Multimodal Neural Machine Translation
with Search Engine Based Image Retrieval

Zhenhao Tang   Xiaobing Zhang
College of Application and Technology,
Shenzhen University,
Shenzhen, China

Zi Long*   Xianghua Fu
College of Big Data and Internet,
Shenzhen Technology University,
Shenzhen, China

Abstract

Recently, numbers of works shows that the performance of neural machine translation (NMT) can be improved to a certain extent with using visual information. However, most of these conclusions are drawn from the analysis of experimental results based on a limited set of bilingual sentence-image pairs, such as *Multi30K*. In these kinds of datasets, the content of one bilingual parallel sentence pair must be well represented by a manually annotated image, which is different with the actual translation situation. Some previous works are proposed to addressed the problem by retrieving images from exiting sentence-image pairs with topic model. However, because of the limited collection of sentence-image pairs they used, their image retrieval method is difficult to deal with the out-of-vocabulary words, and can hardly prove that visual information enhance NMT rather than the co-occurrence of images and sentences. In this paper, we propose an open-vocabulary image retrieval methods to collect descriptive images for bilingual parallel corpus using image search engine. Next, we propose text-aware attentive visual encoder to filter incorrectly collected noise images. Experiment results on *Multi30K* and other two translation datasets show that our proposed method achieves significant improvements over strong baselines.

1 Introduction

With the development of NMT, the role of visual information in machine translation has attracted researchers' attention(Elliott et al., 2017; Barrault et al., 2018; Specia et al., 2016). Although we are still not clear about the specific role of visual information in NMT(Caglayan et al., 2019; Elliott, 2018), visual information can assist NMT model to achieve better translation performance (Calixto and Liu, 2017; Calixto et al., 2017; Su et al., 2021). Different with those text-only NMT(Bahdanau et al.,

Data source	Image	Sentence
Multi30K		**EN:** A dog is running in the snow.
		DE: Ein Hund rennt im Schnee.
UN News		**EN:** Marine plastic debris has impacted over 600 marine species.
		DE: Plastik aus dem meer betroffen mehr als 600 meerestiere.

Table 1: Comparison of example from *Multi30K* dataset and United Nations News.

2014; Gehring et al., 2016), a bilingual parallel corpora with manual image annotations are used to train a multimodal NMT model by an end-to-end framework, and therefore, most of the previous conclusions are drawn from the analysis of experimental results based on a limited set of manually annotated bilingual sentence-image pairs, specifically, *Multi30K*(Elliott et al., 2016).

In *Multi30K*, as shown in 1, the sentences consists mostly of common and simple words, and the content of each bilingual parallel sentence pair is well represented by a single image. Table 1 also shows an example of bilingual sentence-image pair from an actual news report of United Nations News[1]. It is obviously that there is a dramatic difference between the data of *Multi30K* and the real-world multimodal translation situations. Therefore, results and evidences based on *Multi30K* can hardly proved the effectiveness of multimodal NMT model in an actual translation situation, in which sentences contain rare and uncommon words and are partially described by images.

To address the problem, Zhang et al. (2019) proposed to transform the existing sentence-image pairs into a topic-image lookup table, and a group

---

*Corresponding author

[1] https://news.un.org/en/

of images with similar topics to the source sentence is retrieved from the topic-image lookup table. However, the topic-image lookup table is made from a limited collection of sentence-image pairs, such as *Multi30K* and MS COCO image caption dataset (Lin et al., 2014), their image retrieval method is difficult to deal with the out-of-vocabulary words. Besides, results from Zhang et al. (2019) can hardly prove that the performance of NMT is improved by visual information rather than the co-occurrence of images and sentences. Their model may suffer problems in translating sentences with images that are not contained in the topic-image lookup table.

