International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1518-1529

https://doi.org/10.70517/ijhsa464136

Standardized Research on English Translation Teaching Assessment through the Integration of Intelligent Algorithms and AI

Sufang Yu[1,*], Haifeng Li[1] and Minhui Lian[2]
[1] School of Humanities, Nanchang Vocational University, Nanchang, Jiangxi, 330500, China
[2] Fujian Yixue Education Technology Group Co., Ltd., Xiamen, Fujian, 361000, China
Corresponding authors: (e-mail: 18739090586@163.com).

Abstract Multimodal machine translation can improve the accuracy and fluency of machine translation-generated translations, while overcoming the ambiguity problem that exists in traditional text-only machine translation tasks. In this paper, after preprocessing the original text information, the visual information features of the image are extracted using contextual information and convolutional neural network, and then the visual information features are deeply interacted and jointly encoded with the text information features. The text information and visual information can be more closely and accurately fused, so as to improve the comprehension ability of the English multimodal machine translation method. The experimental results show that the English multimodal machine translation method fusing visual information proposed in this paper can alleviate the problem of insufficient resources for real-time translation tasks with fewer samples by virtue of its own good multimodal comprehension ability, and the BLUE score of the model in this paper is improved by 1.04 compared with that of Transformer. It also improves the focus on noun-verb and acquires more semantic features. At the same time, the application of multimodal machine translation also provides richer data support for teaching assessment, which is conducive to the construction of more scientific assessment standards.

Index Terms multimodal machine translation, visual information, contextual information, convolutional neural network, semantic features

I. Introduction

In the English teaching system, in addition to listening, reading and writing, English translation is also a key teaching goal, as well as an important aspect of comprehensively improving the comprehensive ability of English [1], [2]. English language proficiency has become more important as society develops, and ensuring the level of quality course teaching by carrying out effective course design is not only the basis for improving the level of English translation teaching, but also the key to improving the level of English teaching [3]-[6]. However, due to the lack of teaching assessment, there are many problems in the practice of English translation teaching that have not been able to take reasonable countermeasures, resulting in the overall level of teaching is not high, resulting in students' comprehensive ability is weak, etc., and has not yet efficiently accomplished the goal of English teaching, based on which it is of great significance to assess English translation teaching with the help of intelligent technology [7]-[10].

Assessment technology in English translation teaching has always been a necessary tool for teachers and students, which can help teachers better understand the learning situation of students, and then develop a more effective teaching plan to improve the teaching effect [11]-[13]. Intelligent assessment technology refers to the use of artificial intelligence technology to achieve automatic assessment of students' learning, which is mainly used in the field of language education, using technical means to automatically assess students' language skills, oral fluency and other aspects, and provide students with personalized learning suggestions [14]-[17]. Intelligent assessment technology can automate the assessment of students' learning and provide students with immediate feedback and guidance during the learning process, which greatly improves the teaching efficiency of teachers and also enables students to learn more efficiently [18]-[21]. At the same time, using intelligent assessment technology, students can be provided with personalized learning suggestions according to their language level and learning habits, so that they can be more targeted to complete the English learning task [22]-[24]. In addition, with the help of intelligent assessment technology, teachers can have a more comprehensive understanding of the students' learning situation, including the learning process, learning difficulties and learning needs, which helps to improve the accuracy of the assessment, reduce the interference of subjective factors, and improve the quality of teaching [25]-[28].

Aiming at the shortcomings of traditional machine translation that lacks semantic interaction, this paper proposes an English machine translation method that adds visual image processing to the decoder. Firstly, relevant visual features are extracted using a convolutional neural network, and the pre-processed textual information features are co-integrated into a multimodal translation model. Then, based on Transformer's multi-attention mechanism, the sequence of elements is reconstructed to form a brand-new sequence to complete the fusion of visual image information and text information. Finally, the performance of the fusion method in this paper is verified on VATEX and MSVD-Turkish datasets. Experiments are conducted on a variety of English-French and English-German translation tasks to verify the specific effect of the method in real English translation.

II. English multimodal machine translation incorporating visual information

Traditional English teaching assessment criteria mainly rely on the accuracy of the text content, the evaluation standard is relatively single, while the English multimodal machine translation method integrating visual information introduces visual modality as an auxiliary means of assessment, which can more comprehensively examine the students' English translation level.

II. A. Transformer

The Transformer model abandons the previous recurrent and convolutional neural networks in favor of a modular neural network model with an encoder-decoder structure consisting of a self-attention mechanism and a feed-forward neural network (FNN) [29], and after each self-attention and feed-forward neural network layer, it undergoes residual connectivity and layer normalization. The basic structure of Transformer is shown in Fig. 1.

