IEEE Xplore ®
Notice to Reader

“Multimodal Machine Translation”
by Jiatong Liu
published in IEEE *Access* Early Access
Digital Object Identifier: 10.1109/ACCESS.2021.3115135

It is recommended by the Editor-in-Chief of IEEE *Access* that this article will not be published in its final form.

We regret any inconvenience this may have caused.

Derek Abbott
Editor-in-Chief
IEEE *Access*

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Multimodal Machine Translation

JIATONG LIU[1]

[1]School of Informatics, Xiamen University, Xiamen, 10384 China (e-mail: liujiatong@stu.xmu.edu.cn)

This work was supported by the Key Project of National Key R&D Project(No.2017YFC1703303); External Cooperation Project of Fujian Province, China(No.2019I0001); Science and Technology Guiding Project of Fujian Province, China(2019Y0046);Industry University Research Collaboration Project between Ningde city and Xiamen university (No.2020C001).

ABSTRACT In recent years, neural network machine translation, especially in the field of multimodality, has developed rapidly. It has been widely used in natural languages processing tasks such as event detection and sentiment classification. The existing multimodal neural network machine translation is mostly based on the autoencoder framework of the attention mechanism, which further integrates spatial-visual features. However, due to the ubiquitous lack of corpus and the semantic interaction between multimodalities, the quality of machine translation is difficult to guarantee. Therefore, this paper proposes a multi-modal machine translation model that integrates external linguistic knowledge. Specifically, on the encoder side, we adopt the pre-trained Bert model to be used as an additional encoder to integrate with the original text encoder and picture encoder. Under the cooperation of the three encoders, a better text representation and picture representation at the source end is generated. Besides, the decoder decodes and generates a translation based on the image and text representation of the source. To sum up, this paper studies the visual-text semantic interaction on the encoder side and the visual-text semantic interaction on the decoder side, and further improves the quality of translation by introducing external linguistic knowledge. We compared the performance of the multimodal neural network machine translation model with pre-trained Bert and other baseline models in English German translation tasks on the multi30k data sets. The results show that the model can significantly improve the quality of multimodal neural network machine translation, which also verifies the importance of integrating external knowledge and visual text semantic interaction.

INDEX TERMS Multi-domain, Machine Translation, Semantic Interaction, External Knowledge

I. INTRODUCTION

THE real world that human beings live in is a space where text, sound, image, and video coexist. For thousands of years, humans have exchanged information with each other in a variety of ways, such as language, text, and images, and using multiple modalities at the same time can clearly convey information more fully and accurately. Multi-information fusion is an important research trend. However, most of the existing machine translation models only use text data for translation. How to integrate text, image, and video information to improve the quality of translation is a topic worthy of study. As Kalchbrenner et al. proposed the concept of neural network machine translation in 2013, it soon achieved results comparable to, or even better than, traditional statistical machine translation, and it has gradually become a research hotspot.

At present, the mainstream models can be divided into three categories: the first type of model only uses the text attention mechanism, and the image is only used as auxiliary information to improve the generation of text representation. Huang et al.(2016) by integrating the semantic representation of the image as an additional input into the encoder. Calixto et al(2017b) further studied how to use the semantic representation of the image to initialize the hidden state of the decoder. In the framework of multi-task learning, Elliott and Kadar (2017) decompose multi-modal translation into learning translation models and visual representations. In this way, multi-modal models can be trained on parallel text or external data sets describing images, making it possible to use existing resources. Qian et al. (2018) proposed a new algorithm based on the advanced actor-critical algorithm (Bahdanau et al., 2017) to study the effectiveness of reinforcement learning in multi-modal NMT.

The second type of model believes that both text and image information are crucial in multimodal neural network translation. Therefore, two attention mechanisms are simultaneously used to capture text and image contexts for translation. In this regard, Caglayan et al. (2016a, b) first proposed an end-to-end attention multi-modal NMT model, which effectively integrates text and image information into the existing machine translation framework by sharing parameters. In addition, Calixto et al. (2017a) will introduce two independent attention mechanisms for text and image information. Delbrouck et al. (2017) empirically studied the effectiveness of enhanced visual and textual representation to improve the quality of multimodal neural machine translation. The third type of model uses semantic interaction to refine the learned image semantics. Delbrouck and Dupont (2017b) apply a multi-modal compressed bilinear pooling operation to remove the noise information represented by the image based on the text representation. Recently, Yin et al. (2020) proposed a graph-based multi-modal fusion encoder, which is based on a unified graph representing various semantic relations between multi-modal semantic units. Lin et al. (2020) introduced a capsule network to better dynamically extract translation image features. Yang et al. (2020) jointly trained source-to-target and target-to-source translation models, and encouraged these models to share visual information when generating semantically equivalent visual words. However, these models only use visual information to optimize the semantic representation of text, ignoring the strong semantic association between text and image.

In this paper, according to the characteristics of the multi-modal neural network machine translation model, the BERT model is introduced as an additional encoder to encode the input sentence, and then it is used for decoding with the original visual encoder and text encoder of the multi-modal neural network model. , So that the visual context vector and text context vector can learn more external linguistic knowledge to improve their representation ability. In summary, this article has the following innovations:

i) Aiming at the existing problem of lack of semantic interaction in multimodal neural network machine translation, the visual-text semantic interaction on the encoder side and the visual-text semantic interaction on the decoder side are studied separately.

ii) In response to the lack of multimodal machine translation corpus, the introduction of external linguistic knowledge further improves the quality of translation.

ii) Experiments were performed on multiple language pairs on the Multi30k data set, and the results all show that the model in this paper can significantly improve the quality of multimodal neural network machine translation.

