IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 6, December, 2023, pp. 2835-2843
Journal homepage: http://iieta.org/journals/ts

Automated Generation of Chinese Text-Image Summaries Using Deep Learning Techniques

Meiling Xu[1,2], Hayati Abd Rahman[1*], Feng Li[1,2]

[1] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam 40450, Malaysia
[2] College of Computer and Information Engineering, Hebei Finance University, Baoding 071051, China

Corresponding Author Email: hayatiar@tmsk.uitm.edu.my

https://doi.org/10.18280/ts.400644

Received: 17 June 2023
Revised: 26 September 2023
Accepted: 12 October 2023
Available online: 30 December 2023

Keywords:
Chinese text-image summaries, automated summary generation, deep learning, MaliGAN, cross-modal similarity retrieval, adaptive fusion strategy

ABSTRACT

In the era of the internet, an abundance of Chinese text-image content is continuously produced, necessitating effective automated technologies for processing and summarizing these materials. Automated generation of Chinese text-image summaries facilitates rapid comprehension of key information, thereby enhancing the efficiency of information consumption. Due to the unique characteristics of the Chinese language, traditional automatic summarization techniques are inadequately transferable, prompting the development of text-image summary generation technologies tailored to Chinese features. Research indicates that while existing natural language processing and deep learning techniques have made strides in text summarization, deficiencies remain in the deep semantic mining and integration of text-image content. This study primarily focuses on two aspects: Firstly, a generative approach based on an enhanced *MaliGAN* model, employing deep learning models to improve text generation quality. Secondly, a retrieval-based approach, utilizing cross-modal similarity retrieval to extract text information most relevant to the image content, guiding summary generation. Additionally, this study innovatively proposes a model architecture comprising segmentation, cross-modal retrieval, and adaptive fusion strategy modules, significantly augmenting the accuracy and reliability of text-image summary generation.

1. INTRODUCTION

With the explosive growth of internet information, the abundance of text-image data has become increasingly rich, leading to more complex and diverse demands for information processing [1-3]. Among numerous information processing tasks, the automated generation of text-image summaries is particularly significant. This technology assists users in rapidly grasping the core essence of text-image content, significantly enhancing the efficiency of information retrieval [4-7]. Especially for Chinese content, due to its unique language structure, relevant technological research and development hold significant practical significance and application value.

Related studies indicate that automatic summary generation technology not only provides users with rapid information condensation but also plays an important role in various fields such as news reporting, literature retrieval, and social media management [8-10]. In the Chinese context, research and implementation of this technology profoundly impact the advancement of Chinese informatization and the enhancement of Chinese text information processing capabilities [11, 12].

However, current research methods in the field of automated Chinese text-image summary generation face multiple challenges [13-16]. Rule-based and template-based methods lack flexibility and struggle to adapt to diverse text types and styles; singular natural language processing techniques find it difficult to accurately grasp the deep semantic connections of text-image content [17, 18]; and early deep learning models also fall short in cross-modal understanding and information fusion, limitations that restrict the quality and applicability of generated summaries [19, 20].

This paper is grounded in deep learning technology, focusing on researching and improving automated Chinese text-image summary generation. It first proposes a generative method, utilizing the enhanced *MaliGAN* text generation model to construct more accurate and fluent summary texts. Secondly, the paper explores a retrieval-based method, selecting text fragments highly relevant to the images through a cross-modal similarity retrieval mechanism to guide the summary generation process. To enhance the model's interaction and information fusion capabilities across different modalities, this study also designs three core sub-modules: segmentation, cross-modal retrieval, and adaptive fusion strategy. Overall, this paper's research not only propels forward the technology of Chinese text-image summary generation but also provides new research ideas and technical frameworks for the field of multi-modal understanding and generation.

2. GENERATIVE APPROACH

The explosive growth of Chinese text-image data on the internet has created an urgent need to rapidly and accurately extract key content from massive information and generate

useful and information-dense summaries. The technology of automated Chinese text-image summary generation emerges as a solution, significantly improving the accessibility and understandability of information, especially for users who need to quickly grasp the essence of information. This paper researches both generative and retrieval-based approaches. By integrating these two methods, the paper aims to combine the advantages of generative and retrieval-based approaches, using deep learning techniques and natural language processing to construct an efficient and accurate automated Chinese text-image summary generation system. This not only enhances the quality of summaries but also expands the application scope of summary generation technology in the field of Chinese information processing.

The generative summary approach aims to create a model capable of autonomously producing new summary texts. Such a method can capture the deep semantic meaning of the text and is more flexible and diverse in generating summaries, producing results akin to human-written summaries. The retrieval-based summary strategy involves finding the most representative fragments from existing texts to compose summaries. The advantage of this method is that it retains the accurate expression of the original text, reducing the potential distortion of information in generated summaries. When dealing with Chinese text-image information, considering the complexity of Chinese and the relationship between text and image, the retrieval-based approach ensures the precision and completeness of summary content.

Given the high complexity of Chinese grammatical structure and semantic expression, and the necessity to effectively fuse visual and textual information in text-image summary generation, this paper constructs an *LFMGAN* model architecture with a generator and a discriminator for adversarial training. The generator learns the mapping relationship between image features and text features, while the discriminator evaluates whether the generated summary accurately reflects the image content, thereby achieving effective cross-modal fusion. Figure 1 presents a schematic diagram of the *LFMGAN* model architecture.

