Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Research on intelligent language translation algorithm based on computer vision

Xiaojing Dong[1], Li Yuan[2,✉]

[1] *Jilin Engineering Normal University, Changchun, Jilin, 130000, China*
[2] *Northeast Normal University, Changchun, Jilin, 130000, China*

ABSTRACT

The rapid growth of multilingual information online has made traditional translation insufficient, highlighting the need for intelligent language translation. This study employs a convolutional neural network to extract visual features from translated images and uses region-selective attention to align text and image features. The fused information is then processed through a sequence model to develop a computer vision-based translation algorithm. Results show that the proposed algorithm excels in key evaluation metrics, improving translation quality. It maintains a low leakage rate (1.30%), a mistranslation rate of 2.64%, and an average response time of 67.28ms. With strong generalization and applicability in multilingual translation, the algorithm demonstrates high performance and promising real-world applications.

Keywords: convolutional neural network, attention mechanism, transformer, sequence model, intelligent language translation

1. Introduction

In recent years, with the rapid improvement of artificial intelligence technology, the translation tasks of translation robots are becoming more and more complex, the translation demand gradually tends to be intelligent and technological, and computer vision is widely used in the field of language translation. Computer vision is a technology that uses computers to perceive and judge the physical world, which involves computers, cameras, digital image processing, and recognition techniques for examining and analyzing images of physical objects [24, 1]. It captures images by virtue of a camera or other image sensor, reduces image noise and enhances image contrast during preprocessing, and

✉ Corresponding author.
E-mail address: peteryuan2024@126.com (L. Yuan).

Received 10 June 2024; Accepted 11 December 2024; Published Online 19 March 2025.
DOI: 10.61091/jcmcc124-42

subsequently recognizes and extracts meaningful information or features from the image through feature extraction algorithms [13, 16]. In the field of text translation, computer vision realizes accurate recognition of text characters through corresponding character matching, localization and segmentation algorithms to further enhance the language translation effect, which has certain practical application significance in the field of machine translation [18, 22, 2].

As people's requirements for the accuracy and translation quality of automatic language translation are getting higher and higher, grammatical error correction of language translation has gradually become an important way for people to improve the quality of language translation [6, 7]. Using computer vision to accurately recognize and classify the text characters of the translated language can realize English grammatical error correction and accurate translation, providing effective data support for language machine translation [11, 19, 3]. In addition, no matter machine translation tools or auxiliary translation tools, they usually pay more attention to the improvement of translation efficiency, as well as the lightweight upgrade of the product and better user experience when translating in English [25, 21]. The machine vision system for language translation can realize fully automatic proofreading of translations and improve the efficiency and quality of language translation by incorporating semantic understanding, word frequency statistics, multi-level metric algorithms, similarity algorithms and other technologies [8, 15, 12].

[23] proposed an end-to-end Visual Translation Embedding Network (VTransE) designed for image visual relation localization and prediction. The model represented images in a low-dimensional space as simple syntactic vectors and facilitated object-relationship knowledge transfer by introducing a feature extraction layer, which supported both training and inference of linguistic data. [14] developed a visual-semantic embedded long short-term memory (LSTM) network framework that effectively leveraged the relationship between sentence semantics and visual content by constructing a visual-semantic embedding space. This approach addressed the issue where previous visual translation systems generated sentences with true context but incorrect semantics and demonstrated improved performance in computer vision-based language translation tasks. [10] enhanced the neural machine translation model for computer vision applications by proposing a top-down feedback approach, which provided a simpler and more effective encoding representation of translated text features compared to traditional feed-forward feature encoding. Additionally, a noise-reducing self-encoder strategy was introduced to strengthen the encoder's ability to capture latent features of the source sentence, thereby improving neural machine translation performance. [5] utilized a language model to compute a weighted combination of multiple semantically similar words and embedded words with optimal semantics based on contextual information. This method achieved high accuracy on a machine translation dataset related to computer vision tasks. [17] demonstrated that on-the-fly image translation enabled real-time translation of source language text in images into the target language. The study further showed that establishing contextual links between images and text through computer vision techniques enhanced user comprehension of the translated text.

In this paper, we use convolutional neural network to extract the visual features of the images in the translated graphic content, and apply the attention mechanism to extract the correlation features of the image and text semantics, get the attention matrix of the text semantics-visual feature similarity, and realize the alignment operation of the image and text features. Subsequently, the information of each modal feature is fused and processed, input into the sequence translation model based on encoding and decoding, and output the translated language results through dynamic real-time adjustment. In this study, the performance of the intelligent language translation algorithm is tested by using datasets containing content in different translated languages, and the translation

accuracy and response feedback time of the intelligent language translation algorithm are analyzed through the application in real scenarios, so as to explore the application effect of the intelligent language translation algorithm.

