-language pivot, mono-lingual based and visual pivot methods.

The third-language pivot methods [Johnson *et al.*, 2017; Chen *et al.*, 2017] demand the source-to-pivot and pivot-to-target parallel corpus to enable zero-resource translation from source to target. [Johnson *et al.*, 2017] trains a universal encoder-decoder with multiple language pairs, which can perform well in novel language combinations. However, it is not trivial to obtain the pivot language parallel data.

The recent mono-lingual based methods [Artetxe *et al.*, 2018; Lample *et al.*, 2018a] only utilize large-scale mono-lingual corpora for translation. [Lample *et al.*, 2018b] summarizes three key elements for mono-lingual based methods: careful initialization, strong language model, and back-translation, which achieved promising results for both phrase-based and NMT models. The reason we use image-pivot for translation is that images can help reduce ambiguity of texts especially for visual-related sentences such as commercial product descriptions [Zhou *et al.*, 2018; Calixto *et al.*, 2017].

The image-pivot approaches leverage images to connect unpaired source and target languages. [Kiela *et al.*, 2015; Chen *et al.*, 2019] have shown the effectiveness of image pivots for bilingual lexicon induction. [Su *et al.*, 2018] follow mono-lingual based methods but utilize images to enhance decoding performance. The 3-way model [Nakayama and Nishida, 2017] maps source sentences and images into common space and employs an image-to-target caption model for translation. However, it cannot embrace attention mechanism and results in noisy translations since information in images and sentences is not equal. [Chen *et al.*, 2018; Lee *et al.*, 2018] propose a multi-agent communication game with a captioner and translator. The captioner generates source sentence to describe an image, and the translator is trained to maximize rewards from relevancy between image and translated target sentence. [Chen *et al.*, 2018] utilizes log-likelihood as the reward while [Lee *et al.*, 2018] utilizes image retrieval performance. However, since captions related to images are not necessarily good mutual translations, such learning approaches also suffer from noisy rewards. In this work, we propose to suppress such noises and progressively learn the translation in an easy-to-advanced way.

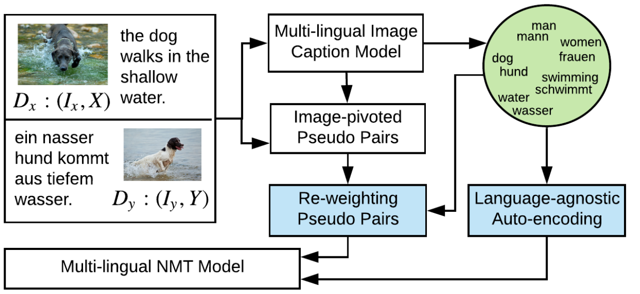

the dog walks in the shallow water.
$D_x : (I_x, X)$
ein nasser hund kommt aus tiefem wasser.
$D_y : (I_y, Y)$
Multi-lingual Image Caption Model
Image-pivoted Pseudo Pairs
Re-weighting Pseudo Pairs
Language-agnostic Auto-encoding
Multi-lingual NMT Model
man
mann
women
frauen
dog
hund
swimming
schwimmt
water
wasser

Figure 2: The overall progressive learning framework for image-pivoted zero-resource machine translation. We firstly learn word translations from image pivot in the green module and then we advance to more challenging sentence translation in blue modules.

3 The Proposed Approach

The goal of zero-resource machine translation is to learn source to target translation without any source-target sentence pairs. We propose to utilize images as pivots to enable zero-resource machine translation. Assume we have two mono-lingual image caption datasets: $D_x = \{(I_x^{(i)}, X^{(i)})\}_{i=1}^{N_x}$ in the source language and $D_y = \{(I_y^{(i)}, Y^{(i)})\}_{i=1}^{N_y}$ in the target language, where I denotes the image, and captions X and Y consist of word sequences $\{x_1, \cdots, x_{T_x}\}$ and $\{y_1, \cdots, y_{T_y}\}$ respectively. We omit the superscript i for simplicity hereinafter. The image sets I_x and I_y do not overlap, which means that an image has only one caption, either in source language or target language. The images are used as pivots during the training stage, but are not involved during the test stage.

