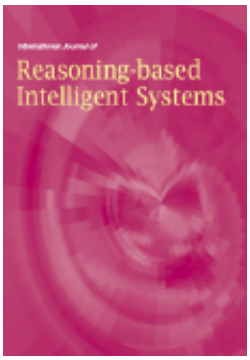


International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

A new double attention decoding model based on cascade RCNN and word embedding fusion for Chinese-English multimodal translation

Haiying Liu

DOI: 10.1504/IJRIS.2023.10052939

Article History:

Received:	06 November 2022
Accepted:	22 November 2022
Published online:	19 March 2024

A new double attention decoding model based on cascade RCNN and word embedding fusion for Chinese-English multimodal translation

Haiying Liu

School of Foreign Languages,
Zhengzhou University of Science and Technology,
Zhengzhou, 450064, China
Email: liuhaiying2006@tom.com

Abstract: Traditional multimodal machine translation (MMT) is to optimise the translation process from the source language to the target language with the help of important feature information in images. However, the information in the image does not necessarily appear in the text, which will interfere with the translation. Compared with the reference translation, mistranslation can be appeared in the translation results. In order to solve above problems, we propose a double attention decoding method based on cascade RCNN to optimise existing multimodal neural machine translation models. The cascade RCNN is applied to source language and source image respectively. Word embedding is used to fuse the initialisation and the semantic information of the dual encoder. In attention computation process, it can reduce the focus on the repetitive information in the past. Finally, experiments are carried out on Chinese-English test sets to verify the effectiveness of the proposed method. Compared with other state-of-the-art methods, the proposed method can obtain better translation results.

Keywords: multimodal machine translation; MMT; double attention decoding; cascade RCNN; word embedding fusion.

Reference to this paper should be made as follows: Liu, H. (2024) 'A new double attention decoding model based on cascade RCNN and word embedding fusion for Chinese-English multimodal translation', *Int. J. Reasoning-based Intelligent Systems*, Vol. 16, No. 1, pp.26–36.

Biographical notes: Haiying Liu is with School of Foreign Languages, Zhengzhou University of Science and Technology. Her majors are English Analysis, MT.

1 Introduction

The multimodal neural machine translation (MNMT) model mainly takes image and text as input and adopts the end-to-end neural network method to achieve the translation result from source language to target language. In the process of translation, the important information features of images are extracted to optimise the translation results. Multimodal neural machine translation has been widely used in many translation tasks, such as picture title translation, literature translation, online news article interpretation, multi-language hidden subtitle translation for international TV and movies, etc. (Heo et al., 2019). Therefore, this technique has attracted extensive interest and attention from many scholars in the field of multimodal image and text. When parallel corpora for training are lacking, better translation performance can be achieved by using images as external data sources.

For example, in Figure 1, the source language description of the image is '一名男子正在图书馆读书', translated as 'a man is reading a book in the library'. However, most translation systems will translate '书' into 'reserve' without adding images. In this scene, the translation of 'reserve' is wrong.

Figure 1 Translation 'a man is reading a book in the library' (see online version for colours)

In this paper, the multimodal machine translation (MMT) system makes full use of image information to translate source language into target language. The information that can be obtained from the image includes: a man holding a book in his hand, the location is in the library, and the action is reading. Therefore, important features such as characters, behaviours and scenes in images

can be extracted to better assist source language translation into target language.

At present, most of the researches on multimodal neural machine translation models mainly include three parts: first, the image encoder is convolved by convolutional neural network (CNN) to extract semantic information representation in the image (Shi et al., 2021); secondly, neural translation network is used to translate source language into target language during decoding. Finally, a cross-modal learning network with visual features (He et al., 2019) is used to improve the learning of linguistic features, including semantic representation and cross-linguistic semantic relevance. Some scholars have applied the single-layer attention network and multi-task learning model to the multimodal neural machine translation integrating image visual features, and achieved better translation performance than the neural machine translation model by assisting source text translation with image features to a certain extent.

Recently, a large number of studies have applied attention mechanism methods to neural machine translation (NMT) and image description generation (IDG) tasks, achieving better performance. Su et al. (2020) proposed a method of dual attention mechanism. Through two independent attention mechanisms, source language information and spatial visual features were respectively concerned in attention calculation, and the experimental performance was improved to some extent compared with the previous MMT model. However, the existing multimodal neural machine translation models do not take into account the historical concerns of text and image information. When decoding, the attention mechanism will re-focus on the source language words and part of the source image at the past moment when calculating the weight of the current moment, and the information in the image may bring interference. So the target language of translation may be over-translated or under-translated. For example, '多个银行都被迫关闭了' is translated to 'many banks were closed to close'. As can be seen, 'close' has been translated twice, it is over-translated situation. '被迫' should be translated as 'were forced to', but it is not translated in the target language, it is under-translated situation.

The proposal of multiple training unsupervised cross-lingual word embedding training methods makes it possible to train unsupervised neural machine translation models using only monolingual corpora. The effect of unsupervised neural machine translation is also significantly improved by combining the denoising auto-encoder and back-translation. Bidirectional encoder representation from transformers (BERT) uses transformer model to build a framework. The model uses transformer model to build a framework, and uses unsupervised method to pre-train and generate language representation model on monolingual data, which can better mine the relationship between words. Subsequently, XLM (cross-lingual language model) model is further extended and trained to generate a cross-lingual language model, which enables words with similar meanings in different languages to have relatively consistent positions in word embedding, finally effectively improving the training effect of unsupervised machine translation model.