2. Method

2.1. Visual information feature extraction methods

Due to representation or modeling limitations, the models are not able to take maximum advantage in integrating image visual semantics into the model. The source text can perform the translation task, but using only linguistic information fails to accurately extract the data features, in this paper, the inputs of each modality are feature extracted to obtain the respective feature vector representation, the embedding vectors of the text sequences are extracted using pre-trained text encoders, and the local features in the image are extracted using convolution [20], with the following formula:

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

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

First set the size of each convolution to 8×8, assuming that the step size $s = 1$, h represents the image height, p represents the image fill. Convolution process, the use of step size to control the size of the output data and the ability to extract, the more the convolution kernel can be extracted to the more features, of course, here also need to pay attention to the network complexity is too high caused by the phenomenon of overfitting.

Convolutional stereo image there are three RGB channels, each channel of the image dimension is $6 \times 6 \times 3$, the image of the convolution process, the number of channels of the filter must be consistent, this paper uses $3 \times 3 \times n_{mg}$ multi-filter to get 4×4 convolution of the image dimension, the output dimension is $3 \times 3 \times 3$.

2.2. Visual semantic and textual information fusion strategies

2.2.1. Text and image visual semantic alignment methods. In this paper, the region selective attention mechanism [9] is applied to the extraction of feature information in the image with the degree of semantic relevance to the text, to minimize the attention of the intelligent language translation model to irrelevant information in the image. When using the region selective attention mechanism to calculate the visual representation of the perceived text, the random sampling method will be applied, so that $X = \{x_1, x_2, ..., x_3\}$ represents the text vector representation of the text sequence after pre-training and linear mapping, and $F = \{f_1, f_1, ..., f_n\}$ represents the visual representation of the text vector mapped to the text vector in the asked dimension after the image feature extraction, and then the text vector and the visual representation are subjected to the multi-attention computation to obtain the text semantic-visual feature similarity attention matrix $Matrix_{sim}$.

where each element s_i in the attention matrix $Matrix_{sim}$ is represented as the similarity score between each source language word vector x_i and image feature vector F. After that, based on the similarity score s_i, the image feature index that is most relevant to each source language word vector x_i is selected as shown in Eq. (3):

$$j_i = argmax(s_i). \tag{3}$$

Obtaining the sampling results of each element in the attention matrix $Matrix_{sim}$ means sampling the image regions with high correlation with the word vectors at the current moment in the text, and finally all the sampling results form the region selection matrix $Matrix_{select}$. After that, the attention matrix $Matrix_{sim}$ is normalized to transform the similarity into a probability distribution, and the normalized attention matrix is multiplied by the region selection matrix $Matrix_{select}$ element by element, to get the final graphic and textual cross-modality Attention Matrix $Matrix_{text-vision}$, as shown in Eq:

$$Matrix_{text-vision} = Matrix_{select} \times softmax(Matrix_{sim}). \tag{4}$$

The visual feature $image_{featspatial}$ is multiplied with the graphic cross-modal attention matrix $Matrix_{text-vision}$ to obtain the final textual visual alignment representation $text-vision_{representation}$.

2.2.2. Information integration strategy. Intelligent language translation algorithms use a sequence translation model based on encoding and decoding. The input is taken as a sequence of source language words $p = \frac{f-1}{2}$ and the output is a sequence of target language words $Y = (y_1, y_2, ..., y_n)$. The goal of the NMT model is to learn the model that maximizes the probability of using Y given X, i.e., $P(X|Y)$. The method of combining other modal information is by manipulating the sequence of elements in a Transformer-based multi-head mechanism to form a new sequence. Each layer of the encoder layer consists of two sub-layers, Multihead Attention and Dot Feedforward Network. For a given source text sentence x, each word source is the sum of word embedding and positional encoding. The computational formula is as follows:

$$PE(pos, 2i) = \sin(pos/10000^{2i/d_{model}}), \tag{5}$$

$$PE(pos, 2i+1) = \cos(pos/10000^{2i/d_{model}}), \tag{6}$$

can be encoded using the same multilayer stacking approach, where each layer consists of two fully connected sublayers. In addition, batch residual connectivity and normalization are introduced for each sublayer. It should be noted that when iterating the customized layers, the list cannot be used directly, otherwise it may lead to the model weights and data inputs not being on the same device during the training process; meanwhile, when performing the model training, the updating of the parameters of each layer can easily lead to the instability of the computed values due to the drastic changes of the output layer. At this point, it is necessary to introduce batch normalization and use a small batch of mean and standard deviation to output the intermediate layers of the visual network to stabilize the stability of the output values of each layer of the neural network.