In this paper, we propose an open-vocabulary image retrieval methods to collect images for bilingual parallel corpus using image search engine, thus addressing the problems caused by limited collection of sentence-image pairs in Zhang et al. (2019). In detail, to focus on the major part of the sentence, we apply the term frequency-inverse document frequency (TF-IDF). Instead of a single keyword, we use multiple words as search query for image retrieval to ensure that the contents of collected images are partially consistent with the given sentences. Since the quality of images from search engine may be varied, we propose to apply a simple but effective attention layer, and introduce a text-aware attentive visual encoder to filter incorrectly collected noise images. The proposed method is then evaluated on three translation datasets, including the *Multi30K* English-to-German, WMT'16 English-to-German, Global Voices (Tiedemann, 2012) English-to-German. Experiment results show that our proposed method achieves significant improvements over strong baselines. To summarize, out contributions are primarily three-fold:

(1) We present an open-vocabulary image retrieval methods with image search engine that overcomes the shortcomings of Zhang et al. (2019) caused by limited image collection.

(2) The proposed method enables the text-only NMT to use visual information from the collected images that are partially consistent with input sentences, which is more close to the actual translation situations.

(3) We further discuss the influence of visual information in the proposed multimodal NMT model , which verified the effectiveness and generality of the proposed approach.

2 Related Work

Recently, multimodal NMT models have gradually become a hot topic in machine translation research. They use image information to improve the translation effect of NMT models through different methods.

In some cases, visual features are directly used as supplementary information to the text presentation. For example, Huang et al. (2016) takes global visual features and local visual features as additional information for sentences. Calixto and Liu (2017) initializes the encoder hidden states or decoder hidden states through global visual features. (Calixto et al., 2017) uses an independent attention mechanism to capture visual representations. (Caglayan et al., 2016) incorporates spatial visual features into the multimodal NMT model via an independent attention mechanism. On this basis, Delbrouck and Dupont (2017) employs Compact Bilinear Pooling to fuse two modalities. Su et al. (2021) introduces image-text mutual interactions to refine their semantic representations. Lin et al. (2020) attempts to introduce the capsule network into multimodal NMT, they use the timestep-specific source-side context vector to guide the routing procedure.

All the above work is performed on the *Multi30K* dataset. However, some recent studies indicate that the visual features may play a less important role in the NMT model than previously thought. (Ive et al., 2019; Zhang et al., 2017; Grönroos et al., 2018). Such problems are mainly caused by the limitations of the *Multi30K* dataset. Zhang et al. (2019) presents a universal visual representation method that overcomes the shortcomings of *Multi30K* dataset. However, all their image information still comes from *Multi30K*, which is obviously not enough to represent complex machine translation corpus.

3 Background

In this section, we give a simple description of the multimodal NMT model proposed by Calixto et al. (2017). The multimodal NMT model is composed of one text encoder, one visual encoder and one decoder with two attention mechanisms. The multimodal NMT aims to construct an end-to-end neural network to model $P = (Y|X, I)$ as follows:

$$\log p(Y|X, I) = \sum_{i=1}^{M} \log p(y_t|y_{<t}, C, A)$$

90

where I represents visual features, $X = (x_1, x_2, \ldots, x_L)$ is the source sentence, and $Y = (y_1, y_2, \ldots, y_M)$ is the target sentence. The text encoder is a Bi-directional Recurrent Neural Network (RNN) with Gated Unit(GRU)(Cho et al., 2014) and learn a time-dependent text hidden states $C = (h_1, h_2, \ldots, h_N)$ for the source sentence. The visual encoder is a pretrained convolutional neural network (CNN) and a visual representation A for the given image.

The decoder is a conditional GRU (cGRU)[2] with two separate attention mechanisms. The text attention mechanism generates a time-dependent context vector c_t based on the text hidden states C and the hidden state proposal s_t' as follows:

$$c_t = f_{att\_text}(C, s_t') \tag{1}$$

Meanwhile, the visual attention computes a time-dependent context vector i_t based on the visual feature maps A and the hidden state proposal s_t' as follows:

$$i_t = f_{att\_img}(A, s_t') \tag{2}$$

Where s_t' is calculated by the previous hidden state s_{t-1} and the previously generated target word y_{t-1}.

4 Our Proposed Method

Figure 1 shows the 4 components of our proposed method, consisting of image retrieval, text-aware attentive visual encoder, RNN text encoder and translation decoder with co-attention & bi-attention.

