Hindawi
International Journal of Digital Multimedia Broadcasting
Volume 2023, Article ID 2813702, 11 pages
https://doi.org/10.1155/2023/2813702

Research Article

On the Principles and Decisions of New Word Translation in Sino-Japan Cross-Border e-Commerce: A Study in the Context of Cross-Cultural Communication

Gaowa Sulun

School of Foreign Languages, Weifang University, Weifang, Shandong, China

Correspondence should be addressed to Gaowa Sulun; 20110540@wfu.edu.cn

Received 5 December 2022; Revised 27 February 2023; Accepted 28 March 2023; Published 18 April 2023

Academic Editor: Sayyouri Mhamed

In the context of the rapid development of multimedia and information technology, machine translation plays an indispensable role in cross-border e-commerce between China and Japan. However, due to the complexity and diversity of natural languages, a single neural machine translation model tends to fall into local optimality, leading to poor accuracy. To solve this problem, this paper proposes a general multimodal machine translation model based on visual information. First, visual information and text information are used to generate a visual representation of perceptual text information. Then, the two modal information are encoded separately, and the proportion of visual information in the whole multimodal information is controlled by a gating network. Finally, experiments are conducted on the image description datasets MSCOCO, Flickr30k, and video dataset VATEX, respectively. The results show that the algorithm in this paper achieves the best performance on both the BLEU and METEOR evaluation metrics.

1. Introduction

With the continuous development of information network technology and the acceleration of economic globalization, cross-border e-commerce in China has been developed significantly [1]. 2018 is the 40th anniversary of the establishment of diplomatic relations between China and Japan, and due to the influence of geography and other factors, the trade between China and Japan has increased year by year, and the number of cross-border e-commerce between China and Japan has also shown a trend of increasing year by year [2]. In the context of cross-border e-commerce, the two sides of the transaction from different countries need to clear customs by Internet, mail, or express, so there are higher requirements for translation. The overall translation and public awareness of small- and medium-sized cross-border e-commerce enterprise sellers are insufficient, and the quality of translation varies [3]. Often because enterprises do not pay enough attention to the translation of products, the translation of Japanese and Japanese products appear as incoherent words, logical structure confusion, and other phenomena, resulting in the inability to effectively let potential customers search for their own products.

New words and phrases are an important part of language life and a vivid record of social development [4]. In this context, the foreign translation of Chinese neologisms has also become an important window and an important part of China's foreign propaganda, and has contemporary significance [5]. As a close neighbour of China, Japan has preserved a large number of Chinese characters and absorbed the essence of Chinese culture, so the Japanese translation of new words is of great significance, and their dissemination and influence in Japan cannot be underestimated.

With the deepening of economic globalization, more and more Internet and e-commerce enterprises rely on translation systems, and machine translation has become an important tool to break through the barriers of different language communication [6]. Multimodal machine translation has better performance in terms of adequacy and accuracy when realizing system translation of more complex topics or scene descriptions [7].

Based on the characteristics of multimedia, the research of multimodal machine translation integrating visual information of text and images has received attention from researchers at home and abroad in recent years. In literature [8], two independent attention mechanisms are used to process word and image regions in the source language separately for enhancing the translation results of the model. Literature [9] is one of the few papers that incorporate local visual features of images for multimodal machine translation. The researchers extracted the local and whole image regions of the pictures and projected them into the vector space, respectively, and regarded them as pseudowords to be added into the input sequence of the model. It initially explores end-to-end multimodal machine translation incorporating local visual information and attention-based mechanisms. Literature [10] provided an in-depth study on whether multimodal information contributes to machine translation. In literature [11], it showed that image features extracted by convolutional neural networks (CNN) can have better results compared with the previously used synthetic image features.

Based on this, literature [12] improved the network by decomposing large convolutional features into multiple small convolutional feature layers, so that deeper networks can be trained to obtain better image feature representations. In addition, to select the optimal image representation layer, literature [13] performed an image classification task to study the accuracy of different image features. It extracts features from each layer of ResNet-50 [14] and evaluates the classification performance. The results based on the generated English description show a more gradual improvement in the performance from VGG-19 to ResNet-152 models. Three strategies are used in literature [15] to fuse picture features with text features. The first strategy is to add the global picture features extracted by the convolutional neural network to the head or tail of the original text sequence at the encoder side. A transformation matrix is used to solve the problem of mismatch between image and text embedding dimensions. The second strategy is to add multiple local image features to the head or tail of the original text sequence to help the model generate a more accurate representation on top of the first strategy. The main problem that needs to be solved is how to identify multiple local regions from a single image and how to extract and rank these visual features. In literature [16], the weighted sum of image-space representations was used as image features and combined with text features in a decoder based on a two-layer attention mechanism. Literature [17] used the gate mechanism [18] to apply image features to the encoder or decoder to improve the model's ability to understand words with duality.

Although the above studies incorporate image local features, they do not fully consider the different contributions of different parts of the image to the translated text and lack a translation framework that can handle multiple subtasks simultaneously. Therefore, this paper proposes a universal multimodal machine translation model based on visual information, aiming to handle two different multimodal machine translation tasks with a universal model.

The contributions of this paper are as follows:

(1) We migrate the methods in text-image machine translation to text-video machine translation and model the two multimodal machine translation subtasks in a unified way and handle the two multimodal translation tasks with a common model

(2) The model in this paper is based on the visual representation of perceptual text and introduces a multimodal gating network to selectively fuse visual information, so as to achieve the full utilization of multimodal advantages in specific scenes and accurately identify the semantic information of new words

(3) In the decoder part, this question model improves the encoder part of plain text translation. The whole model can model the task from multiple perspectives and multilevel data perspectives, which improves the translation effect and quality of the daily translation of new words

This paper consists of four main parts: the first part is the introduction, the second part is the Methodology, the third part is the Result Analysis and Discussion, and the fourth part is the Conclusion.