II. RELATED WORK

A. BERT

Pre-training technology has a long history in the field of machine learning and natural language processing, and its related applications can be traced back to (Erhan et al., 2010). Since then, Mikolov et al. (2013) and Pennington et al. (2014) pioneered Xindi proposed a word embedding representation, and this pre-training technique was widely used at that time. Dai & Le (2015) trained an autoencoder using unlabeled data and then used the model for downstream tasks. As the scale of data is getting larger and larger and deep neural network models are widely used, pre-training technology has been widely used and has achieved remarkable results, but it has also received more and more attention. Peters et al. (2018) designed ELMo based on the two-way cyclic long- and short-term memory unit, and input the pre-trained ELMo as global information into downstream tasks. In 2018, Radford designed the language model GPT based on Transformer, which uses unlabeled data for pre-training and fine-tuned through specific downstream tasks. Drawing lessons from the design ideas of Transformer model encoder, Devlin et al. designed the BERT model in 2019, which is widely used for the initialization of downstream task models. On the basis of BERT, many variant models have been derived, such as the multilingual pre-training model XLM (Lample & Conneau, 2019), which introduces more unlabeled data and removes the "NSP (predict next sentence)" module of RoBERTa (Liu et al., 2019), and XLNet based on permutation modeling method (Yang et al., 2019b). In recent years, with a large number of pre-training techniques/models, such as: ELMo (Peters et al., 2018), GPT/GPT-2 (Radford et al., 2018) , BERT (Devlinet al., 2019) and cross-language language XLM (Lample & Conneau, 2019), XLNet (yang et al., 2019b) ,RoBERTa (Liu et al., 2019) and other models have refreshed the performance records in the corresponding field time and time again, and the pre-training technology has attracted widespread attention from the machine learning and natural language processing communities. These models are pre-trained on a large amount of unlabeled data to better learn the representation of the model input. These models are then used to provide context-aware word embedding representations of the input sequence for downstream tasks (Peters et al., 2018) or to initialize model parameters for downstream tasks. Practice has shown that in natural language understanding tasks, effective use of this type of pre-trained model can effectively improve the performance of the model. BERT and its variant models have been widely used in tasks such as natural language understanding tasks and text classification, and have greatly promoted the development of corresponding fields. Multimodal neural network machine translation aims to simultaneously use source language sentences and corresponding visual information to obtain high-quality target language translations. In this process, the input sentence and image need to be encoded first, and then decoded to generate the target language sentence. Today, BERT has been proven to improve the model performance of many natural language processing tasks, so it is of great significance to study the application of BERT in the direction of multimodal machine translation. Limited by equipment and computing power, the cost of retraining BERT is unbearable. Therefore, this article mainly focuses on: by introducing the pre-trained BERT model, the multi-modal machine translation model can learn more external linguistic knowledge to improve the translation quality of the model.

III. METHOLOGY

In this section, we design a multi-modal neural network machine translation model incorporating pre-trained BERT,

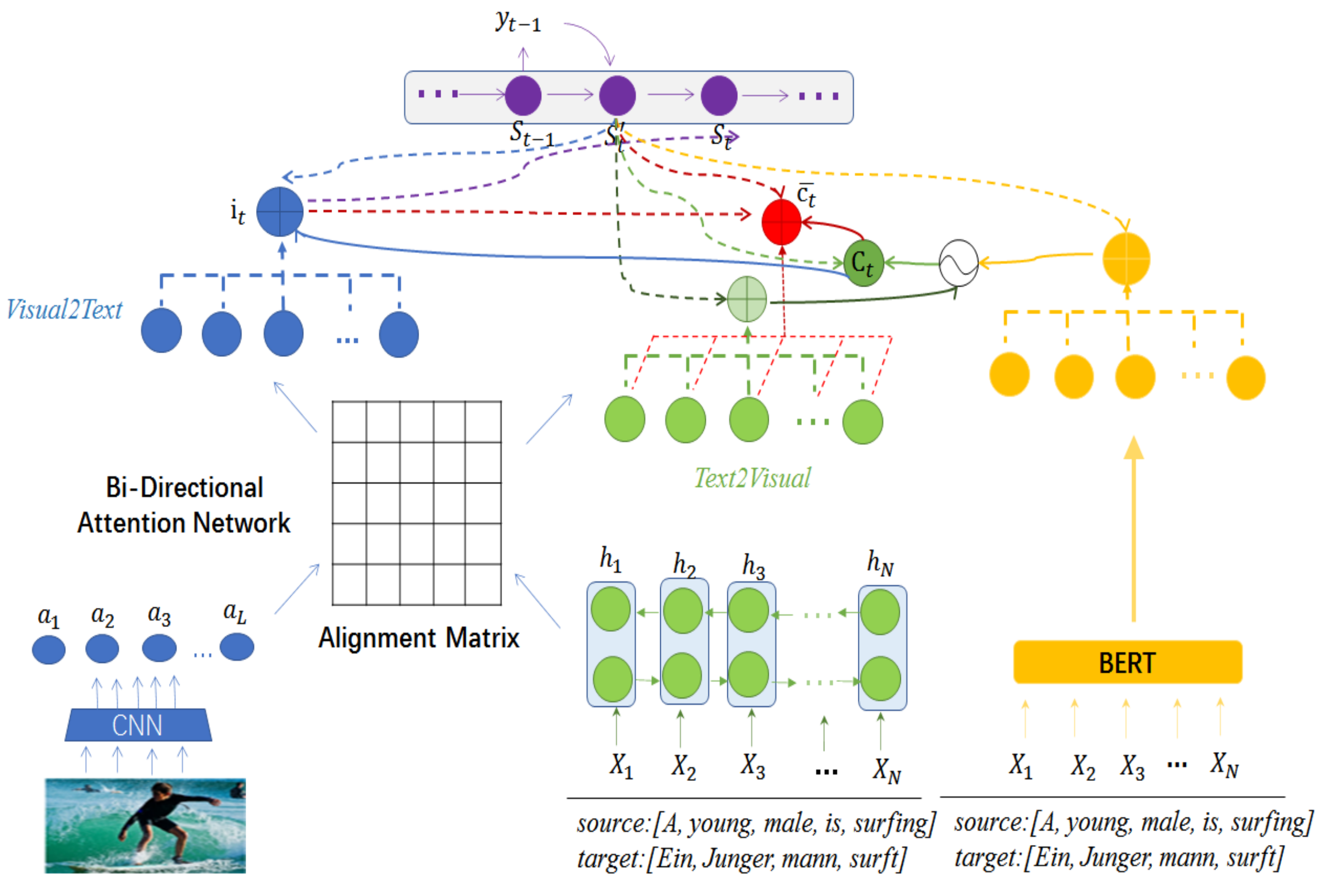

y_{t-1}
S_{t-1}
S'_t
S_t
i_t
$\bar{c}_t$
C_t
Visual2Text
Text2Visual
Bi-Directional Attention Network
Alignment Matrix
a_1 a_2 a_3 ... a_L
CNN
h_1 h_2 h_3 h_N
X_1 X_2 X_3 ... X_N
BERT
X_1 X_2 X_3 ... X_N
source:[A, young, male, is, surfing]
target:[Ein, Junger, mann, surft]
source:[A, young, male, is, surfing]
target:[Ein, Junger, mann, surft]

