Journal of Computers Vol. 33 No. 2, April 2022, pp. 37- 50
doi: 10.53106/199115992022043302004

Image Visual Attention Mechanism-based Global and Local Semantic Information Fusion for Multi-modal English Machine Translation

Xiaobin Guo*

Zhengzhou Railway Vocational and Technical College, Zhengzhou 450000, China
publicgj@163.com

Received 27 May 2021; Revised 7 September 2021; Accepted 7 October 2021

Abstract. Machine translation is a hot research topic at present. Traditional machine translation methods are not effective because they require a large number of training samples. Image visual semantic information can improve the effect of the text machine translation model. Most of the existing works fuse the whole image visual semantic information into the translation model, but the image may contain different semantic objects. These different local semantic objects have different effects on the words prediction of the decoder. Therefore, this paper proposes a multi-modal machine translation model based on the image visual attention mechanism via global and local semantic information fusion. The global semantic information in the image and the local semantic information are fused into the text attention weight as the image attention. Thus, the alignment information between the hidden state of the decoder and the text of the source language is further enhanced. Experimental results on the English-German translation pair and the Indonesian-Chinese translation pair on the Multi30K dataset show that the proposed model has a better performance than the state-of-the-art multi-modal machine translation models, the BLEU values of English-German translation results and Indonesian-Chinese translation results exceed 43% and 29%, which proves the effectiveness of the proposed model.

Keywords: multi-modal machine translation, image visual attention mechanism, global and local semantic information fusion, alignment information

1 Introduction

Multi-modal machine translation (MMT) is a hybrid machine translation model that fuses multi-modal information from text, speech, video, and image [1]. Compared with the plain text machine translation model, the multi-modal machine translation model can make up for the deficiency of single mode machine translation and improve the accuracy of machine translation with other modal information to some extent [2]. This paper focuses on multi-modal machine translation by fusing text and image mode information. Intuitively, the visual semantic information of images can assist and resolve the difficult semantic ambiguity in the text to a certain extent [3]. For example, when translating the word “bank” (Chinese word) in the plain text mode, it is necessary to infer from its context information whether it is translated as “bank” (financing institution) or “bank” (levee) two different semantic information. When the visual information of the image is used, that is, when the image contains the bank (financing institution), the semantic information of the word “bank” can be determined as “bank” (financing institution) rather than “embankment” with a high probability. Libovický [4] proposed two novel approaches to combine the outputs of attention mechanisms over each source sequence, flat and hierarchical for multi-source sequence-to-sequence learning. Zhou [5] introduced a novel multimodal machine translation model that utilized parallel visual and textual information. The model jointly optimized the learning of a shared visual-language embedding and a translator. The model leveraged a visual attention grounding mechanism that linked the visual semantics with the corresponding textual semantics. Yao [6] introduce the multimodal self-attention in Transformer to solve the problem without considering the relative importance of multiple modalities. Nishihara [7] proposed a supervised visual attention mechanism for multimodal neural machine translation (MNMT), trained with constraints based on manual alignments between words in a sentence and their corresponding regions of an image. Zhao [8] proposed the application of semantic image regions for MNMT by integrating visual and textual features using two individual attention mechanisms (double attention). However, it had been suggested that the visual modality was only marginally beneficial. Conventional visual attention mechanisms had been used to select the visual features from equally-sized grids generated by convolutional neural networks (CNNs) and would have had modest effects on aligning the visual concepts associated with textual objects, because the grid

* Corresponding Author

visual features did not capture semantic information.

Combining the advantages of the multi-modal machine translation model, the research on the fusion of text and image visual information for multi-modal machine translation has attracted the attention of researchers in recent years.

In images description generation task, Vinyals et al. [9] utilized the end-to-end framework in machine translation to translate the encoder of the source language text sentence in the traditional translation architecture as the image vector output by pre-trained convolution neural network. Then it is sent to the decoder as the initial hidden state vector. In this way, in the process of image description sentence generation, the decoder can fully use the semantic information in the image and improve the effect of image description generation task. Calixto et al. [10] integrated the visual information of images into the encoder and decoder of the translation framework respectively based on the encoder-decoder end-to-end machine translation framework to enhance the effect of images on text machine translation. Calixto [11] used two independent attention mechanism frameworks to process the word region and image region separately in the source language to improve the translation results of the model.

Previous works have fused image visual information from different perspectives and achieved good results in multi-modal machine translation. However, these works extract the whole visual semantic information of the image from the global perspective and integrate it into the text translation model as the hidden representation of the image. Since the images may contain many different semantic objects as shown in Fig. 1 (three different semantic objects: man, horse and bull), they have the different role in text translation model. Therefore, extracting local images and obtaining different local visual semantic information can improve the results of multi-modal machine translation from different perspectives.

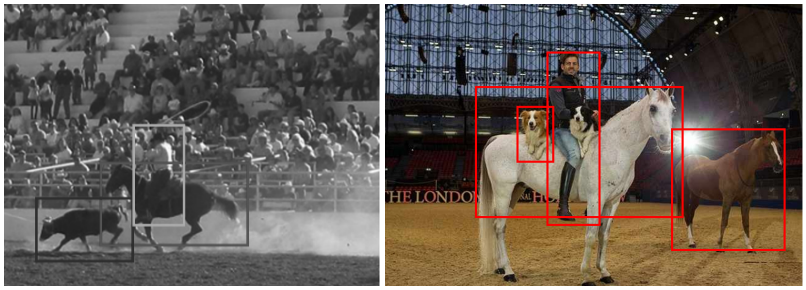


Fig. 1. Different semantic objects in an image

Huang et al. [12] integrated local visual features of images for multi-modal machine translation. They extracted the local area and the global area of the image respectively and projected them into the vector space, which was added into the input sequence of the model as pseudowords. An end-to-end multi-modal machine translation with local visual information and attention mechanism was explored. After extracting the local area representation and the global image representation features, Caglayan et al. [13] fused these global and local images with text in different ways to improve the overall effect of machine translation.