2. Methodology

This section introduces the techniques underlying the model in this paper, including the transformer-based machine translation model and the pretrained convolutional network. The model proposed in this paper is mainly improved based on the transformer model, and the visual information in multimodal information fusion is in the form of pretrained features.

2.1. Transformer-Based Machine Translation Model. Since deep learning became popular, some scholars began to try to introduce deep models into machine translation. One of the early attempts was the neural network machine translation model on a specific corpus in the 1990s. However, it did not receive wide attention due to the limitation of computer computing power. In 2013, Cambridge University proposed an end-to-end neural machine translation model [19], and then neural network-based translation models became the mainstream of machine translation.

The mainstream neural machine translation models adopt the "encoder-decoder" structure. Usually, the encoder or decoder consists of recurrent neural networks [20], CNN [21], and so on. Figure 1 shows a simple machine translation model with an "encoder-decoder" structure. The encoder encodes the source language text into an intermediate state vector, and the decoder decodes the intermediate state vector into the target language text.

The "encoder-decoder" structure enables text conversion of different lengths, which is not only suitable for machine translation tasks but also widely used in the fields of text summarization and text retelling. The equation for the

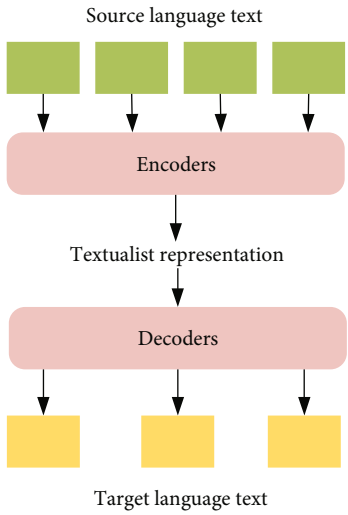

Figure 1: A machine translation model with a simple "encoder-decoder" structure.

"encoder-decoder" structure is as follows:

$$C = F(i_1, i_2, \cdots, i_w), \tag{1}$$

$$j_x = A(C, j_0, \cdots, j_{x-1}), \tag{2}$$

where $i_1, i_2, \cdots, i_w$ represents the input source language text sequence, F represents the encoder, C represents the sentence vector of the source language text; that is, the encoder is used to encode the source language sentence to obtain the sentence vector. In decoding, the input of decoder A is the sentence vector C of the source language and the already translated sequence of words in the target language $j_0, \cdots, j_{x-1}$. Since the "encoder-decoder" structure was proposed, neural machine translation models have almost always been explored, improved, and optimized around this structure.

In 2017, Google's research team proposed the groundbreaking transformer model. The model not only achieves optimal results on multiple datasets of machine translation but also excels on other tasks. The main structure of transformer is shown in Figure 2.

The main difference between the transformer and previous networks is the use of attention mechanisms. When encoding text, the transformer uses a self-attentive mechanism instead of the commonly used recurrent neural network. The self-attentiveness is calculated as follows:

$$\text{Attention}(V, Z, Q) = \text{Softmax}\left(\frac{VZ^N}{\sqrt{d_z}}\right)Q, \tag{3}$$

where V, Z, and Q are all outputs of the previous layer. In the initial layer, V, Z, and Q are the input word representation vectors. d_z is the dimension of V, Z, and Q. In self-attentiveness, the dimensions of the three are the same. In the specific implementation, the model uses multiheaded attention to improve the modeling ability of attention.

$$\begin{aligned}\text{MultiHead}(V, Z, Q) &= \text{Concat}(\text{head}_1, \cdots, \text{head}_b)M^0, \text{head}_x \\ &= \text{Attention}\left(VM_x^V, ZM_x^Z, QM_x^Q\right).\end{aligned} \tag{4}$$

In which, the multiheaded self-attended V, Z, Q are mapped to different spaces using a fully connected network, and then the self-attended operations are done separately for each head. The final results are stitched together, and then the final output is obtained using the fully connected network.

The use of the residual connection is the key to the transformer's ability to achieve multilayer network stacking. The specific implementation of residual connection is as follows:

$$i_{l+1} = i_l + F(i_l, M_l), \tag{5}$$

where i_l and i_{l+1} denote the hidden state of the current layer and the hidden state of the next layer, respectively. $F(i_l, M_l)$ is a layer in the model; M_l is model parameters of the current layer. It generally uses 4-6 layers of codec structure in the field of machine translation, and even tens of layers of encoders or decoders in pretraining networks. The ability to achieve deep network structure is largely due to the use of residual connections.

In the decoder, the modeling of the source language text relies equally on multiheaded self-attentiveness. Since the model does not know the future words in advance during the inference prediction phase, unlike the multiheaded self-attentiveness in the encoder, the self-attentiveness in the decoder requires the use of a hiding matrix to hide the words that have not yet been translated. The semantic interaction between the decoder and the encoder is also implemented in the form of multiheaded attention, and the implementation is similar to that of self-attentiveness. The difference is that Z and Q come from the output of the upper layer of the encoder, while V comes from the encoded output of the target language text. The model in this paper takes the transformer as the base model and makes corresponding improvements and innovations for the characteristics and difficulties of multimodal machine translation tasks.

2.2. Pretrained CNN. In multimodal machine translation, directly using the original image or video increases the difficulty of model construction. CNN is the most commonly used deep model in computer vision. 2012 Alexnet [22], proposed by a scholar, improved the original convolutional neural network by deepening the network structure and introducing the dropout module and won the ImageNet competition. Since then, there has been a lot of research on CNN. CNN is composed of a convolutional layer, a pooling layer, and a fully connected layer. Figure 3 shows a classical convolutional neural network structure.