In view of the above problems, a multimodal neural machine translation model based on double attention decoding method and cascade RCNN is proposed in this paper. The model firstly combines two independent attention mechanisms to focus on the relevant information in the source language and the target image respectively to enhance the semantic relevance of text and image features. In addition, the model integrates the cascade RCNN method, which is used to record the text features and image features that the model has paid attention to before when calculating the attention of the current moment, so as to avoid the situation that some repetitive features are continuously paid attention to. In this paper, an adversarial evaluation experiment is conducted to verify that using images as an additional input can bring effective information to the model and help optimise the translation of the source language. Finally, the evaluation results of the proposed model on the Chinese-English test set verify the effectiveness of the method and reduce the over-translation or under-translation.

The organisational structure of this paper is as follows: Section 1 introduces related work. Section 2 introduces the research background. Section 3 introduces multimodal neural machine translation based on double attention decoding method and cascade RCNN. The experimental results and analysis are given in Section 4. Section 5 summarises the work of this paper.

2 Related works

Multimodal neural machine translation is proposed by the machine translation committee through a shared task. The multimodal network proposed by Jelicic et al. (2013) in the early stage integrated the features of text and vision, and applied it to the task of IDG and image description sorting. In their work, the authors combined text representations and global image features in recurrent neural networks (RNN) in independent multimodal layers. Yu et al. (2015) proposed a neural IDG model based on sequence-to-sequence framework, which was based on end-to-end training. Shi et al. (2020) proposed a model to generate multilingual description of images, which learned and converted image features in two independent, non-attention IDG models. Qu et al. (2017) proposed the first IDG model based on attention mechanism. In this model, when generating natural language description of images, attention mechanism would pay attention to different regions of images.

For neural machine translation, Xiao et al. (2020) proposed a multi-language neural machine translation model based on attention mechanism, which translated two different source languages into one target language and achieved better performance than the single-language reference system. Feng et al. (2016) applied reproduction model and coverage method to NMT respectively, and

optimised translation results by introducing linguistic coverage and neural network-based coverage. Singh et al. (2021) proposed a multi-task learning method, which could translate one source language into multiple target languages during model training. However, in the above model, not each group of languages had an independent attention mechanism, but only a shared attention mechanism, that is, each target language had an attention mechanism shared by all the source languages. Qian and Tian (2022) proposed a multi-channel model, which translated a variety of source languages into a variety of target languages during training. They trained the model through two tasks (main task and auxiliary task) and a shared decoder. The main task was to translate German into English, and the auxiliary task was to generate English image description. Their experimental results showed that the performance of the main translation task were also improved when training the auxiliary task of IDG.

Although there is no fixed neural multimodal model so far, it can significantly improve the pure text NMT (Bahdanau et al., 2014) and statistical machine translation (SMT) model performance. Different research groups have proposed to reorder the global features and spatial visual features of images generated by SMT or NMT systems, and then take *n*-best features, which has achieved some success. The multimodal neural machine translation model proposed by Huang et al. achieved good experimental performance. They obtained global features of images by using VGG19 network (Yin et al., 2019) and selected different regions in images by using RCNN network. Their model offered a significant performance improvement over the text-only NMT benchmark system.

At present, a large number of studies are trying to integrate image features into neural machine translation in different ways to improve translation performance. For example, Snell et al. (2018) proposed the graded attention connection method. The above multimodal models verify that images can bring effective information to the neural machine translation model.

The contribution of this paper is to propose a multimodal neural machine translation model based on double attention decoding method and cascade RCNN. The characteristic of this paper is that the cascade RCNN method is integrated on the basis of the double-attention mechanism method. Cascade RCNN acts on both the text decoding and the image decoding at the same time, so as to improve the over-translation and under-translation by reducing the attention to the previous repeated information. Compared with the previous multimodal models, the performance of proposed method is better. In addition, in this paper, combining text and visual features into a multimodal shared space can improve the performance of the model at a certain range.

3 Preliminaries

3.1 Neural machine translation based on attention mechanism

This section introduces a neural machine translation model based on attention mechanism. Given source language sequence $X = (x_1, x_2, \ldots, x_N)$ and the target language sequence $Y = (y_1, y_2, \ldots, y_M)$ to establish the mapping function from the source language into the target language, the mapping function is usually expressed as conditional probability $P(Y|X)$.

The whole network consists of an encoder and a decoder based on attention mechanism, which are implemented by two RNNs and a multilayer perceptron respectively. Where x_i and y_i are the i^{th} word of source sequence and target sequence respectively. $E_x \in R^{|V_x| \times d_x}$ and $E_y \in R^{|V_y| \times d_y}$ are word embedding matrices respectively. Where V_x and V_y are word tables of source language and target language respectively. d_x and d_y are the word vector dimensions of source word and target word respectively.

The encoder of this paper is a bidirectional RNN based on gated recurrent unit (GRU). Forward RNN $\vec{\Phi}_{enc}$ reads the words in the source sequence X sequentially from left to right. Its function is to capture the previous information of the current word and generate the source language forward hidden state sequence $\vec{H} = (\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N)$. The reverse RNN $\overleftarrow{\Phi}_{enc}$ reads the words in the source sequence X in reverse order from right to left. Its function is to capture the following information of the current word and generate the backward hidden state sequence of the source language $\overleftarrow{H} = (\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_N)$. The state corresponding to the position in the forward and reverse hidden state sequence is spliced into the hidden layer of the word at the position to represent $h_i = [\vec{h}_i, \overleftarrow{h}_i]$, which contains the context information of the word at the position. The final interpretation vector containing all the information of the source sequence is represented as $C = (h_1, h_2, \ldots, h_N)$.