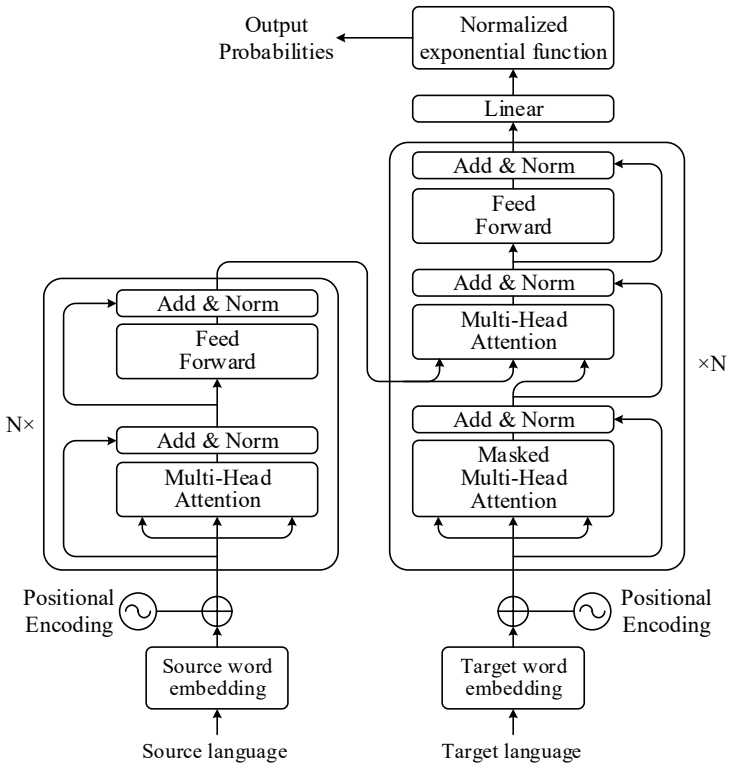


Figure 1: Transformer model structure

II. A. 1) Location coding

The recurrent neural network inputs only one word per time step so that the model can automatically recognize the position information of each word of the sequence, but in Transformer is to provide all the input utterance representations to the attention mechanism, due to the insensitivity of the attention mechanism to the position information, both the encoder and decoder need to cue the position information of each word to the model, so it is important to input the Each word is embedded with a vector representing the position information, this is the positional encoding of Transformer [30], there are many ways of positional encoding, in Transformer the sine cosine function with different frequencies is used:

$$PE(pos,2i)=\sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{1}$$

$$PE(pos,2i+1)=\cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{2}$$

II. A. 2) Self-attention mechanisms

The self-attention mechanism uses the same idea of query, key as well as value, and Transformer uses a scaled dot product attention mechanism [31]:

$$Attention(Q,K,V)=Soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

In fact Transformer uses not just an attention mechanism but a multi-head attention mechanism [32].

The multi-head attention mechanism slices Q, K, and V evenly according to the number of heads of attention, denoted as h, then $\{Q_1,Q_2,\cdots Q_h\}$, $\{K_1,K_2,\cdots K_h\}$, and $\{V_1,V_2,\cdots V_h\}$ will be obtained, and the attention result of the i th head is shown below:

$$head_i=Attention(Q_i,K_i,V_i) \tag{4}$$

Linear transformations are required for different attention heads using different parameter matrices i.e. $Q_i=QW_i^Q$, $K_i=KW_i^K$, $V_i=VW_i^V$ respectively and finally information fusion is done by right-multiplying another parameter matrix W^O, computed for multiple heads of attention, as shown below:

$$Multihead(Q,K,V)=Contact(head_1,head_2,\cdots head_h)W^O \tag{5}$$

The Transformer decoder is similar in structure to the encoder, the difference is that the decoder has an additional multi-attention module that adds a mask, this is because during model training, the decoder can only rely on outputs prior to the current timestep to make a prediction, this method is called future information masking, and the mask is added to the self-attention mechanism as follows:

$$Attention(Q,K,V)=Soft\max\left(\frac{QK^T}{\sqrt{d_k}}+Mask\right)V \tag{6}$$

II. A. 3) Residual connectivity and layer normalization

After each self-attention module and feedforward neural network of Transformer, it needs to be processed including residual connectivity and layer normalization. Transformer has multiple encoder layers and decoder layers, and each encoder and decoder contains many sub-modules, and each module contains many network layers, which makes Transformer's structure huge and forms a deep network , which leads to complex information transfer in it and produces gradient explosion or gradient vanishing in the process of back propagation, for this reason it is necessary to introduce residual connectivity as shown below:

$$x^{l+1}=F(x^l)+x^l \tag{7}$$

x^l represents the output vector of layer l, $F(\ )$ represents the computation of the sublayer, and the residual connection is the sum of the output of the previous layer and l. But at the same time, it will lead to the instability of the training process, and it is necessary to add the layer normalization process as follows:

$$LN(x)=g\cdot\frac{x-\mu}{\sigma}+b \tag{8}$$

Layer normalization normalizes the data into a standard distribution with mean 0 and variance 1. μ and σ represent the mean and variance, respectively, and g and b are learnable parameters.

II. A. 4) Feedforward Neural Networks

The feed-forward neural network will extend the output of the self-attention layer to higher dimensions, which is beneficial for operations such as nonlinear transformations afterward, and Transformer relies heavily on the gain brought by this high-dimensional extension, which is computed as shown below:

$$FNN(x)=\max(0,xW_1+b_1)W_2+b_2 \tag{9}$$

It contains two linear transformations and one layer of ReLU activation function $\max(0,x)$, W_1, W_2, b_1, b_2 are learnable parameters, and the vectors are expanded to high dimensions by feed-forward neural network, but too high dimensionality will lead to excessive computational overhead, which requires the support of high-performance hardware.

As can be seen from the above introduction to the Transformer, the Transformer model is modular, each module has a specific function and role, the benefits of such a design in addition to easier to understand the structure and working principle of the entire model, improve the efficiency of the model's parallel computation, but also according to the task requirements of the model can be more convenient to modify and optimize the various modules targeted. This project is based on the Transformer model, and combines the characteristics and problems of different English multimodal machine translation tasks to carry out targeted design and innovation.

II. B.Text data preprocessing

The samples in the parallel sentence pairs first need to be screened to filter out samples that are too short or too long in length. Sentences with high repetition need to be removed as well. The general method is to calculate the local hash value between every two sentences and remove the samples whose similarity is higher than a certain threshold. The sequence then needs to be cut into lexical elements using a lexer.