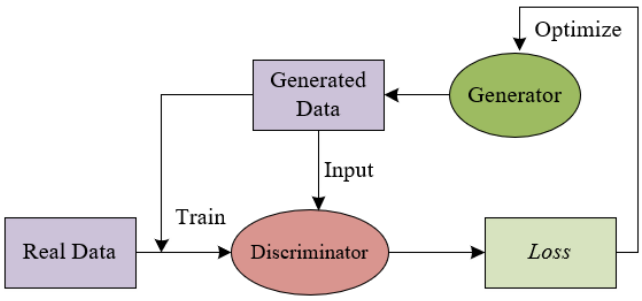


Figure 1. Architecture of the *LFMGAN* model

2.1 Model generator

In the *LFMGAN* model for the automated generation of Chinese text-image summaries based on generative methods, the design of the generator is a core component. It comprises three main layers: the *Embedding* layer, the *GPT*-2 network layer, and the *softmax* layer. The *Embedding* layer primarily functions to convert the model's discrete textual identifiers into continuous vector representations, capturing the semantic and syntactic relationships of words. The *GPT*-2 network layer acts as the heart of deep semantic processing and structural generation. The *softmax* layer, positioned at the end of the network, transforms the output vectors from the *GPT*-2 network into a probability distribution, representing the likelihood of each word in the vocabulary being the next word.

Due to the *Transformer* model not inherently processing sequence information as naturally as recurrent neural networks, positional encoding is added to the input word vectors to provide position information for each word in the sequence. After positional encoding, the word embeddings are projected into a more appropriate representation space, obtaining the corresponding Q, K, and V vectors for each word vector. To determine the importance of each word in generating attention, the scores for each word vector are calculated through the product of the word vector and a weight matrix. The formula is:

$$SC = W^{Y} J \tag{1}$$

Using the normalized scores and corresponding word vectors, a weighted average is calculated, which serves as the attention output, represented by the following formula:

$$ATT(W, J, C) = \text{Softmax}\left(\frac{W^{Y} J}{\sqrt{f}}\right) C \tag{2}$$

To reduce internal covariate shift and accelerate training speed, each neuron's activation output on every mini-batch is standardized, involving subtraction of the mean and division by the standard deviation, followed by a learnable scaling and shifting transformation. Suppose the unnormalized word vector is represented by t_b, the mean of word vectors by ω_N, the variance of word vectors by δ^2_N, and the batch-normalized word vector for the b-th word by $\hat{t}_b$, minor numbers are represented by γ and α, then the process is:

$$\hat{t}_b = \frac{t_b - \omega_N}{\sqrt{\delta_N^2 + \gamma}} + \alpha \tag{3}$$

The data involved in automated Chinese text-image summary generation can be highly diverse and contain substantial noise and irregularities. In such tasks, models are prone to overfitting on training data, failing to generalize well to new samples. Therefore, to prevent over-reliance on any single feature during training, this study incorporates a *dropout* layer, thereby enhancing the model's generalization capability. Moreover, when processing Chinese text-image summaries, deep network structures may be required to capture complex linguistic features and text-image information. By integrating *LayerNormalization*, training is stabilized, convergence speed is accelerated, and model performance is improved. Unlike *BatchNormalization*, *LayerNormalization* is independent of other samples in the batch, maintaining performance even on small batch sizes, which is particularly crucial for text-image summary tasks that may operate on batches of varying sizes.

2.2 Model discriminator

The complexity of Chinese text is high, especially in text-image summaries, where an understanding of not only textual information but also text within images is required. In the *LFMGAN* model for automated Chinese text-image summary generation, the *RoBERTa* model is chosen for the

discriminator. Optimized *RoBERTa* exhibits enhanced contextual understanding capabilities, enabling more accurate judgments of the authenticity of generated summaries in complex Chinese contexts.

After feature extraction, data is typically transformed into an embedding space, a low-dimensional space capable of representing data's dense features. In text processing, this means mapping words, sentences, or paragraphs into a vector space. Here, the *RoBERTa* model can act as an embedding layer, converting high-dimensional textual data into a form more amenable to machine processing. Following embedding representation, the discriminator classifies generated and real text-image summaries, discerning whether they are authentic or produced by the generator. The *RoBERTa* model outputs a probability value indicating whether the input data is real or fake. In the *GAN* framework, the discriminator's output is often used as a loss function to guide the training of the generator.

Automated Chinese text-image summary generation requires not only grammatically correct text but also semantically accurate and rich summaries. To help the model learn complex data distributions more effectively and focus the generator on producing text semantically close to real summaries, this study constructs a *Loss* value based on *MaliGAN*'s *Reward* and the semantic similarity between generated and real text to guide the generator's optimization. The calculation formula is as follows:

$$LOSS = s \cdot RW1 + (1-s) \cdot RW2 \tag{4}$$

Suppose *MaliGAN*'s *Reward* is represented by $RW1$, the generator's distribution by $o_h(z)$, and the real distribution by $o_{DA}(z)$, then the expression is:

$$o_{DA}(z) = o_h(z) \frac{F_H^*(z)}{1-F_H^*(z)} \tag{5}$$

$RW2$ is determined by the semantic similarity *COS.SIM* between generated and real text, expressed as:

$$RW2 = 1 - COS.SIM \tag{6}$$

COS.SIM is calculated using *Sentence-BERT*. *Sentence-BERT* first transforms the input Chinese sentences into fixed-size vector representations. Sentences are passed through a pre-trained *BERT* model, and a fixed-length sentence vector is obtained through pooling operations. This sentence vector represents the overall semantics of the sentence. Once sentence vectors for generated and real texts are computed, a similarity measure, often cosine similarity, is used to calculate the similarity between these two vectors. The cosine similarity measures the directional similarity of two vectors, unaffected by their length. Its expression is:

$$COS(i,c) = \frac{i \cdot c}{|i||c|} \tag{7}$$

For generated text-image summaries, lower *loss* values typically indicate that the generated summaries are closer to real summaries in semantics and structure. High *loss* values suggest significant differences between generated and real summaries, possibly in grammar, semantics, or relevance. The generator iteratively optimizes based on the discriminator's *loss* values. An optimized generator should produce summaries with lower *loss* values, i.e., summaries that are more "real" under the discriminator's evaluation criteria. This process continues until a preset performance standard is reached or there is no significant decrease in *loss* values.