Although the above methods fused the local features of images, they did not fully consider the different contributions of different image parts for the translation text, namely, the semantic interaction information between different parts of the image and the text words in the source language. Different interactive attention information between different local visual objects in images and the words in the source language can be used to enhance the pure text attention in the neural machine translation model and assist the alignment between decoder and source language text, so as to better improve the text translation results [14]. Therefore, this paper proposes an end-to-end multi-modal machine translation model based on image attention.

The different local image visual information, global visual information and the interaction of the source language text word are as the image attention and fused to the text attention, which makes the decoder observe the source text decoder information, and visual information different from the region visual information in the source text, and thus it can better output the decoded translation text.

The main contributions of this paper are as follows:

1) An image attention fusion mechanism is proposed to enhance the plain text attention, so that the decoder can better align with the source language text when decoding the target words, so as to improve the results of multi-modal machine translation. Experimental results on multi-modal machine translation data set Multi30K show that the proposed image attention fusion mechanism can effectively improve the performance of multi-mod-

al machine translation.

2) In order to test the translation effect of the proposed model on small datasets with scarce resources, we manually annotate the bilingual sentence pairs of the test set and verification set in the MULTI30K dataset respectively, and label the English pairs as Indonesian-Chinese translation sentence pairs. The proposed model is tested on this dataset, and it is found that the improvement effect is higher than that of the English-German experiment.

The upcoming parts of this paper are organized as follows. Section 2 introduces the related works. In Section 3, a multi-modal machine translation model with image attention mechanism is constructed and analyzed. In Section 4, we analyze the proposed machine translation model with abundant experiments. Section 5 concludes this paper.

2 Related Works

Multi-modal machine translation was proposed by the Machine Translation Committee through a shared task. In the early stage, the multi-modal network proposed by Mao et al. [15] integrated the features of text and vision, and applied them to the image description generation task and image description sorting task. In their work, the authors merged the recurrent neural network in the independent multi-modal layer. Kanerva et al. [16] proposed a neural image description generation model (IDG) based on a sequence-to-sequence framework, which was based on end-to-end training. Dinh et al. [17] proposed a model for generating multilingual description of images. The model learned and transformed image features in two independent, non-attentional IDG models. Qu et al. [18] proposed the first image description generation model based on the attention mechanism. In this model, the attention mechanism would pay attention to different regions of the image when generating a natural language description of the image.

More recently, multi-modal translation has seen renewed interest due to combined efforts of the computer vision and natural language processing (NLP) communities and the recent availability of large multimodal datasets. A particularly popular problem is visual scene description, also known as image and video captioning, which acts as a great test bed for a number of computer vision and NLP problems. To solve it, we not only need to fully understand the visual scene and to identify its salient parts, but also to produce grammatically correct and comprehensive yet concise sentences describing it.

Kano et al. [19] proposed a multi-task learning method, which could translate one source language into multiple target languages during model training. However, in Kano's model, not each language had an independent attention mechanism, but it had a shared attention mechanism. That is, each target language had an attention mechanism that was shared by all source languages. Firat et al. [20] proposed a multi-channel model, which translated a variety of different source languages into different target languages during training. Chu et al. [21] proposed a multi-task approach, in which two tasks (primary task, auxiliary task) and a shared decoder were used to train the model. The main task was to translate German into English, and the auxiliary task was to generate an English image description. Their experimental results showed that the performance of the main translation task was also improved when the auxiliary task of image description generation was trained.

Although, as yet, there is no fixed neural multi-modal that can significantly improve the performance of plain text NMT and Statistical Machine Translation (SMT) models. Different research groups have proposed to reorder the global features and spatial visual features of the images generated by the SMT system or NMT system, and then take n-best features. This method has achieved some success. The multi-modal neural machine translation has achieved good experimental performance. By using the VGG19 network [22], they obtain the global features of the image and used the RCNN network to extract different regions in the image. Their model offered a significant performance improvement over the plain text NMT benchmark system.

This paper fully combines global and local features, which makes the decoder observe the source text decoder information, and visual information different from the region visual information in the source text, and thus it can better output the decoded translation text. This method can improve the over-translation and under-translation situation.

3 A Multi-modal Machine Translation Model with Image Attention Mechanism

The framework of classical neural machine translation (NMT) is sequence-to-sequence translation model based on encoder-decoder. The input and output are source language word sequences $X = (x_1, x_2, \cdots, x_M)$ and

target language word sequences $Y=(y_1,y_2,\cdots,y_N)$, respectively. NMT model is expected to obtain the maximum probability $P(Y\mid X)$ of X translated to Y, which can learn the conditional probability distribution of the training set data. In this paper, we also adopt this end-to-end framework. The bidirectional LSTM [23] is used in the coder of the model to encode the source language sentences, and the LSTM is used in the decoder of the model to decode the target language sentences. Meanwhile, attention mechanism is introduced so that the decoder can get the alignment information between the hidden state and the source language text when decoding and predicting the next target word. In this paper, we propose an image attention fusion mechanism, interacting the different local and global information of image and semantic information of source language text word, which is as the image attention mechanism and integrated into the proposed text attention. An enhanced fusion attention is obtained, so that the decoder can obtain richer context information and further improve the translation effect of the model. The overall structure of the model is shown in Fig. 2. The parameters in Fig. 2 are explained in the following sections.