Byte pair coding is a commonly used technique for slicing into lexemes.The first step of BPE records the words in the text and the frequency of occurrence of each word, then the characters used in the word are added to the vocabulary and the frequency of occurrence of each character is recorded. The second step looks for the character pairs with the highest frequency of occurrence to combine and merge together, and then repeats the first two steps for iteration until a human-set number of iterations or number of lexical elements is reached.

II. C.Image processing techniques

In the English multimodal machine translation task, in addition to the preprocessing of the text, the visual features of the image need to be extracted and transformed into a feature representation before the English multimodal translation model can utilize the semantic information present in the image. The current best method is to use a pre-trained convolutional neural network, which can be roughly divided into four steps:

(1) First build a convolutional neural network model.

(2) Train the convolutional neural network model.

(3) In the actual process of English multimodal translation task, generally do not train the convolutional network model from scratch, but use the trained model for feature extraction, which can save valuable resources and time.

(4) When using the pre-trained model to extract the visual features of an image, usually only the part before the convolutional layer is used, and the visual features are obtained by convolving the image with the convolutional kernel of the convolutional layer, and it is common to use the pooling layer to downsize the features before transforming them into vectors through the fully connected layer.

Finally it is necessary to design the model structure for feature fusion of text features and image feature vectors to help the model understand the semantic information in them.

II. D.Multimodal Machine Translation Methods for Fusing Visual Information

Traditional machine translation mainly improves translation quality by introducing static images from other modal information, and has obvious drawbacks in dealing with linguistic richness, cultural sensitivity, context-dependency, syntactic complexity, and overall coherence of the text, and needs to rely on sample distance algorithms to control the sensitivity to the data. As a result, traditional machine translation can cause problems.

II. D. 1) Visual Information Feature Extraction

In this paper, the inputs of each modality are feature extracted to obtain the respective feature vector representation:

$$h = \frac{h - f + 2p}{s} + 1 \tag{10}$$

$$p = \frac{f - 1}{2} \tag{11}$$

II. D. 2) Image visual semantic and textual information fusion strategy

Neural Machine Translation (NMT) uses a sequence-to-sequence translation model based on encoding and decoding.

$$PE\left(pos, 2i\right) = \sin\left(pos / 10000^{2i/d_{model}}\right) \tag{12}$$

$$PE\left(pos, 2i+1\right) = \cos\left(pos / 10000^{2i/d_{model}}\right) \tag{13}$$

Then:

$$Attention\left(Q, K, V\right) = soft\max(k)\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

The model adds a key component to the architecture, i.e., capturing and fusing the correlations between text and images through a specially designed component to extract effective contextual visual information guidance vectors from the input visual content, the structure of which is shown in Figure 2.

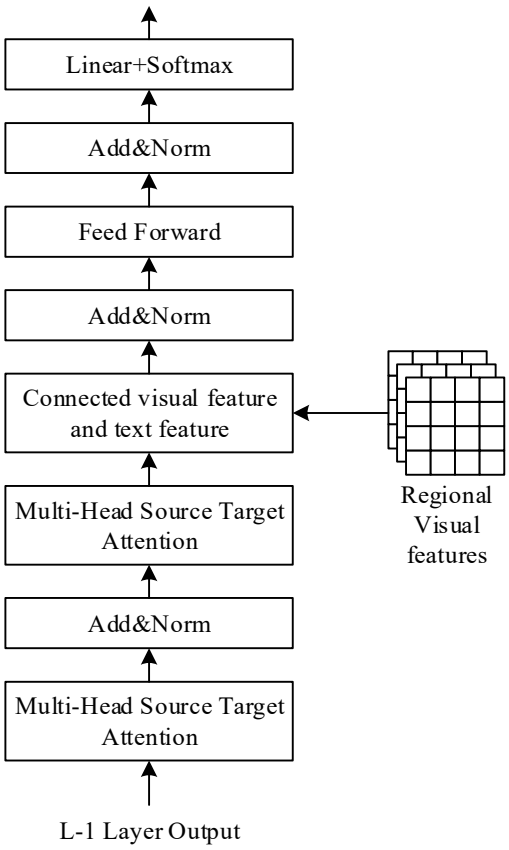


Figure 2: Model architecture

The decoder in this paper is an extension of the Transformer decoder:

$$H^{(l)}e = MultiHead\left(D^{(l-1)}, D^{(l-1)}, D^{(l-1)}\right) \tag{15}$$

$$H^{(l)}d = MultiHead\left(T^{(l-1)}, T^{(l-1)}, T^{(l-1)}\right), 1 \le l \le Ld \tag{16}$$

A trained model is introduced for experimentation and then predictions are made on the specified data:

$$MultiHead\left(K, Q, V\right) = Concat\left(head_1, \cdots, head_h\right)W_o \tag{17}$$

$$wherehead_i = Attention\left(QW_{iQ}, KW_{iK}, VW_{iV}\right) \tag{18}$$

III. Comprehensive model performance analysis and experimentation

III. A. Data sets

VATEX is a large-scale multilingual video description dataset. The dataset contains 600 human activities, each with Chinese and English descriptions provided by worker annotators. The official division of this dataset divides the dataset into training, validation and test sets. In this study, the dataset is divided into training set, validation set and test set according to the ratio of 80%, 10%, 10%.

MSVD-Turkish is an extension of the parallel corpus based on the MSVD dataset by adding Turkish descriptions to the original. It consists of more than 1000 videos with an average length between 10 and 25 seconds. The dataset follows the official standard segmentation and is partly used for training, partly for validation and the remaining part is used for testing.