The decoder in this paper is a one-way RNN based on conditional gated recurrent unit (CGRU), and the hidden state sequence C obtained after encoder acts on the decoder. Decoder is a neural language model, which uses the source sequence information C and the information y_{t-1} of the generated words in the decoding stage of the previous moment to predict the target word y_t at the current moment under the action of the attention mechanism. It initialises the decoder's hidden state s_0 at time $t = 0$ with a multilayer perceptron. The initial input to the decoder is a vector that connects the forward and reverse hidden states of the encoder at the last moment $[\vec{h}_N, \overleftarrow{h}_1]$. At the time step t of decoder, the calculation of the source context vector C is based on the interpretation vector C and the hidden state s_{t-1} at the previous time of decoder.

3.2 Conditional gated recurrent unit

This section introduces the specific calculation method of CGRU at decoding time (Zhao et al., 2020). At time t of decoder time step, the conditional loop network is divided into three parts.

1 According to the decoding of hidden state s_{t-1} at the previous moment and the predicted generated target word y_{t-1}, the candidate hidden state s_t' at the current t moment is calculated.

2 On the attention mechanism of implicit state information in source language RNN, context vector c_t is calculated by source interpretation vector C and candidate implicit state s_t'.

3 The final hidden state s_t is calculated by candidate hidden state s_t' and text attention vector c_t.

The following is the overall calculation. First of all, the weight $\alpha_{t,i}^{src}$ of the attention model is obtained from the target language implicit state s_t and source language implicit state sequence h_i at time t through linear transformation and then activation function, as shown in equations (1) and (2).

$$e_{t,i}^{src} = \left(v_a^{src}\right)^T \tanh\left(U_a^{src} s_t + W_a^{src} h_i\right). \tag{1}$$

$$\alpha_{t,i}^{src} = \frac{\exp\left(e_{t,i}^{src}\right)}{\sum_{j=1}^{N} \exp\left(e_{t,i}^{src}\right)}. \tag{2}$$

where tanh is the nonlinear activation function. v_a^{src}, U_a^{src}, W_a^{src} are parameters used for learning in the model. The weight $e_{t,i}^{src}$ can be interpreted as the correlation degree between the target word generated by the decoder at time t and the source sequence word x_i. $\alpha_{t,i}^{src}$ means to normalise the obtained similarity score.

Secondly, the text attention vector c_t at time t is obtained by the weighted sum of the source language implicit state sequence h_i and the weight $\alpha_{t,i}^{src}$ obtained by the text attention model, as shown in equation (3).

$$c_t = \sum_{i=1}^{N} \alpha_{t,i}^{src} h_i. \tag{3}$$

Finally, the prediction of the target word y_t at the t moment is related to the implicit state s_t of the target word at the current moment, and the target word y_{t-1} generated by the prediction at the previous moment and the text attention vector c_t, as shown in equation (4).

$$\begin{aligned} &P\left(y_t \left| y_{t-1}, C\right.\right) = \operatorname{softmax}\left(f\left(s_t, y_{t-1}, c_t\right)\right) \\ &\propto \exp\left(L_o \tanh\left(L_s s_t + L_w E_y\left[y_{t-1}\right] + L_c c_t\right)\right). \end{aligned} \tag{4}$$

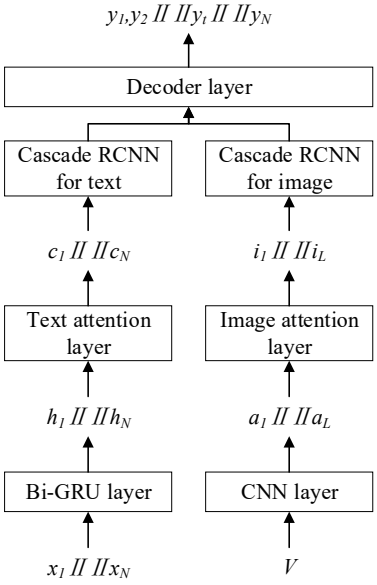
where f and softmax are nonlinear activation functions. L_o, L_s, L_w and L_c are the parameters used by the model for learning.

4 Proposed method

4.1 Overall structure of new model

Multimodal neural machine translation can be regarded as the addition of image features on the basis of neural machine translation. The image contains information about the source sentence. Image features are used to optimise the translation results. The multimodal neural machine translation model in this paper combines the cascade RCNN method on the basis of text feature and image feature. The overall structure of the model in this article is presented as in Figure 2.

Figure 2 Proposed structure (see online version for colours)



Source:马儿在草地上奔跑
Target: The horse is running on the grass

First, for a source language sentence $S = (s_1, s_2, \ldots, s_N)$, a pre-trained word vector is used as a distributed representation of the sense of a word. During initialisation, it takes a vector representing $X = (x_1, \ldots, x_N)$ for each word. On this basis, the learning process of feature information is as follows:

1 Input $X = (x_1, \ldots, x_N)$ into bi-GRU layer. Through bi-GRU layer encoding, the hidden layer $H = (h_1, \ldots, h_N)$ containing context information of each word is obtained. Input the corresponding source image V into CNN layer, and extract the hidden layer $A = (a_1, \ldots, a_N)$ of important feature information in the image through CNN convolution.

2 On this basis, two independent gated dynamic attention mechanisms are used to dynamically assign different attention weights to the source text context H and the source image context A respectively. Then the source text attention vector $C = (c_1, \ldots, c_N)$ and the source image attention vector $I = (i_1, \ldots, i_N)$ are calculated by the attention weight and context vector respectively.