3. RETRIEVAL-BASED METHOD

Figure 2 illustrates the principle framework of the retrieval-based method for the automated generation of Chinese text-image summaries. In the context of automated Chinese text-image summary generation, the retrieval-based model constructed includes segmentation, cross-modal retrieval, and adaptive fusion strategy modules. The segmentation module is designed to separate the image from the corresponding textual content within the text-image material. This separation allows for the individual processing of each modality, facilitating more precise feature extraction. The cross-modal retrieval module aims to connect the semantic spaces of images and texts, enabling the model to retrieve text content that is similar or relevant to a given image input. The purpose of the adaptive fusion strategy module is to integrate information from different modalities to generate the final summary. Through such a modular design, the model ensures that the text in the summary is highly relevant to the image content. By combining segmentation, retrieval, and fusion strategies, the generated summaries are made to be accurate, comprehensive, and well-formatted. Given a Chinese text-image description dataset represented by $Z=\{z^o_u, t_u, x_u\}^B_{u=1}$, the grayscale value of the Chinese image is represented by z^o_u, the analytical text corresponding to the image by t_u, and the diagnostic result corresponding to the image by x_u. The size of the dataset is denoted by B.

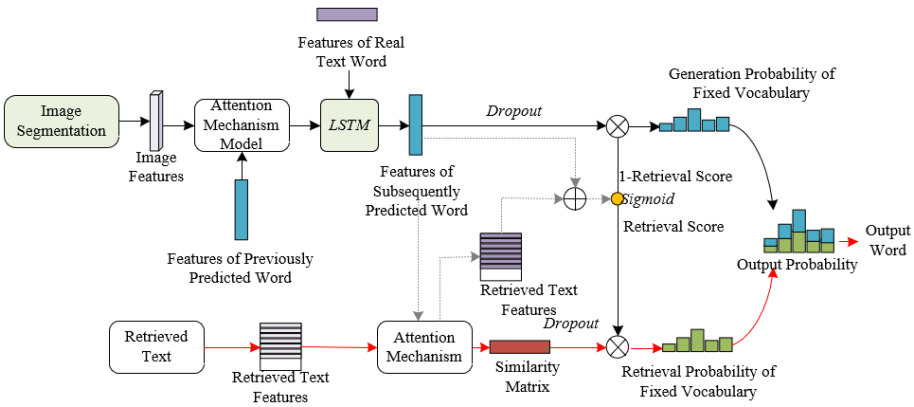


Figure 2. Principle framework of the retrieval-based method for automated generation of Chinese text-image summaries

This paper ensures that the generation process is guided by the semantic features of textual content by applying the image and text features from the intermediate layers of the retrieval model to the generative model. This approach enhances the semantic relevance and accuracy of the generated summaries, strengthens alignment between different modalities, reduces modal discrepancies between images and texts, and improves the overall quality of the final summary. This method promotes more effective cross-modal operation of the model and is a valuable supplement to existing summary generation technologies. Specifically, let the u-th Chinese image be represented by z^o_u, the intermediate layer feature by c_u, and the generation probability of each vocabulary in the vocabulary by h^m_{uj}. After preprocessing, z_u is fed into a pre-trained model. On one hand, c_u is extracted as input for the generative model, then fed into an Long Short Term Memory (*LSTM*) decoder to generate h^m_{uj}.

Direct generation might lead to content repetition or insufficient relevance to the image, while a purely retrieval-based model might fail to cover all details of the image or capture all important information. To ensure that the final summary contains precise information relevant to the image while maintaining the fluency and coherence of the text, this paper proposes an adaptive fusion strategy. This strategy calculates retrieval scores using the similarity between the retrieval features and fixed vocabulary generated, using this score as a retrieval weight for calculating the weighted sum of the retrieval and generation probabilities of fixed vocabulary. Specifically, let the retrieval probability of the vocabulary corresponding to the retrieval feature be represented by c^m_{uj}, and the generation probability by h^m_{uj}. Finally, c^m_{uj} and h^m_{uj} are combined through an adaptive fusion strategy, serving as the output probability of the vocabulary.

3.1 Image segmentation

In the model for the automated generation of Chinese text-image summaries, the task of the segmentation module is to separate relevant visual content from Chinese images. This step is foundational for subsequent text-image fusion and summary generation. The module takes into account variations in image brightness, structural changes under scaling, and linear relationships after scaling. This allows for the accurate extraction of text or other visual elements requiring summarization from complex image backgrounds. Let the grayscale value of the Chinese image obtained from each *DICOM* file be represented by z^o_u, the rescaling slope by z^a_u, and the rescaling intercept by z^y_u, then the calculation formula is:

$$z^v_u = z^o_u * z^a_u + z^y_u \tag{8}$$

The preprocessed Chinese image $z_u \in E^{V\times A\times Q\times G}$ is then fed into a pre-trained cross-modal retrieval model.

3.2 Cross-modal retrieval

In the context of the automated Chinese text-image summary generation model, the purpose of the cross-modal retrieval module is to establish effective connections between images and text. This enables the model to retrieve relevant text information based on image content, or vice versa. To achieve this, the module needs to address two main issues: one is understanding and expressing the relevance between text and image, and the other is ensuring the discriminability of information within a single modality. Therefore, a pre-trained 3D convolutional neural network specifically designed for Chinese images is used for the image modality. For the text modality, a pre-trained natural language processing model customized for the Chinese text-image domain is employed, capable of understanding and extracting deep information from Chinese text. Figure 3 shows the framework of the cross-modal retrieval module.