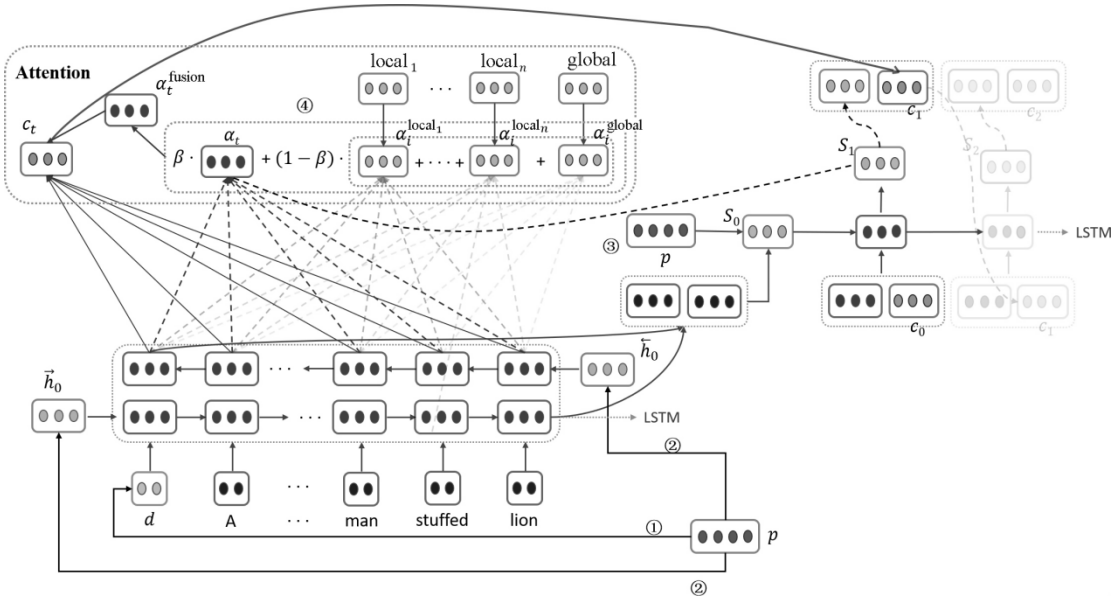


Fig. 2. Proposed multi-modal machine translation structure diagram

The new model contains five aspects: source language text encoding, image attention fusion mechanism, visual semantic representation of different regions, different fusion methods for image visual information and model training. The following sections will give the detailed explanation.

3.1 Source Language Text Encoding

Proposed model uses the bidirection LSTM to encode the source language text. The forward LSTM receives the word vector of each word and the output hidden vector from left to right according to the word order, the output sequence $(\vec{h}_1,\vec{h}_2,\cdots,\vec{h}_N)$ is obtained. The backward LSTM receives the word vector of each word and the output hidden vector from right to left, the output sequence $(\overleftarrow{h}_1,\overleftarrow{h}_2,\cdots,\overleftarrow{h}_N)$ is obtained.

They are calculated by formulas (1) and (2), where W_x is the word vector look-up matrix of the source language word x_i converted into the word vector. Finally, the output of encoder on each timestamp is a mosaic of forward and backward hidden vectors for each word, namely, $h_i=[\vec{h}_i;\overleftarrow{h}_i]$. The source language input sequence output after encoding is $h=(h_1,h_2,\cdots,h_N)$.

$$\vec{h}_i = f_{enc}(W_x[x_i], \vec{h}_{i-1}). \tag{1}$$

$$\overleftarrow{h}_i = f_{enc}(W_x[x_i], \overleftarrow{h}_{i-1}). \tag{2}$$

3.2 Image Attention Fusion Mechanism

By using the attention mechanism in the decoder of the model, the neural machine translation model based on plain text attention can better obtain the alignment between the target word and a word in the source language text when the decoder translates the next target word. The implementation method of the pure text attention mechanism is shown in equations (3)-(6). Firstly, the hidden state s_t at the moment t and the alignment information $e_{t,i}$ of the hidden state h_i of the encoder are calculated, as shown in equation (3). Secondly, the weight $\alpha_{t,i}$ of different hidden state h_i of the encoder corresponding to the hidden state of the decoder at time t is calculated by the Softmax function, as shown in equation (4). Thirdly, the context vector c_t of the hidden state at the moment t of the decoder is calculated, as shown in equation (5). Finally, the results of the hidden state s_t at the time t of the decoder are calculated from three parts as shown in equation (6). 1) The previous hidden state of the decoder s_{t-1}; 2) output predicted target word $\widetilde{y}_{t-1}$ at time $t-1$; 3) the context vector c_t of the decoder hidden state at time t.

$$e_{t,i} = h_i^T W_a S_{t-1}. \tag{3}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{N} \exp(e_{t,i})}. \tag{4}$$

$$c_t = \sum_{i=1}^{N} \alpha_{t,i} h_i. \tag{5}$$

$$S_t = f_{dec}(W_y[\tilde{y}_{t-1}], s_{t-1}, c_t). \tag{6}$$

In this paper, the visual semantic information of images is integrated into the translation model, and the semantic features of different image regions and the semantic interaction information of source language text can help the model to better align with the source language words at the decoder. Therefore, image attention is integrated into plain text attention, and an enhanced fusion attention is obtained. Similar to the text attention calculation method, the image attention calculation method is shown in equations (7)~ (12). First, the alignment information e_i^{global} and e_i^{localk} (k = 1, 2, …, L) denotes the semantic information of the k-th local image) between the encoder hidden state and the global, local images as shown equations (7)-(8). L is the total number of local area images. p_{global} and p_{localk} are the global and local image features, respectively. They are extracted from the pre-trained CNN network, and the specific extraction method is shown in Fig. 3.