FIGURE 1. The overall framework, which consists of a image feature extraction module, an attention based LSTM module and a joint leading re-position relation network for figure question answering.

as shown in Figure .1, using BERT to encode the input text sequence. Due to the vocabulary size and sub-word unit division of the BERT model, there may be differences from the existing multi-modal model. Solve the problem of embedding dimension and sentence length alignment by introducing two mapping matrices. Then, the hidden layer on the decoder side is used to pay attention to the hidden layer sequence encoded by BERT to obtain an additional context vector; the context vector and the original text context vector are merged through a gate to generate an integrated linguistic knowledge Text context vector. The model designed in this paper includes a visual encoder, an RNN text encoder, an additional pre-trained BERT text encoder, and a decoder. We will introduce these modules in detail.

A. VISUAL ENCODER

Given the picture I and the source language description sentence $X = (x_1, x_2, \cdots, x_N)$ of the picture, where n is the length of the source language sentence. And the corresponding target language translation $Y = (y_1, y_2, \cdots, y_M)$, where m is the length of the target language sentence. The goal of multimodal neural network machine translation is to construct an end-to-end neural network model to model $P = (Y \mid X, I)$. In this model, a pre-trained Bert model is added to the multimodal translation model. The source language sentences are encoded by the original coder in **MNMT** model and the pre trained Bert model respectively to obtain the hidden layer sequence C and the hidden layer sequence Q. In addition, the visual encoder encodes the picture to represent A. Then, the decoder decodes the encoded text sequence representation C and Q and the encoded visual sequence a according to the conditional probability formula 5:

$$\log p(Y \mid X, I) = \sum_{i=1}^{M} \log \left(y_{<t}, C, A, Q\right) \tag{1}$$

In practice, researchers often use gated recurrent neural network (Gru) (CHO et al., 2014) as the implementation of recurrent neural network: specifically, the network uses forward encoder $\vec{\Phi}_{enc}$ and reverse encoder $\overleftarrow{\Phi}_{enc}$ to encode the input sentences from two directions to generate forward hidden layer sequence $\left(\vec{h}_1, \vec{h}_2, \cdots, \vec{h}_N\right)$ and reverse hidden layer sequence $\left(\overleftarrow{h}_1, \overleftarrow{h}_2, \cdots, \overleftarrow{h}_N\right)$ respectively. The specific generation process is shown in formulas 2 and 3:

$$\overrightarrow{\mathrm{h}}_{\mathrm{i}} = \vec{\Phi}_{\mathrm{enc}}\left(\mathrm{E_x}\left[\mathrm{x_i}\right], \overrightarrow{\mathrm{h}}_{\mathrm{i-1}}\right) \tag{2}$$

$$\overleftarrow{h}_i = \overleftarrow{\Phi}_{enc}\left(E_x\left[x_i\right], \overleftarrow{h}_{i-1}\right) \tag{3}$$

Where $\vec{\Phi}_{enc}$ and $\overleftarrow{\Phi}_{enc}$ are GRU activation functions in two directions respectively, and $E_x\left[x_i\right]$ represents the word vector corresponding to the source word x_i. The final hidden layer vector in a given time step is composed of forward and reverse hidden layer vectors $h_i = \left[\vec{h}_i; \overleftarrow{h}_i\right]$. Based on this, we can use the hidden layer vector sequence $C = (h_1, h_2, \cdots, h_N)$ to represent the input sentence.

As shown in Figure 2, the visual encoder adopts the pre trained convolutional neural network, and the parameters of the encoder do not participate in the update during training. Specifically, the encoder is a 50 layer residual network (resnet-50) (he et al., 2016) to encode the visual semantic information into a matrix $A = (a_1, a_2, \cdots, a_{196}), a_i \in R^{1024}$ and each line is composed of a 1024 dimensional feature vector encoding a specific image region. Since the purpose of acquiring visual representation is to initialize the hidden layer state (vector with dimension of 256) of decoder, a two-layer full connected layer is used to transform the dimension of visual representation. In addition, the forgetting layer is added to the network to improve the robustness of the model and make it have stronger generalization ability.

After generating text hidden layer sequence and visual representation by using bidirectional recurrent neural network text coder and visual coder, fine-grained semantic interaction between text and vision is realized under the action of bidirectional attention mechanism, and the improved text hidden layer sequence is represented as $\bar{C}$ and $\bar{A}$

IV. PRE-TRAINING BERT

As shown in Figure 3, the Bert model is mainly composed of bidirectional transformers. In the Bert model, the context information on the left and the context information on the right are considered in the process of generating the representation of each layer. Given the input source language sentence sequence $(E_1, E_2, \cdots, E_N)$, where represents the i-th subword unit in the input sentence. The pre-training Bert model encodes the input sequence into hidden layer sequence Q according to formula 4

$$Q = \Phi_{enc_{BERT}}\left(E_1, E_2, \cdots, E_N\right) \tag{4}$$

1) Decoder

The decoder is a conditional threshold control unit (cGRU) with four independent attention mechanisms, three of which are used to process text information and the other is used to process visual information. Specifically, cGRU consists of two stacked GRU activation units REC_1 and REC_2. At time t, REC_1 employ the hidden layer vector s_{t-1} of the previous time and the target word y_{t-1} to generate the target word y_t by using the formulas 5

$$\begin{aligned} s_t' &= (1 - z_t') \bigodot \underline{s}_t' + z_t' \odot s_{t-1} \\ \underline{s}_t' &= \tanh\left(W' E_y\left[y_{t-1}\right] + r_t' \odot \left(U' s_{t-1}\right)\right) \\ r_t' &= \sigma\left(W_r' E\left[y_{t-1}\right] + U_r' s_{t-1}\right) \\ z_t' &= \sigma\left(W_z' E\left[y_{t-1}\right] + U_z' s_{t-1}\right) \end{aligned} \tag{5}$$

where s_{t-1} represents the hidden layer state of GRU unit REC_1 at the previous time, $\underline{S}_t'$ means the new memory of GRU unit REC_1. Z_t' is the update gate of REC_2 which determines the fusion of the newly generated memory and the hidden state of REC_1 mode. r_t is the REC_2 reset gate which determines the importance of the hidden state of REC_2 to the generation of new memory. W_r, U_r, W_Z, U_z are parameter sets which are used to generating reset gate r_t and updating gate z_t of REC_2.