Each decoder consists of three sublayers, namely, masked multi-head attention, multi-head attention, and pointwise feedforward network. In this case, the query matrix Q, the key matrix K and the value matrix V in the multi-head self-attention function receive the output of the masked multi-head attention sublayer and the output of the encoder as inputs, respectively. After the target is embedded, the output is embedded and positional coding is added to obtain the input of the decoder layer. The formula is:

$$Attention(Q, K, V) = softmax(k)(\frac{QK^T}{\sqrt{d_k}})V. \tag{7}$$

Intelligent language translation algorithms can more accurately translate "source text" into "target language" through images. The most important feature of the intelligent language translation model based on computer vision in this paper is the introduction of contextual visual information guiding vectors, which are used to dynamically learn to generate multimodal context vectors for translation,

a mechanism that allows the model to take full advantage of multimodal data when dealing with translation tasks. Specifically, the model is architecturally augmented with a key component that captures and fuses the correlations between text and images through a specially designed component to distill valid contextual visual information guidance vectors from the input visual content.

The process of generating this guidance vector is dynamic and adaptive, and it can be adjusted in real time according to specific input scenarios (e.g., graphic content), resulting in a highly relevant multimodal contextual representation. This representation not only reflects the grammatical structures and lexical meanings in plain text, but also integrates the unique contextual information conveyed by visual elements, such as character expressions, object locations, action indications, and other non-textual descriptions of important content.

In practice, the model guides translation decisions based on this multimodal context vector when generating translation results, ensuring that the translated text not only accurately corresponds to the literal meaning of the original text, but also aptly reflects the specific context in which the original text is embedded, thus greatly enhancing the quality of translation and the efficiency of cross-lingual and cross-cultural communication.

The decoder in this paper is an extension of the Transformer decoder [4], where the generated sequences are used as inputs, and the hidden states of the target layer are used to generate multiple heads of attention through the stacking of the LDs in the same layer, enabling the model to represent the subspace information at different locations as input $D(l-1) \in \mathrm{R}$, where $D(0)$ is the dimension of the source sentence in layer $L-1$, d_w is the dimension of the model, and $D(0)$ denotes the concatenation of all the source words in the corpus:

$$H^{(l)}e = MultiHead(D^{(l-1)}, D^{(l-1)}, D^{(l-1)}), \tag{8}$$

$$H^{(l)}d = MultiHead(T^{(l-1)}, T^{(l-1)}, T^{(l-1)}), 1 \leq l \leq Ld. \tag{9}$$

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

$$MultiHead(K, Q, V) = Concat(head_1, \cdots, head_h)W_o, \tag{10}$$

where

$$head_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{11}$$

2.3. Intelligent language translation modeling

Based on the above image visual feature extraction and text and image visual feature fusion strategy, this paper constructs an intelligent language translation model based on computer vision, and the basic flow of the model is shown in Figure 1. After extracting the textual and image visual features in the text to be translated, a textual-visual semantic alignment similarity matrix is formed by calculating the similarity between the textual information and the visual information. Then, according to this similarity matrix, redistribute the weights to the visual information to get the textual visual semantic alignment representation, which effectively combines the visual information with the textual information to better reflect the correlation between the text and the image, so as to provide richer visual context and information. In the process of generating visual information, this paper mainly focuses on the object location and semantic information in the image, which is more inclined to the local features to supplement the scene context for the intelligent language translation process. In the first step, the text input is encoded by a parallel pre-training module to obtain the text representation $text_{represemtation}$. In the second step, the image input is feature extracted by the FasterR-CNN

convolutional neural network model to obtain the image feature vector $image_{featspatial}$. It can be seen from the third and the fourth steps that when the dimensions of the image features are unequal to the dimensions of the text representation, the image features are mapped by a linear layer to make their dimension matches with the textual representation. In the fifth and sixth steps, the dot product of the transpose of the textual representation and the image features is used to obtain the textual visual semantic alignment attention matrix, which is normalized using the softmax function to map the attention weights to probability distributions. In the seventh step, the textual visual semantic alignment attention matrix is utilized and multiplied with the image features to obtain the visual textual semantic alignment representation $vision - text_{representation}$. The role of the attention matrix is to weight the different image features for better fusion with the textual representation. In the ninth and tenth steps, the textual representation is fed into the text encoder of the Transformer to get the textual hidden representation $text_{hidden}$. In the twelfth and thirteenth steps, the textual visual semantic alignment representation is fed into the visual encoder of the Transformer to get the visual hidden representation $image_{hidden}$. Finally, the textual hidden representation and the visual hidden representation are used as the outputs of the final translation result.