$$e_i^{global} = h_i^T p_{global}. \tag{7}$$

$$e_i^{localk} = h_i^T p_{localk}. \tag{8}$$

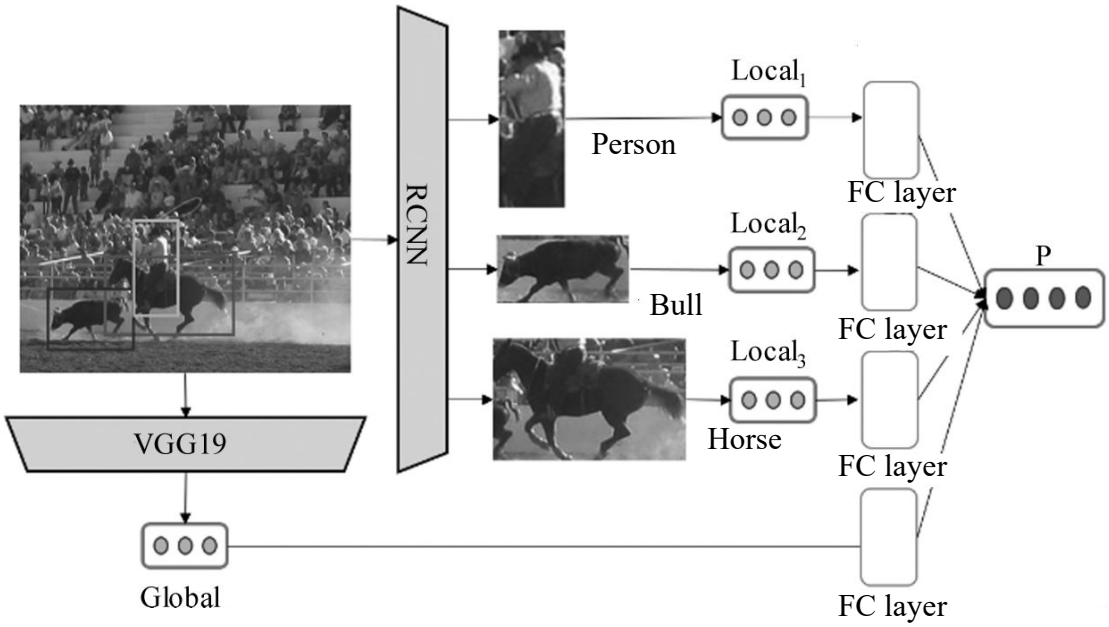


Fig. 3. Fusion hidden representation of global and local region image

Softmax function is used to calculate the sum a_i^{img} of the influence weights between global, local images and the hidden state h_i of the encoder, as shown in equation (9), where *k*= 1, 2, …, *L*. It is noted that the proposed image attention a_i^{img} in this paper is not correlated with the time t of the decoder, which is different from the plain text attention. Because the words associated with the image or appearing in the image in the source language do not change over time during encoding. Therefore, the alignment information calculated by the proposed image attention in this paper is independent of the decoding time t.

$$a_i^{img} = \frac{\exp(e_i^{global})}{\sum_{i=1}^{N}\exp(e_i^{global})} + \frac{\exp(e_i^{global1})}{\sum_{i=1}^{N}\exp(e_i^{global1})} + \cdots + \frac{\exp(e_i^{globalL})}{\sum_{i=1}^{N}\exp(e_i^{globalL})}. \tag{9}$$

Considering that the image semantic information has the same importance degree as the predicted next word for the alignment relation of the word in the source language text, this paper obtains the enhancement effect of image attention through the weight sum calculation. Finally, at hidden state of time t, and the fusion attention weight $\alpha_{t,i}^{fusion}$ is the sum of text attention weight $\alpha_{t,i}$ and image attention weight a_i^{img}, as shown in equation (10), where β is the adjusting parameter, $\beta \in (0,1)$.

$$\alpha_{t,1}^{fusion} = \beta\alpha_{t,i} + (1-\beta)a_i^{img}. \tag{10}$$

$$c_t' = \sum_{i=1}^{N}\alpha_{t,i}^{fusion}h_i. \tag{11}$$

$$s_t' = f_{dec}(W_y[\tilde{y}_{t-1}], s_{t-1}', c_t'). \tag{12}$$

The output s'_t of the model in this paper at the time t is calculated by three parts: 1) The previous hidden state of the decoder s'_{t-1}; 2) output predicted target word $\widetilde{y}_{t-1}$ at time $t-1$; 3) the context vector c'_t of the decoder hidden state at time t. The calculation of c'_t is shown in equation (11), and the calculation of overall s'_t is shown in equation (12).

3.3 Visual Semantic Representations of Fused Different Image Areas

In the multi-modal machine translation task, some important semantic objects in images often appear in the corresponding text sentences. Therefore, extracting the local image features can effectively obtain the interactive correspondence between the image and the text. For this reason, we extract the global feature representation and the local feature representation of different image regions, and send them into the fully connection network respectively. The compressed representation is obtained and spliced, so the visual semantic vector representation p of the image integrating the complete semantic information of the image and the different local image information is obtained, as shown in equation (13). The extracted case diagram is shown in Fig. 3.

$$p = [\tanh(W_g p_{global}); \tanh(W_{l1} p_{local1}); \ldots; \tanh(W_{lN} p_{localN}). \quad \textbf{(13)}$$

3.4 Image Visual Information Fusion Mode

In addition to improving the attention mechanism, this paper also fuses the local and global visual semantic information of images in the encoder and decoder of the model [24]. We integrate the hidden representation vector p of the local and global visual semantic information into the encoder and decoder of the model respectively. There are three methods for visual information fusion. The detailed descriptions are as follows.