During the above process, based on the temporary hidden layer vector s_t' and the improved text hidden layer state sequence $\bar{C}$, text attention mechanism uses the following formula to generate a time independent temporary text context vector C_{t_temp}.

$$c_{t-temp} = f_{att-text}\left(C, S_t'\right) \tag{6}$$

Meanwhile, based on the temporary hidden layer vector s_t' and the hidden layer state sequence Q, the text attention mechanism adopts the formula 7 generate a time independent temporary text context vector C_{t_BERT}:

$$c_{t-BERT} = f_{att-text}\left(Q, S_t'\right) \tag{7}$$

Then, a threshold unit $g(*)$ is generated by using the temporary hidden layer vector under the action of the forward neural network. The threshold unit is used to fuse the temporary context vector generated by the bidirectional cyclic neural network encoder c_{t-temp}. And the temporary context vector generated based on the pre trained Bert encode c_{t-BERT} to get the text context vector C_t. The calculation process is shown in formula 8

$$c_t = g(c_{t-temp}, c_{t-BERT}) \tag{8}$$

At the same time, the visual attention mechanism uses the temporary hidden layer vector s_t' and the visual feature matrix A to adopt the formula 9 generate time independent visual context vector i_t:

$$i_t = f_{att-img}(A, s_t') \tag{9}$$

Then, under the use of collaborative attention mechanism, the temporary hidden layer vector s_t', the text context vector C_t, and the text context vector i_t. And the visual up and down vectors realize the high-level semantic interaction between text and vision, and generate the hidden layer of the current moment vector S_t. Finally, based on the hidden layer state y_{t-1}, the calculation process of context vector C_t and visual context vector i_t is get by $p\left(y_t \mid y_{<t}, C, A\right) \propto$ $\exp\left(L_o \tanh\left(L_s s_t + L_w E_y\left[y_{t-1}\right] + L_{cs} c_t + L_{ci} i_t\right)\right)$, where

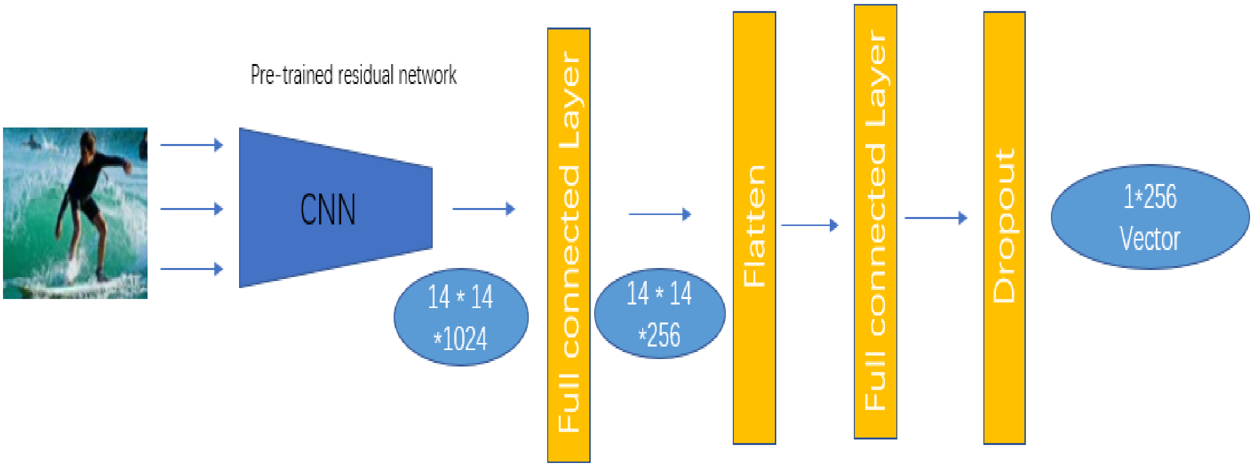


FIGURE 2. The overall framework, which consists of a image feature extraction module, an attention based LSTM module and a joint leading re-position relation network for figure question answering.

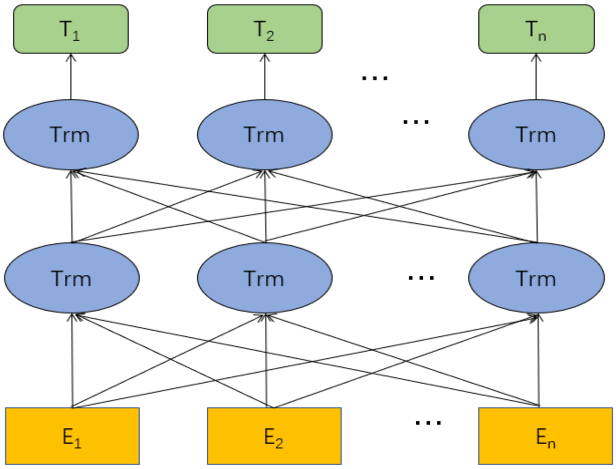


FIGURE 3. The overall framework, which consists of a image feature extraction module, an attention based LSTM module and a joint leading re-position relation network for figure question answering.

$L_o, L_s, L_w, L_{cs}, L_{ci}$ are hyper parameters corresponding to the model.

V. EXPERIMENTS

A. EXPERIMENTAL SETTING

The experiment in this paper also uses the M30k data set, and the model parameter settings are the same as the multimodal neural network machine translation based on deep semantic interaction described above. Each instance in M30KC consists of one image, five English descriptions and five German descriptions in a triad, where the English and German descriptions are independent of each other. For the experiments, the data set was divided as follows: training set of 29,000 triples, validation set of 1014 triples and test set of 1,000 triples.