Among them, the main role of the convolution layer is to extract the features of the image and generate a new feature map by filter calculation. After generating the new feature map, it is necessary to apply a modified linear

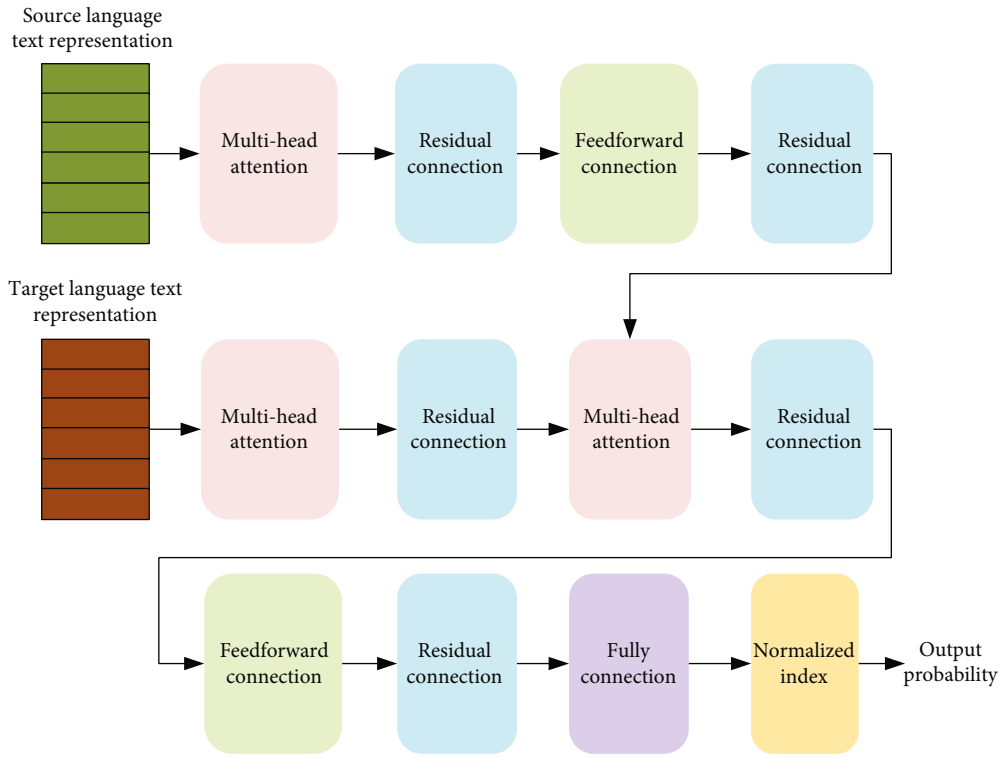

Figure 2: Main structure of transformer model.

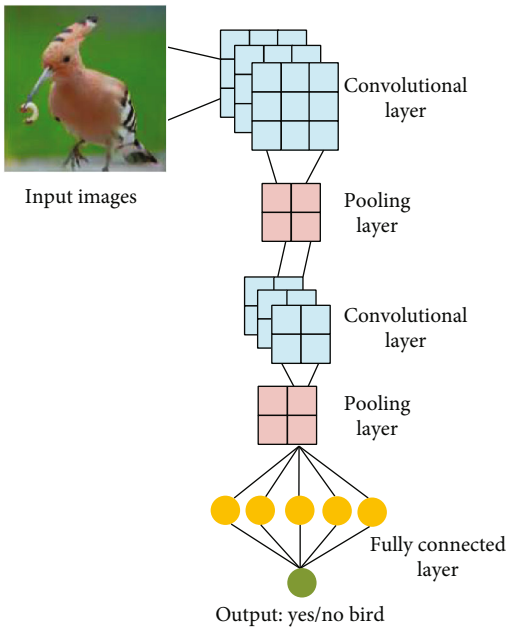

Figure 3: Classical convolutional neural network structure (for recognizing multimedia information such as images).

unit (ReLU) to introduce nonlinearity to the model and give it nonlinear representation capability. The modified linear unit is calculated as follows:

$$F(i) = \max(0, i). \tag{6}$$

That is, the modified linear unit will return itself for all inputs greater than 0 and vice versa output 0. After the convolution layer, due to the large dimensionality of the feature map, it is necessary to reduce the feature dimensionality and retain most of the key information. The generally chosen method is pooling operation, including maximum pooling and average pooling. The new feature map after sampling is obtained by the pooling layer. Usually, the deep convolution model repeats the convolution and pooling operations several times.

Finally, in order to get the probability distribution of each class and classify the image according to the features, the convolutional features are connected using a fully connected layer. Finally, the probability value of each category is predicted by a normalized exponential function (Softmax).

The convolutional network model trained on the classification task can be used as a generic vision model when the training set is large enough and the data is distributed over a generic scene. The intermediate features of the model can be used in a variety of machine vision scenarios. Such trained convolutional networks are called pretrained CNN. In this topic, for image data in text-image multimodal machine translation, this paper uses VGG pretrained network to extract visual features. For video data in text-to-video multimodal machine translation, the I3D pretraining network is used to extract visual features.

The VGG network is proposed by Oxford University and Google, and the model uses a series of small convolutional kernels and maximum pooling layers to deepen the network. It enhances the model generalization ability while reducing the error rate and is a commonly used pretraining model

for extracting image features. Specifically, the model uses a 3×3 sized convolution kernel and a 2×2 sized pooling kernel, followed by three fully connected layers after a series of convolutional layers. The first two fully connected layers have a fixed number of channels, 4096, and the last layer is used for classification, which has 1000 channels because the categories are 1000.