Method 1. The image vector p is projected into the same space as the source language word, using it as a pseudo word and an input to the encoder, which is called $Local\_Global_W$. So $d = W_I^2(W_I^1 \cdot p + b_I^1) + b_I^2$, where W_I^1 and W_I^2 are the trainable transformation matrix to transform images into the same space same as the words. b_I^1 and b_I^2 are the bias vectors. d is regarded as the first word of the source language word sequence.

Method 2. The image visual information is used as the initialization vectors of the two directions of the source language sentence encoder LSTM, which are $\vec{h}_0 = \tanh(W_f p + b_f)$ and $\overleftarrow{h}_0 = \tanh(W_b p + b_b)$ respectively. Where W_f and W_b are parameter matrices converting the image vector p into the dimension of the encoder hidden vector. b_f and b_b are the bias vectors. This process is called $Local\_Global_E$.

Method 3. The image visual information is used as the initial hidden vector of the decoder as additional input. The global and local features of the image are spliced through a single-layer feed-forward neural network, and the vector p is obtained. p is projected to a vector with the same dimension as the decoder's hidden state s_i through the parameter matrix W_{img}, which is used to initialize s_0, as shown in equation (14). This fusion way is denoted as $Local\_Global_D$. In the following experiment, we also use this method to initialize the hidden state of the image attention fusion mechanism model.

$$s_0 = \tanh(W_{di}[\vec{h}_N; \overleftarrow{h}_1] + W_{img} p + b_{di}). \quad \textbf{(14)}$$

3.5 Model Training

In this paper, the maximum likelihood conditional probability is used as the loss function to train and optimize

the neural machine translation model based on the attention mechanism [25]. The calculation formula of the loss function is shown in equations (15) and (16).

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} -\log p_{\theta}(y_n \mid x_n) . \tag{15}$$

$$p(y_t \mid y_1, \cdots, y_{t-1}, x) = g(y_{t-1}, s_t, c_t) . \tag{16}$$

Where θ is the model parameter, N is the sample number of the training set. x_n represents the input sequence of the source language. y_n represents the target language output sequence. c_t is the context vector of the hidden state calculated by the encoder at the time t. y_t is the target word. s_t is the hidden state of the decoder at time t.

4 Experiment and Analysis

4.1 Experiment Data

Multi30k [26] as a standard data set in multimodal machine translation is selected as the experimental data. This dataset is an extended version of Flickr30k@ for the image description generation task. Each image in the dataset consists of an English sentence and a German sentence translated by a professional translator. The training set, verification set and testing set have 29,000, 1024 and 1000 images and their corresponding English-German bilingual sentence pairs respectively. In the data preprocessing, we use Moses statistical machine translation library to preprocess scripts for German and English, including word segmentation, punctuation normalization and correct handling of capital and lower-case letter. In the experiment, the Multi30k is used for training, and the models are selected on the verification set. The optimal model is selected to test the testing set. The evaluation index used in the experiment is BLEU4 index commonly used in machine translation.

4.2 Experiment Set

In terms of image semantic extraction, the 4096-dimensional vector in the penultimate fully connection layer of pre-trained VGG19 [27] on the ImageNet [28] is used as the global feature representation of the image. For semantic extraction of local images, this paper uses the RCNN method to extract the local region corresponding to each image. The regional proposal network in the RCNN is pre-trained on the VOC20070 and VOC 20120 datasets. In this paper, the local regions extracted from the regional proposal network are then extracted by VGG19 using the same extraction method of global image features to obtain semantic vector representations of all local regions. Both the forward and backward LSTM hidden state vectors in the encoder bidirectional LSTM are 512 dimensions. The word vector dimension of the source language and target language is 620-dimension. The decoder single directional LSTM hidden state vector is 512-dimension.

This model adopts Adam optimizer, initial learning rate=0.001, batch size=32, training epoch= 26. If the confusion degree value of the 4-round on the verification set does not decrease, then the learning rate attenuation is performed once, and the attenuation rate is set as 0.5. If this situation occurs for 5 times, the training is stopped by using the Early Stop strategy. In addition, the parameter of the fused image attention formula in all experiments is set to 0.9. In the test set, beam search is used inspired by reference [29], and the beam size in the experiment is set to 10.

4.3 Comparison Results and Analysis

In order to verify the effectiveness of the proposed model in this paper, we select the multi-modal machine translation models Huang et al., Calixto et al., and Caglayan et al., as three benchmark methods. At the same time, we also compare the results of different fusion methods using text attention mechanism and proposed image attention mechanism in this paper. In Table 1, $Local\_Global_W$, $Local\_Global_E$, $Local\_Global_D$

are stated in section 2.3. $Local\_Global_{W+E}$, $Local\_Global_{W+D}$, $Local\_Global_{W+E+D}$ are the fused methods. The Fusion_Attention_Global and Fusion_Attention_Global_Local are image attention fusion mechanisms, the former only uses global image features, while the latter uses local and global image features.

Experiment results of English-German translation pairs.

Firstly, we conducted experiments on the English-German translation in the Multi30K data set, and the experiment results are shown in Table 1. It can be seen from Table 1 that the proposed model in this paper has been improved compared with the existing different image fusion models, such as local image fusion model and global image fusion model. Where, the optimal value of (fusion_attention_Global_Local) in this paper improves by 6.1 BLEU, 5.7 BLEU, 5.0 BLEU than Huang et al, Calixto et al and Caglayan et al. respectively. In the proposed image attention fusion model, image features are used to initialize the vector of the decoder. As can be seen from the experimental results in Table 1, both Fusion_Attention_Global and Fusion_Attention_Global_Local are better than Calixto et al.'s work only by fusing Global images and Local_Global method, which verifies the improvement effect of image attention mechanism.