4.1 Image Retrieval

In this section, we will introduce the proposed open-vocabulary image retrieval methods using image search engine.

Similar with Zhang et al. (2019), to focus on the major part of the sentence and suppress the noise such as stopwords and low-frequency words, we apply the term frequency-inverse document frequency (TF-IDF) (Witten et al., 2005) to create search queries for image search engines. Specifically, given the ith ($i = 1, 2, \ldots, N$, N represents the number of samples in the training set) source language sentence $X_i = \{x_i^1, x_i^2, \ldots, x_i^L\}$ of length L, X_i is first filtered by as stopword list[3], and the filtered input sentence X_i^f is obtained. We then

regard X_i^f as a document d_i, and compute the TF-IDF score $TI_{i,j}$ for each word x_i^j ($j = 1, 2, \ldots, L$) in d_i. The formula is as follows:

$$TI_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,k}} \times \log \frac{|D|}{1 + |\{k|x_i^j \in d_k\}|}$$

where $n_{i,j}$ is the number of occurrences of the word x_i^j in document d_i, $\sum_k n_{i,k}$ represents the total number of words in document d_i. $|D| = N$ represents the total number of source language sentences in the training data, and $|\{k|x_i^j \in d_k\}|$ represents the number of sentences including x_i^j in the dataset. For input sentence X_i, words are then listed in descending order by $TI_{i,j}$ score, represented as $Q_i = (x_i^{t_1}, x_i^{t_2}, \ldots, x_i^{t_L})$ ($TI_{i,t_1} \geq TI_{i,t_2} \geq \ldots \geq TI_{i,t_L}$).

Instead of using the top-k high TF-IDF words separately, we concatenate several words from the top-k high TF-IDF words as search query. Specifically, for the sorted words list Q_i, the mth search query q_m is defined as following:

$$q_m = \text{concat}(x_i^{t_1}, x_i^{t_2}, \ldots, x_i^{t_m})$$

Where $\text{concat}(\cdot)$ means that words are concatenated with blanks as separator. search query q_m is then applied in image search engine and the first available image is collected as the mth image for input sentence X_i, represented as A_i^m. According to the results of preliminary experiment, we build 5 search queries and collect 5 images for each sentence[4].

4.2 Text-Aware Attentive Visual Encoder

For each collected image, we employ a 50-layer Residual Network (ResNet-50) (He et al., 2016) to represent the visual semantic information as a 196×1024 feature vector.

As described in Section 4.1, for source language sentence X_i, we collect 5 images $A_i^1, A_i^2, \ldots, A_i^5$ using image search engine. In order to filter the incorrectly collected noise images, we apply a simple but effective scaled dot-product attention in visual encoder, where the visual representation A_i of input sentence X_i is defined as the following formula:

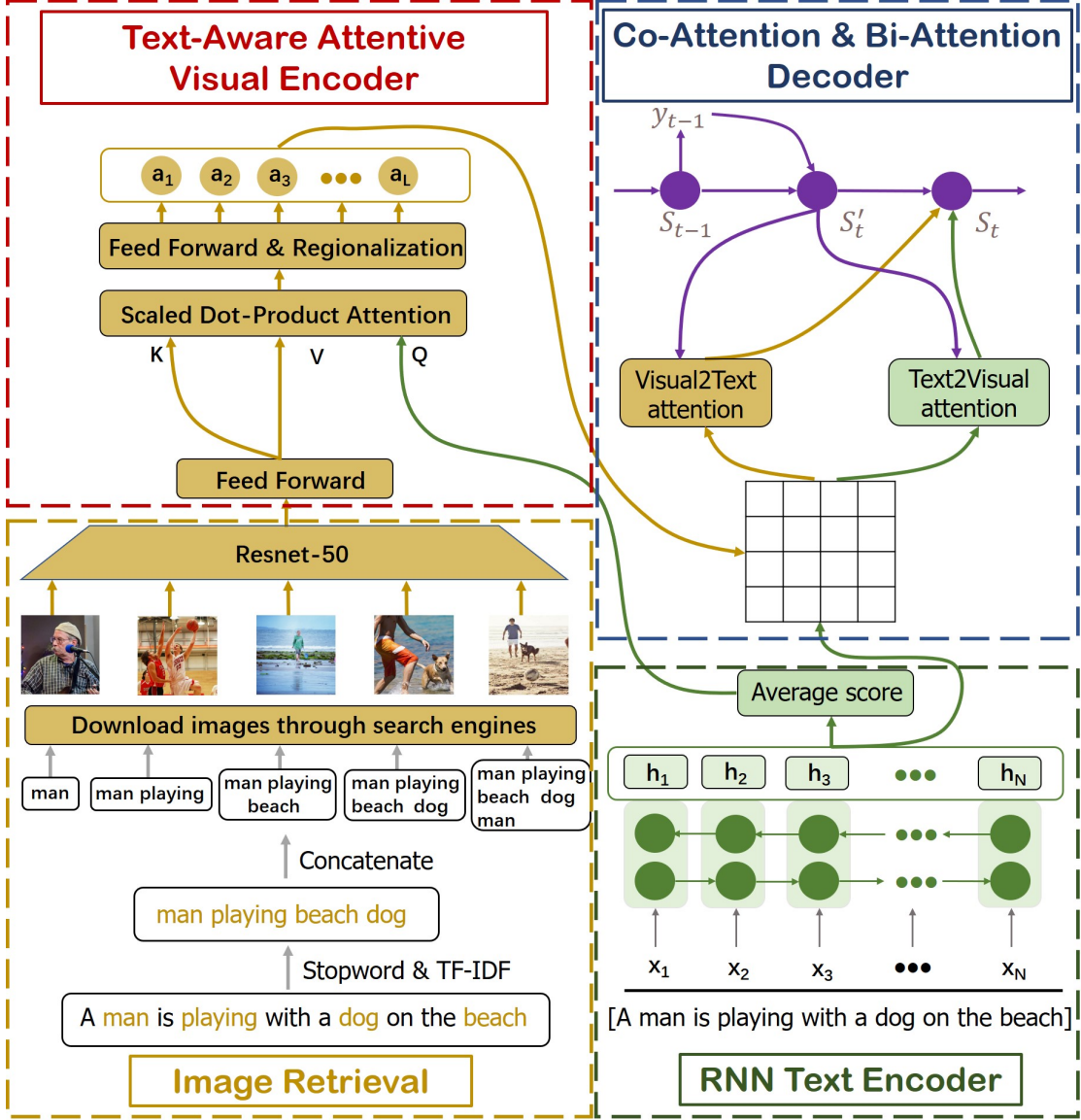$$A_i = \sum_{m=1}^{5} \alpha_{i,m} A_i^m$$

Figure 1: The overview of the framework of our proposed method

where $\alpha_{i,m}$ represents the weight of mth images for input sentence X_i. The $\alpha_{i,m}$ is then computed as follows:

$$\alpha_{i,m} = \text{softmax}(W(A_i^m) \cdot C_i')$$

$$C_i' = \frac{1}{N}\sum_{t=1}^{N} h_i^t$$

where $\text{softmax}(\cdot)$ stands for softmax activation function, and C_i' represents an average pool of the hidden states $C_i = (h_i^1, h_i^2, \ldots, h_i^N)$ for input sentence X_i.

Finally, the obtained 196×1024D visual representation is considered as a matrix $A_i = (a_i^1, a_i^2, \ldots, a_i^L)$, $a_i^l \in R^{1024}$. Each of the $L = 196$ rows consists of a 1024D feature vector that

represents a specific image region. Visual representation $A_i = (a_i^1, a_i^2, \ldots, a_i^L)$ and text representation $C_i = (h_i^1, h_i^2, \ldots, h_i^N)$ are then used as the inputs of tanslation decoder .

4.3 Translation Decoder

As shown in figure 2, we apply a bi-directional attention network[5] and a co-attention network (Su et al., 2021) to model underlying semantic interactions between text and image.