The VGG network provides pretraining models with different numbers of layers, the most commonly used ones are VGG16 with 16 layers and VGG19 with 19 layers. In this project, the VGG19 is chosen as the pretraining model, and the output before the fully-connected layer is taken as the spatial features of the image with the feature dimension of $7 \times 7 \times 512$. The obtained visual features retain the semantic and positional information of the input image.

In the text-image machine translation task, the features of the VGG pretrained network are stored in binary files in this paper. When the model is trained, the corresponding visual feature vectors are returned by the training data serial number.

Unlike the image-oriented pretraining model, the video-oriented pretraining model is still in the development stage, and the video dataset is relatively small. To address the above problems, this paper adopts the I3D video pretraining model, which is used to migrate the model trained on massive image data to video data.

The network structure of the I3D model is optimized based on the image network structure. The convolution kernel and pooling kernel are transformed by repeating the $N \times N$ 2D convolution kernel or pooling kernel N times, and by increasing the dimensionality in the temporal direction, the 3D convolution can obtain the features in the temporal dimension. Meanwhile, to further improve the model performance, optical flow optimization is applied by using a dual flow network, where two networks are trained with RGB data and optical flow data, respectively, and the predicted results are averaged at test time. The structure of the dual-stream network is shown in Figure 4.

In the text-to-video machine translation task, this paper also stores the features of the I3D pretraining network into binary files so that the video features are involved in the training process of the model.

2.3. Construction of the Proposed Model. Based on the transformer machine translation model and pretrained convolutional networks mentioned in the previous section, this section proposes a universal multimodal machine translation model based on visual information. This method solves the problem of the lack of a universal multimodal information fusion framework in multimodal machine translation and solves multiple problems with one model.

The universal multimodal machine translation (MMT) model framework proposed in this section is shown in Figure 5. The framework organically combines visual information and text information to generate a visual representation of perceptual text through a multimodal attention mechanism. And the text representation and the visual representation of the perceptual text are encoded using independent encoders. After the implicit state output of the

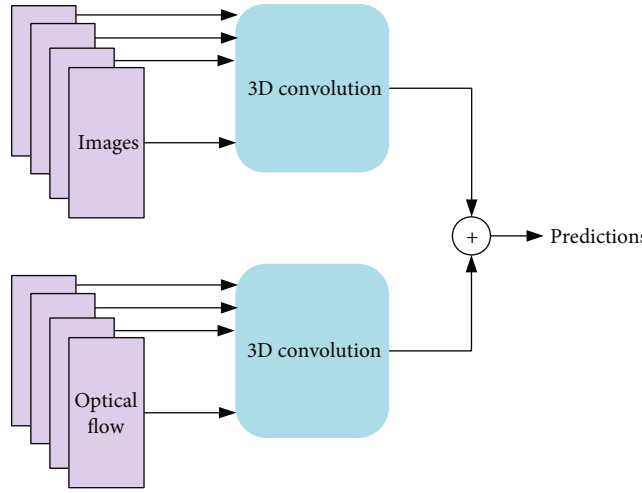

Figure 4: Dual-stream I3D model based on 3D convolution.

encoder is obtained, the percentage of visual information is automatically controlled by the independent multimodal gating network designed in this paper, and the multimodal contextual representation is finally obtained.

The method proposed in this paper can be applied to a variety of multimodal machine translation scenarios, e.g., text-image multimodal machine translation. A randomly initialized text representation and an image representation obtained from a pretrained convolutional network are combined to generate an image representation of the perceived text. In text-video multimodal machine translation, the text representation and the video representation obtained from the 3D convolutional pretraining network produce the video representation of the perceptual text. The proposed method can be used not only in the field of multimodal machine translation but can also be applied to areas involving other multimodal feature fusions.

Based on the above characteristics, the proposed method is named universal MMT (universal multimodal machine translation model) based on visual information. The word "universal" highlights the fact that the proposed method can be applied to a variety of multimodal scenarios. The improvement of this method over the baseline model is mainly in the encoder. In order to use visual information wisely, a visual representation of the perceptual text and a multimodal gating network is introduced in this paper, which is described separately in this section.

2.3.1. Visual Representation of Perceptual Text. The visual representation of perceptual text is to align the visual information with the text information, calculate the similarity score between the text information and the visual information to obtain the attention matrix, and then further reassign the weights to the visual representation according to the attention matrix. For each word of the text sequence, the visual representation of the perceptual text is obtained by weighting and summing all regions of all visual representations. This subsection describes in detail the computation of the visual representation of the perceptual text and its application in this paper. In the previous section, this paper

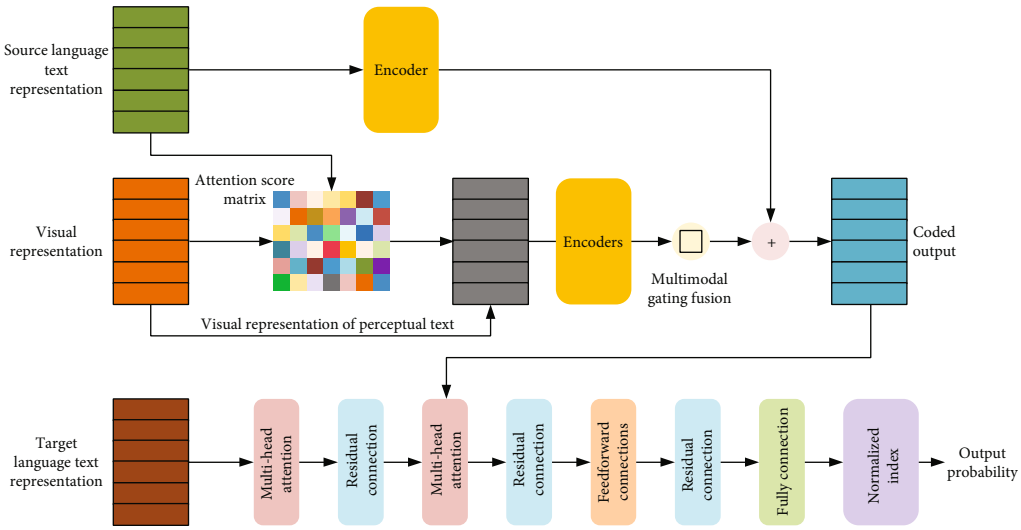

Figure 5: A framework for a generic multimodal machine translation model based on multimedia visual information.