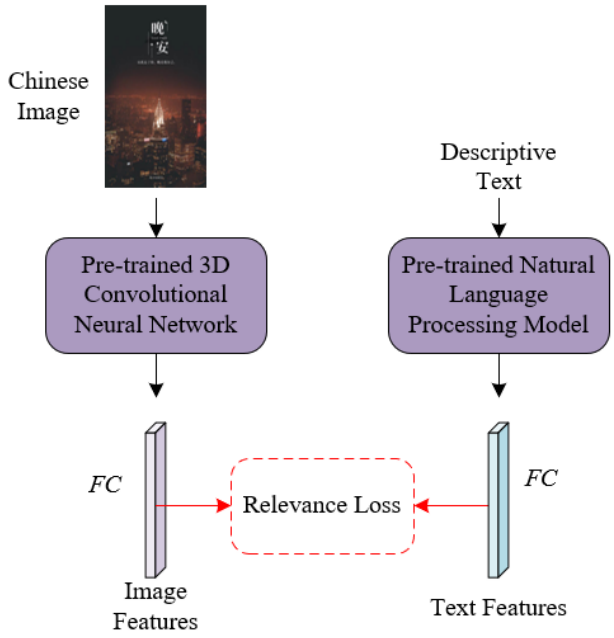


Figure 3. Framework of the cross-modal retrieval module

Specifically, for the image modality, the preprocessed $z_u \in E^{V\times A\times Q\times G}$ is fed into a pre-trained 3D model, retaining the last convolutional layer's features as the image feature z'_u. The image features directly output from the convolutional layer might have a high dimension, while text features may come from a different dimensional space. A fully connected layer can transform image features into dimensions matching the text features, facilitating comparison and fusion in subsequent steps. Suppose the activation function is represented by *GELU* and the fully connected layer functions by d_1 and d_2, as follows:

$$c_u = d_2\left(GELU\left(d_1\left(z'_u\right)\right)\right) \tag{9}$$

Similarly, the same operation is performed for the text modality. Suppose the common text representation is represented by i_u, and the fully connected layer functions by h_1 and h_2, as follows:

$$i_u = h_2\left(GELU\left(h_1\left(y_u\right)\right)\right) \tag{10}$$

After feature mapping is completed, it is necessary to ensure effective alignment of image and text features in a common space. This typically involves similarity measurement and optimization algorithms, so that the feature vectors of related text-image pairs are closer in space, while unrelated ones are farther apart. To ensure relevance, the model needs to implement certain constraints to ensure high correlation when matching text and image, thereby reducing modal heterogeneity. This study constructs the following cross-

modal retrieval relevance loss function. Suppose the common image is represented by c_u, the common text by i_u, the image feature vector by $c'_u \in E^f$, and the text feature vector by $i'_u \in E^f$. The feature center of the corresponding class x' for the image is represented by $x'_u \in E^f$, and the expression for the loss function is as follows:

$$M_1 = -\frac{1}{B}\sum_u^B \log\left(\frac{\exp\left(c'_u i'^{Y}_u\right)}{1+\sum_u^B \exp\left(c'_u i'^{Y}_u\right)}\right) + \frac{1}{B}\sum_u^B \left\|c'_u - x'_u\right\|_2^2 + \frac{1}{B}\sum_u^B \left\|i'_u - x'_u\right\|_2^2 \tag{11}$$

3.3 Adaptive fusion of modal information

In the task of automated generation of Chinese text-image summaries, the text generated by the decoder predominantly relies on image features and may lack specific domain expertise or contextual information. In contrast, the text retrieved typically originates from actual datasets, containing rich professional knowledge and real context. The fusion of these two types of text can enhance the semantic coherence and accuracy of the summary.

An adaptive fusion strategy is proposed in this study, wherein the combination of the attention mechanism decoder and the adaptive fusion module effectively utilizes visual information from images and textual information from retrieved texts. The attention mechanism dynamically focuses on different parts of the image to generate related text, while the adaptive fusion module intelligently decides the weight distribution between text generated by the decoder and retrieved text based on the context. Figure 4 illustrates the framework of the employed visual attention module.