The bi-directional attention network is used

---

[5]According to the result of the preliminary experiment, we found that Transformer-based model can hardly produce an advantage in performance on such small dataset as *Multi30K*. Therefore, we chosed LSTM as our basic model. As a future work, we are going to integrate Transformer into our proposed method and evaluate it on some larger datasets.
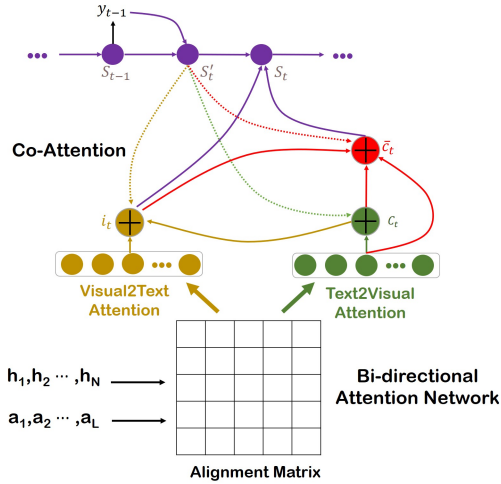
Figure 2: multimodel NMT model with deep semantic interactions

to enhance text and image representations. Specifically, we use text representation $C_i = (h_i^1, h_i^2, \ldots, h_i^N)$ and visual representation $A_i = (a_i^1, a_i^2, \ldots, a_i^L)$ for bi-direction attention network to obtain a shared alignment matrix $S \in R^{N \times L}$. The alignment matrix is computed as follows:

$$S_{n,l} = g(h_i^n \cdot a_i^l)$$

where $g(\cdot)$ is a scalar function. The $S_{n,l} \in R^{N \times L}$ measures how well the n-th row vector in C_i semantically matches the l-th row vector in A_i. After that, *Text-to-Visual Attention* $\overline{h_i^n}$ and *Visual-to-Text Attention* $\overline{a_i^l}$ will be calculated respectively according to the alignment matrix S. The $\overline{h_i^n}$ calculation formula is as follows:

$$
\begin{aligned}
w_n^{t2v} &= \text{softmax}(S_{n:}) \\
\overline{h_i^n} &= h_i^n + \sum_l w_{nl}^{t2v} a_i^n
\end{aligned}
$$

The $\overline{a_i^l}$ calculation formula is as follows:

$$
\begin{aligned}
w_l^{v2t} &= \text{softmax}(S_{:l}) \\
\overline{a_i^l} &= a_i^l + \sum_i w_{ln}^{v2t} h_i^n
\end{aligned}
$$

Among them, w_n^{t2v} signifies which image regions are most relevant to each source word. w_l^{v2t} signifies which source words semantically match each visual region mostly. Thus, we can get the final visual feature maps $\overline{A_i} = (\overline{a_i^1}, \overline{a_i^2}, \ldots, \overline{a_i^L})$, and the vectors for the whole source sentence $\overline{C_i} = (\overline{h_i^1}, \overline{h_i^2}, \ldots, \overline{h_i^N})$. Finally, we substituted $\overline{C_i}$

and $\overline{A_i}$ into formulas (1) and (2) in Section 3 to obtain the time-dependent context vector c_t and the time-dependent visual vector i_t.

5 Experiments

5.1 Data

To evaluate our approach, we experimented with three commonly used machine translation dataset, including multimodal machine translation dataset *Multi30K* (Elliott et al., 2016) English-to-German (EN-DE), Global Voices English-to-German (EN-DE)(Tiedemann, 2012), and WMT' 16 (100k) English-to-German (EN-DE).

Multi30K *Multi30K* dataset consists of about 31k bilingual sentence-images pairs, . In this paper, we use 29K English to German parallel sentence pairs with visual annotations as the training set. The 1,014 English to German sentence pairs visual annotations are used as dev set. Finally, the test2016 test dataset is used for evaluation.

Global Voices Global Voices (EN-DE) dataset consists of more than 70k bilingual sentence pairs from summaries of news articles. We randomly sample 2000 data as dev set, 2000 as test set, and use the remained as training set.