Figure 2 illustrates the proposed progressive learning approach for image-pivoted zero-resource machine translation. Firstly, from the mono-lingual image caption datasets D_x and D_y, we can build image caption models $f_{i \to x}$ and $f_{i \to y}$ to translate an image into sentence descriptions in source and target languages respectively. Therefore, for each image $I_x \in D_x$, we can obtain triplet captions $(X, \tilde{X}, \tilde{Y})$ where X refers to the groundtruth caption, $\tilde{X}$ and $\tilde{Y}$ refer to the generated captions from the image caption models $f_{i \to x}$ and $f_{i \to y}$ in source and target languages respectively. Similarly, for image $I_y \in D_y$, we can obtain triplet captions $(Y, \tilde{X}, \tilde{Y})$. We then induce $(X, \tilde{Y})$, $(\tilde{X}, Y)$ and $(\tilde{X}, \tilde{Y})$ from above triplets as image-pivoted source-target pseudo sentence pairs, which can be utilized to train NMT models.

However, since the diversity of descriptive sentences is quite large, such pseudo pair may not be precisely semantic matched, which would greatly hinder the translation performance. In order to suppress noises in pseudo pairs, we propose a progressive learning approach which firstly learns word translation from image pivots. Since the diversity of words is better constrained when the word is grounded in image regions, word translation can be more effectively learned with the image-pivot approach. The word translation encodes multi-lingual words into a common semantic space, which

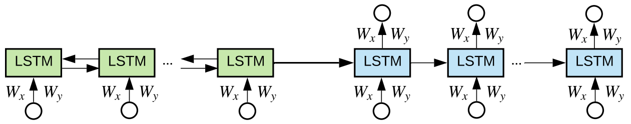


Figure 3: The structure of the NMT model. The encoder and decoder both contain source and target embedding matrices W_x and W_y, so we can encode and decode source and target sentences in one model.

is then utilized to assist image-pivoted sentence translation from two aspects: 1) to re-weight image-pivoted pseudo pairs and 2) to learn language-agnostic sentence representations. In such progressive manner, the NMT model is able to alleviate noises from image-pivot learning.

In the following, we will firstly describe the structure of NMT model in Section 3.1, and then introduce the progressive learning strategy to train the NMT model with image pivots in details in Section 3.2.

3.1 Neural Machine Translation Model

We utilize a shared multi-lingual encoder-decoder architecture as our NMT model, which can perform source-to-target and target-to-source translation in one model. Given the source sentence $X = \{x_1, ..., x_{T_x}\}$ and the target sentence $Y = \{y_1, ..., y_{T_y}\}$, the encoder converts the source sentence into a sequence of vectors, and then the decoder sequentially predicts target word y_t conditioning on the encoded vectors and previously generated words $y_{<t}$.

Specifically, our encoder is a bi-directional LSTM [Hochreiter and Schmidhuber, 1997] with a source word embedding matrix W_x and a target word embedding matrix W_y. The source and target sentences share the same encoder parameters except the word embedding matrix as follows:

$$z_t^x = \text{biLSTM}(W_x x_t, z_{t-1}^x, z_{t+1}^x; \Theta_e) \quad (1)$$

$$z_t^y = \text{biLSTM}(W_y y_t, z_{t-1}^y, z_{t+1}^y; \Theta_e) \quad (2)$$

where Θ_e are parameters of the bi-directional LSTM.

The decoder is a LSTM with both source word embedding W_x and target word embedding W_y. Suppose $z = \{z_1, \cdots, z_T\}$ is the encoded input sentence, the decoder can translate z into the source sentence X by:

$$p(x_t|x_{<t}, z) = \text{softmax}(W_x h_t) \quad (3)$$

$$h_t = \text{LSTM}([W_x x_{t-1}, c_t], h_{t-1}; \Theta_d) \quad (4)$$

where Θ_d are parameters in the decoder LSTM, h_0 is initialized as z_T, $[\cdot]$ is the vector concatenation operation and c_t is a context vector to employ relevant input vector z_i to predict the target word x_t via attention mechanism [Bahdanau *et al.*, 2015]. The computation of c_t is formulated as:

$$c_t = \sum_{i=1}^{T} a_{i,t} z_i \quad (5)$$

$$a_{i,t} = \frac{\exp(f_a([h_{t-1}, z_i]))}{\sum_j \exp(f_a([h_{t-1}, z_j]))}, \quad (6)$$

where f_a is a feed forward neural network to compute the attention weight a_i for each h_i. Therefore, the probability of generating X conditioning on z is:

$$p(X|z) = \prod_{t=1}^{T_x} p(x_t|x_{<t}, z) \quad (7)$$

Similarly, the decoder can also translate the input z into the target sentence Y by replacing W_x in Eq (3) and Eq (4) with W_y. Figure 3 presents the structure of the NMT model.