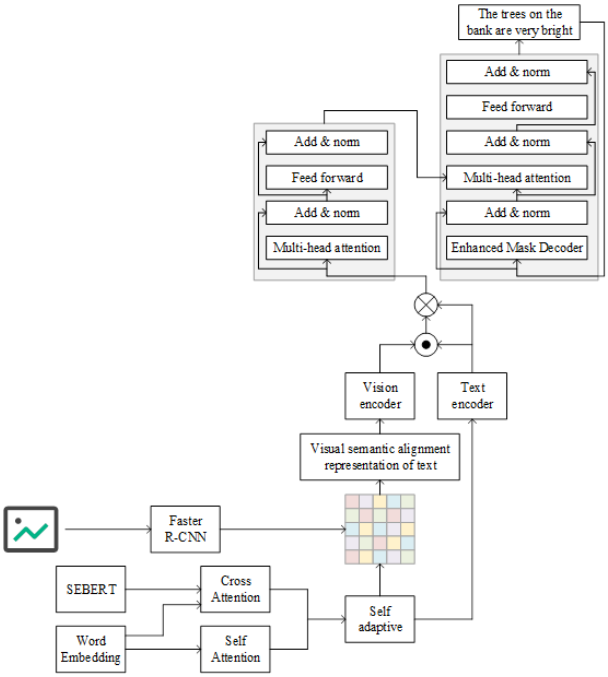


Fig. 1. Intelligent language translation algorithm process

3. Results and discussion

3.1. Intelligent translation algorithm performance test

3.1.1. Experimental data set.

1) Test Set. In order to verify the effectiveness of the intelligent language translation algorithm built based on computer vision in this paper, the algorithm is tested on each language translation task in WMT 21 Metrics Task and WMT 22 Metrics Task respectively.The statistics of the number of systems and the number of sentence pairs contained in a single system in five translation language pairs, namely, German-Chinese, English-Chinese, Korean-Chinese, French-Chinese, and Japanese-

Chinese, are shown in Table 1. Chinese, Korean-Chinese, French-Chinese, and Japanese-Chinese, the statistics of the number of systems and the number of sentence pairs contained in a single system are shown in Table 1, where the total number of sentence pairs = the number of systems × the number of sentence pairs contained in a single system. The total number of sentence pairs in the German-Chinese language in the WMT 21 Metrics Task dataset and the WMT 22 Metrics Task dataset are 41364 and 22935, respectively.

2) Training Set. For German-Chinese and Japanese-Chinese language translation tasks, this paper uses the WMT 15-17 Metrics Task German-Chinese and Japanese-Chinese language pairs of sentence-level task datasets for training. For English-Chinese, Korean-Chinese and French-Chinese translation tasks, this paper uses the sentence-level task dataset of WMT 17-19 Metrics Task for training.

Table 1. The test set is calculated according to the statistics

Data set	Data	German- Chinese	English - Chinese	Korean - Chinese	French - Chinese	Japanese - Chinese
WMT 21 Metrics Task	System number	27	10	20	16	10
	The log of a single system	1532	1209	1517	1275	1766
	Sentence to total	41364	12090	30340	20400	17660
WMT 22 Metrics Task	System number	15	19	10	11	20
	The log of a single system	1529	1659	1771	1444	1163
	Sentence to total	22935	31521	17710	15884	23260

3.1.2. Baseline methodologies and evaluation indicators. Comparison benchmarking models include several key approaches.

1) SCST directly optimizes evaluation metrics through a policy gradient approach to reinforcement learning, using the model's self-generated sequences as benchmarks against non-differential evaluation criteria (e.g., BLEU or CIDEr), with a view to directly boosting the scores of these evaluation metrics during the training process.

2) The Up-Down model uses a dual attention mechanism - the Bottom-Up mechanism focuses on recognizing key object features in the image, while the Top-Down mechanism directs visual attention based on the context of the generated text.

3) The RFNet model uses a recurrent neural network structure to fuse and refine features from different convolutional layers layer by layer, which enhances the representation of features in spatial and channel dimensions.

4) GCN-LSTM combines Graph Convolutional Networks (GCN) and Long Short-Term Memory Networks (LSTM) with the aim of capturing complex object relationships within an image by GCN to construct a high-level visual graph representation, while LSTM is responsible for generating coherent and closely related textual descriptions using this representation.

5) M2Transformer effectively integrates visual and textual information through multimodal information fusion and multi-attention mechanism to generate more accurate and detailed image descriptions.

6) The bilinear attention mechanism introduced by X-Transformer strengthens the interaction between image and linguistic features, and promotes the significant improvement of the image description generation process.

In this paper, standard description evaluation metrics, including BLEU-4 and METEOR, are used to evaluate the intelligent language translation algorithm proposed in this paper.

BLEU is a performance evaluation method widely used in the field of machine translation, and has

been gradually extended to other tasks of natural language processing, including image description generation, in recent years. The core of this metric is to evaluate the quality of the generated text through the similarity between the machine-generated text and a series of reference texts, with the basic idea that "the closer the machine translation is to the human translation, the better".The calculation of BLEU is based on the degree of overlap between the generated descriptions and the reference descriptions with different granularities (i.e., n-grams). The calculation of BLEU is based on the degree of overlap of different granularities (i.e. n-grams) between the generated description and the reference description. Common values of n are 1, 2, 3, and 4, which correspond to different levels of granularity.