WMT'16 (100k) WMT dataset (EN-DE) consists of more than 4.5M bilingual sentence pairs mainly from the proceedings of the European Parliament. In order to focus on evaluating the effectiveness of the retrieved visual information, we attempt to exclude the influence of data size, and randomly sampled 100k sentence pairs as our training set instead of the total 4.5M sentence pairs, which is similar to the number of sentences in the *Multi30K* dataset and Global Voices. We used Newstest2016 as the test set.

5.2 System Setting

Image Retrieval Implementation We used the Microsoft Bing[6] as image search engine. As described in Section 4.1, for each source language sentence, we build 5 search queries and collect 5 images for each sentence. Specifically, if the number of words is less than 5 after stopword filtering, we simply repeat the keyword list several times

---

[6] https://global.bing.com/images

93

to ensure that the number of remained words is enough for creating 5 search queries.

Model Implementation: We implemented our proposed model on the top of Su et al. (2021), which was developed based on OpenNMT (Klein et al., 2017). We used *MOSES*[7] scripts to tokenize, normalize, and lowercase both source and target sentences. For text encoder, we used bi-directional RNN with GRU to extract text features. One 256D single-layer RNN was used for both forward and backward. For visual encoder, we used the res4f layer of pre-trained ResNet-50 (He et al., 2016) to extract visual features. We used Adam optimizer with mini-batches size of 32 to train all models, and set the learning rate as 0.001.

We trained the model up to 15 epochs, and the training was early-stopped if BLEU (Papineni et al., 2002) score of dev set did not improve for 3 epochs. The model with highest BLEU score of the dev set was selected to evaluate the test set. In order to reduce the influence of random seeds on the experimental results and ensure the stability of the final experimental results, we repeated the experiment 5 times with fixed 5 random seeds and used the macro average of BLEU scores as the final result.

Baseline For each dataset, we used the text-only LSTM (Graves, 2012) as a baseline.

For *Multi30K* dataset, we quantitatively compared the proposed method with the following models:

- Zhang et al. (2019) used a text-only Transformer and proposed a universal visual representation method by retrieving images from a topic-image lookup table.

- Su et al. (2021) used a bi-direction attention network and a co-attention mechanism to enhance semantic interaction of text and images.

- Zhao et al. (2021) proposed a novel integration strategy Word-Region Alignment(WRA) of the MNMT model that leverages the WRA to guide the model to translate certain source words into target words while attending to semantically relevant image regions.

We trained these models by employing the same training set and the same training parameters as the proposed model, and report the 4-gram BLEU score (Papineni et al., 2002) for all baselines as well as the proposed method.

Method	BLEU Score
Text-only NMT	
Bi-LSTM (Calixto et al., 2017)	33.7
Transformer (Zhang et al., 2019)	36.86
Multimodal NMT with Original Images	
Zhang et al. (2019)	36.86
Zhao et al. (2021)	38.40
Su et al. (2021)	39.20
The proposed method	38.14
Multimodal NMT with Retrieved Images	
Zhang et al. (2019)	36.94
The proposed method	**38.43**

Table 2: Results on Multi30K

System	BLEU Score	
	Global Voices	WMT'16 (100k)
Text-only LSTM	9.22	7.99
The proposed Method	**9.81**	**8.41**

Table 3: Results on Global Voices and WMT'16 (100k)

5.3 Experimental Results

Table 2 shows the experimental results on *Multi30K* dataset. The proposed method obtains a BLEU score of 38.43. Compared with the text-only NMT (Calixto et al., 2017; Vaswani et al., 2017), the proposed method obtains a significantly higher BLEU score. Compared with the multimodal NMT methods with original images (Zhang et al., 2019; Zhao et al., 2021; Su et al., 2021), our proposed method obtains a comparable BLEU score[8]. Compared with the multimodal NMT method with retrieved images (Zhang et al., 2019), the performance gain of the proposed method is approximately 1.5 BLEU.

Futhermore, we quantitiatively compared our study with text-only NMT (Calixto et al., 2017) on two dataset, i.e., Global Voices and WMT'16 (100k), which consist of bilingual sentence pairs without visual annotation. As shown in Table 3, the proposed method achieved a higher BLUE score, demonstrating the effectiveness of the proposed search engine based image retrieval. More experi-

---

[8]For Su et al. (2021), we trained the multimodal NMT model using the same parameters with our proposed method, and got a comprable BLEU score of 38.1 with our proposed method.