3.2 Progressive Learning

In this section, we describe in details the progressive learning procedure, from learning the word-level translation to more challenging sentence-level translation.

Learning Word-level Translation

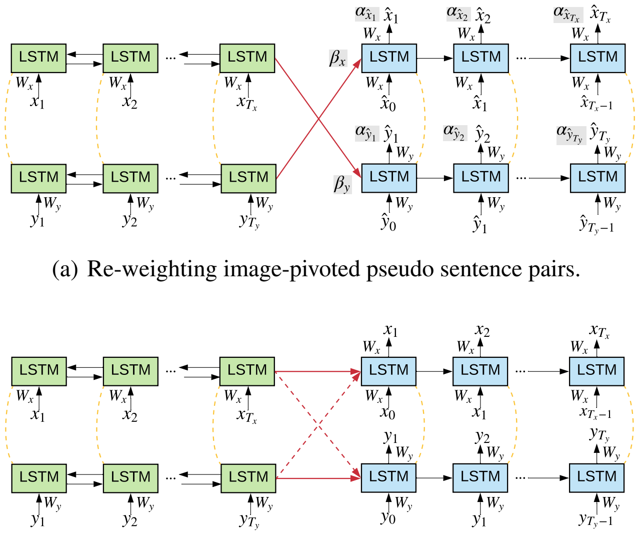
In order to translate multi-lingual words, similar to work in [Chen *et al.*, 2019], we build a multi-lingual image caption model to encode the source and target words into a joint semantic space. The multi-lingual image caption model is based on the encoder-decoder framework, where the encoder converts an image into a set of visual features, and the decoder generates sentences conditioning on the visual features with attention mechanism. For image captioning in the source and target languages, the encoder and decoder are shared except the word embedding matrices. Therefore, the word embedding matrices of the source and target languages are enforced to be in a common space constrained by the image pivots. We denote the learned word embedding matrices for the source and target words as W_x and W_y, which are employed in our NMT model and remain fixed.

Re-weighting Image-pivoted Pseudo Sentence Pairs

We employ above learned W_x, W_y to assess the qualities of image-pivoted pseudo sentence pairs. We formulate the semantic distance of the sentence pair as a special case of Earth Mover's Distance (EMD) [Rubner *et al.*, 2000; Kusner *et al.*, 2015], which is to find the minimal transportation solution from one sentence to another. For source sentence $X = \{x_1, \cdots, x_{T_x}\}$ and target sentence $Y = \{y_1, \cdots, y_{T_y}\}$, their EMD distance d is:

$$\begin{aligned} d = \min_{\mathrm{A} \geq 0} \quad & \sum_{i,j} \mathrm{A}_{i,j}\, D_{i,j} \\ \text{subject to:} \quad & \sum_{j} \mathrm{A}_{i,j} = 1/T_x,\ i = 1, \ldots, T_x \\ & \sum_{i} \mathrm{A}_{i,j} = 1/T_y,\ j = 1, \ldots, T_y \end{aligned} \quad (8)$$

where $D_{i,j}$ is the cosine distance of x_i and y_j based on W_x and W_y. The optimal matrix A can be solved efficiently via an off-the-shelf toolkit proposed by [Pele and Werman, 2009], which can be viewed as words alignment in the pseudo pair. Besides the sentence distance d, we could also derive token-level distances d_{x_i} and d_{y_j} for x_i and y_j respectively based on the optimal A, where $d_{x_i} = \sum_j \mathrm{A}_{i,j}\, D_{i,j}$ and $d_{y_j} = \sum_i \mathrm{A}_{i,j}\, D_{i,j}$. Based on the sentence-level and token-level distance, we propose to re-weight the i-th pseudo pair


LSTM
W_x
W_y
x_1
x_2
x_{T_x}
y_1
y_2
y_{T_y}
$\alpha_{\hat{x}_1}$ $\hat{x}_1$
$\alpha_{\hat{x}_2}$ $\hat{x}_2$
$\alpha_{\hat{x}_{T_x}}$ $\hat{x}_{T_x}$
β_x
β_y
$\hat{x}_0$
$\hat{x}_{T_x-1}$
$\alpha_{\hat{y}_1}$ $\hat{y}_1$
$\alpha_{\hat{y}_2}$ $\hat{y}_2$
$\alpha_{\hat{y}_{T_y}}$ $\hat{y}_{T_y}$
$\hat{y}_0$
$\hat{y}_1$
$\hat{y}_{T_y-1}$
x_0
x_{T_x-1}
y_0
y_{T_y-1}

(a) Re-weighting image-pivoted pseudo sentence pairs.