describes the application method of visual information, where image data are transformed into spatial visual features and video data are transformed into spatiotemporal visual features. In this paper, we define the spatial visual features of an image as $\text{feat}_{\text{spatial}} = [\text{feat}^0_{\text{spatial}}, \text{feat}^1_{\text{spatial}}, \text{feat}^2_{\text{spatial}}, \cdots, \text{feat}^w_{\text{spatial}}]$, where W is the number of image regions. $\text{fea t}_{\text{temporal}} = [\text{feat}^0_{\text{temporal}}, \text{feat}^1_{\text{temporal}}, \text{feat}^2_{\text{temporal}}, \cdots, \text{feat}^T_{\text{temporal}}]$ is the spatiotemporal feature of the video, where T is the spatiotemporal visual feature. Define text sequence as $\text{text} = [\text{text}^0, \text{text}^1, \text{text}^2, \cdots, \text{text}^V]$, where V is the text length. Table 1 explains the specific method for computing the visual representation of the perceptual text using text-image machine translation as an example.

In Table 1, step 1 first transforms the input text into word embeddings according to the word embedding table. The word embedding table can be obtained from pretrained word embedding or by random initialization. In this paper, random initialization is used as the generation method of the word embedding table. Step 2 is to get the spatial visual features of the image by image pretraining model VGG19. Steps 3 and 4 indicate that if the visual feature dimension is different from the dimension of the text representation, the visual feature dimension needs to be converted to the same dimension as the text representation by transforming the dimension through a linear, fully connected network. Step 5 indicates that the attention matrix is obtained by dot-producing the text representation and visual features and by normalizing the normalized index. Step 6 is to obtain the visual representation of the perceptual text using the attention matrix obtained in step 3 and the spatial visual features in step 2. In steps 7 and 8 of the algorithm, two independent transformer encoders are used to encode the text representation and the visual representation of the perceptual text, respectively. After obtaining the two textual and visual hidden states, the two hidden states are returned in step 9. Later, the visual hidden states are assigned weights. and the hidden states of the two modalities are fused in a multimodal gated network.

For the sake of formal brevity, steps 4 and 5 of Table 1 can be expressed in the same form as the attention mechanism in the transformer.

$$\text{Attention}(V, Z, Q) = \text{Softmax}\left(\frac{VZ^N}{\sqrt{d_z}}\right)Q, \tag{7}$$

$$\begin{aligned}\text{MultiHead}(V, Z, Q) &= \text{Concat}\left(\text{head}_1, \cdots, \text{head}_b\right)M^0, \text{head}_x \\ &= \text{Attention}\left(VM^V_x, ZM^Z_x, QM^Q_x\right),\end{aligned} \tag{8}$$

where V is the textual representation and Z and Q are both visual representations. Since the visual representation has been converted to the same dimension as the textual representation, d_z is the dimension of the textual representation and the visual representation. In order to enhance the representation capability of the model and strengthen its robustness, the attention in this section is the same as in the transformer, which also uses multihead attention. As in Equation (5), different, fully connected matrices are applied to each attention head mapped to different spaces, calculated separately, and the output is stitched as the final output, which is the visual representation of the perceptual text.

2.3.2. Multimodal Gating Network. In different multimodal machine learning tasks, visual information plays different roles. For example, in the image description generation (image caption) task, the model needs to understand the semantics in the image and decode it into text. At this point, the visual information plays a decisive role in the whole task, and the description text cannot be obtained without the input of images. However, in multimodal machine translation, visual information does not always play a decisive role. In multimodal machine translation, pictures are usually used as an additional supplement to the bilingual parallel corpus, and the model relies on information from the source language text to understand the complete semantics as well. It has been shown that visual information plays a greater role

Table 1: Visual representation algorithm for generating perceptual text.

Step no.	Specification
1	Input: image, text
2	$\text{Text} - \text{representation} = \text{embedding\_lookup\_table}(\text{text})$
3	$\text{feat}_{\text{spatial}} = \text{VGG19}(\text{image})$
4	$\text{If dim}\left(\text{feat}_{\text{spatial}}\right) \neq \text{dim}(\text{text} - \text{representation})\ \text{then}$
5	$\text{feat}_{\text{spatial}} = \text{visual} - \text{linear}\left(\text{feat}_{\text{spatial}}\right)$
6	$\text{Attention} - \text{matrix} = \text{softmax}\left(\text{text} - \text{representation} \bullet \text{feat}_{\text{spatial}}^{N}\right)$
7	$\text{Text} - \text{aware visual representation} = \text{attention} - \text{matrix} \bullet \text{feat}_{\text{spatial}}$
8	$\text{Text} - \text{hidden} = \text{transformer} - \text{text encoder}(\text{text} - \text{representation})$
9	$\text{Visual} - \text{hidden} = \text{transformer} - \text{visual encoder}(\text{text} - \text{aware visual representation})$
10	Return text-hidden, visual-hidden
11	Output: text-aware visual representation

when there is ambiguity in the text and the semantic expression of words is unclear. In contrast, in general scenarios, the text alone can express the semantics of the whole sentence more accurately.