Source(En):	**Guitar player performs at a nightclub red guitar .**
Target(De):	**Gitarrist spielt** in einem **Nachtklub** auf einer roten Gitarre .

NMT:	ein Gitarrespieler spielt auf einer Reifenschaukel .

Mistranslated "Guitarist" for "Musician"

MNMT(Multi30k Image): ein **Musiker spielt** auf einer beigefarbenen Gitarre .

Our method: ein **Gitarrespieler spielt** in einer **Nachtclub** Gitarre .

The "Guitarist" and the "Nightclub" are correctly translated

Search keywords: Guitar | Guitar Nightclub | Guitar Nightclub Performs | Guitar Nightclub Performs Player | Guitar Nightclub Performs Player Red

Figure 3: Example of correct translation by the proposed method

mental results and discussions for the influence of collected images are described in Section 6.

Figure 3 shows an example of correct translation by the proposed method. In this example, English words "nightclub" is failed to be translated by the model of Su et al. (2021), as well as the text-only NMT. It is mainly because that the text infomation is not enough for translating while the original image from *Multi30K* is ambiguous and misleading. In the proposed method, we collected 5 images with image search engine according to the method described in Section 4.1, among which 3 images provide effective visual information about "nightclub", and therefore, the proposed method correctly translate "nightclub" into "Nachtclub". Besides, benefit from visual information about "guitar player", the proposed method generates a partially correct translation "Gitarrespieler spielt", while is the model of Su et al. (2021) incorrectly translate "guitar player" into "Musiker spielt" (musician).

6 Analysis and Discussion

6.1 Influence of the Number of Images

For each sentence, several images can be obtained by following the image retrieval method in section 4.1. To evaluate the influence of the number of paired images m, we constrained m in $\{1, 2, 3, 4, 5, 6, 7, 8\}$ for experiments on the *Multi30K* dataset. As shown in figure 4, for dif-

Dataset	images	BLEU
Multi30K	Text-only	37.77
	Random Images	37.65
	Blank Images	37.79
	Retrieved Images	**38.43**
Global Voices	Text-only	9.22
	Random Images	9.29
	Blank Images	9.46
	Retrieved Images	**9.81**
WMT'16 (100k)	Text-only	7.99
	Random Images	8.11
	Blank Images	8.31
	Retrieved Images	**8.41**

Table 4: Translation effect of different data sets under different image conditions

ferent m, we used the images retrieved by search engine and the original images in *Multi30K* dataset respectively for experiments. For images retrieved based on search engines, as the number of images increases, the BLEU score also increased at the beginning(from 37.96 to 38.43) and then decreased when m exceeds 5. The reason might be that retrieving too many images through search engines will lead to an increase in the number of noise images. Therefore, we set $m = 5$ in our models, and drawn a same conclusion as Zhang et al. (2019).

For the original *Multi30K* image, it only has the visual features of an image, so as the number of

95

Dataset	Sentence	Retrieved image
Multi30K	The person in the striped shirt is mountain climbing.	
Global Voices	Now the city is under a siege from the security forces.	
WMT'16	In the future, integration will be a topic for the whole of society even more than it is today.	

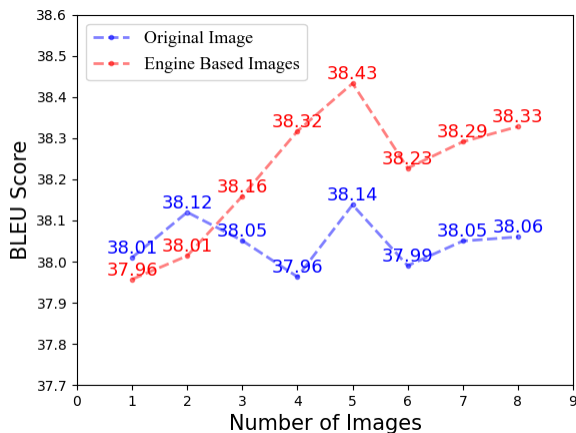Table 5: Examples of retrieved image from different datasets

Figure 4: Influence of number of images on the BLEU score.