3 The double attention model in the previous step does not consider the historical attention information of the text and image, but refocuses on the source language words and part of the source image at the past moment. Therefore, the cascade RCNN is introduced in this paper to punish the repeated attention to historical information, and the attention weight obtained by the previous layer passes through the cascade layer of text and image respectively to obtain the vector of text and image.

4 The hidden states of vector, source text and image, and target word at the previous moment are input to the decoding layer, and the corresponding predicted target sentence $Y = (y_1, y_2, \ldots, y_N)$ for each word is obtained.

4.2 Unsupervised machine translation

Unsupervised machine translation usually follows three steps: initialisation, language modelling, and reverse translation.

1 Initialisation: For the machine translation task, it is hoped to use the model pre-training initialisation to establish the initial connection between two different languages, so as to facilitate the later training. By combining the training data of two languages, joint BPE is carried out to make them share the same thesaurus, and then monolingual data is used to train the encoder to establish the potential connection.

2 Establishment of language model: in the process of machine translation, it is essential to master the information of language sentences. Noise reduction auto-encoder can be used to train a large number of monolingual data to master the information contained in sentences. The noise reduction auto-encoder is a variant of the basic auto-encoder, which generates $\hat{x}$ by adding noise 'damage' to the input sample x by random mapping. Then, it is trained as the input of the auto-encoder. Its output result is compared with the original sentence to calculate the loss, as shown in equation (5):

$$L_{auto} = E_{x\sim D,\hat{x}\sim C(x)}\left[-\log P(x|\hat{x})\right]. \tag{5}$$

where D is monolingual dataset. $C(x)$ represents the process of adding noise to sample x. Through training, the model can master the basic construction information of each language sentence.

DIV method training translation model is divided into one-way training and two-way training. In one-way training, the denoising auto-encoder is used to build the language model and the model is optimised through back-translation. Similarly, the source training data is replaced by the dictionary during training, and then the replaced source and target data are used for training. In addition to the preliminary training, it will replace the data with the original data as one-way training outside. The word embedded fusion is proposed to initialise double encoder training, it makes full use of learned knowledge in training in all directions, which can better use the dictionary to replace the data, to reduce the gap between different languages.

4.3 Attention layer based on bi-attention mechanism

This section introduces the calculation method of dynamic attention mechanism. In a single-layer RNN decoder, related information in source language words and image features is integrated in the decoding stage through two independent attention mechanisms. The double attention GRU network in this paper is related to the hidden state s_{t-1} at the previous moment of decoding, the target word y_{t-1} generated by the previous prediction, and the information in the source sentence and image that is concerned by two independent attention mechanisms.

In this paper, CGRU is extended to double CGRU network. In addition to the source language attention mechanism ATT_{src}, we also introduce a new image attention ATT_{img}. The visual attention mechanism calculates the image attention vector i_t by decoding the hidden state s_t at time t and the image interpretation vector $A = (a_1, \ldots, a_L)$ obtained by using the soft attention mechanism on the image.

The image attention mechanism is similar to the source language attention mechanism, except that it has an additional gated scalar. Firstly, a single-layer feed-forward neural network is used to calculate the similarity score of decoded hidden state s_t at the time of each image interpretation vector a_l and t. This represents which part of the image features should be paid more attention to when decoding the target word at the current moment, as shown in equations (6) and (7).

$$e_{t,l}^{img} = \left(v_a^{img}\right)^T \tanh\left(U_a^{img} s_t + W_a^{img} a_l\right). \tag{6}$$

$$\alpha_{t,l}^{img} = \frac{\exp\left(e_{t,l}^{img}\right)}{\sum_{j=1}^{L}\exp\left(e_{t,j}^{img}\right)}. \tag{7}$$

where tanh is a nonlinear function. $v_a^{img}, U_a^{img}, W_a^{img}$ are parameters used for learning in the model. $\alpha_{t,l}^{img}$ means to normalise the obtained similarity score.

Then a gated scalar $\beta \in [0, 1]$ is calculated to weigh the importance of the image context vector relative to the next target word, as shown in equation (8).

$$\beta_t = \sigma\left(W_\beta s_{t-1} + b_\beta\right). \tag{8}$$

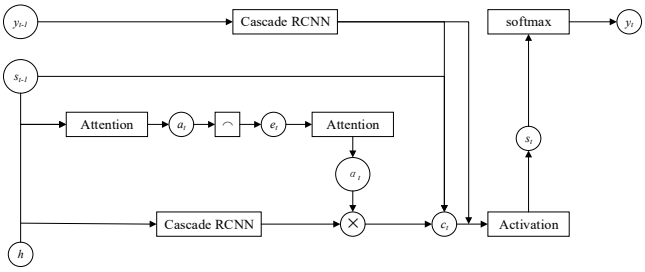
where W_β, and b_β the parameter of the model. β_t will be used in the calculation of image semantic vector i_t. i_t is obtained by the weighted average of image interpretation vector a_l and image feature attention weight $\alpha_{t,l}^{img}$, as shown in equation (9).

$$i_t = \beta_t \sum_{l=1}^{L} \alpha_{t,l}^{img}. \tag{9}$$

4.4 Cascade RCNN

This section introduces how to obtain the coverage vector and the updated attention vector of each target word in the sentence to be predicted through the cascade RCNN layer. The structure is shown in Figure 3.