In order to make visual information be applied more rationally in multimodal machine translation, a gating network is designed in this section. By controlling the proportion of visual information in the overall multimodal information, the weights of visual information are dynamically assigned so as to fully utilize multimodal information in specific scenes and reduce information redundancy in general scenes.

In neural networks, deciding some information retention or forgetting is usually done using gating networks. For example, in long short-term memory (LSTM) networks, input gates, forgetting gates, and output gates are introduced to control the flow of information in the network in order to dynamically control the effect of historical information on the current state. The introduction of the three gating networks in the long short-term memory network controls the proportion of current state and historical information on the one hand and solves problems such as gradient disappearance explosion in the traditional recurrent neural network on the other hand. The input gates, forgetting gates, and output gates in the long- and short-term memory networks are calculated as follows:

$$x_n = \text{Sigmoid}(M_1 i_n + M_2 b_{n-1}), \tag{9}$$

$$f_n = \text{Sigmoid}(M_3 i_n + M_4 b_{n-1}), \tag{10}$$

$$o_n = \text{Sigmoid}(M_5 i_n + M_6 b_{n-1}). \tag{11}$$

The form of all three is consistent in that the weight sums of the input of the current time step and the hidden state output of the previous time step are obtained first, and then the probabilistic outputs with values between 0 and 1 are obtained by the activation function Sigmoid.

In order to allow visual information to dynamically obtain its share in multimodal information, this paper uses a multimodal gating network to dynamically assign the weights of visual information. The specific approach is similar to that of gating networks in long- and short-term memory networks.

$$\lambda = \text{Sigmoid}\left(M_1 b_{\text{text}} + M_2 b_{\text{image}}\right), \tag{12}$$

$$b_{ww} = b_{\text{text}} + \lambda b_{\text{image}}. \tag{13}$$

In Equation (12), b_{text} and b_{image} represent the outputs of the text and visual independent encoders, respectively; M_1 and M_2 represent the two independent parameter matrices. After obtaining the output of the multimodal gating network, the output λ is used as the weight of the visual information. The textual hidden state output and the visual hidden state output with weights are summed as the final multimodal contextual representation b_{ww}. The contextual representation b_{ww} incorporates both textual and visual information and automatically assigns the percentage of visual information according to the semantics, thus, at certain degree, reducing the redundancy of information. Compared with a single-machine translation model, the multimodal-machine translation model can better fuse and recognize multimedia text-image information.

3. Result Analysis and Discussion

3.1. Experimental Datasets. To verify the effectiveness of the model in this paper on text-image multimodal machine translation tasks, experiments are conducted using two mainstream image description datasets, MSCOCO and Flickr30k. The Flickr30k dataset contains 31,000 images. Referring to the literature [11], 29,000 of these images were randomly used for training, 1000 images for testing, and 1000 images for validation. The MSCOC0 dataset is the most commonly used benchmark dataset for testing image description tasks, with 123,287 images covering the vast

majority of natural life scenes, containing 82,783 training images and 40,504 validation images. Following the setup method widely used in the literature [6], 113,287 images were divided for training, 5,000 images for validation, and 5,000 images for testing.

In the video-text machine translation task, the dataset used for text is the VATEX dataset, which is the only video-text machine translation dataset available. 28,991 videos are provided in the VATEX dataset.

3.2. Evaluation Method. In the experimental stage, two evaluation metrics, BLEU and METEOR, are used to evaluate the generated translations. BLEU is the most commonly used evaluation metric in machine translation, and the main idea of the BLEU evaluation metric is to calculate the N -gram cooccurrence degree of the generated translation and the reference translation. First, the N-gram accuracy is calculated as follows:

$$U_t = \frac{\sum_x \sum_z \min\left(b_z(c_x), \max_{y \in w} b_z(s_{xy})\right)}{\sum_x \sum_z \min\left(b_z(c_x)\right)}, \quad (14)$$

where z and x denote the z-th N-gram fragment and the x -th sentence. w and y denote the total w-th reference translation and the y-th reference translation. c_x, s_{xy}, and b_z denote the currently processed sentence, the standard reference translation, and the number of occurrences of the current N-gram sentence, respectively. N-gram precision is biased for phrase translation, so the final output of BLEU contains a multiplicative factor BP.

$$\mathrm{BP} = \begin{cases} 1 & xf\ l_c > l_s, \\ e^{1-l_s/l_c} & xf\ l_c \le l_s, \end{cases} \quad (15)$$

where l_c is the length of the translation model-generated translation and l_S is the length of the standard translation. The final BLEU is calculated as follows, where the upper value of N is generally 4.

$$\mathrm{BLEU} = \mathrm{BP} \times e^{\sum_{t=1}^{T} m_t \log U_t}, m_t = \frac{1}{T}. \quad (16)$$

METEOR index improves some deficiencies in the BLEU index and also considers recall rate and accuracy rate. METEOR also introduced WordNet, with the aim of counting synonyms to get closer to how humans judge a translation. The higher the score, the better the resulting sentence quality.

The text-to-image machine translation task will use both BLEU and METEOR multimodal machine translation evaluation metrics. The text-video image machine translation will be evaluated using the BLEU metric only due to the limitation of the test set.

Figure 6 shows the results of different models on the MSCOCO (Microsoft COCO) dataset. The ordinal axes in Figure 6 represent the quantitative metrics of METEOR for BLEU in percentage (%). Among them, this model achieves 75.7%, 60.3%, 46.6%, 37.3%, and 29.7% on BLEU1, BLEU-2, BLEU-3, BLEU-4, and METEOR, respectively. It can be seen that the performance of the proposed model is significantly better than that of other methods. The methods used for comparison are the latest methods from literature [23], literature [24], and literature [25], respectively.