Table 1. English-German translation results

Model	Translation model	BLEU4
RNN-based	Huang et al	37.8
	Calixto et al	38.2
	Caglayan et al	38.9
Proposed	$Local\_Global_{W}$	39.2
	$Local\_Global_{D}$	39.4
	$Local\_Global_{W+D}$	39.7
	$Local\_Global_{W+E+D}$	40.9
	Fusion_Attention_Global	42.6
	Fusion_Attention_Global_Local	43.9

Meanwhile, the proposed method is by fusing the global semantic features and the local semantic features of the image, the overall translation effect is improved. This shows that the fusion of local semantic information in different regions of images can indeed provide more semantic information for the text machine translation model, which is used to help improve the translation quality.

Experiment results of Indonesian-Chinese translation pairs.

This paper also tests the effect of the experiment on the data of small languages with scarce resources. For testing, we manually label all the English-German translation pairs in Multi30K. All data are marked as one-to-one corresponding Indonesian-Chinese translation pairs. For the Indonesian-Chinese training set data, Google tool is used to translate English-German translation sentence pairs into corresponding Indonesian-Chinese translation sentence pairs respectively, so as to obtain the corresponding training set data. The experimental results of the proposed model in this paper and the different fusion methods for Indonesian-Chinese translation pairs are shown in Table 2. The model corresponding to the first line in Table 2 (Text-only NMT) adopts neural machine translation based on plain text attention. The difference between Text-only NMT and the proposed model in this paper is that it does not fuse any visual information of images and adopts the plain text attention mechanism.

Table 2. Indonesian-Chinese translation results

Method	Translation model	BLEU4
Baseline model	Text-only NMT	27.59
Proposed	$Local\_Global_{W}$	28.26
	$Local\_Global_{D}$	27.94
	$Local\_Global_{W+D}$	27.91
	Fusion_Attention_Global	29.82
	Fusion_Attention_Global_Local	29.86

As can be seen from the experimental results in Table 2, compared with the neural machine translation model with plain text attention mechanism, the improvement of the proposed model in this paper is more obvious. The optimal results of the model presented in this paper improved by 2.27 BLEU values on the Indonesian-Chinese translation pair data test set compared to the plain text attention model. However, the model without image attention, but with image visual information (visual semantic information integrating local and global information), has a certain improvement compared with the plain text attention translation model without any image visual information. For example, the image vector p is projected into the same space as the source

language word. It is treated as a pseudo word as input to the encoder, the result is improved by 0.67 BLEU values compared to the plain text attention model without the visual semantic information of the image.

At the same time, we find that the model in Indonesian-Chinese small language data sets is below the results of the English-German translation in terms of BLEU4 value. On the one hand, there may be big noise in the training data sets after Google processing. On the other hand, because Indonesian and Chinese belong to two different language families and their data distribution is different.

4.4 Case Study

In order to better demonstrate the effect of the proposed model in this paper, a case study is conducted on the translation results. This paper selects the translation results of Indonesian-Chinese translation sentence pairs to carry on the case analysis. Two images, the model based on pure text attention and the reference manual translation sentence are selected respectively, and the translation results of the proposed model in this paper are compared. Fig. 4 and Fig. 5 are the corresponding pictures of the two cases respectively. Table 3 and Table 4 are the translation results of different models in the two pictures.

Fig. 4. Image of case 1

Fig. 5. Image of case 2

Table 3. Machine translation effects of different models on case 1

Translation model	Sentence
Original Indonesian	wanita yang berjubah merah menari dengan pria berjas
Manual Chinese translation	身穿红色连衣裙的女士正与穿着西装的男子**共舞**
Text-only NMT	穿着红色长袍的女士和穿着西装的男人
Local_Global$_w$	穿着红色长袍的女人与穿着西装的男人一起**跳舞**
Fusion_attention_global	穿着红色长袍的女人与穿着西装的男人**共舞**
Fusion_attention_global_local	红色长袍的妇女**跳舞**与穿着西装的男人

Table 4. Machine translation effects of different models on case 2

Translation model	Sentence
Original Indonesian	polisi anti huru hara berdiri di belakang sementara seorang pria muda dengan syal merah menutupi wajahnya berjalan
Manual Chinese translation	防暴警察站在后面，一名戴着红色围巾的年轻男子**捂着脸**走路
Text-only NMT	防暴警察站在后面，而一名戴着红色围巾的年轻人在她的脸上 <unk>
Local_Global$_w$	警察防暴警察站在后面，而一个年轻人戴着红色围巾 <unk> 走
Fusion_attention_global	防暴警察站在后面，而一个戴着红色围巾的年轻人正在走路
Fusion_attention_global_local	防暴警察站在后面，而一个年轻人用一条红色的围巾遮住了走去

Fig. 4 corresponding to Case 1 is taken as an example, and the experimental results of different models are shown in Table 3. From the results of the experiment, it can be seen that the original Indonesian text of the image is “wanita yang berjubah merah menari dengan pria berjas.” The artificial Chinese translation is “身穿红色连衣裙的女士正与穿着西装的男子**共舞**”. The text-only attention model (which does not incorporate visual semantic information from images) translates it as “穿着红色长袍的女士和穿着西装的男人”. We can see that the translation results do not translate the movement “**共舞**”, but the proposed model in this paper translates it well.