(b) Language-agnostic auto-encoding.

Figure 4: Training of the NMT model. The red solid lines denote translation in training; red dashed lines denote implicitly learned cross-language translation; yellow dashes lines denote tied weights.

at both sentence-level and fine-grained token-level via the inverse distance weight as follows:

$$\alpha_{x_t}^{(i)} = \frac{1}{(1 + d_{x_t}^{(i)} - \min_k d_{x_k}^{(i)})^{\lambda_{token}}} \quad (9)$$

$$\alpha_{y_t}^{(i)} = \frac{1}{(1 + d_{y_t}^{(i)} - \min_k d_{y_k}^{(i)})^{\lambda_{token}}} \quad (10)$$

$$\beta^{(i)} = \frac{1}{(1 + d^{(i)} - \min_k d^{(k)})^{\lambda_{sent}}} \quad (11)$$

where $\lambda_{token}, \lambda_{sent}$ are hyper-parameters to punish noisy tokens and sentences. So $\alpha_{x_t}^{(i)}$ and $\alpha_{y_t}^{(i)}$ represent the relative importance of token x_t and y_t in the source and target sentence to form a good mutual translation pair, and $\beta^{(i)}$ is the relative quality of pair i among all pseudo pairs. Figure 4(a) illustrates the re-weighting training process. Therefore, the objective function for image-pivoted sentence pairs is:

$$\begin{aligned} L_{pivot} = &-\sum_{i=1}^{N} \beta^{(i)} \sum_{t=1}^{T_y} \alpha_{y_t}^{(i)} \log p(y_t^{(i)} | y_{<t}^{(i)}, z^{x^{(i)}}) \\ &-\sum_{i=1}^{N} \beta^{(i)} \sum_{t=1}^{T_x} \alpha_{x_t}^{(i)} \log p(x_t^{(i)} | x_{<t}^{(i)}, z^{y^{(i)}}) \end{aligned} \quad (12)$$

Language-agnostic Auto-encoding

We propose to employ auto-encoding to learn a language-agnostic sentence representation for cross language translation as illustrated in Figure 4(b). We fix source and target word embedding matrices in our NMT model with W_x and W_y which are in the common space, so that the sentence representation encoded by the shared encoder are constrained in a common latent space. Since it is trivial to auto-encode a sentence, we apply the corruption operation $C(\cdot)$ on the original sentence as [Lample *et al.*, 2018a], which includes word order jitter, word insertion and deletion. The corrupted sentence is used as the input and the NMT model is trained to reconstruct the original sentence from the corrupted one. We apply the denoising auto-encoding for both source and target languages, so the objective is to minimize:

Pairs	Train I-En	Train I-De	Val I-En	Val I-De	Test En-De
IAPR-TC12	9k	9k	500	500	1k
Multi30k	14.5k	14.5k	507	507	1k

Table 1: Data splits for the IAPR-TC12 and Multi30k datasets.

$$\begin{aligned} L_{ae} = &-\sum_{i=1}^{N_x} \log p(X^{(i)} | C(X^{(i)})) \\ &-\sum_{i=1}^{N_y} \log p(Y^{(i)} | C(Y^{(i)})) \end{aligned} \quad (13)$$

The full objective function to train the NMT model is:

$$L = L_{pivot} + \lambda L_{ae} \quad (14)$$

where λ is the weight to balance the two losses.

4 Experiments

4.1 Experimental Setup

We utilize two benchmark image-pivot translation datasets IAPR-TC12 and Multi30k. The IAPR-TC12 dataset [Grubinger *et al.*, 2006] contains 20K images and each image is annotated with multi-sentences in English and its translation in German. We follow [Chen *et al.*, 2018] to use the first sentence since it describes the most salient image content. We randomly select 18K images for training, 1K for validation and 1K for testing. Since in the realistic image-pivot setting, images in different languages are mostly non-overlapped, we randomly split the training and validation set into two parts of equal size. One part is constructed only with image-English pairs and the other only with image-German pairs. The Multi30k dataset [Elliott *et al.*, 2016] contains 30K images with two task annotations, one for machine translation and the other for multi-lingual captioning. We utilize the former task annotation, where each image is annotated with one English description and its German translation. We follow the standard split in [Nakayama and Nishida, 2017] with 29K, 1,014 and 1K in the training, validation and test sets respectively. We also apply the similar non-overlapping split operation as in IAPR-TC12 to simulate the non-overlap setting. Table 1 presents the data split in our experiments.