III. B. Realization details

The work in this paper is implemented using PyTorch framework. In this paper, the model is trained using ADAM optimizer with an initialized learning rate of 0.001 and a batch size of 256. All experiments were performed on a GTX 3060Ti GPU computer.

III. C. Baseline model

Since there is less related work on video-guided multimodal machine translation, only the following six models are used as baseline standards in this paper:

(1) LSTM: As one of the most classical machine translation models, this approach is reimplemented in this paper. Since this model cannot take both modalities as inputs, this paper only uses text as input to the model.

(2) Transformer: As one of the most mainstream machine translation models and the main model used in this paper, this algorithm is re-implemented in this paper in order to compare the effectiveness of the method in this paper and only use text as the input of the model.

(3) VMT: Video-guided multimodal machine translation system uses I3D features as input, bidirectional LSTM as video encoder, LSTM as language decoder, and introduces a simple attention mechanism.

(4) HAPE: This model is based on hierarchical attention and location coding and uses visual features other than I3D features as input.

(5) SHAN: The Spatial HAN network uses visual features other than I3D features as input. The network is migrated to video feature extraction and video features are extracted at object, frame and video level.

(6) DEAR: This approach uses reverse translation techniques to enhance the semantics at different levels of granularity, integrating dynamic semantics at sentence and concept level.

III. D. Evaluation indicators

This paper uses BLEU-4 as the final evaluation index. At the same time, in order to improve the persuasiveness of the experimental results, this paper also uses the indicator METEOR.

III. E. Experimental results and analysis

III. E. 1) Comprehensive performance comparison experiments

Table 1 shows the corpus-level BLEU-4 scores for different models on the VATEX and MSVD-Turkish datasets. The model in this paper obtained BLEU scores of 36.20 and 36.54 on VATEX and MSVD-Turkish, respectively. On the VATEX dataset, it is 0.35 points higher than the highest model available. On the smaller dataset MSVD-Turkish, on the other hand, the model proposed in this paper obtains a BLEU score of 36.54, which is 0.31 points lower compared to the best performing model.

Meanwhile, on the METEOR evaluation index, this paper's model all achieved the highest scores, 0.78 and 0.44 points higher than the state-of-the-art model on the two data and respectively. The experimental results show that the Transformer model, which uses only text as input, has approached or even surpassed the multimodal approach in terms of score. This shows that Transformer has a clear advantage in machine translation, which is one of the reasons for the higher scores of the models in this paper. This result shows the effectiveness of the method proposed in this paper.

Table 1: Comparison of model translation results

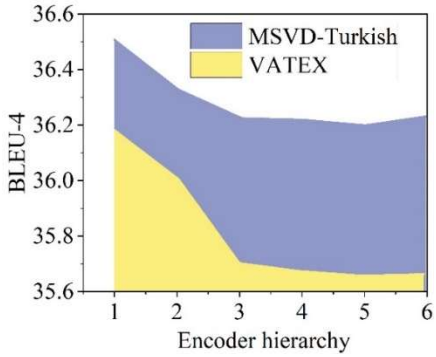
Model	Based model	Input mode	VATEX		MSVD-Turkish	
			BLEU	METEOR	BLEU	METEOR
LSTM	LSTM	Text	26.14	24.53	24.59	21.48
Transformer	Transformer	Text	35.17	33.14	35.5	32.81
SHAN	GRU	Text and video	35.85	--	36.18	--
HAPE	GRU	Text and video	35.25	31.40	35.48	30.60
HAPE^{+}	GRU	Text and video	35.35	31.65	35.48	31.69
DEAR	GRU	Text and video	35.80	33.39	36.85	34.22
VMT	Transformer	Text and video	31.06	29.60	27.14	27.16
This model	Transformer	Text and video	36.20	34.17	36.54	34.66

III. E. 2) Superparameters and other experiments

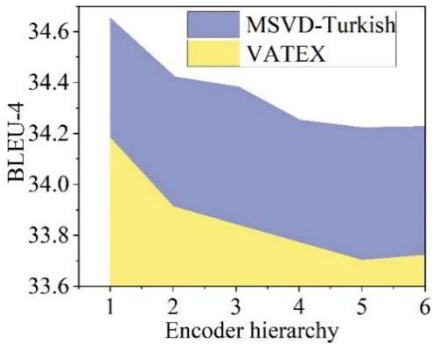
In this paper, the outputs of the six layers of the encoder are used as inputs for contrast learning respectively, and the final results obtained are shown in Fig. 3.

In the model mentioned in Chapter 2, this paper uses 8 attention heads, therefore, this paper assigns 0 to 8 attention heads for word level and phrase level respectively, in order to compare the effects of different methods of attention head assignment on translation quality. The experimental results are shown in Figure 4.

At the same time, this paper has experimented on the value of the hyperparameter λ. The experimental results are shown in Fig. 5, and the final value of λ reaches the best effect when it is 0.8.