METEOR is a metric for evaluating the quality of machine translation, and has also been widely used in the evaluation of image description generation in recent years. It was originally designed to make up for the shortcomings of traditional evaluation metrics such as BLEU, especially in assessing the match between the translated text and the reference text in a more fine-grained and comprehensive way. The calculation formula of the assessment metrics is as follows:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{12}$$

$$F_{mean} = \frac{10PR}{R + 9P}, \tag{13}$$

$$Meteor = F_{mean}\left(1 - \frac{0.5(c)^3}{m}\right). \tag{14}$$

3.1.3. Experimental results and analysis. This section will mainly show the experimental results and analysis of intelligent language translation algorithms based on computer vision, which are mainly divided into the following three kinds, performance experimental results comparison with existing methods and ablation experimental analysis.

1) Performance comparison analysis. The results of the performance comparison between the intelligent language translation algorithm proposed in this study and the current state-of-the-art methods on the WMT 21 Metrics Task and WMT 22 Metrics Task test datasets are shown in Figure 2, with 2a-2e representing the performance in German-Chinese, English-Chinese, Korean-Chinese, French-Chinese, and Japanese-Chinese translation, respectively. It can be seen that the proposed method obtains an average of 85.42, 79.71, 65.71, 54.50, and 34.52 on the metrics B@1 (BLUE-1), B@2 (BLUE-2), B@3 (BLUE-3), B@4 (BLUE-4), and M (Meteor), which are all superior to all the compared methods. Specifically, on the Meteor metrics score, the model in this paper (34.52) achieves a +1.14% increase compared to the best performing X-Transformer (34.13). This result not only demonstrates the robustness of this paper's intelligent language translation model in understanding the translated content of images and its relevance, but also emphasizes the superior performance in generating high-quality and relevant language translated text. Similarly, on the BLEU-4 score, compared to the best method X-Transformer (53.76), this paper's model (53.76) also shows an improvement of +1.38%, further validating the effectiveness of this paper's proposed method in accurately capturing key information and converting it into an accurately described translation language. Combining the results of these quantitative metrics comparisons, we can see the overall advantages of this paper's model in several key evaluation metrics, highlighting its innovative contribution to improving the quality of translation language generation.

2) Results of ablation experiments. In order to evaluate the effectiveness of the proposed visual feature extraction module (M1) and image text visual feature fusion module (M2), this study explores the effect of each component on the model performance by constructing ablation models with different settings. The experiments consisted of the following three configurations: a base model without any modules, the integration of the M1 module on top of the baseline model, and the integration of both the M1 and M2 modules into the baseline model. The results of the ablation experiments are shown in Table 2. In the experiments on the WMT 21 Metrics Task and WMT 22 Metrics Task datasets, the addition of the M1 module alone resulted in a 1.52%, 3.45%, 1.75%, and 5.26% improvement of the B@4 and Meteor indicators, respectively, which emphasizes the effectiveness of the visual feature extraction module in guiding the intelligent language translation algorithm to understand and translate the content more accurately. Further, when the M2 module is integrated on top of the M1 module, the performance of the Intelligent Language Translation Model is enhanced even further, reaching new heights in all evaluation metrics. In particular, the B@1 and Meteor scores were increased by 3.94%, 3.74%, 8.61% and 8.64% respectively compared with the baseline model, which fully proved the significant role of the text and image visual feature fusion module in promoting the deep integration of visual information and text information, and then improving the quality of model language translation generation. This experimental result not only verifies the superiority of the intelligent language translation model proposed in this paper on the translation task, but also reveals the key role of the two modules in improving the performance of the model.