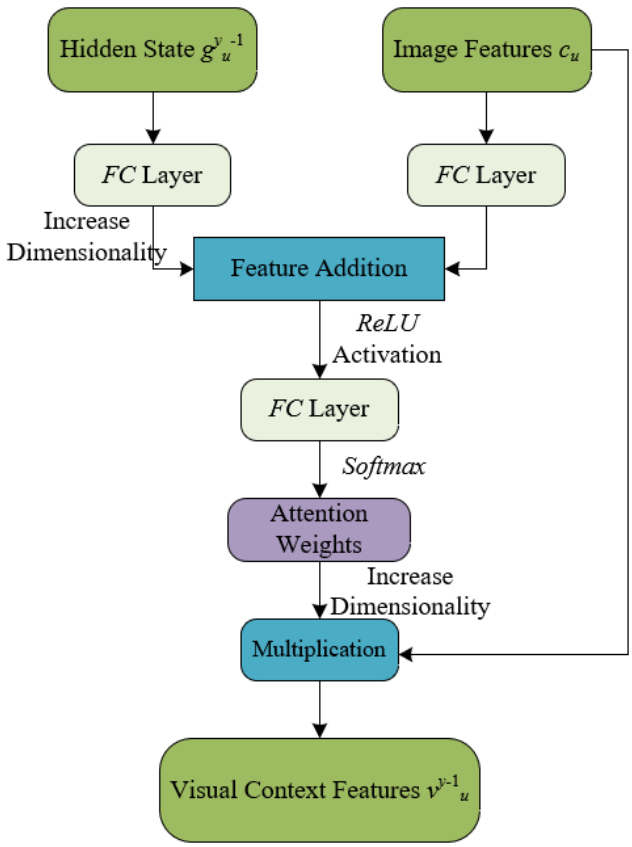


Figure 4. Framework of the visual attention module

Initially, visual features are extracted from the image, accomplished through a pre-trained convolutional neural network (*CNN*). The *CNN* captures both local and global features of the image, serving as visual context to support the text generation process. Assuming the previous hidden state is represented by g^{y-1}_u, the image feature by c_u, the visual context feature by v^{y-1}_u, and the standard attention mechanism module with parameters ϕ_c represented by $ATT(Q,K,V;\phi)$, the calculation formula is:

$$v_u^{y-1} = ATT\left(g_u^{y-1}, c_u, c_u; \varphi_c\right) \tag{12}$$

Subsequently, the *LSTM*, through its recurrent structure, remembers the composite representation of previously generated words and visual features, thus considering previous textual information and visual context in generating each new word. Assuming the embedding feature of the previous generated text word is represented by r^{y-1}_u, the corresponding visual context feature by v^{y-1}_u, the current hidden state by g^{y}_u, and the parameters of *LSTM* by γ, the calculation formula is:

$$g_u^y = LSTM\left(\left[r_u^{y-1}, v_u^{y-1}\right], g_u^{y-1}; \gamma\right) \tag{13}$$

At each time step, the *LSTM* predicts the next most likely word based on the aggregated features. Assuming learnable parameters are represented by Q_β and n_β, the formula for generating the current text word is:

$$o_u^u = Q_\beta g_u^y + n_\beta \tag{14}$$

At each decoding step y, the pre-trained cross-modal retrieval model is utilized to encode the retrieved relevant text, extracting its semantic features i_u. For the retrieved text, it is also necessary to compute the probability distribution of the next word. This typically involves transforming the semantic features of the retrieved text through one or more fully connected layers and a *softmax* layer into a probability distribution, reflecting the likelihood of each potential word being the next. Besides calculating the probability of individual words, the overall context of the retrieved text must also be considered. This may involve using an attention mechanism or other context-encoding strategies to utilize the global features of the retrieved text e^y_u in generating the summary. Assuming the parameters of the standard attention mechanism module are represented by ϕ_i, the specifics are as follows:

$$a_u^y, e_u^y = ATT\left(g_u^y, i_u, i_u; \varphi_i\right) \tag{15}$$

Finally, the word probability distribution retrieved is integrated with the word probability distribution generated through *LSTM*. The adaptive fusion module dynamically adjusts the weights of these two probability distributions based on the current context and strategy, ensuring that the generated word reflects both the context directly derived from the image and the integrated information from the retrieved text. Assuming the retrieved word probability distribution is represented by a^y_u and the learnable parameters by Q_σ and n_σ, the word probability distribution mapped to the vocabulary space is:

$$w_u^y = Q_\sigma a_u^y + n_\sigma \tag{16}$$

Assuming learnable parameters are represented by Q_α, n_α, Q_ε, and n_ε, and the activation function by *Sigmoid*, the information retrieval score available in the retrieved text can be calculated as follows:

$$\partial_u^y = \text{Sigmoid}\left(\left(Q_\alpha e_u^y + n_\alpha\right) + \left(Q_\varepsilon g_u^y + n_\varepsilon\right)\right) \tag{17}$$

The probability distribution of the retrieval-based generated text word in the vocabulary space at the current time step y can be obtained through the following calculation:

$$p_u^y = \left(1 - \partial_u^y\right) o_u^y + \partial_u^y w_u^y \tag{18}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an index used to assess the quality of machine translation. It evaluates translation accuracy by considering the matching degree of synonyms and sentence structure. In the task of automated text summarization, *METEOR* can be used to measure the quality of generated summaries, including the accuracy of content and the fluency of language. From Figure 5, it is observed that the *METEOR* scores of the model proposed in this study are compared with three other models (*BERT*, *Seq2Seq*, *AB-LSTM*) across three different datasets.

On Dataset 3 (Huawei Noah Open Dataset), the *METEOR* score of the proposed model is 0.351, showing a significant improvement over the *BERT* model (0.255), *Seq2Seq* model (0.314), and *AB-LSTM* model (0.224). On Dataset 2 (*QBSUM*), the *METEOR* score of the proposed model is 0.204, also surpassing the scores of the other three models (*BERT* at 0.166, *Seq2Seq* at 0.191, *AB-LSTM* at 0.158). On Dataset 1 (*NLPCC* 2017 *Shared Task*), the proposed model continues to lead with a score of 0.22, compared to the *BERT* model (0.179), *Seq2Seq* model (0.204), and *AB-LSTM* model (0.174). Based on the comparison of *METEOR* scores, it can be concluded that the proposed model, based on the improved *MaliGAN* text generation approach, outperforms the other generative models in all three datasets. These results indicate that the proposed method not only improves the quality of generated summaries but also demonstrates broad applicability and effectiveness across different types of datasets.