We use Moses SMT Toolkit [Koehn *et al.*, 2007] to normalize and tokenize descriptions. For IAPR-TC12 dataset, we keep words appeared more than 3 times, which results in 1,621 words for English and 2,102 words for German. For Multi30K dataset, we employ a joint byte pair (BPE) [Sennrich *et al.*, 2016] with 10k merge operations, which results in 5,202 tokens for English and 7,065 tokens for German.

		IAPR-TC12		Multi30k	
		De-En	En-De	De-En	En-De
3-way model		13.9	8.6	8.4	8.0
Multi-agents[1]		18.6	14.2	-	-
Emergent model[2]		-	-	6.5	7.4
S-txt-img[2]		-	-	7.5	7.7
ours	L_{pivot}	38.0	30.3	9.9	8.4
	L_{ae}	58.9	39.8	21.9	17.6
	$L_{pivot} + L_{ae}$	**61.3**	**47.1**	**23.0**	**18.3**

Table 2: The BLEU4 performance of different methods for image-pivoted zero-resource machine translation.

		IAPR-TC12		Multi30k	
		De-En	En-De	De-En	En-De
L_{pivot}	w/o	31.0	25.5	8.5	7.7
	with	**38.0**	**30.3**	**9.9**	**8.4**
L_{ae}	w/o	39.6	26.8	7.9	6.8
	with	**58.9**	**39.8**	**21.9**	**17.6**

Table 3: Translation performance comparison of without and with the progressive learning approach for different sentence losses.

For the multi-lingual image caption model, we leverage the Resnet152 pretrained on the ImageNet [He *et al.*, 2016] as the encoder and a single-layer LSTM with 512 hidden units as the decoder. We utilize beam search with beam width of 5 to generate one description for each image. For the NMT model, the encoder is a one-layer bidirectional LSTM with 256 hidden units and the decoder is a one-layer LSTM with 512 hidden units. We set hyper-parameters $\lambda_{token} = 10, \lambda_{sent} = 5$ and $\lambda = 1$ based on validation performance. We utilize the Adam algorithm to train models with learning rate of 0.0005 and batch size of 64. The best model is selected by the loss on validation set. We evaluate the machine translation performance with BLEU4 metric [Papineni *et al.*, 2002].

4.2 Comparison with State-of-the-Art Methods

We compare the proposed progressively learned NMT model with state-of-the-art image-pivoted models as follows:

1. 3-way model [Nakayama and Nishida, 2017]: It utilizes an target image caption model to translate a learned modality-agnostic feature of the source sentence.

2. Multi-agents [Chen *et al.*, 2018]: The NMT model is trained to translate a generated source caption of I_y with image-relevance rewards from the log probability of predicting groundtruth Y.

3. Emergent model [Lee *et al.*, 2018]: Similar to [Chen *et al.*, 2018], but the model utilizes an image retrieval task to calculate the image-caption relevance as rewards.

4. S-txt-img [Su *et al.*, 2018]: The model is similar to mono-lingual based methods with auto-encoding and cycle-consistency loss, but employs an additional image encoder in training which can be absent in testing.

Table 2 presents the performance of different methods for zero-resource machine translation with image pivots. Since our approach utilizes word translation learned in the first progressive step to benefit sentence translation from two losses, we compare each of the loss and their combination with previous methods. The L_{pivot} achieves superior performance to previous image-pivoted methods on both datasets and translation directions. It demonstrates the effectiveness of the re-weighting approach to suppress noises in image-pivoted sentence pairs, while such noises are ignored in previous methods. The auto-encoding loss L_{ae} brings more significant performance gains, which shows that re-using word embedding matrices from the image-pivoted word translation is effective for the NMT model to encode language-agnostic sentence representation. The two losses are also complementary with each other and the combination of them achieves the best translation performance on the two datasets.