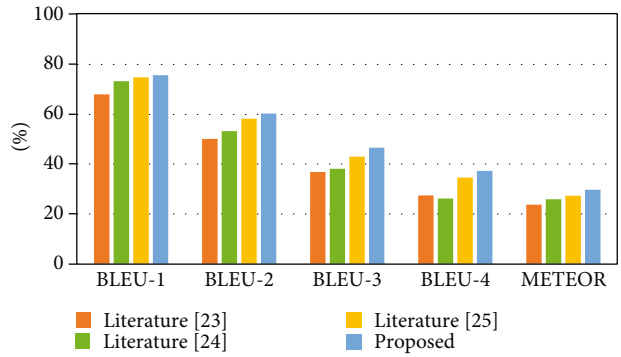

Figure 6: Comparison of experimental results of different models on the Microsoft COCO dataset.

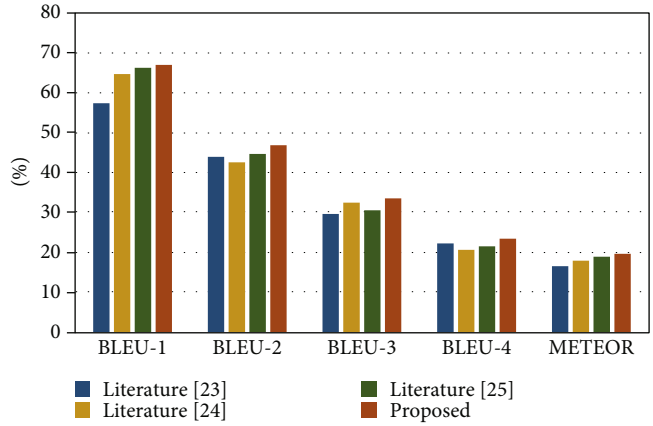

Figure 7: Comparison of experimental results of different models on the Flickr30k dataset.

Table 2: Comparison results of the proposed model with other methods on the VATEX dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Literature [23]	60.45	42.24	26.85	19.12	53.95
Literature [24]	66.31	40.55	29.34	17.43	52.37
Literature [25]	62.08	39.16	33.02	22.02	51.54
Proposed	67.73	42.85	34.95	23.83	55.86

The results of different models on the Flickr30k dataset are shown in Figure 7. The ordinal axes in Figure 7 represent the quantitative metrics of METEOR for BLEU in percentage (%). Similarly, the best performance of BLEU and METEOR metrics is obtained for the models in this paper on the Flickr30k dataset.

Literature [23]: Lead the way
Proposed: One belt one road

Literature [23]: Blue's defense war
Proposed: Blue sky defense

Literature [23]: A place for free trade
Proposed: Pilot free trade zone

Figure 8: Comparison of recognizing multimedia image information between the proposed model and the model in literature [23].

Literature [24]: Performance guarantee insurance for consumer loans
Proposed: Personal loan insurance

Literature [24]: Sharing economy
Proposed: Global sharing economy

Literature [24]: Tagline for app package purchase
Proposed: Banners for app purchases

Figure 9: Comparison of recognizing multimedia image information between the proposed model and the model in literature [24].

Product A

Literature [25]: You can put it in your bag
Proposed: Foldable, small and light, can be put in the bag

Product B

Literature [25]: The apron can be adjusted
Proposed: H-shape to reduce shoulder stiffness; adjustable neck strap

Figure 10: Comparison of recognizing multimedia image information between the proposed model and the model in literature [25].

From the results in Figures 6 and 7, it can be seen that the proposed model improves on the encoder of the transformer model which relies on plain text data only and utilizes picture data, and the BLEU and METEOR metrics are substantially improved on all datasets. This shows that the image information can help the model understand the contextual semantics and improve the translation effect.

In order to verify the effectiveness of the proposed universal MMT model on text-video multimodal machine translation tasks, the results were tested on the text-video dataset VATEX as shown in Table 2.

From Table 2, it can be seen that the model in this paper has the highest scores on BLEU and METEOR, which proves the superior performance of the proposed model.

Figures 8–10 show the qualitative comparison of the description results generated by the model in this paper with the methods in the literature [23], literature [24], and literature [25]. From Figures 8 and 9, it can be seen that the description of new words generated by the model in this paper contains richer image information and are better for long sequence words. Moreover, the model in this paper can describe new words appearing in images such as FTA and blue sky defense war, while other models cannot aptly translate the image contents. Compared with the translations generated by other models, the method in this paper contains richer semantic information and can achieve an accurate and overall description of the image content with higher translation quality.

Table 3: Summary of comparative methods.

Model	Summary	Strengths	Limitations
Literature [23]	Based on the encoder-decoder framework, it aims to guide the model to generate a more descriptive sentence for a given image by introducing reference information.	The sentences it generates sound more natural.	Sentence expressions are on the rigid side and performance is weak.
Literature [24]	It mimics the cortical lateral inhibition mechanism in the human visual system. Then, each image feature is identified by global saliency.	It can accurately detect the most prominent areas in the image while ignoring other local interferences, thus obtaining high results.	The algorithm has a high false alarm rate for scenes that are cluttered and have no obvious prominent areas in the scene.
Literature [25]	In order to extract the keywords from the original documents, it proposes a keyword extraction algorithm based on a probabilistic neural network and visual attention mechanism.	It has a strong ability to extract context-rich information about keywords.	The algorithm is time-consuming, and the algorithm complexity needs to be optimized.
Proposed	Perceptual text information is generated using visual information and text information. A generic machine translation model is implemented by controlling the proportion of visual information in the overall multimodal information.	The multimodal text information is fully utilized and the accuracy of identifying semantic information of new words is higher.	The dataset it uses suffers from a small size, and the available data needs to be expanded to enhance the expressive power of the model.