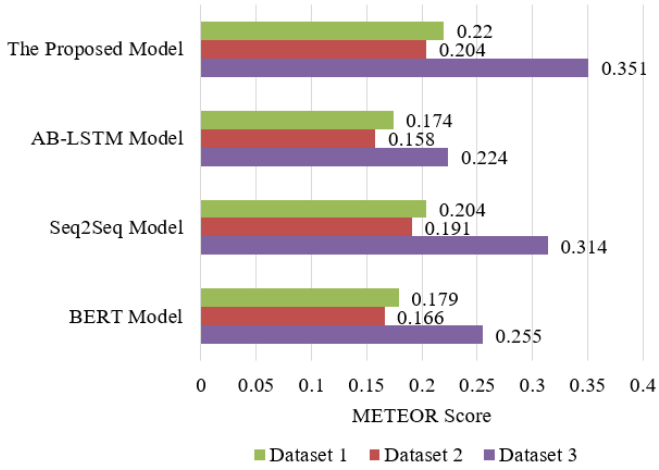


Figure 5. Variations in *METEOR* metrics across different datasets for various generative methods

Table 1. Ablation experiment results of the retrieval-based Chinese text-image summary automatic generation method

Method	*BLUE-1*	*BLUE--2*	*BLUE--3*	*BLUE--4*	*METEOR*	*ROUGE_L*	*CIDEr*
Without Retrieval Text	61.2	51.7	45.9	42.1	32.1	55.8	63.1
Without Pretrained Image Features in Retrieval Model	59.8	51.6	45.2	38.9	28.9	28.9	61.4
Without Adaptive Fusion Strategy	58.8	53.1	47.8	43.2	32.5	32.1	62.8
Final Model of This Study	61.2	55.3	49.8	46.1	31.4	58.9	78.9

Table 2. Computational costs of different Chinese text-image summary automatic generation methods

Model Type	Method	*FLOPs*	*Parameters*
Generative Method	*CNN+LSTM*	523.2546	15.2347
	Generative Model propose in this study	345.2368	111.2536
Retrieval-Based Method	*Text-Summarizer-Pytorch-Chinese*	412.2358	101.2458
	Retrieval-Based Model propose in this study	478.3694	102.3651

Ablation study is an experimental method for evaluating the importance of each component of a model, by observing the change in model performance when a component is removed. Table 1 shows the results of the ablation experiment on the retrieval-based Chinese text-image summary automatic generation method. By comparing the performance of each variant with the final model, the contribution of different components can be assessed. The analysis is based on several commonly used evaluation metrics in the table: *BLEU* (a series of metrics for assessing translation quality, with *BLEU*-1 to *BLEU*-4 corresponding to different *n-gram* lengths), *METEOR*, *ROUGE_L* (considering the longest common subsequence in the summary), and *CIDEr* (a metric designed for image description, emphasizing the importance of common *n-grams*). The final model of this study outperforms the other variants in almost all evaluation metrics, especially showing a significant improvement in the *CIDEr* metric, indicating that the model integrating retrieved text, pretrained image features, and adaptive fusion strategy can generate more accurate and relevant summaries.

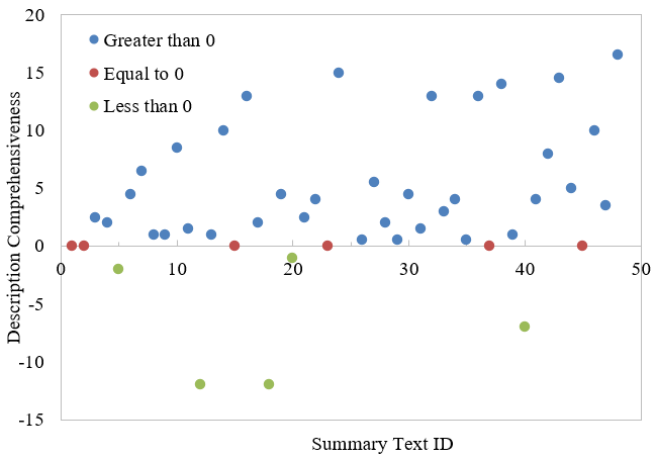
Table 2 lists the computational costs of four different Chinese text-image summary automatic generation methods, with specific metrics including the number of floating-point operations (*FLOPs*, i.e., the number of floating-point operations required for each inference) and the number of model parameters (in millions). *FLOPs* is an indicator of model computational complexity, while the number of parameters relates to the model's storage needs and potential overfitting risk. As per the table, the generative model proposed in this study significantly reduces computational

costs in terms of *FLOPs*, making it more suitable for scenarios with limited computational resources. However, this efficiency comes at the cost of an increased number of model parameters, which may require more storage space and could increase the risk of overfitting, although more parameters usually provide stronger learning capacity. In retrieval-based methods, despite higher *FLOPs* for the model proposed in this study, the slight increase in the number of parameters indicates that storage requirements have not significantly increased, and the high *FLOPs* may be due to complex computations to improve retrieval effectiveness and summary quality.

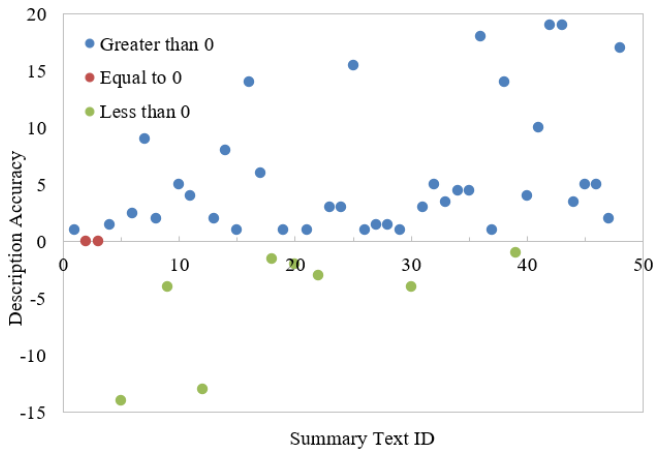
Table 3. Performance comparison of different Chinese text-image summary automatic generation methods