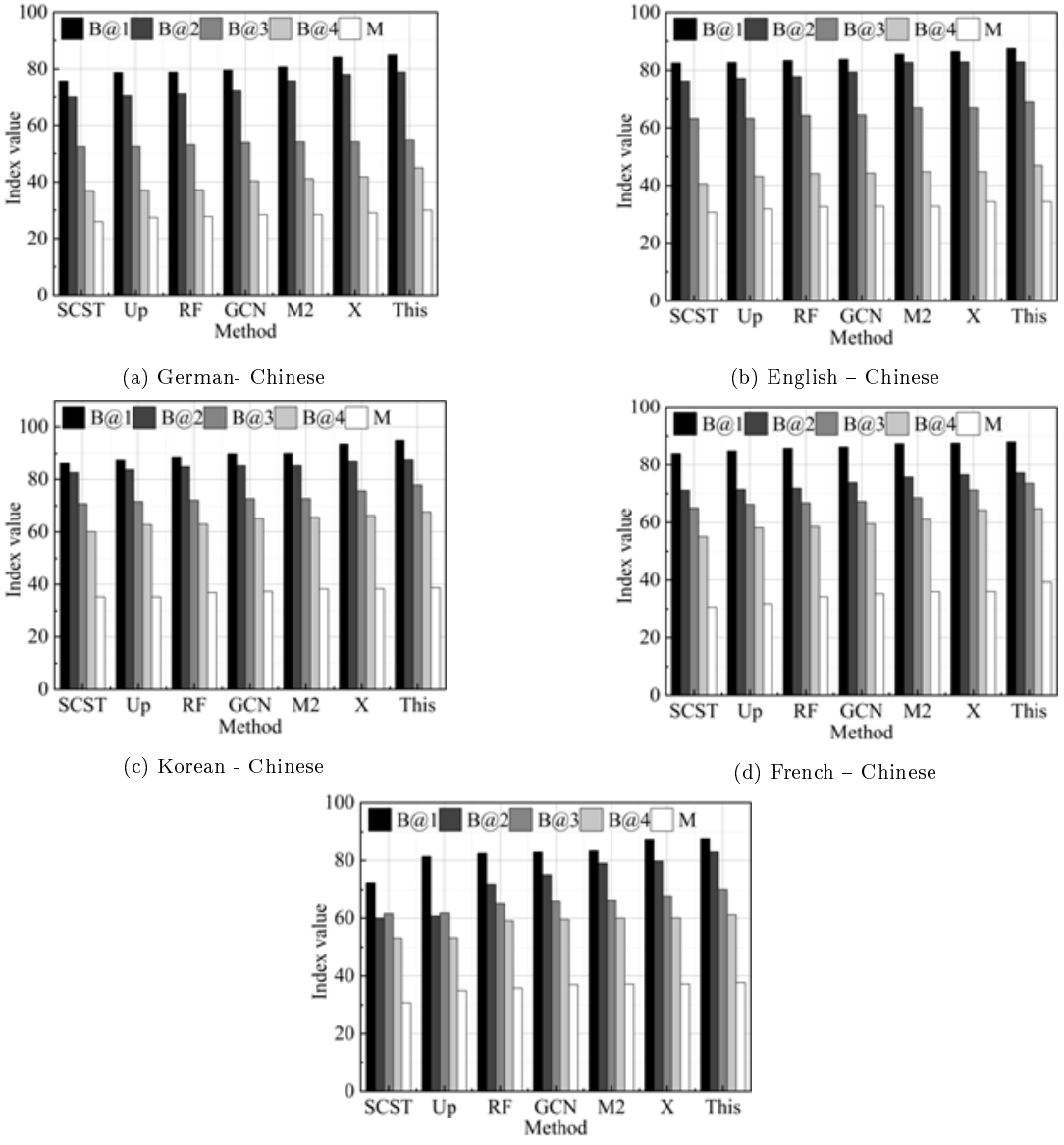
Table 2. Ablation experiment analysis results

M1	M2	WMT 21 Metrics Task				
		B@1	B@2	B@3	B@4	Meteor
×	×	83.26	75.94	62.31	52.49	33.69
√	×	85.69	76.23	64.59	53.29	34.28
√	√	86.54	78.94	66.59	55.67	36.59
M1	M2	WMT 22 Metrics Task				
		B@1	B@2	B@3	B@4	Meteor
×	×	81.26	78.56	62.39	49.52	29.87
√	×	82.63	79.88	64.21	51.23	31.44
√	√	84.3	80.48	64.83	53.33	32.45

3.2. Example analysis of intelligent language translation algorithm

3.2.1. Analysis of the effectiveness of multilingual translation. This paper verifies the application effect of this paper's intelligent language translation model in translating the content containing graphic information through example analysis, and utilizes manual proofreading to examine the translation results obtained by the intelligent language translation model, analyzing them from four aspects, namely, omission, mistranslation, semantics, and sequencing, respectively. In order not to lose the generality, three language translation sentences with large differences in length are chosen as a comparison. In order to more quantitatively analyze the effect of the relative distance method on the translated sentences of different lengths, this paper divides the language translated texts containing graphic information into three categories in total, short, medium and long. Among them, sentences containing less than 15 words are considered short sentences, sentences containing 15 to 25 words

are considered medium sentences, and sentences containing more than 25 words are long sentences. The comparison of the number of short, medium and long sentences in each language category in the graphic content set to be translated is shown in Table 3, with short sentences accounting for the largest proportion (67.66%-79.54%).