For visual information, this paper uses a pre-trained 50-layer residual neural network to extract the local features of the image. As shown in Figure 4, this paper performs a series of preprocessing operations on the text data before the model training.

It is worth noting that this paper mainly studies the English German translation based on the m30k public data set. Because the model proposed in this paper is based on the encoder to represent the context information of sentences, and the low data resource language output integrating external semantic information is realized through the decoder, so as to establish the end-to-end sequence mapping from the source language to the target language. Therefore, the model can also be applied to other language translations, such as Slavic ones, as long as there is a corresponding corpus for training.

B. BASELINE

Next, in order to verify the effectiveness of the introduction of external linguistic knowledge for multi-modal neural network machine translation, this chapter designs multiple sets of comparative experiments. The following describes the models involved in the experiments Baseline Model: In this paper, to verify the advantages of the deep semantic interaction-based MNMT model designed in this paper, we compare it with the following mainstream models: Parallel RCNN [37]: this model uses an encoder that contains multiple encoding threads with long- and short-term memory units in each encoding thread [38] share parameters. MNMT [39]: this model introduces two separate attentional mechanisms that utilize image features and text sequences to decode and generate translations. IMG [40]:This model uses image features as additional input to initialise the implicit units of the decoder. Soft-Attention [41]: this model uses an encoder-decoder framework that not only considers the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3115135, IEEE Access

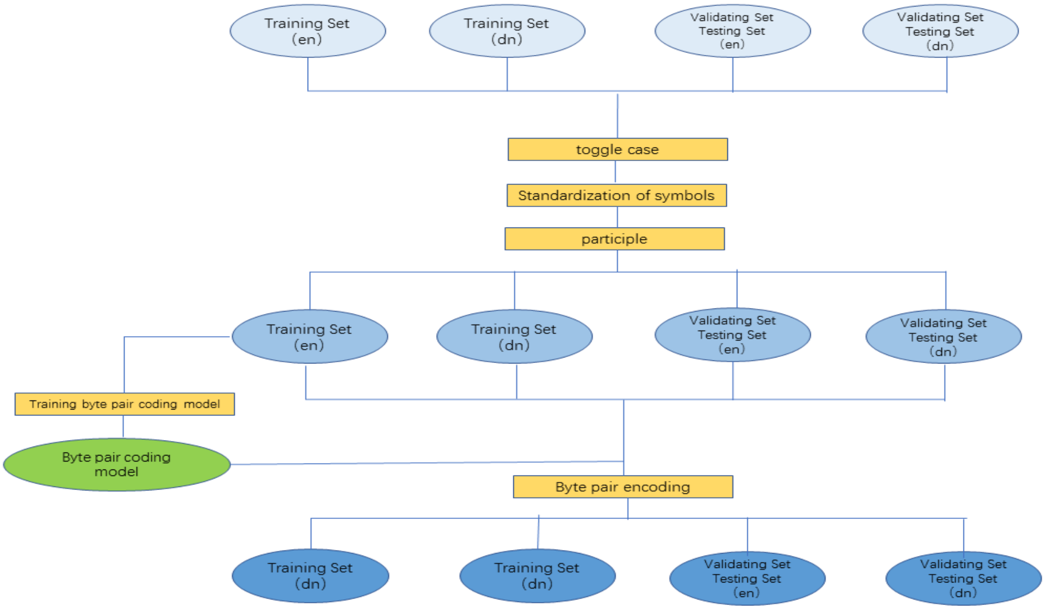

Training Set (en)
Training Set (dn)
Validating Set Testing Set (en)
Validating Set Testing Set (dn)
toggle case
Standardization of symbols
participle
Training Set (en)
Training Set (dn)
Validating Set Testing Set (en)
Validating Set Testing Set (dn)
Training byte pair coding model
Byte pair coding model
Byte pair encoding
Training Set (dn)
Training Set (dn)
Validating Set Testing Set (en)
Validating Set Testing Set (dn)

FIGURE 4. The overall framework, which consists of a image feature extraction module, an attention based LSTM module and a joint leading re-position relation network for figure question answering.

sequence of text representations on the source side when generating the context vector on the decoder side, but also introduces an additional local attention mechanism to extract image features to The model not only takes into account the sequence of text representations at the source side when generating the context vector at the decoder side, but also introduces an additional local attention mechanism to extract image features to assist in generating better context vectors. Hard-Attention [41] uses two separate attention mechanisms for generating image and text context vectors, one of which weights all text representations and the other considers only one image feature at each moment.

C. VISUAL ANALYSIS

Since the translation quality of neural network machine translation models is closely related to sentence length, the sentence length of the dataset was visually analysed to guide the experimental setup (e.g., length penalty terms etc.), as is shown in Fig.5 and 6. The best scores are achieved through multiple rounds of parameter optimization of models, and the length of original word sentences will also affect the result of translation in the way that too long sentences may lead to the weakening or even loss of relevant information between words with large spacing, while too short sentences may not be able to learn effective sentence representation and become phrase translation.

Figure.7 show the translation results generated by each model for a sample English->German multimodal translation. It is worth noting that the German word "klatscht" in blue means "claps" in the source language. In order to explore how the multimodal neural network machine translation model based on deep semantic interaction designed in this paper can improve the translation results. Here, the translation results of this model are compared with those of other baseline models. A sample English->German multimodal translation from the M30K_T test set is shown in Table x. In the sample, it can be seen that the key word "clap" in the source sentence is missing from the MNMT_CO_ATT_BIATT translation results. Although both Soft-Attention and Hard-Attention use two separate attention mechanisms to generate text and image context vectors, they also lose the keyword "clap" in the source sentence during the translation process. It is worth noting that the model designed in this paper and its variants produce correct translations. The results confirm the effectiveness of introducing text-visual semantic interactions, while pointing out the shortcomings

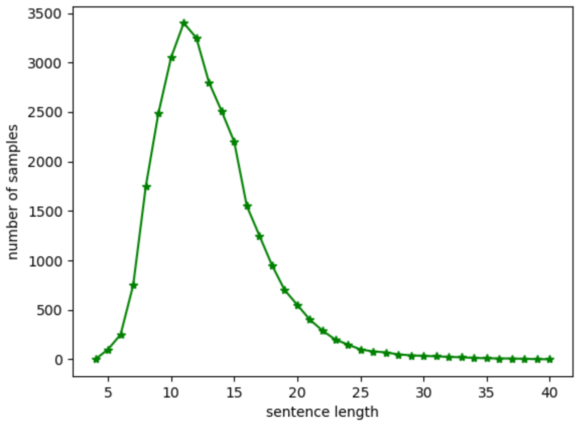


FIGURE 5. Distribution of sentence lengths in the training set(English).