Figure 3 Cascade RCNN



During decoding, the attention weight vectors α_t^{src} and α_t^{img} of text and image are obtained through the double attention mechanism layer according to the hidden state s_{t-1} of the previous moment, the hidden state sequence H of source language and the image interpretation vector sequence A. The key point of the cascade RCNN layer is to maintain a vector C during the prediction process. This vector is the sum of attention distribution of all previous prediction steps, which records the historical information that the model has paid attention to and avoids paying attention to repetitive information, as shown in formulas (10) and (11).

$$C_t^{src} = \sum_{\hat{t}=0}^{t-1} \alpha_{\hat{t}}^{src}. \tag{10}$$

$$C_t^{img} = \sum_{\hat{t}=0}^{t-1} \alpha_{\hat{t}}^{img}. \tag{11}$$

where C_t^{src} and C_t^{img} are source language vector and source image vector respectively.

The obtained vector is applied to the attention layer to get the updated attention weight, as shown in equations (12) and (13).

$$e_{t,i}^{src} = \left(v_a^{src}\right)^T \tanh\left(U_a^{src} s_t + W_a^{src} h_i + V_a^{src} C_{t,i}^{src}\right). \tag{12}$$

$$e_{t,i}^{img} = \left(v_a^{img}\right)^T \tanh\left(U_a^{img} s_t + W_a^{img} h_i + V_a^{img} C_{t,i}^{src}\right). \tag{13}$$

The calculation method is similar to formulas (1) and (5), but the difference is that vectors are added as additional inputs to jointly affect the prediction of target language. Then, the updated contextual attention vector c_t, i_t is obtained according to the calculation methods of equations (2), (3) and (6)~(8).

Next, this paper uses the updated i_t as an additional input to update the calculations in Subsection 3.2. By using candidate hidden state s_t', source language attention vector c_t and image attention vector i_t, the final hidden state s_t at time t is calculated, as shown in equations (14)~(17).

$$z_t = \sigma\left(W_z^{src} c_t + W_z^{img} i_t + U_z s_j'\right). \tag{14}$$

$$r_t = \sigma\left(W_r^{src} c_t + W_r^{img} i_t + U_r s_j'\right). \tag{15}$$

$$\hat{s}_t = \tanh\left(W^{src} c_t + W^{img} i_t + r_t \Theta\left(U s_t'\right)\right). \tag{16}$$

$$s_t = \left(1 - z_t\right) \Theta \hat{s}_t + z_t \Theta s_t'. \tag{17}$$

where z_t is the update gate. r_t is the reset gate. $\hat{s}_t$ is the candidate hidden state. s_t is the final hidden state. W_z^{src}, W_z^{img}, U_z, W_r^{src}, W_r^{img}, U_r, W^{src}, W^{img}, U are parameters.

Finally, in the output part of the model, the prediction of the target word y_t at time t is related to the implicit state s_t of the target word at the current time, the target word y_{t-1} generated by the prediction at the previous time, text attention vector c_t, and image attention vector i_t, as shown in equation (18).

$$\begin{aligned} P\left(y_t \left| y_{t-1}, C, A\right.\right) &= \text{softmax}\left(f\left(s_t, y_{t-1}, c_t, i_t\right)\right) \\ &\propto \exp\left(L_o \tanh\left(L_s s_t + L_w E_y\left[y_{t-1}\right] + L_{cs} c_t + L_{ci} i_t\right)\right). \end{aligned} \tag{18}$$

where f and softmax are nonlinear activation functions. L_o, L_s, L_w, L_{cs} and L_{ci} are the parameters used by the model for learning.

4.5 Dual encoder fusion training

Although the word embedding fusion initialisation method can use the training results of the replacement words in the model pre-trained in the other direction to promote the training of this direction, it still loses some information learned from the training of the replacement words and other words. In order to make better use of the training results of the training model with two directions, a dual encoder fusion training method is proposed in this paper. Taking Chinese-English training as an example, after pre-training, when training the model in the Chinese-English direction, the encoder model in the Chinese-English direction and the encoder model in the Anglo-Chinese direction obtained in the pre-training are loaded at the same time. The training data are input into the two encoders for training respectively, and the output results are processed using equation (19):

$$enc_{out} = enc_{out-src2rgr} + \sigma \times enc_{our-src2svc} \tag{19}$$

where $enc_{out-src2rgr}$ is the output of the encoder from the source end to the target end, $enc_{out-src2svc}$ is the output of the encoder from the target end to the source end, σ is used to control the proportion of the output, and then the final result enc_{out} is input into the decoder model from the target end to the source end for training.

In this way, both Chinese-English and English-Chinese pre-trained encoder models can be used for training in the Chinese-English training process. The encoder can output two different views of the coding results. It is processed according to a certain proportion (Li, 2022; Ying, 2022; Deena and Raja, 2022), and some neglected places in the

Chinese-English encoder perspective can be corrected by using the Anglo-Chinese encoder perspective, so as to achieve better results. It is processed according to a certain proportion, and some neglected places in the Chinese-English encoder perspective can be corrected by using the Anglo-Chinese encoder perspective, so as to achieve better results. Since the source of training data is also the dictionary replacement data combined with the original data for one-way training in all directions, the two encoders can assist each other in training and make use of the knowledge learned by each other in pre-training.