Model Type	Method	*BLUE*-1	*BLUE*--2	*BLUE*--3	*BLUE*--4	*METEOR*	*ROUGE_L*	*CIDEr*
Generative Method	*BERT*	22.3	11.9	8.6	6.7	-	31.2	28.9
	Seq2Seq	21.5	12.3	8.8	6.9	-	31.8	28.7
	Attention-based LSTM	27.8	15.6	11.2	7.3	-	23.3	26.4
	CNN+LSTM	46.3	27.8	21.4	14.9	-	35.8	26.3
	The proposed model	46.2	32.1	23.5	15.8	-	34.9	26.1
Retrieval-Based Method	*RAG*	31.2	22.3	15.9	13.2	14.8	23.4	13.1
	FiD	44.8	28.5	21.4	14.5	-	31.8	33.8
	Ret-Gen	47.9	31.7	23.7	16.1	-	32.6	27.8
	Text-Summarizer-Pytorch-Chinese	48.5	33.6	22.5	15.8	-	37.8	33.5
	The proposed model	51.3	38.9	31.5	28.3	24.3	39.2	34.5

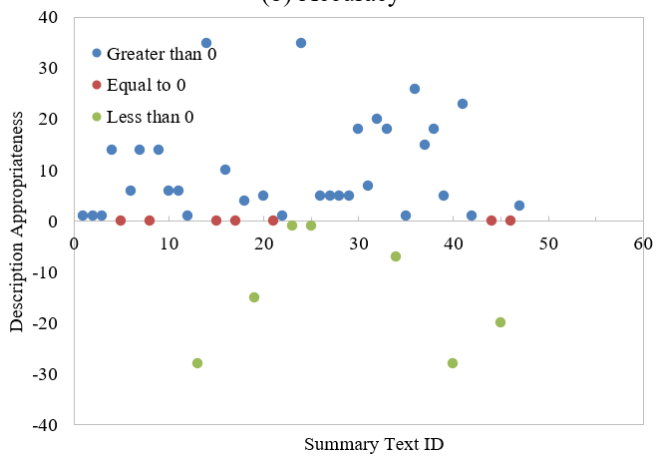
Table 3 displays the performance of different Chinese text-image summary automatic generation methods across multiple performance metrics, including *BLEU*-1 to *BLEU*-4, *METEOR*, *ROUGE_L*, and *CIDEr*. It is evident that the generative model proposed in this study is comparable to *CNN+LSTM* in all *BLEU* metrics and surpasses it from *BLEU*-2 to *BLEU*-4, particularly in longer *n-gram* matches (such as *BLEU*-4), suggesting its generated summaries may be more coherent and complete. However, its score in *CIDEr* is slightly lower, which might be an aspect for further optimization. The *Ret-Gen* and *Text-Summarizer-Pytorch-Chinese* models show strong performance across all metrics, but the retrieval model of this study achieves higher scores in all metrics, especially in *CIDEr*, indicating its ability to generate more relevant and information-rich summaries. Overall, the model proposed in this study demonstrates superior performance across multiple key performance metrics, particularly in the retrieval-based method, where its scores in *BLEU*-1 to *BLEU*-4, *METEOR*, *ROUGE_L*, and *CIDEr* are the highest, showcasing its excellent overall performance. These results emphasize the effectiveness of the proposed method in the field of Chinese text-image summary automatic generation, especially in presenting content highly relevant to images. Moreover, while maintaining high-quality summary outputs, the proposed method also provides more coherent, complete, and information-rich text, which is highly valuable for practical applications.



(a) Comprehensiveness



(b) Accuracy



(c) Appropriateness

Figure 6. Score differences of the proposed method in various evaluation criteria

Figure 6 displays the scoring of the proposed Chinese text-image summary automatic generation method in various evaluation criteria. The distribution of blue points (representing samples with scores greater than zero) indicates positive evaluation results. The concentration and location of these blue points illustrate the method's effectiveness in terms of description comprehensiveness. Red points (equal to zero) and green points (less than zero) represent neutral or negative evaluations. If these points are relatively few in number and

mainly clustered around zero, it might suggest that most summaries receive positive evaluations. By comparing the number and distribution of blue points against red and green ones, and seeing that blue points in all three graphs are significantly more numerous and distributed in higher score areas, it can be concluded that the method performs well in terms of description comprehensiveness, accuracy, and appropriateness.

Table 4. *BLEU* metric results on different datasets

Dataset	*Model*	*BLEU*-1	*BLEU*-2	*BLEU*-3	*BLEU*-4
NLPCC 2017 Shared Task	*CNN+LSTM*	0.897	0.865	0.714	0.693
	Proposed Generative Model	0.932	0.889	0.824	0.731
	Text-Summarizer-Pytorch-Chinese	0.914	0.885	0.789	0.715
	Proposed Retrieval-Based Model	0.962	0.923	0.832	0.77
QBSUM	*CNN+LSTM*	0.914	0.895	0.825	0.798
	Proposed Generative Model	0.956	0.923	0.895	0.812
	Text-Summarizer-Pytorch-Chinese	0.937	0.914	0.87	0.784
	Proposed Retrieval-Based Model	0.985	0.963	0.923	0.825
Huawei Noah Open Dataset	*CNN+LSTM*	0.693	0.589	0.285	0.223
	Proposed Generative Model	0.734	0.623	0.374	0.235
	Text-Summarizer-Pytorch-Chinese	0.715	0.589	0.326	0.236
	Proposed Retrieval-Based Model	0.795	0.634	0.389	0.268