To be noted, it might be unfair to directly compare our results with the best model in the recent mono-lingual based method [Lample *et al.*, 2018a]. The best model in [Lample *et al.*, 2018a] utilizes word vectors pretrained on a large-scale mono-lingual corpus and multiple iterations of back-translation to achieve good performance on Multi30k dataset (De-En 26.26 and En-De 22.74). Without pretraining and back-translation, it only achieves translation performance of 7.52 for De-En and 6.24 for En-De on Multi30k dataset. However, our approach which only employs very limited mono-lingual image caption data and single round training without back-translation can still achieve comparable performance of the best model in [Lample *et al.*, 2018a]. It suggests that image-pivoted approaches could be more effective to translate visually relevant sentences. What is more, our approach is orthogonal to mono-lingual based methods.

4.3 Ablation Study

In Table 3, we evaluate improvements from progressive learning for the two proposed losses. The loss L_{pivot} without progressive learning denotes utilizing all image-pivoted pseudo pairs for training without re-weighting by the learned word translation. We can see that it suffers from the noisy pairs and is inferior to our re-weighting model on both datasets. To gain an intuition on the effects of re-weighting strategy, we illustrate some image-pivoted pseudo sentence pairs in Table 5 as well as their sentence and token weights. The clean pseudo sentence pairs are ranked higher than the noisy ones, and the token-level re-weighting provides a more fine-grained supervision from noisy pairs. For the loss L_{ae}, the lack of progressive learning means we do not employ the learned multi-lingual word representation in the first word translation step. As shown in Table 3, the image-pivoted word embedding plays an important role for the effectiveness of auto-encoding loss. It demonstrates that performance boost from L_{ae} mainly contributes to the progressive learning strategy.

Since the proposed translation learning relies on results from image caption models, we investigate the relation be-

[1] Multi-agents method [Chen *et al.*, 2018] utilizes different training and evaluation setup on Multi30k dataset compared with others.

[2] The two works did not report results on the IAPR-TC12 dataset.

Image	Source and groundtruth target sentence	Translated sentences
	ein schwarz-weißer hund läuft zu einem kaputten ball im schnee. a black-and-white dog goes for a flattened ball on the snow.	**ours**: a black and white dog runs to a catch a ball in the snow. **supervised:** a black and white dog runs towards a broken ball in the snow.
	eine ältere person überquert die straße mit einem regenschirm in der hand. an elderly person is crossing a street with an umbrella in their hands.	**ours**: a elderly man crossing the street with a umbrella in the hand. **supervised:** an elderly person is walking across the street with an umbrella.

Table 4: Translation examples from our progressive learning approach and the fully supervised NMT model. Images are only for visualization.

β_i	Image-pivoted Pseudo En-De Sentence Pairs
0.97	a man in a blue shirt and blue shorts playing tennis . ein mann in einem blauen oberteil und blauen shorts spielt tennis .
0.70	a brown dog jumps over a fence . ein weißer hund springt über eine hürde .
0.35	a man carving a pumpkin in his boxers . ein mann in einem blauen hemd hält einen hammer.
0.13	a woman in a white shirt is preparing a meal . peperoni kochen im winter !

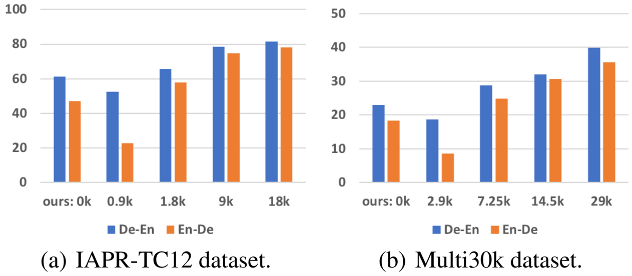
Table 5: Image-pivoted pseudo sentence pairs ranked by sentence-level weights. The tokens in the German sentence with high token-level weights are colored in blue. Best viewed in color.

	IAPR-TC12		Multi30k	
	I/De-En	I/En-De	I/De-En	I/En-De
Caption	23.0	18.9	6.6	5.6
Translation	61.3	47.1	23.0	18.3

Table 6: The BLEU4 scores of image caption model and image-pivoted translation model on IAPR-TC12 and Multi30k datasets.

tween image caption performance and image-pivoted zero-resource translation performance in Table 6. Firstly, the translation performance is much higher than the image captioning performance. This is mainly because the translation output is more constrained by the input while the caption output is more diverse. So the image caption model directly used in previous image-pivot translation methods might be a main bottleneck. Secondly, the translation performance is proportional to the image captioning performance, which indicates that better image captioning can further improve our zero-resource machine translation.