5 Experimental results and analysis

The dataset Flickr30k (Young et al., 2014) contains 30k images. Each image corresponds to five English descriptions of the image content. This article uses the Multi30k dataset, which is extended from two multilingual versions of the original dataset Flickr30k: one is the translation dataset M30kT and the other is the comparison dataset M30kC. For each image in Flickr30k, dataset M30kT contains an English description and a Chinese description automatically translated by a professional translator. Among them, the training set, verification set and test set contain 29k, 1,014 and 1k images respectively. Each image corresponds to a sentence pair (original English description and translated Chinese description). For each image in Flickr30k, dataset M30kC contains five English descriptions and corresponding translated Chinese descriptions. Among them, the training set, verification set and test set contain 29k, 1,014 and 1k images respectively. Each image corresponds to five sentences in English and five sentences in Chinese.

In this paper, multimodal neural machine translation models are trained using Multi30K-16 and Multi30K-17 training sets. The validation set is used to select the model by BLEU4 value, and the generalisation ability of the final model is evaluated by the test set.

The encoder of the model in this paper is a bi-directional circulating neural network based on GRU, and the output dimension of single-layer forward and backward RNN is 500. The word vector dimension of source word and target word is 500 dimension. The word vector is initialised by random sampling of a Gaussian distribution $N(0, 0.012)$. The recurrent evaluation method is initialised by random orthogonal and all bias vectors are initialised to 0. The model decoder uses a single-layer GRU network.

Image features are obtained by inputting images into the pre-trained VGG19 CNN model, and the final full-connection layer output is taken as image features. Before generating a target word, we set the dropout value to 0.5. In this paper, Gal and Ghahramani methods use the same mask method when applying dropout to encoder bidirectional RNN and decoder RNN.

ADAM stochastic gradient descent algorithm is used for all models during training. Each batch size of data (batch-size) is 80 (text-based NMT) and 40 (MNMT). In MNMT model, each training sample contains one English sentence, one Chinese sentence and one corresponding image. Model selection is based on BLEU4 value and early stop method is adopted, that is, if a model does not continue to improve its BLEU4 value after more than 20 rounds in the verification set, it will be stopped in the training phase. The evaluation indexes of translation quality of the MMT model in this paper are BLEU4, METEOR and TER (Denkowski and Lavie, 2014)

In order to compare with the double attention decoding method proposed in this paper and verify the effectiveness of the proposed model, the following comparison models are setup in this paper.

- *NMT:* A text-only neural machine translation model that translates source language into target language under the attention mechanism.
- *Delbrouck et al. (2017):* The MMT model of the image is added, and the image is convolved by CNN to extract global features and local features, and then the encoding end is added, so that feature information of both text and image can be paid attention to during decoding.
- *Merritt et al. (2020):* It includes two independent attention mechanism MMT models, which focuses on source text features and source image features respectively, and uses image features to assist text translation in decoding.
- *Hirasawa and Komachi (2019):* It is a MMT model based on self-attention mechanism. Encoders and decoders replace traditional RNN structure by Transformer architecture. It fully utilises attention mechanism, which can carry out parallel operation and better solve long time dependence problem.
- *Yu et al. (2020):* The multimodal model does not simply extract image features and apply them to the encoding end, but attempts to efficiently utilise visual information from multiple angles and types by activating functions or dot multiplying image features.
- *Song et al. (2021):* This multimodal model maps image features and text features into a shared space, and connects visual semantics and corresponding text semantic information through a visual attention mechanism during decoding.

In the following section, we conduct comparison from three aspects.

Table 1 Comparison between the proposed model and the benchmark system (%)

Chinese → English	*Multi30K-16*			*Multi30K-17*		
	BELU	*METEOR*	*TER*	*BELU*	*METEOR*	*TER*
Baseline	36.2	54.8	45.0	28.8	48.6	53.3
Text-based	37.0	55.5	44.2	29.3	49.3	52.8
Image-based	36.9	55.4	44.4	29.2	49.0	53.0
Proposed	*37.4*	*55.9*	*43.8*	*29.5*	*49.6*	*52.5*

Table 2 Comparison between the proposed model and other models (%)

Chinese → English	*Multi30K-16*			*Multi30K-17*		
	BELU	*METEOR*	*TER*	*BELU*	*METEOR*	*TER*
NMT	33.9	52.5	46.9	19.5	42.1	72.4
Delbrouck	35.6	53.1	45.7	27.8	45.6	66.1
Merritt	35.8	54.7	45.2	27.9	46.1	63.4
Hirasawa	36.2	54.9	44.9	28.5	47.9	57.1
Yu	37.0	55.5	44.2	29.3	49.3	52.8
Song	36.9	55.4	44.4	29.2	49.0	53.0
Proposed	*37.4*	*55.9*	*43.8*	*29.5*	*49.6*	*52.5*

Table 3 Comparison of translation results between the model and the reference system (see online version for colours)

	Source description	两辆车发生交通事故，其中一人受伤
	NMT	Translation: two cars have a traffic accident and one person is injured.
	Proposed	Translation: two cars were involved in an accident and one of them was injured.
	Source description	两个小朋友光着脚丫子在沙滩玩足球
	NMT	Translation: two children are playing football without shoes on the beach.
	Proposed	Translation: two children are barefoot playing football on the beach.
	Source description	一位女士在书店预定了一本书
	NMT	Translation: one lady booked one book at a bookstore.
	Proposed	Translation: a lady ordered a book at a bookstore.

5.1 Comparative analysis with benchmark model

Table 1 shows the performance comparison between the proposed model in this paper and the benchmark when Chinese is translated into English on Multi30K-16 and Multi30K-17 datasets respectively. It is worth noting that the performance of the open source system based on this method is first obtained according to the dual-attention mechanism. Based on the parameters provided by the benchmark system, Table 1 reports the comparison between the proposed method and the benchmark system.