The comparative analysis of the above qualitative experimental results shows that the model in this paper is able to detect and accurately describe the target objects contained in the images. It has some guiding significance for the translation strategy of Japanese translation of new words in cross-border e-commerce.

To facilitate understanding and comparison, the summary, strengths, and limitations of these comparative methods are represented in Table 3.

4. Conclusion

With the rapid development of multimedia and e-commerce, the data volume of cross-border e-commerce platforms is large, and various new words are appearing every day. To improve the accuracy of machine translation, this paper proposes a general multimodal machine translation model based on visual information. The method in text-image machine translation is migrated to text-video machine translation, and the two multimodal machine translation subtasks are modeled uniformly. The two multimodal translation tasks are handled by a generic model. The model in this paper is based on the visual representation of perceptual text and selectively incorporates visual information so as to take full advantage of multimodality and accurately identify the semantic information of new words. However, the dataset used in this paper suffers from the problem of small size. Therefore, in future work, we will consider using techniques such as semisupervised unsupervised to expand the available data to enhance the expressive power of the model. In the meantime, the "K-fold cross-validation" technique will be used to validate the algorithm for systematic analysis of its performance.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The author declares no competing interests.

Acknowledgments

This study is supported by Weifang University.

References

[1] S. Ma, Y. Chai, and H. Zhang, "Rise of cross-border E-commerce exports in China," *China & World Economy*, vol. 26, no. 3, pp. 63–87, 2018.

[2] X. D. Shen, X. Chen, R. Ji, and R. H. Wu, "The new ecosystem of cross-border e-commerce among Korea, China and Japan based on blockchain," *Journal of Korea Trade*, vol. 24, no. 5, pp. 87–105, 2020.

[3] Y. Wang, M. Agyemang, and F. Jia, "Resource orchestration in supply chain service-based business model: the case of a cross-border E-commerce company," *Sustainability*, vol. 13, no. 21, p. 11820, 2021.

[4] Y. Qian, Y. Du, and X. Deng, "Detecting new Chinese words from massive domain texts with word embedding," *Journal of Information Science*, vol. 45, no. 2, pp. 196–211, 2019.

[5] C. Jiahuan, "The translation of Chinese diplomatic neologisms: a case study of[J]," *Journal of Sociology and Ethnology*, vol. 3, no. 6, pp. 64–70, 2021.

[6] K. Xu, J. Ba, and R. Kiros, "Show, attend and tell: neural image caption generation with visual attention[C]," *International conference on machine learning*, 2015, pp. 2048–2057, Miami, Florida, USA, 2015, PMLR.

[7] U. Sulubacak, O. Caglayan, S. A. Grönroos et al., "Multimodal machine translation through visuals and speech," *Machine Translation*, vol. 34, no. 2-3, pp. 97–147, 2020.

[8] R. Wang, "Research on intelligent English translation method based on the improved attention mechanism model[J]," *Scientific Programming*, vol. 2021, Article ID 9667255, 8 pages, 2021.

[9] I. Calixto and Q. Liu, "An error analysis for image-based multi-modal neural machine translation," *Machine Translation*, vol. 33, no. 1-2, pp. 155–177, 2019.

[10] Y. Ren, N. Xu, M. Ling, and X. Geng, "Label distribution for multimodal machine learning," *Frontiers of Computer Science*, vol. 16, no. 1, pp. 1–11, 2022.

[11] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions[C]," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137, Boston, MA, USA, 2014.

[12] M. Wang, J. Guo, and Y. Chen, "Make the blind translator see the world: a novel transfer learning solution for multimodal machine translation[C]," *Proceedings of the 18th biennial machine translation summit Virtual, online*, vol. 1, 2021no. 1, pp. 139–149, 2021.

[13] C. Ju, A. Bibaut, and M. van der Laan, "The relative performance of ensemble methods with deep convolutional neural networks for image classification," *Journal of Applied Statistics*, vol. 45, no. 15, pp. 2800–2818, 2018.

[14] C. Coleman, D. Kang, D. Narayanan et al., "Analysis of dawnbench, a time-to-accuracy machine learning performance benchmark," *ACM SIGOPS Operating Systems Review*, vol. 53, no. 1, pp. 14–25, 2019.

[15] B. Ren, "The use of machine translation algorithm based on residual and LSTM neural network in translation teaching," *PLoS One*, vol. 15, no. 11, article e0240663, 2020.

[16] B. Zhang, D. Xiong, and J. Su, "Neural machine translation with deep attention[J]," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 154–163, 2020.

[17] B. Zhang, D. Xiong, J. Xie, and J. Su, "Neural machine translation with GRU-gated attention model," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4688–4698, 2020.

[18] F. Meng and J. Zhang, "DTMT: a novel deep transition architecture for neural machine translation[C]," *Proceedings of the AAAI conference on artificial intelligence, 2019*, vol. 33, 2019no. 1, pp. 224–231, Honolulu, Hawaii USA, 2018.

[19] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation[J]," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–38, 2021.

[20] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[21] J. Lu, L. Tan, and H. Jiang, "Review on convolutional neural network (CNN) applied to plant leaf disease classification," *Agriculture*, vol. 11, no. 8, p. 707, 2021.

[22] R. A. Minhas, A. Javed, A. Irtaza, M. T. Mahmood, and Y. B. Joo, "Shot classification of field sports videos using AlexNet convolutional neural network," *Applied Sciences*, vol. 9, no. 3, p. 483, 2019.

[23] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han, and Q. Liu, "Neural image caption generation with weighted training and reference," *Cognitive Computation*, vol. 11, no. 6, pp. 763–777, 2019.

[24] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3264–3274, 2018.

[25] X. Wu, Z. Du, and Y. Guo, "A visual attention-based keyword extraction for document classification," *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 25355–25367, 2018.