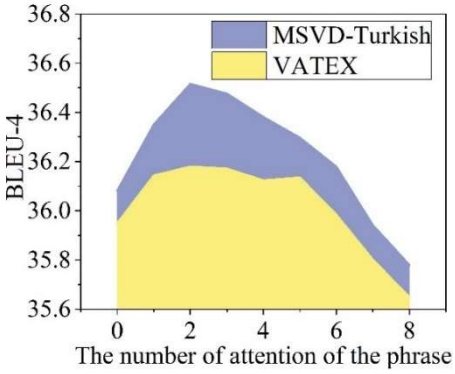


(a) BLEU-4 Value

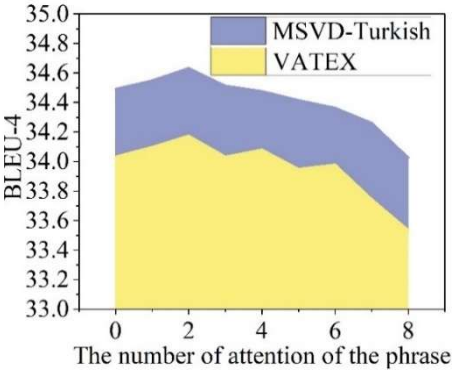


(b) METEOR-4 Value

Figure 3: Different encoder levels are used to compare learning effects

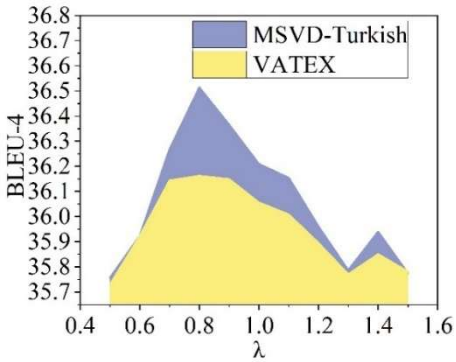


(a)BLEU-4 Value

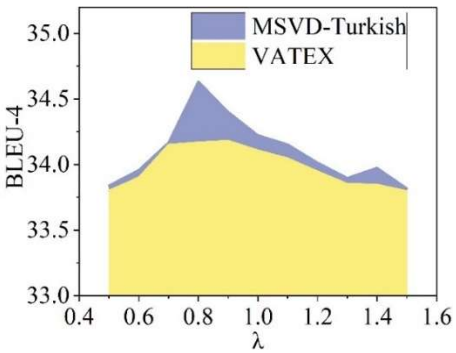


(b)METEOR-4 Value

Figure 4: The number of attention headers in the phrase level



(a)BLEU-4 Value



(b)METEOR-4 Value

Figure 5: Compare the superparametric λ

In order to explore the similarity between the samples, the exploration results are shown in Fig. 6, where 3000 samples are first randomly selected and their semantic similarity is calculated. In addition, in this paper, the computed results are sorted and normalized, where a value of zero means no relationship at all. The semantic similarity heatmap on the left shows that most values are zero, thus indicating that there is no semantic similarity between most samples. Therefore, the comparison learning method in this paper requires a large number of samples to obtain semantically related positive samples.

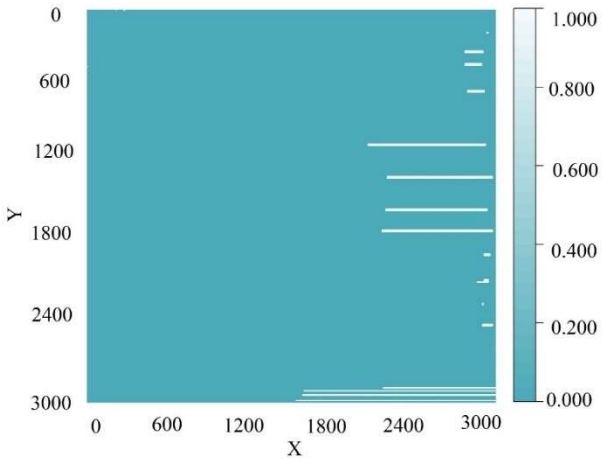


Figure 6: Semantic similarity between samples

IV. Example analysis of translation by multimodal machine translation methods

IV. A. Data set and experimental design

In order to compare with existing studies, this paper conducts experiments on the topic-enhanced translation model on the manually labeled dataset Multi30k.

For the Transformer-based topic enhancement model, the latest multimodal machine translation studies using Transformer as the computational unit are selected for comparison, including:

(1) VAR-MMT, a model that exploits visual coherence regularization on visual entity attention.

(2) GMMT, a graph model using object-level image features.

(3) Generative Imagination, an Imagination model that incorporates an adversarial neural network GAN.

(4) OVC, an object-oriented visual context modeling framework for efficiently capturing and exploring visual information for multimodal machine translation.

(5) DCC, a translation model incorporating dynamic capsule networks.

IV. B. Experimental results of translation effect comparison

The model tested the intertranslation from English to French and German using both BLEU and METEOR as measures. All the results are averaged after 5 experiments, and the specific results are shown in Table 2-Table 5.

From the results, it can be seen that this paper's translation model maintains the lead in both BLEU and METEOR metrics in the Multi30k dataset, and the METEOR metric is 1.1 ahead in the EN-DE task compared to the latest research, and the BLEU improves by 0.8. In MSCOCO, this paper's model based on the Transformer is better in English-German translation than that of the latest machine translation studies. The extraction of semantic features of the encoder is strengthened by fusing visual information, which improves the processing capability of the translation model itself on sentence semantics. Has a wider scope of application.