As can be seen from Table 1, when translating Chinese into English on the Multi30K-16 dataset, compared with text-based, image-based, the proposed method has a higher performance with 37.4% BLEU, 55.9% METEOR and 43.8% TER, respectively. It shows that cascade RCNN in this paper has better translation fluency when applied to both text and image. Compared with the benchmark system, the BLEU value of the proposed method is improved by 1.2%, indicating that the proposed system can effectively improve the syntactic continuity of the translated sentences. The TER value of the proposed method decreases by 1.2%, indicating that the editing distance between the cover-all translation result and the manual translation is shorter, and the translation error rate is lower (Karim et al., 2022; Wang et al., 2020). However, the METEOR value of image-based system increases by 0.6%, because image features can optimise the expression of semantic correlation of sentences, and the integration of cascade RCNN method can improve the synonym matching degree of translated sentences. In the Multi30K-17 dataset, the overall performance of the proposed method is better than that of text-based and image-based. Compared with the benchmark system, the BLEU value of proposed increases by 0.7% and the METEOR value increases by 1.0%. The performance improvement is comparable to the Multi30K-16 test results. However, the performance of Multi30K-17 test results is generally lower than that of Multi30K-16, because the test data of Multi30K-17 is new and the difficulty of sentence description is increased compared with that of Multi30K-16. When translating Chinese to English, the performance of proposed method is high, its BLEU value and METEOR value are increased by 0.6 and 0.9% respectively, and TER value is decreased by 1.0%. It is superior to the reference system in translation fluency, sentence semantic expression and translation error rate. From the above experimental results, it can be seen that the proposed model in this paper is superior to the benchmark system in both the translation of source language Chinese into English with relatively complex grammar and the new test set with relatively complex sentence description difficulty. The effectiveness of the proposed method is verified by the above experimental comparison and analysis.

5.2 Comparison with other models

Table 2 shows the performance comparison between the proposed model and different models when translating Chinese into English on Multi30K-16 and Multi30K-17 datasets respectively.

As can be seen from Table 2, when translating Chinese to English, compared with NMT, BLEU value and METEOR value of proposed model in the Multi30K-16 dataset increases by 3.5% and 3.4% respectively, and TER value decreases by 3.1%. The reason is that the NMT model only uses the relevant information of the source language, and NMT will refocus on the information of the past moment in the attention calculation, while the proposed model solves the above problems well. However, in the Multi30K-17 dataset, compared with the Multi30K-16 dataset, the maximum BLEU value and METEOR value of the proposed model increase by 10.0% and 7.5% respectively, and the TER value decreases by 19.9%. The reason is that the test corpus in Multi30K-17 is more difficult than that in Multi30K-16, and the NMT model is weaker in terms of translation fluency and sentence semantic relevance expression. The cascade RCNN model in this paper first incorporates image features in attention calculation, providing linguistic information that can optimise sentence fluency and semantic expression in translation and reduce translation error rate.

5.3 Comparative analysis of translation results with reference system

In order to verify that the application of the proposed method can reduce the over-translation or under-translation, Table 3 compares the Chinese translation results between proposed method and the benchmark system NMT through three examples. In example 1, the proposed model can correctly capture the occurrence tense of the accident, using the ‘past tense’. The benchmark model uses the ‘present tense’. In example 2, the reference system translates the description word ‘光脚丫’ in the source language into ‘without shoes’, while the proposed model describes the source language ‘光脚丫’ as ‘barefoot’, because the dual attention method in this paper can better capture context dependence and give some weight to important entities, thus improving the situation of under-translation. In example 3, our model correctly translates the word ‘预定’ as ‘order’ in the source language description. The benchmark model translates ‘book’ twice. The reason is that the double attention method in this paper can reduce the repeated attention to the generated words in the attention calculation, so it can reduce the over-translation and under-translation to achieve better translation results.

6 Conclusions

This paper proposes a Chinese-English translation method combining cascade RCNN and double attention decoding. The proposed method combines two independent attention mechanisms to focus on text features and spatial visual features respectively. In addition, the method incorporates cascade RCNN, which is used to record the text features and image features that have been paid attention to by the model during the calculation of current moment attention to avoid paying attention to repeated feature information. Experimental results show that the proposed method performs better than the benchmark system on the open test set and improves the translation performance. If we encounter complex sentences, our method may not guarantee validity when extracting features. In the future, we will make improvements in the image-assisted text task, try to use the image description generation model to obtain the hidden state of image features, and use the information of the hidden state of image features to further optimise the source text translation through the attention mechanism.

References

Bahdanau, D., Cho, K. and Bengio, Y. (2014) 'Neural machine translation by jointly learning to align and translate', *Computer Science*, DOI: 10.48550/arXiv.1409.0473.

Deena, G. and Raja, K. (2022) 'Keyword extraction using latent semantic analysis for question generation', *Journal of Applied Science and Engineering*, Vol. 26, No. 4, pp.501–510.

Delbrouck, B., Dupont, S. and Seddati, O. (2017) 'Visually grounded word embeddings and richer visual features for improving multimodal neural machine translation', DOI: 10.21437/GLU.2017-13.

Denkowski, M. and Lavie, A. (2014) 'Meteor universal: language specific translation evaluation for any target language', in *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, p.376380, https://aclanthology.org/W14-3348.pdf.

Feng, S., Liu, S., Li, M. et al. (2016) 'Implicit distortion and fertility models for attention-based encoder-decoder NMT model', DOI: 10.48550/arXiv.1601.03317.