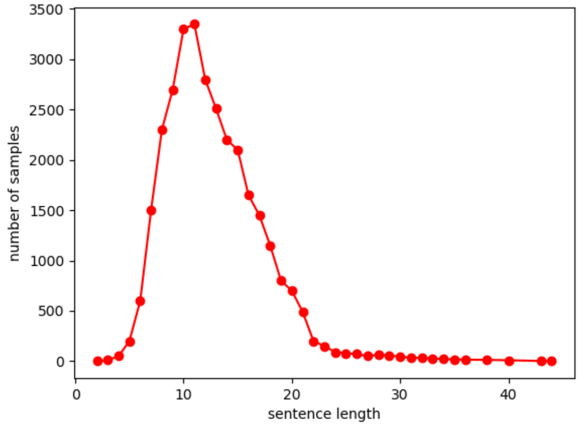
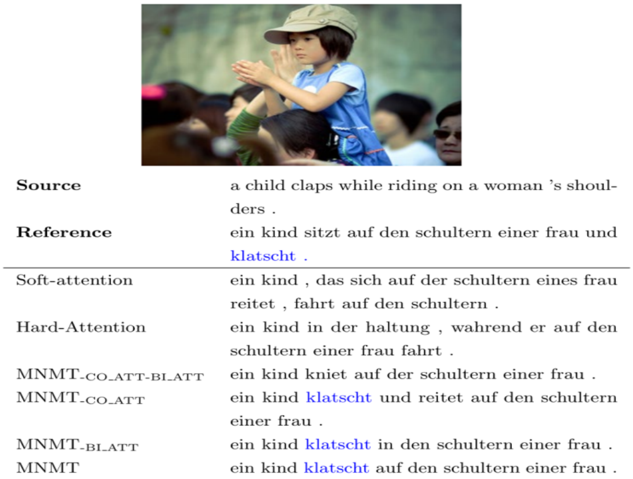


FIGURE 6. Distribution of sentence lengths in the training set(German).



Source	a child claps while riding on a woman 's shoulders .
Reference	ein kind sitzt auf den schultern einer frau und klatscht .
Soft-attention	ein kind , das sich auf der schultern eines frau reitet , fahrt auf den schultern .
Hard-Attention	ein kind in der haltung , wahrend er auf den schultern einer frau fahrt .
MNMT-CO_ATT-BI_ATT	ein kind kniet auf der schultern einer frau .
MNMT-CO_ATT	ein kind klatscht und reitet auf den schultern einer frau .
MNMT-BI_ATT	ein kind klatscht in den schultern einer frau .
MNMT	ein kind klatscht auf den schultern einer frau .

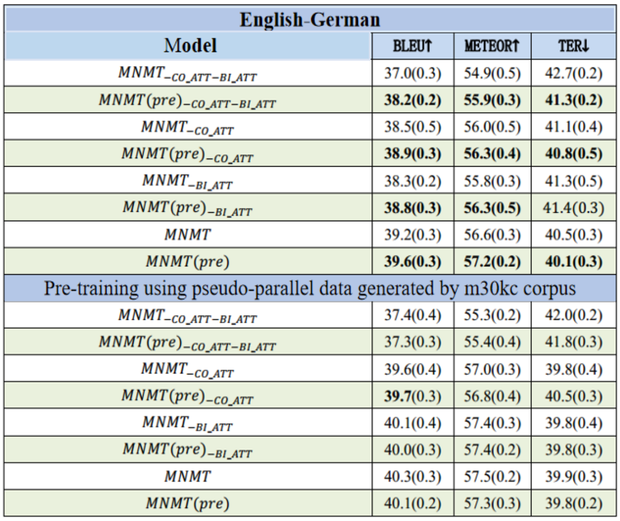
FIGURE 7. Sample English->German translations corresponding to language descriptions and pictures.

of traditional multimodal neural network machine translation models (modelling text and picture semantics separately, neglecting the link between them).

D. ACCURACY ANALYSIS

Figure 8 and Figure 9 respectively list the performance of each multimodal model in $English- > German$ translation and $German- > English$ translation after introducing external linguistic knowledge. The experimental results of the upper half of the two tables are obtained without pre-training using the pseudo-parallel corpus generated by the M30Kc corpus. The experimental results of the lower half of the two tables are obtained by using the M30Kc corpus to generate the pseudo-parallel corpus. Parallel corpus is obtained after pre-training the model.

Without pre-training the model using the back-flip corpus, except that the $MNMT(pre)-BI_{ATT}$ model has no performance improvement in $German- > English$ translation, the other models $MNMT(pre)$, $MNMT(pre)-CO_{ATT}-BI_{ATT}$ and $MNMT(pre)-CO_{ATT}$, $MNMT(pre)-CO_{ATT}-BI_{ATT}$ that introduce external linguistic knowledge have varying degrees Performance improvement. The experimental results show that by adding the pre-trained BERT as an additional encoder to the multimodal neural network machine translation model, the translation quality of the model can be significantly improved.