Table 2: The results on the Multi30k data set (GRU)

Method	EN-DE		DE-EN		EN-FR		FR-EN	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Text-Only	31.2	52.6	38.5	37.2	53.6	69.8	48.3	41.5
Imagination	31.5	52.7	38.5	37.5	54.1	70.4	49.0	41.7
VAG	32.0	52.6	39.1	37.0	53.8	70.6	48.0	41.7
Ours	32.8	53.8	39.5	37.8	54.4	70.6	49.7	41.9

Table 3: The results from the Multi30k data set (transformer)

Method	EN-DE		DE-EN		EN-FR		FR-EN	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Text-Only	31.2	52.4	38.2	37.4	54.3	70.7	48.3	41.8
VAR-MMT	29.6	51.1	35.7	35.4	53.3	70.5	48.9	41.7
GMMT	31.7	51.7	--	--	53.6	69.6	--	--
DCC	31.4	49.8	--	--	54.1	69.9	--	--
Generative Imag	32.7	52.7	--	--	52.4	68.4	--	--
OVC	32.4	52.2	--	--	54.5	71.0	--	--
Ours	33.6	53.9	40.3	38.5	54.9	71.9	49.5	42.0

Table 4: Translation results in the MSCOCO data set (GRU)

Method	EN-DE		DE-EN		EN-FR		FR-EN	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Text-Only	27.7	47.8	29.3	33.0	44.8	64.7	44.6	40.1
Imagination	28.2	48.1	29.5	33.0	44.8	64.0	44.8	39.8
VAG	28.4	48.1	30.1	33.7	45.1	64.8	45.2	40.5
Ours	28.9	48.6	30.7	33.9	45.8	65.1	45.9	41.1

Table 5: Translation results in the MSCOCO data set (transformer)

Method	EN-DE		DE-EN		EN-FR		FR-EN	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Text-Only	27.5	48.0	30.4	33.5	44.5	64.1	44.5	40.3
VAR-MMT	26.5	45.5	28.9	32.3	43.4	63.2	42.9	38.6
GMMT	28.4	47.8	--	--	--	--	--	--
DCC	26.9	45.5	--	--	45.7	65	--	--
Generative Imag	28.6	29.2	--	--	45.5	65.2	--	--
OVC	29.2	48.2	--	--	45.1	64.4	--	--
Ours	29.5	48.5	32.9	34.8	45.5	65.7	46.5	41.3

Further, the T-SNE is utilized to project the feature embeddings in the semantic space in a downscaled manner, taking English-French as an example, and the downscaled projection results are shown in Fig. 7 for the MSCOCO's test set, and the images, English, and French show good and more consistent distributions in the semantic space as a whole after the projection, with distributions ranging from -30 to 30 and from -40 to 40 on the x-axis and the y-axis, respectively.

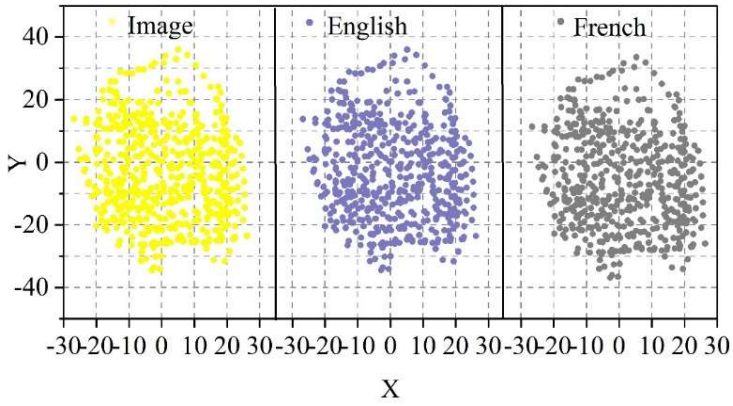


Figure 7: Reduced projection

Meanwhile, the instances of the three in the semantic space are projected into the same graph, and the projection results are shown in Fig. 8, where it can be observed that in details, there is a close pairing relationship between the source language, target language, and image subject of the 20 pairs of instances shown in the image.

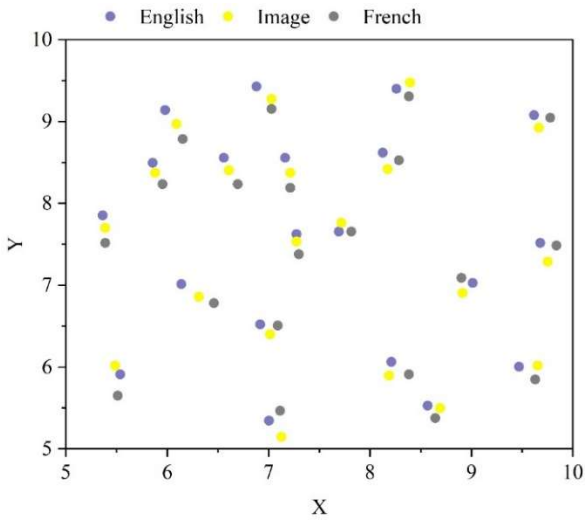


Figure 8: The descending dimension of the original space and semantic space

IV. C. The role of visual information in translation

Visual information in translation guides the model to focus on the semantic information of the text through the semantic space construction task. The Omission score is again used to observe the distribution of the attention of this paper's translation model to each part of the sentence, and the distribution of the Omission score with the synergetic attention value is shown in Fig. 9.Before constructing the semantic space, the model pays more attention to the syntactic structure under the guidance of the translation task, and thus the prepositions have higher Omission scores as compared to the part of the nouns and verbs.Visual Information After the intervention, it can be seen from the Omission score and parallel attention weights that the model in this paper pays higher attention to nouns and verbs under the guidance of the semantic space construction task, and the Omission score and parallel attention weights for the word waterfall reach 0.27 and 0.33 respectively.