He, Y., Xiang, S., Kang, C., Wang, J. and Pan, C. (2019) 'Cross-modal retrieval via deep and bidirectional representation learning', *IEEE Transactions on Multimedia*, Vol. 18, No. 7, pp.1363–1377.

Heo, Y., Kang, S. and Yoo, D. (2019) 'Multimodal neural machine translation with weakly labeled images', *IEEE Access*, Vol. 7, pp.54042–54053, https://ieeexplore.ieee.org/abstract/document/9265176.

Hirasawa, T. and Komachi, M. (2019) 'Debiasing word embeddings improves multimodal machine translation', DOI: 10.48550/arXiv.1905.10464.

Jelicic, V., Magno, M., Brunelli, D., Paci, G. and Benini, L. (2013) 'Context-adaptive multimodal wireless sensor network for energy-efficient gas monitoring', *IEEE Sensors Journal*, Vol. 13, No. 1, pp.328–338.

Karim, S., Tong, G., Li, J., Qadir, A., Farooq, U. and Yu, Y. (2022) 'Current advances and future perspectives of image fusion: a comprehensive review', *Information Fusion*, Vol. 90, pp.185–217, https://doi.org/10.1016/j.inffus.2022.09.019.

Li, F. (2022) 'Research on the management system of college students' innovation and entrepreneurship education based on B/S architecture', *Journal of Applied Science and Engineering*, Vol. 26, No. 5, pp.597–604.

Merritt, A., Chu, C. and Arase, Y. (2020) 'A corpus for English-Japanese multimodal neural machine translation with comparable sentences', DOI: 10.48550/arXiv.2010.08725.

Qian, K. and Tian, L.L. (2022) 'A topic-based multi-channel attention model under hybrid mode for image caption', *Neural Computing and Applications*, Vol. 34, No. 11, pp.2207–2216.

Qu, S., Xi, Y. and Ding, S. (2017) 'Visual attention based on long-short term memory model for image caption generation', in *2017 29th Chinese Control and Decision Conference (CCDC)*, pp.4789–4794, DOI: 10.1109/CCDC.2017.7979342.

Shi, Q., Yin, S., Wang, K., Teng, L. and Li, H. (2021) 'Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation', *Evolving Systems* [online] https://doi.org/10.1007/s12530-021-09392-3.

Shi, X. (2020) 'Image description generation in Chinese based on keywords guidance', *Computer Science and Application*, Vol. 10, No. 6, pp.1087–1097.

Singh, A., Saha, S., Hasanuzzaman, M. et al. (2021) 'Multitask learning for complaint identification and sentiment analysis', *Cognitive Computation*, Vol. 6, pp.1–16, https://doi.org/10.1007/s12559-021-09844-7.

Snell, J., Math, S., Mirault, J. et al. (2018) 'Parallel graded attention in reading: a pupillometric study', *Rep*, Vol. 8, No. 1, pp.1–9.

Song, Y., Chen, S., Jin, Q., Luo, W., Xie, J. and Huang, F. (2021) 'Enhancing neural machine translation with dual-side multi-modal awareness', *IEEE Transactions on Multimedia*, DOI: 10.1109/TMM.2021.3092187.

Su, J., Chen, J., Jiang, H., Zhou, C., Lin, H., Ge, Y., Wu, Q. and Lai, Y. (2020) 'Multi-modal neural machine translation with deep semantic interactions', *Information Sciences*, Vol. 554, pp.47–60, https://doi.org/10.1016/j.ins.2020.11.024.

Wang, X., Yin, S., Li, H. et al. (2020) 'A modified homomorphic encryption method for multiple keywords retrieval', *International Journal of Network Security*, Vol. 22, No. 6, pp.905–910.

Xiao, Q., Chang, X., Zhang, X. and Liu, X. (2020) 'Multi-information spatial-temporal LSTM fusion continuous sign language neural machine translation', *IEEE Access*, Vol. 8, pp.216718–216728, https://aclanthology.org/W14-3348.pdf.

Yin, S., Meng, L. and Liu, J. (2019) 'A new apple segmentation and recognition method based on modified fuzzy C-means and hough transform', *Journal of Applied Science and Engineering*, Vol. 22, No. 2, pp.349–354.

Ying, W. (2022) 'Gated recurrent unit based on feature attention mechanism for physical behavior recognition analysis', *Journal of Applied Science and Engineering*, Vol. 26, No. 3, pp.357–365, https://ieeexplore.ieee.org/abstract/document/9265176.

Young, P., Lai, A., Hodosh, M. et al. (2014) 'From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions', *Transactions of the Association for Computational Linguistics*, Vol. 2, pp.67–78, https://doi.org/10.1162/tacl_a_00166.

Yu, J., Li, J., Yu, Z. and Huang, Q. (2020) 'Multimodal transformer with multi-view visual representation for image captioning', *IEEE Transactions on Circuits and Systems for Video Technology*, December, Vol. 30, No. 12, pp.4467–4480, DOI: 10.1109/TCSVT.2019.2947482.

Yu, L., Park, E., Berg, A.C. and Berg, T.L. (2015) 'Visual Madlibs: fill in the blank description generation and question answering', in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp.2461–2469, DOI: 10.1109/ICCV.2015.283.

Zhao, B., Tao, J., Yang, M. et al. (2020) 'Deep imitator: handwriting calligraphy imitation via deep attention networks', *Pattern Recognition*, Vol. 104, p.107080, https://doi.org/10.1016/j.patcog.2019.107080.