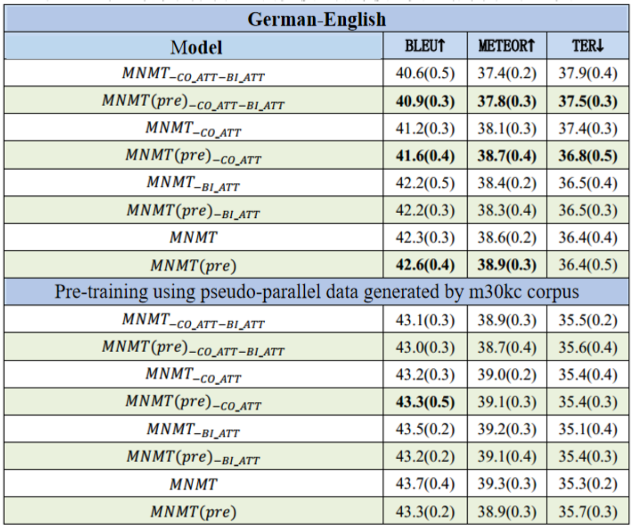
English-German			
Model	BLEU↑	METEOR↑	TER↓
$MNMT_{-CO\_ATT-BI\_ATT}$	37.0(0.3)	54.9(0.5)	42.7(0.2)
$MNMT(pre)_{-CO\_ATT-BI\_ATT}$	**38.2(0.2)**	**55.9(0.3)**	**41.3(0.2)**
$MNMT_{-CO\_ATT}$	38.5(0.5)	56.0(0.5)	41.1(0.4)
$MNMT(pre)_{-CO\_ATT}$	**38.9(0.3)**	**56.3(0.4)**	**40.8(0.5)**
$MNMT_{-BI\_ATT}$	38.3(0.2)	55.8(0.3)	41.3(0.5)
$MNMT(pre)_{-BI\_ATT}$	**38.8(0.3)**	**56.3(0.5)**	41.4(0.3)
$MNMT$	39.2(0.3)	56.6(0.3)	40.5(0.3)
$MNMT(pre)$	**39.6(0.3)**	**57.2(0.2)**	**40.1(0.3)**
Pre-training using pseudo-parallel data generated by m30kc corpus			
$MNMT_{-CO\_ATT-BI\_ATT}$	37.4(0.4)	55.3(0.2)	42.0(0.2)
$MNMT(pre)_{-CO\_ATT-BI\_ATT}$	37.3(0.3)	55.4(0.4)	41.8(0.3)
$MNMT_{-CO\_ATT}$	39.6(0.4)	57.0(0.3)	39.8(0.4)
$MNMT(pre)_{-CO\_ATT}$	**39.7**(0.3)	56.8(0.4)	40.5(0.3)
$MNMT_{-BI\_ATT}$	40.1(0.4)	57.4(0.3)	39.8(0.4)
$MNMT(pre)_{-BI\_ATT}$	40.0(0.3)	57.4(0.2)	39.8(0.3)
$MNMT$	40.3(0.3)	57.5(0.2)	39.9(0.3)
$MNMT(pre)$	40.1(0.2)	57.3(0.3)	39.8(0.2)

FIGURE 8. Experimental Results of Multimodal Translation Model with External Lingistic Knowledge (English-German)

However, after pre-training the model using the pseudo-parallel corpus generated from the M30Kc corpus, in addition to $MNMT(pre)-CO_{ATT}$ slightly improved performance in $English- > German- > English$ translation, the others use pre-trained BERT as an additional encoder The performance of all models has declined. Especially in the $German- > language$ translation, the BLEU of the $MNMT$ translation result dropped by 0.4 points. The experimental results show that after pre-training the model using the pseudo-parallel corpus of the M30Kc corpus, the multimodal neural network machine translation model has learned more knowledge of external linguistics. At this time, because the knowledge in the pre-trained BERT contains more noise (compared to the $M30K_c$ flip-back corpus, the corpus used to pre-train the BERT is much different from the $M30K_T$ domain and has more noise).

VI. CONCLUSION AND FUTURE

This paper designs a multi-modal neural network machine translation model that incorporates pre-trained BERT. By introducing external linguistic knowledge, the neural network machine translation model can generate better translations. It also introduces the four main components in the design model of this article: visual encoder, text encoder, BERT pre-training model, and decoder. Next, this article briefly introduces the multi-modal experimental data set and experimental settings. Since the sentence length is closely related to the model translation quality, this article visually analyzes the



German-English			
Model	**BLEU↑**	**METEOR↑**	**TER↓**
$MNMT_{-CO\_ATT-BI\_ATT}$	40.6(0.5)	37.4(0.2)	37.9(0.4)
$MNMT(pre)_{-CO\_ATT-BI\_ATT}$	**40.9(0.3)**	**37.8(0.3)**	**37.5(0.3)**
$MNMT_{-CO\_ATT}$	41.2(0.3)	38.1(0.3)	37.4(0.3)
$MNMT(pre)_{-CO\_ATT}$	**41.6(0.4)**	**38.7(0.4)**	**36.8(0.5)**
$MNMT_{-BI\_ATT}$	42.2(0.5)	38.4(0.2)	36.5(0.4)
$MNMT(pre)_{-BI\_ATT}$	42.2(0.3)	38.3(0.4)	36.5(0.3)
$MNMT$	42.3(0.3)	38.6(0.2)	36.4(0.4)
$MNMT(pre)$	**42.6(0.4)**	**38.9(0.3)**	36.4(0.5)
Pre-training using pseudo-parallel data generated by m30kc corpus			
$MNMT_{-CO\_ATT-BI\_ATT}$	43.1(0.3)	38.9(0.3)	35.5(0.2)
$MNMT(pre)_{-CO\_ATT-BI\_ATT}$	43.0(0.3)	38.7(0.4)	35.6(0.4)
$MNMT_{-CO\_ATT}$	43.2(0.3)	39.0(0.2)	35.4(0.4)
$MNMT(pre)_{-CO\_ATT}$	**43.3(0.5)**	39.1(0.3)	35.4(0.3)
$MNMT_{-BI\_ATT}$	43.5(0.2)	39.2(0.3)	35.1(0.4)
$MNMT(pre)_{-BI\_ATT}$	43.2(0.2)	39.1(0.4)	35.4(0.3)
$MNMT$	43.7(0.4)	39.3(0.3)	35.3(0.2)
$MNMT(pre)$	43.3(0.2)	38.9(0.3)	35.7(0.3)

FIGURE 9. Experimental Results of Multimodal Translation Model with External Lingistic Knowledge (German-English)

sentence length in the data set to better guide the experiment.