Finally, we compare our progressively learned zero-resource NMT model with fully supervised NMT model with respect to variable amount of training data in Figure 5. Our proposed model with zero-resource reaches 75% of the best performance from the supervised NMT model with 18k training pairs for De-to-En, 60% for En-to-De on IAPR-TC12 dataset. Similarly, it reaches 58% of the best performance from the supervised NMT model with 29k training pairs for De-to-En and 51% for En-to-De on Multi30k dataset. Table 4 presents some randomly selected examples. The proposed approach can generate promising translation results simply based on image pivots without parallel texts.



(a) IAPR-TC12 dataset.

(b) Multi30k dataset.

Figure 5: Comparison with the fully supervised NMT model trained on variable amount of parallel texts. The x-axis denotes number of training pairs and y-axis denotes the BLEU4 score.

5 Conclusion

In this paper, we address the zero-resource machine translation problem by exploiting visual images as pivots. Due to the nature that a picture tells a thousand words, description sentences for an image may be semantically nonequivalent, which leads to noisy supervisions to train the NMT model in previous works. In this work, we propose a progressive learning approach which consists of progressive easy-to-advanced steps towards learning effective NMT models under zero resource settings. The learning starts with word-level translation with image pivots and then progresses to sentence-level translation assisted by the word translation and image pivots. Experiments on IAPR-TC12 and Multi30k datasets prove the effectiveness of the proposed approach, which significantly outperforms previous image-pivot methods.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61772535), Beijing Natural Science Foundation (No. 4192028), National Key Research and Development Plan (No. 2016YFB1001202) and Research Foundation of Beijing Municipal Science & Technology Commission (No. Z181100008918002).

References

[Artetxe *et al.*, 2018] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *ICLR*, 2018.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[Calixto *et al.*, 2017] Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. Using images to improve machine-translating e-commerce product listings. In *EACL*, pages 637–643, 2017.

[Castilho *et al.*, 2017] Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120, 2017.

[Chen *et al.*, 2017] Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. A teacher-student framework for zero-resource neural machine translation. In *ACL*, pages 1925–1935, 2017.

[Chen *et al.*, 2018] Yun Chen, Yang Liu, and Victor OK Li. Zero-resource neural machine translation with multi-agent communication game. *AAAI*, 2018.

[Chen *et al.*, 2019] Shizhe Chen, Jin Qin, and Alexander Hauptmann. Unsupervised bilingual lexicon induction with mono-lingual multimodal data. *AAAI*, 2019.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[Elliott *et al.*, 2016] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016.

[Grubinger *et al.*, 2006] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *Int. Workshop OntoImage*, volume 5, 2006.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Johnson *et al.*, 2017] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5(1):339–351, 2017.

[Kiela *et al.*, 2015] Douwe Kiela, Ivan Vulić, and Stephen Clark. Visual bilingual lexicon induction with transferred convnet features. In *EMNLP*, pages 148–158, 2015.

[Koehn *et al.*, 2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180, 2007.

[Kusner *et al.*, 2015] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *ICML*, pages 957–966, 2015.

[Lample *et al.*, 2018a] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *ICLR*, 2018.

[Lample *et al.*, 2018b] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. Phrase-based & neural unsupervised machine translation. In *EMNLP*, pages 5039–5049, 2018.

[Lee *et al.*, 2018] Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. Emergent translation in multi-agent communication. *ICLR*, 2018.

[Nakayama and Nishida, 2017] Hideki Nakayama and Noriki Nishida. Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[Pele and Werman, 2009] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *ICCV*, pages 460–467. IEEE, September 2009.

[Rubner *et al.*, 2000] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *ICCV*, 40(2):99–121, 2000.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, volume 1, pages 1715–1725, 2016.

[Sharma *et al.*, 2018] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, volume 1, pages 2556–2565, 2018.

[Su *et al.*, 2018] Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. Unsupervised multimodal neural machine translation. *arXiv preprint arXiv:1811.11365*, 2018.

[Zhou *et al.*, 2018] Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. A visual attention grounding neural model for multimodal machine translation. In *EMNLP*, pages 3643–3653, 2018.

[Zoph *et al.*, 2016] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *EMNLP*, pages 1568–1575, 2016.