Dataset	Number of noise images
Multi30K	61
Global Voices	228
WMT'16	685

Table 6: Number of noise images in 1000 collected images for each dataset

images increases, the BLEU score has no obvious upward trend. In addition, when m is less than 3, the BLEU score of the image using the original *Multi30K* is higher than that of the retrieved image.

6.2 Influence of the Quality of Images

To evaluate the influence of the quality of collected images, we train the proposed model with randomly retrieved unrelated images, blank images, and retrieved images from image search engine, respec-

tively. The evaluation results are shown in table 4. It is obvious that proposed method achieves the highest BLEU score on all *Multi30K* and Global Voices, demonstrating the effectiveness of visual information from collected images.

Compared with the model with random images and blank images, the performance gain of collected images is approximately 0.7 & 0.6 BLEU score on *Multi30K*, and 0.5 & 0.3 BLUE score on Global Voices. However, on the WMT'16 (100k) dataset, model with collected images obtains almost the same BLUE score as the model with blank images.

One of the possible reason is that sentences from WMT dataset contains fewer entity words that can be represented by images, and therefore, the proposed search engine based image retrieval method collects numbers of noise images. Sentences from

96

WMT'16 (100k) describe abstract concepts and complex events, while sentences from *Multi30K* and Global Voices describe real objects and people, which is more reliable for image retrieval. Examples of retrieved images of each dataset are shown in Table 5. For the sentence from *Multi30K* dataset, our method easily retrieves an image that represents "A man is rock climbing". For the sentence from Global Voice dataset, the retrieved image is partially consistent with the source sentence, containing contents of "city","siege" and"forces". However, for the sentence from WMT'16 dataset, it is obvious that the retrieved image contains little effective visual information and can hardly provide assistance to translation.

To verify the hypotheses, we randomly sampled 1,000 images from the collected image set of each dataset, and manually classify the collected images into 2 classes, i.e., class of images that can provide visual information of the search query, and class of images that can not. Images in second class are defined as noise images. As shown in Table 6, for *Multi30K* dataset, only 61 out of 1000 collected images sampled are noise images, and the proportion is 6.1%. However, in the WMT'16 dataset, the number of noise images obtained through retrieval is 685, accounting for more than half of the total number of images.Therefore, our method performs poorly on the WMT'16 dataset. For the Global Voices dataset, the number of noise images is 228, which is between the *Multi30K* and WMT'16 dataset, and the retrieved images also show better performance than the NMT model. It is insteresting to find that collected image set for *Multi30K* has smallest proportion of noise image and achieves the biggest gain of translation performance, while the collected image set has the largest proportion of noise image and achieves the smallest gain of translation performance.

7 Conclusions

In this paper, inspired by problem of Zhang et al. (2019) caused by applying limited collections of sentence-image pairs, we propose an open-vocabulary image retrieval methods to collect descriptive images for bilingual parallel corpus using image search engine, and introduce text-aware attentive visual encoder to filter incorrectly collected noise images. Experiment results show that our proposed method achieves significant improvements over strong baselines, especially on *Multi30K* and

Global Voices. Further analysis shows that the effectiveness of the proposed methods in translating sentences that describe real objects and people.

As one of our future work, we are going to evaluate our proposed method on some larger datasets, such as the entire WMT'16 dataset, and analyze the influence of the number of texts for the task of multimodal NMT.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, volume 2, pages 308–327.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *arXiv preprint arXiv:1609.03976*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North*, pages 4159–4170. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. 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)*, pages 1913–1924.

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

Jean-Benoit Delbrouck and Stephane Dupont. 2017. Multimodal compact bilinear pooling for multimodal neural machine translation. *arXiv preprint arXiv:1703.08084*.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.

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

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

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, et al. 2018. The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. *arXiv preprint arXiv:1906.07701*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Opensource toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. 2021. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences*, 554:47–60.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, and Craig G Nevill-Manning. 2005. Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, pages 129–152. IGI global.

Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. 2017. Nict-naist system for wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 477–482.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2019. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259.