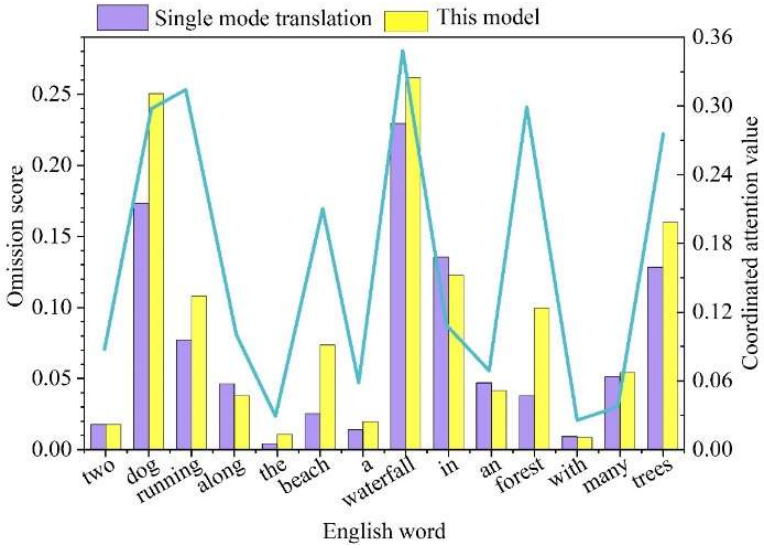


Figure 9: The problem score and the collaborative attention value distribution

V. Application of machine translation for standardization of assessment in the teaching of English translation

V. A. Focus on multi-platform access to resources

The application of English multimodal machine translation method integrating visual information in university English translation teaching requires teachers to obtain teaching resources through a richer teaching platform, which provides diversified materials for the standardization of English translation teaching assessment and makes the assessment content closer to the actual translation scene.

In recent years, the rapid development of advanced information technology has provided more abundant teaching resources for higher education. With the application of information network, teachers can obtain teaching resources from the broad network environment and reasonably apply them in the teaching classroom to carry out the teaching assessment work, which solves the limitations of the venue and time in the teaching in the past, and creates a more efficient interactive platform for students. For example, students can often use wiki encyclopedia, China Knowledge Network and other platforms for online learning. Through the rich translation resources on various Internet teaching platforms, a richer translation classroom environment has been created, and teachers have further improved the theoretical teaching classroom by creating online interactive tasks, online Q&A and many other different ways. Teachers assign translation tasks to students in a timely manner through the Internet teaching platform, organize students to carry out online real-time translation practice activities, and even complete the test and evaluation on the Internet platform, providing students with more convenient and rich teaching resources.

V. B. Comparing machine and human translation

It is more common to apply machine translation in the evaluation of English translation teaching, but there is a certain difference in the correctness rate of different machine translations, so teachers are required to investigate the machine translation software often applied by students in the class at the beginning of the university English translation teaching course.

Subsequently, students are organized to make a comparative analysis of the contents of different machine translations to help them grasp the differences and advantages and disadvantages of different machine translation software in the process of application, summarize the types of articles to which different machine translation software applies in practical application, and ensure that students are able to reasonably choose the translation software to provide translation services according to their own needs. The teacher provides an article and organizes students to choose machine translation software to translate it, and after the translation is completed, the teacher shows the manual translation and organizes students to compare the results of manual translation and machine translation, so as to make the translation results more scientific and standardized.

V. C. Expanding post-translation editorial corrections

Post-translation editing and correction is an important step in applying machine translation in university English translation teaching, which requires students to have the ability to correct the errors on their own after the self-translation is completed. The self-translated English translations should be touched up and modified more deeply according to the translation content of the English multimodal machine translation method that integrates visual information. In a way, post-translation editing is crucial in order to obtain a more accurate translation. The post-translation editing correction is more complicated and specific, including correction of morphology, singular and plural, tense, lexicality, etc. in grammar, while in syntax, it includes correction of problems such as improper connection of vocabulary, irrational sentence breaks, and lack of constituents.

VI. Conclusion

In this paper, we combine the visual image information with the traditional attention mechanism to propose an English multimodal machine translation model that extracts contextual information from vision.

The results of comprehensive performance comparison experiments show that this paper's model can achieve better English translation results. On VATEX and MSVD-Turkish, this paper's model obtains BLEU scores of 36.20 and 36.54, which are 0.35 and 0.31 points higher and lower than the best-performing model, respectively. The lesser performance than DEAR in the dataset MSVD-Turkish is due to the small sample size of this dataset, which limits the training of this paper's model. The model in this paper improves the BLUE score of Transformer, by 1.04 points, which proves the superiority of this paper's method in combining visual image information with the attention mechanism.

Under the guidance of visual image information, the English multimodal machine translation model in this paper obtains the most excellent English translation results on EN-DE, DE-EN, EN-FR and FR-EN tasks. Compared with the high attention to prepositions before the introduction of visual information, the model in this paper exerts higher attention to visually associated nouns and verbs, and the nouns and words contain more semantic features, thus improving the quality of English machine translation.

This paper shows the application prospect of machine translation in the standardization of English translation teaching assessment from three aspects: focusing on the acquisition of multi-platform resources, comparing machine translation with human translation and post-translation editing and correction, and this paper's English multimodal machine translation method integrating visual information reduces the interference of human factors and improves the objectivity and accuracy of the assessment.

References

[1] Ge, S. (2021). Application of translation workshop to college English translation teaching. International Journal of Emerging Trends in Social Sciences, 10(1), 34-40.

[2] Mazenod, A. (2018). Lost in translation? Comparative education research and the production of academic knowledge. Compare: A Journal of Comparative and International Education, 48(2), 189-205.

[3] Jalalzai, N. N. (2023). Unraveling Challenges: Factors Influencing English Translation Competence in Higher Education ESL Students. INTERNATIONAL JOURNAL OF LITERATURE, LINGUISTICS AND TRANSLATION STUDIES, 3(2), 38-54.

[4] Hartono, R. (2015). Teaching translation through the interactive web. Language Circle: Journal of Language and Literature, 9(2).