Although the model studied in this paper has achieved preliminary results, there are still some areas worthy of improvement. The next steps of this article include: this article only explores the method of incorporating image information into the translation process. In the future, the author will further study how to incorporate audio, video and other information into the translation process in order to further improve the translation quality. In addition, due to the large difference between the image classification task and the translation task, and the pre-trained residual neural network model used in this paper is pre-trained on the image classification task, the visual representation obtained in this way has certain problems. We plan to study integrating the convolutional neural network used to extract image features into the translation model and train it during the translation process.

ACKNOWLEDGMENT

Here, I would like to thank Prof. Qingfeng Wu and Mr. Chen for their help in this work.

REFERENCES

[1] Chaudhry, Ritwick, et al. "Leaf-qa: Locate, encode and attend for figure question answering." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 2020.

[2] Clarke, Charles LA, Gordon V. Cormack, and Thomas R. Lynam. "Exploiting redundancy in question answering." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.* 2001.

[3] Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE international conference on computer vision.* 2015.

[4] Teney, Damien, et al. "Tips and tricks for visual question answering: Learnings from the 2017 challenge." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018.

[5] Brill, Eric, Susan Dumais, and Michele Banko. "An analysis of the AskMSR question-answering system." *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).* 2002.

[6] Andreas, Jacob, et al. "Learning to compose neural networks for question answering." *arXiv preprint arXiv:1601.01705(2016).*

[7] Agrawal, Aishwarya, Dhruv Batra, and Devi Parikh. "Analyzing the behavior of visual question answering models." *arXiv preprint arXiv:1606.07356 (2016) .*

[8] Zhou, Bolei, et al. "Simple baseline for visual question answering." *arXiv preprint arXiv:1512.02167 (2015) .*

[9] Miller, John, et al. "The effect of natural distribution shift on question answering models." *International Conference on Machine Learning. PMLR, 2020*

[10] Hu, Ronghang, et al. "Learning to reason: End-to-end module networks for visual question answering." *Proceedings of the IEEE International Conference on Computer Vision.* 2017.

[11] Zhang, Junbei, et al. "Exploring question understanding and adaptation in neural-network-based question answering." *arXiv preprint arXiv:1703.04617 (2017).*

[12] Radev, Dragomir, et al. "Probabilistic question answering on the web." *Proceedings of the 11th international conference on World Wide Web.* 2002.

[13] Gibson, Ronald F. "A review of recent research on mechanics of multifunctional composite materials and structures." *Composite structures 92.12 (2010): 2793-2810.*

[14] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.*

[15] Çiçek, Özgün, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." *International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2016.*

[16] Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." *2016 fourth international conference on 3D vision (3DV). IEEE, 2016.*

[17] Kamnitsas, Konstantinos, et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation." *Medical image analysis 36 (2017): 61-78.*

[18] Ghafoorian, Mohsen, et al. "Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation." *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE, 2016.*

[19] Wang, Changhan, et al. "A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks." *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2015.*

[20] Brosch, Tom, et al. "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation." *IEEE transactions on medical imaging 35.5 (2016): 1229-1239.*

[21] Vaswani, Ashish, et al. "Attention is all you need." *arXiv preprint arXiv:1706.03762 (2017).*

[22] Och, Franz Josef, and Hermann Ney. "The alignment template approach to statistical machine translation." *Computational linguistics 30.4 (2004): 417-449.*

[23] Fukui, Hiroshi, et al. "Attention branch network: Learning of attention mechanism for visual explanation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.*

[24] Sung, Flood, et al. "Learning to compare: Relation network for few-shot learning." *Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.*

[25] Johnson, Justin, et al. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.*

[26] Mertens, Koen C., et al. "Sub-pixel mapping and sub-pixel sharpening using neural network predicted wavelet coefficients." *Remote Sensing of Environment 91.2 (2004): 225-236.*

[27] Kampffmeyer, Michael, et al. "ConnNet: A long-range relation-aware pixel-connectivity network for salient segmentation." *IEEE Transactions on Image Processing 28.5 (2018): 2518-2529.*

[28] Hartmann, Marc, et al. "Relation between baseline plaque burden and subsequent remodelling of atherosclerotic left main coronary arteries: a serial intravascular ultrasound study with long-term follow-up." *European heart journal 27.15 (2006): 1778-1784.*

[29] Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." *Neural computation 12.10 (2000): 2451-2471.*

[30] Wang, Yequan, et al. "Attention-based LSTM for aspect-level sentiment classification." *Proceedings of the 2016 conference on empirical methods in natural language processing. 2016.*

[31] Kahou, Samira Ebrahimi, et al. "Figureqa: An annotated figure dataset for visual reasoning." *arXiv preprint arXiv:1710.07300(2017).*

[32] Kafle, Kushal, et al. "Dvqa: Understanding data visualizations via question answering." *Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.*

[33] Santoro, Adam, et al. "A simple neural network module for relational reasoning." *arXiv preprint arXiv:1706.01427 (2017).*

[34] Reddy, Revanth, et al. "Figurenet: A deep learning model for question-answering on scientific plots." *2019 International Joint Conference on Neural Networks (IJCNN). IEEE.* 2019.

[35] Jialong Zou, Guoli Wu, Taofeng Xue, Qingfeng Wu. "An affinity-driven relation network for figure question answering." *IEEE International Conference on Multimedia and Expo (ICME), Piscataway, NJ: IEEE, 2020: 1-6.*

[36] Chaudhry, Ritwick, et al. "Leaf-qa: Locate, encode and attend for figure question answering." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 2020.

[37] Huang P Y, Liu F, Shiang S R, et al. Attention-based multimodal neural machine translation[C]//Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016: 639-645.

[38] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[39] Calixto I, Liu Q, Campbell N. Doubly-attentive decoder for multi-modal neural machine translation[J]. arXiv preprint arXiv:1702.01287, 2017.

[40] Calixto I, Liu Q, Campbell N. Incorporating global visual features into attention-based neural machine translation[J]. arXiv preprint arXiv:1701.06521, 2017.

[41] Delbrouck J B, Dupont S, Seddati O. Visually grounded word embeddings and richer visual features for improving multimodal neural machine translation[J]. arXiv preprint arXiv:1707.01009, 2017.