(a) German- Chinese

(b) English – Chinese

(c) Korean - Chinese

(d) French – Chinese

(e) Japanese - Chinese

Fig. 2. WMT 21 metrics task data center analysis

Using the intelligent language translation model proposed in this paper to translate various types of language content, the translation results obtained by the model are examined using manual proofreading, and the results obtained are shown in Table 4. It can be seen that the intelligent language translation model established based on computer vision in this paper has an average proportion of 2.64% and 1.30% of mistranslated and omitted content in the actual scenario of multi-language translation, and the translation correctness rate of semantic as well as sequential translations in the translated content reaches 93.59% and 97.49% respectively. From the perspective of translation results in different languages, the intelligent language translation algorithm has the lowest omission rate in Korean-Chinese phrase translation, which is only 0.68%. The correct rate of semantic

translation in French-Chinese language and the correct rate of sequential translation in Japanese-Chinese language are 95.53% and 98.05%, respectively, which have achieved good results in practical application.

Table 3. The number of short and long sentences is compared

Sentence category	Short sentence		Middle sentence		Long sentence	
	Number	Proportion	Number	Proportion	Number	Proportion
German- Chinese	14794	76.01%	4473	22.98%	197	1.01%
English - Chinese	13707	79.54%	3234	18.77%	292	1.69%
Korean - Chinese	11649	75.15%	3639	23.47%	214	1.38%
French - Chinese	14185	75.19%	4413	23.39%	267	1.42%
Japanese - Chinese	9167	67.66%	4160	30.70%	222	1.64%

Table 4. Example translation analysis results

Language	Type	Mistranslation (%)	Omission\newline (%)	Semantic accuracy (%)	Sequential accuracy (%)
German- Chinese	Short sentence	2.49	2.75	93.72	95.08
	Middle sentence	4.06	0.75	93.08	99.27
	Long sentence	3.78	2.72	94.55	99.33
English - Chinese	Short sentence	3.42	0.57	90.34	95.82
	Middle sentence	1.08	1.19	93.69	96.78
	Long sentence	0.93	1.66	97.89	96.05
Korean - Chinese	Short sentence	3.7	0.33	95.42	98.77
	Middle sentence	2.87	1.23	97.08	99.25
	Long sentence	2.37	0.48	90.36	96.13
French - Chinese	Short sentence	4.71	2.78	96.04	97.76
	Middle sentence	1.08	1.98	94.06	96.74
	Long sentence	0.07	0.69	96.48	96.61
Japanese - Chinese	Short sentence	1.94	1.45	90.69	97.37
	Middle sentence	3.57	0.03	91.76	99.67
	Long sentence	3.55	0.93	94.09	97.69
Average		2.64	1.30	93.95	97.49

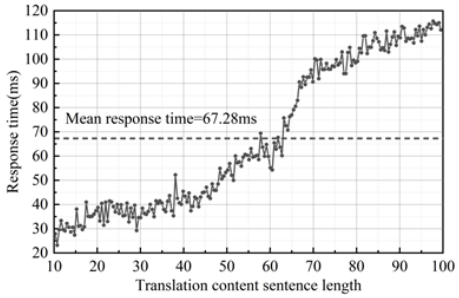


Fig. 3. Algorithm translation response time

3.2.2. Analysis of model translation speed. Finally, this paper analyzes the empirical performance of the intelligent language translation model from the perspective of translation speed. The translation speed performance test of the intelligent language translation algorithm in real scenarios is carried out under a single machine to measure the translation response time of the algorithm under different input contents, and the results of the analysis of the inferred response time of the intelligent language

translation algorithm are shown in Figure 3. The average response time of the intelligent language translation algorithm in real scenario applications stays around 67.28ms. In addition, the overall translation time improves with the increase of the sentence length of the text contained in the graphic information, keeping the response below 69.39ms in the range of normal sentence length (10-60). In summary, the intelligent language translation algorithm based on computer vision is able to meet the performance requirements of users.

4. Conclusion

In this paper, visual feature information is extracted based on convolutional neural network model, and image visual features and textual information features are processed by sequence translation model containing attention mechanism, and the translated language results are output to get the intelligent language translation algorithm based on computer vision. The performance of the algorithm is tested and analyzed, and the application effect of the algorithm is analyzed through empirical applications, and the results show that:

1) In terms of Meteor and BLEU-4 index scores, this paper's algorithm (34.52, 53.76) achieves +1.14% and +1.38% growth compared to the best-performing X-Transformer (34.13, 53.76), which confirms the comprehensive advantages of this paper's model in several key evaluation indexes and highlights its ability to improve the translation language generation performance in terms of quality. Meanwhile, it is found that the visual feature extraction module and the image-text visual feature fusion module proposed in this paper play a key role in improving the performance of the model.

2) In the application of practical scenarios, the intelligent language translation model of this paper has an average ratio of 2.64% and 1.30% of mistranslations and omissions in the actual scenarios of multi-language translation, and the omission rate is the lowest in the translation of Korean-Chinese language, which is only 0.68%, which achieves better results in practical application. In addition, the average response time of the intelligent language translation algorithm in practical scene applications is maintained at about 67.28ms, which can meet the performance requirements.