According to Fig. 5 corresponding to Case 2, the experimental results of different models are shown in Table 4. The experimental results show that the proposed model translates the sentence well for Indonesian sentence,

while the plain text model translates the sentence as "防暴警察站在后面，而一名戴着红色围巾的年轻人在她的脸上 <unk>". It is relatively bad. It is also found from the experimental results that the result of fusing local and local image information in the proposed model is slightly higher than the result of fusing global image information in the BLEU4 index. However, the latter is better than the former in some translation effects.

4.5 Visualization Analysis

To observe if the image information can better align the words of the source language, we separately calculate the weight a_i^{img} of the image for each source language word h_i, as shown in Fig. 6.

The words with the highest weight are highlighted in black bold. It can be seen that in the four examples, words with high weight basically appear in the picture. For example, in Fig. 6(a) "swimming" and "pool" are aligned; Fig. 6(b) "uniform" "palms" are aligned, etc., Fig. 6(c) "rides", "bull" are aligned, etc.; Fig. 6(d) "playing", "soccer" are aligned, etc. Some key words, such as "girl", "boy" and other prominent local objects in the figure, are not well captured. Meanwhile, the weight calculated by some sentences is relatively average. We believe that this has something to do with the features of the extracted local images. Because the pre-training model used in the extraction is pre-trained on a small data set, the accuracy and quantity of the identified objects may be not accurate enough.

(a) **A young girl** swimming **in a** pool

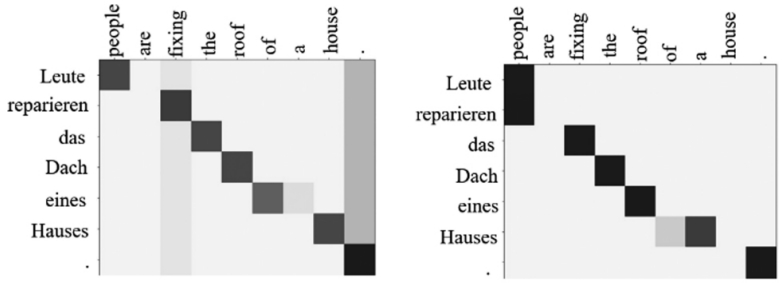
(b) **A young boy in a soccer** uniform crying **into his** palms

(c) **A man** rides **a** kicking bull **in a bullpen**

(d) **Four people are** playing soccer **on a beach**

Fig. 6. The word with the highest weight of the image to source language sentence

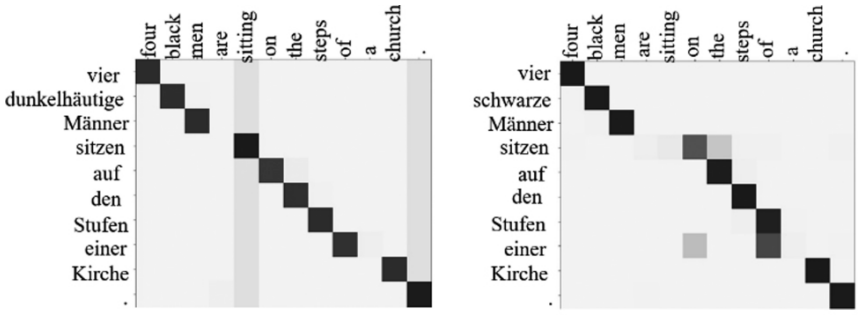
In order to demonstrate the enhanced effect of the proposed image attention on text attention, some examples are visualized to show the comparison of the visualization effect of alignment using text attention and fused image attention. As shown in Fig. 7, the visualization analysis of case 1 is a pair of English-German translated sentences, in which the source language sentence is "people are fixing the roof of a house" and the target language sentence is "Leute reparierendas Dach eines Hauses". We can see that the target language word "Leute" aligns with the source language word "people" in both the text attention and the fused image attention proposed in this paper. And the word "reparieren" misaligns to the source language word "people" with text attention. In the fused image attention model in this paper, the word "reparieren" is correctly aligned to the word "fixing" in the source language. Similarly, the word "Dach" is misaligned to the word "the" in the text attention model, and correctly aligned to the word "roof" in the presented model.



(a) Image fusion attention (b) Text-only attention

Fig. 7. Visual analysis of case 1

Also in the English-German sentence pair "Four black men are sitting on the steps of a church" and "Vier Schwarze Manner sitzen auf den Stufen einer Kirche." In the case (see visualized case study 2 in Fig. 8), the words "Sitzen" and "Stufen" in the target language are both misaligned in the text attention, while in our model, the words "Sitting" and "Steps" are correctly aligned, indicating the effectiveness of our method.



(a) Image fusion attention (b) Text-only attention

Fig. 8. Visual analysis of case 2

5 Conclusion

In recent years, the research on multi-modal machine translation has become one of the new research hotspots, because the multi-modal machine translation model fuses various information of different models and better improves the results of machine translation. In this paper, a multi-modal machine translation model with image attention fusion is proposed for the task of integrating image visual semantic information. By integrating the visual semantic information of different parts of the image, the global visual semantic information and the interaction information of different hidden states (source language text) of the encoder into the pure text attention as the image attention, an enhanced multi-modal machine translation model of image attention is obtained.

The model is tested on Multi30k English-German translation pair and Indonesian-Chinese translation pair respectively. The results show that compared with the existing baseline system (integrating image visual semantic information into the translation model from different perspectives), the proposed model in this paper has a good improvement, especially when translating the dataset of Indonesian and Chinese, the improvement is more obvious, which verifies the effectiveness of the proposed model in this paper.