[5] Su, W. (2021). How to use modern teaching methods to improve English Chinese translation ability. Open Access Library Journal, 8(12), 1-6.

[6] Liang, H., & Li, X. (2018). Research on Innovation Method of College English Translation Teaching Under the Concept of Constructivism. Educational Sciences: Theory & Practice, 18(5).

[7] Siregar, R. (2018). Grammar based translation method in translation teaching. International Journal of English Language and Translation Studies, 6(02), 148-154.

[8] Petrocchi, V. (2014). Pedagogic translation vs. translation teaching: a compromise between theory and practice. Italica, 91(1), 95-109.

[9] Hu, J. (2023). Analysis of the feasibility and advantages of using big data technology for English translation. Soft Computing, 27(16), 11755-11766.

[10] Li, Z. (2022). The construction of university English translation teaching model based on fuzzy comprehensive assessment. Mathematical Problems in Engineering, 2022(1), 7755508.

[11] Bin, W. (2016). Empirical Study on the Computer-aided College English Translation Teaching. International Journal of Emerging Technologies in Learning, 11(12).

[12] Lai, N. (2024). An IoT‐based multiple teaching quality evaluation method for English translation with improved deep learning. Engineering Reports, 6(11), e12896.

[13] Bhatti, M. S., & Mukhtar, R. (2017). Analyzing the utility of grammar translation method & direct method for teaching English at intermediate level. IJAEDU-International E-Journal of Advances in Education, 3(7), 60-67.

[14] Sun, M. H., Li, Y. G., & He, B. (2017). Study on a quality evaluation method for college English classroom teaching. Future Internet, 9(3), 41.

[15] Wang, X. (2022). College English teaching quality monitoring and intelligent analysis based on internet of things technology. Wireless Communications and Mobile Computing, 2022(1), 6567123

[16] Jing, C., Zhao, X., Ren, H., Chen, X., & Gaowa, N. (2022). An approach to oral English assessment based on intelligent computing model. Scientific Programming, 2022(1), 4663574.

[17] Greeni, A., Chitkara, P., Pathak, P., Orosoo, M., Rengarajan, M., & Bala, B. K. (2024, July). Advancing Adaptive Assessment in English Language Teaching: A Deep Learning-based Approach within Intelligent Tutoring Systems. In 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) (pp. 1-7). IEEE.

[18] Zhang, G. (2023). The Evaluation and Development of University English Teaching Quality Based on Wireless Network Artificial Intelligence. Journal of Combinatorial Mathematics and Combinatorial Computing, 117, 77-86.

[19] Li, X., & Huang, C. (2025). Design of an Intelligent Grading System for College English Translation Based on Big Data Technology. Systems and Soft Computing, 200205.

[20] Zhao, X. (2022, June). From Tradition to Innovation: How Intelligent Technology Empowers Educational Evaluation. In Proceedings of the 8th International Conference on Frontiers of Educational Technologies (pp. 36-40).

[21] Yuanyuan, L., Zengzhao, C. H. E. N., Rong, C. H. E. N., Yawen, S. H. I., & Qiuyu, Z. H. E. N. G. (2023). Research on the Application Framework of Intelligent Technologies to Promote Teachers' Classroom Teaching Behavior Evaluation. Frontiers of Education in China, 18(2).

[22] Fang, Y. (2022). Design of oral English intelligent evaluation system based on DTW algorithm. Mobile Networks and Applications, 27(4), 1378-1385.

[23] Susyla, D., & Jaya, S. (2023). Digital Assessment in English Language Teaching (ELT): A Systematic Literature Review. Edu-Ling: Journal of English Education and Linguistics, 7(1), 135-156.

[24] Huang, L., & Ma, L. (2024). Research on the Application of Intelligent Grading Method based on Improved ML Algorithm in Sustainable English Education. Scalable Computing: Practice and Experience, 25(1), 451-463.

[25] Wang, Z. (2023). The Effect of Intelligent Evaluation Technology on Students' Initiative in Post-lecture Evaluation of Online Teaching. International Journal of Emerging Technologies in Learning (iJET), 18(22), 88-99.

[26] Wu, W., Berestova, A., Lobuteva, A., & Stroiteleva, N. (2021). An Intelligent Computer System for Assessing Student Performance. International Journal of Emerging Technologies in Learning, 16(2).

[27] Su, Y., Chen, G., Li, M., Shi, T., & Fang, D. (2021). Design and implementation of web multimedia teaching evaluation system based on artificial intelligence and jQuery. Mobile Information Systems, 2021(1), 7318891.

[28] Yang, G., & Huang, Y. (2023). Application of an intelligent evaluation model of online teaching based on improved BPNN. Systems and Soft Computing, 5, 200065.

[29] Navid Aftabi,Nima Moradi & Fatemeh Mahroo. (2025). Feed-forward neural networks as a mixed-integer program. Engineering with Computers,(prepublish),1-19.

[30] Caizheng Liu,Zhengyu Zhu,Wanming Hao & Gangcan Sun. (2025). Heterogeneous multivariate time series imputation by transformer model with missing position encoding. Expert Systems With Applications,271,126435-126435.

[31] Ma Xiang & Zhang Junsheng. (2020). GSA-Net: gated scaled dot-product attention based neural network for reading comprehension. Automatika,61(4),643-650.

[32] Fugang Liu,Shenyang Liu,Yuan Chai & Yongtao Zhu. (2025). Enhanced Mamba model with multi-head attention mechanism and learnable scaling parameters for remaining useful life prediction. Scientific Reports,15(1),7178-7178.