The intelligent language translation algorithm proposed in this paper has excellent multilingual translation performance, and has great development potential and broad application prospects in the future translation field.

Funding

This research was supported by the Doctoral Research Start-up Fund Project of Jilin Engineering Normal University, under the project titled *"A Study on the Path of College English Curriculum Ideology and Politics"* (Project No. BSSK202301).

References

[1] J. Cho, J. Lei, H. Tan, and M. Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, volume 139, pages 1931–1942. PMLR, 2021.

[2] Y. Ding, Y. Liu, H. Luan, and M. Sun. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, 2017. https://doi.org/10.18653/v1/P17-1106.

[3] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[4] R. Dou, J. Li, X. Wan, H. Chang, H. Zheng, and G. Gao. A decoder structure guided cnn-transformer network for face super-resolution. *IET Computer Vision*, 18(4):473–484, 2024. https://doi.org/10.1049/cvi2.12251.

[5] F. Gao, J. Zhu, L. Wu, Y. Xia, T. Qin, X. Cheng, W. Zhou, and T.-Y. Liu. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, 2019. https://doi.org/10.18653/v1/P19-1555.

[6] J. Guo, H. He, T. He, L. Lausen, M. Li, H. Lin, X. Shi, C. Wang, J. Xie, and S. Zha. Gluoncv and gluonnlp: deep learning in computer vision and natural language processing. *Journal of Machine Learning Research*, 21(23):1–7, 2020.

[7] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[8] K. Kafle and C. Kanan. Visual question answering: datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017. https://doi.org/10.1016/j.cviu.2017.06.005.

[9] A. G. Kakisim and Z. Turgut. Multi-channel convolutional neural network with attention mechanism using dual-band wifi signals for indoor positioning systems in smart buildings. *Internet of Things*, 29:101435, 2025. https://doi.org/10.1016/j.iot.2024.101435.

[10] Y. Li, J. Li, and M. Zhang. Improving neural machine translation with latent features feedback. *Neurocomputing*, 463:368–378, 2021. https://doi.org/10.1016/j.neucom.2021.08.019.

[11] B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: contextualized word vectors. *Advances in Neural Information Processing Systems*, 30:6297–6308, 2017.

[12] L. S. Meetei, T. D. Singh, and S. Bandyopadhyay. Exploiting multiple correlated modalities can enhance low-resource machine translation quality. *Multimedia Tools and Applications*, 83(5):13137–13157, 2024. https://doi.org/10.1007/s11042-023-15721-2.

[13] A. Mogadala, M. Kalimuthu, and D. Klakow. Trends in integration of vision and language research: a survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71:1183–1317, 2021. https://doi.org/10.1613/jair.1.11688.

[14] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4594–4602, 2016.

[15] V. Shelke, R. Dungarwal, V. Makwana, and K. Babariya. Thing translator: an efficient way to classify different things. In *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*, 2021. https://dx.doi.org/10.2139/ssrn.3852083.

[16] S. C. Siu. Revolutionising translation with ai: unravelling neural machine translation and generative pre-trained large language models. In *New advances in Translation Technology: Applications and Pedagogy*, pages 29–54. Springer, 2024. https://doi.org/10.1007/978-981-97-2958-6_3.

[17] S. Suriya, K. Ridhi, S. J. Adwin, S. Sasank, A. Jayabharathi, and G. Gopisankar. Translate and recreate text in an image. In *Intelligent Systems and Applications in Computer Vision*, pages 227–256. CRC Press, 2023.

[18] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, and A. Zadeh. Multimodal research in vision and language: a review of current and emerging trends. *Information Fusion*, 77:149–171, 2022. https://doi.org/10.1016/j.inffus.2021.07.009.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: a neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[20] K. Xu, J. Yao, and L. Yao. Improved convolutional neural network and spectrogram image feature for traffic sound event classification. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 238(13):4230–4244, 2024. https://doi.org/10.1177/09544070231189910.

[21] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902, 2017.

[22] B. Zhang, D. Xiong, J. Su, and J. Luo. Future-aware knowledge distillation for neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2278–2287, 2019. https://doi.org/10.1109/TASLP.2019.2946480.

[23] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5532–5540, 2017.

[24] Y. Zhao, J. Zhang, and C. Zong. Transformer: a general framework from machine translation to others. *Machine Intelligence Research*, 20(4):514–538, 2023. https://doi.org/10.1007/s11633-022-1393-5.

[25] Z. Zhipeng and P. Aleksey. Research on the development of data augmentation techniques in the field of machine translation. *International Journal of Open Information Technologies*, 11(5):33–40, 2023.