In this paper, a multi-modal machine translation model based on the framework of RNN is developed, which integrates the visual semantic information of images. The proposed model is verified on the experimental data set. The follow-up work will further explore the multi-modal machine translation model based on "Transformer" that integrates visual semantic information of images. In future work, we will make improvements in the image-assisted text task using advanced deep learning method. This paper will attempt to use the image description generation model to get the hidden states of image features, and adopt the information of the hidden states of image features to further optimize the translation of source text through the attention mechanism.

References

[1] T. Baltrušaitis, C. Ahuja, L. Morency, Multimodal Machine Learning: A Survey and Taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2)(2019) 423-443.

[2] X. Li, J. Ma, S. Qin, Image Attention Fusion for Multimodal Machine Translation, Journal of Chinese Information Processing 34(7)(2020) 68-78. (In Chinese)

[3] X. Wang, S. Yin, D. Liu, H. Li, Accurate playground localisation based on multi-feature extraction and cascade classifier in optical remote sensing images, International Journal of Image and Data Fusion 11(3)(2020) 233-250.

[4] J. Libovický, J. Helcl, Attention Strategies for Multi-Source Sequence-to-Sequence Learning, in: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017.

[5] M. Zhou, R. Cheng, Y.-J. Lee, Z. Yu, A Visual Attention Grounding Neural Model for Multimodal Machine Translation, in: Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[6] S. Yao, X. Wan, Multimodal Transformer for Multimodal Machine Translation, in: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[7] T. Nishihara, A. Tamura, T. Ninomiya, Y. Omote, H. Nakayama, Supervised Visual Attention for Multimodal Neural Machine Translation, in: Proc. of the 28th International Conference on Computational Linguistics, 2020.

[8] Y. Zhao, M. Komachi, T. Kajiwara, C. Chu, Double Attention-based Multimodal Neural Machine Translation with Semantic Image Regions, in: Proc. of the 22nd Annual Conference of the European Association for Machine Translation, 2020.

[9] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[10]I. Calixto, Q. Liu, N. Campbell, Incorporating Global Visual Features into Attention-Based Neural Machine Translation, 2017.

[11]I. Calixto, Q. Liu, An error analysis for image-based multi-modal neural machine translation, Machine Translation 33(1-2) (2019) 1-23.

[12]P. Huang, F. Liu, S. Shiang, J. Oh, C. Dyer, Attention-based Multimodal Neural Machine Translation, in: Proc. of the First Conference on Machine Translation: Volume 2, 2016.

[13]O. Caglayan, W. Aransa, A. Bardet, M. García-Martínez, F. Bougares, L. Barrault, M. Masana, L. Herranz, J. van de Weijer, LIUM-CVC Submissions for WMT17 Multimodal Translation Task, in: Proc. of the Second Conference on Machine Translation, 2017.

[14]J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, J. Xie, A Hierarchy-to-Sequence Attentional Neural Machine Translation Model, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(3)(2018) 623-632.

[15]J. Mao, W. Xu, Y. Yang, J. Wang, A. L. Yuille, Explain images with multimodal recurrent neural networks. <https://arxiv.org/abs/1410.1090 >, 2014 (accessed 14.10.10).

[16]J. Kanerva, F. Ginte, T. Salakoski, Universal Lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks, Natural Language Engineering 27(5)(2020) 1-30.

[17]L. Dinh, D. Krueger, Y. Bengio, NICE: Non-linear Independent Components Estimation. <https://arxiv.org/abs/1410.8516> , 2014 (accessed 14.10.08).

[18]S. Qu, Y. Xi, S. Ding, Visual attention based on long-short term memory model for image caption generation, in: Proc. of 2017 29th Chinese Control and Decision Conference (CCDC), 2017.

[19]T. Kano, S. Sakti, S. Nakamura, End-to-End Speech Translation with Transcoding by Multi-Task Learning for Distant Language Pairs, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28(2020) 1342-1355.

[20]O. Firat, K. Cho, B. Sankaran, F.T.Y. Vural, Y. Bengio, Multi-way, multilingual neural machine translation, Computer Speech & Language 45(2017) 236-252.

[21]Y. Chu, C. Guo, T. He, Y. Wang, J. Hwang, C. Feng, Inductive Embedding Learning on Attributed Heterogeneous Networks via Multi-task Sequence-to-Sequence Learning, in: Proc. 2019 IEEE International Conference on Data Mining (ICDM), 2019.

[22]Q. Shi, S. Yin, K. Wang, L. Teng, H. Li, Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation, Evolving Systems (2021).

[23]Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, X. Li, Describing Video With Attention-Based Bidirectional LSTM, IEEE Transactions on Cybernetics 49(7)(2019) 2631-2641.

[24]J. Makin, D. Moses, E. Chang, Machine translation of cortical activity to text with an encoder-decoder framework, Nature Neuroscience 23(4)(2020).

[25]H. Choi, K. Cho, Y. Bengio, Fine-Grained Attention Mechanism for Neural Machine Translation, Neurocomputing 284 (2018) 171-176.

[26]D. Elliott, S. Frank, K. Sima'An, L. Specia, Multi30K: Multilingual English-German Image Descriptions, in: Proc. of the 5th Workshop on Vision and Language, 2016.

[27]X. Wang, S. Yin, K. Sun, et al. GKFC-CNN: Modified Gaussian Kernel Fuzzy C-means and Convolutional Neural Network for Apple Segmentation and Recognition, Journal of Applied Science and Engineering 23(3)(2020) 555-561.

[28]S. Yin, H. Li, L. Teng, Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images, Sensing and Imaging 21(2020) 49.

[29]I. Sutskever, O. Vinyals, Q. Le, Sequence to sequence learning with neural networks, in: Proc. of Advances in Neural Information Processing Systems, 2014.