Table 4 shows the *BLEU* metric results of four different models on three distinct Chinese datasets. On the *NLPCC* 2017 *Shared Task* dataset, the proposed retrieval-based model scored the highest in all *BLEU* metrics, particularly in *BLEU*-4 with a score of 0.77, significantly outperforming other models. This indicates that, for this dataset, the proposed retrieval-based model produces the most accurate and coherent summaries. The proposed generative model also outperforms the *CNN+LSTM* and *Text-Summarizer-Pytorch-Chinese* models, showing its improvement in generating accurate text. On the *QBSUM* dataset, similarly, the proposed retrieval-based model leads in all *BLEU* metrics, consistent with the results of the *NLPCC* 2017 *Shared Task* dataset, demonstrating its stability and superiority. The proposed generative model also performs well, scoring 0.812 in *BLEU*-4, surpassing the *Text-Summarizer-Pytorch-Chinese* model. On the Huawei Noah Open Dataset, although all model scores are lower, likely due to the increased difficulty of the dataset, the proposed retrieval-based model still maintains the highest scores in all *BLEU* metrics, especially noticeable in *BLEU*-1 and *BLEU*-2, indicating that the proposed model captures key information more effectively in this dataset. The proposed generative model also shows better performance across all *BLEU* metrics compared to the other non-proposed models.

In summary, whether on any dataset, the proposed retrieval-based model consistently outperforms other models, demonstrating significant advantages in *BLEU* metrics, which proves the model's robust effectiveness in automatic generation of Chinese text-image summaries. Additionally, the proposed generative model also shows better performance than traditional *CNN+LSTM* and *Text-Summarizer-Pytorch-Chinese* models, further emphasizing the superiority of the proposed method. These results indicate that the proposed method can generate high-quality summaries that more accurately match the original text-image content.

5. CONCLUSION

This paper's research is centered on the automatic generation of Chinese text-image summaries, employing deep learning technology as the foundational basis. Two primary methods are proposed: a generative approach and a retrieval-based approach, with the aim of generating more accurate and fluent summary texts while ensuring high relevance of the text content to the images. Initially, a generative method based on the improved *MaliGAN* model is introduced. This method, by optimizing the text generation model, seeks to enhance the accuracy and fluency of the summaries. In various experiments, the generative model developed in this study outperformed the traditional *CNN+LSTM* model in *BLEU* evaluation metrics, demonstrating its effectiveness in generating accurate and coherent texts. The paper further explores a retrieval-based method, utilizing a cross-modal similarity retrieval mechanism to guide summary generation from an image perspective, ensuring high relevance of the summaries to the image content. Experimental results show that the retrieval-based model performed the best in *BLEU* metrics across multiple datasets, indicating its effectiveness in extracting text segments relevant to image content.

The research findings of this paper prove that through deep learning technology, particularly the improved generative model and cross-modal similarity retrieval mechanism, the quality of automatic Chinese text-image summary generation can be significantly enhanced. Experimental results on multiple Chinese datasets demonstrate that both the generative and retrieval-based models proposed in this study surpassed existing techniques in *BLEU* metrics, especially in higher-order *BLEU* metrics, highlighting the coherence and completeness of the summaries generated by this method. The design of core sub-modules is another highlight of this method, enhancing the model's interaction and information fusion capabilities across different modalities, which is crucial in cross-modal tasks.

In conclusion, the methods proposed in this paper are not only effective in generating high-quality Chinese summaries closely related to image content but also demonstrate good stability and generalization capabilities. These research outcomes provide strong technical support for the development of automatic Chinese text-image summary generation technology and have significant implications and impact on research and applications in related fields.

REFERENCES

[1] Kasi, G., Abirami, S., Lakshmi, R.D. (2023). A deep learning based cross model text to image generation

using DC-GAN. In 2023 12th International Conference on Advanced Computing (ICoAC), Chennai, India, pp. 1-6. https://doi.org/10.1109/ICoAC59537.2023.10250086

[2] Bajić, F., Job, J. (2023). Review of chart image detection and classification. International Journal on Document Analysis and Recognition, 26(4): 453-474. https://doi.org/10.1007/s10032-022-00424-5

[3] Zhu, L., Sheng, X. (2022). On image-processing-based identification method of express logistics information. Traitement du Signal, 39(3): 1019-1025. https://doi.org/10.18280/ts.390329

[4] Lee, H., Ullah, U., Lee, J.S., Jeong, B., Choi, H.C. (2021). A brief survey of text driven image generation and maniulation. In 2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Gangwon, Korea, pp. 1-4. https://doi.org/10.1109/ICCE-Asia53811.2021.9641929

[5] Phueaksri, I., Kastner, M.A., Kawanishi, Y., Komamizu, T., Ide, I. (2023). An Approach to Generate a Caption for an Image Collection Using Scene Graph Generation. IEEE Access, 11: 128245-128260. https://doi.org/10.1109/ACCESS.2023.3332098

[6] Mahalakshmi, P., Fatima, N. S. (2022). Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques. IEEE Access, 10: 18289-18297. https://doi.org/10.1109/ACCESS.2022.3150414

[7] Xie, F., Chen, J., Chen, K. (2023). Extractabltive text-image summarization with relation-enhanced graph attention network. Journal of Intelligent Information Systems, 61(2): 325-341. https://doi.org/10.1007/s10844-022-00757-x

[8] Huang, W., Bu, X., Xiao, Y., Wen, Y., Deng, Z. (2022). Research on Chinese summary generation based on pointer key information. In 5th International Conference on Computer Information Science and Application Technology (CISAT 2022), Chongqing, China, pp. 578-586. https://doi.org/10.1117/12.2656552

[9] Xu, C., Liu, D. (2018). Chinese text summarization algorithm based on Word2vec. Journal of Physics: Conference Series, 976(1).

[10] Patel, I., Julka, V., Kudtarkar, P., Sharma, I., Sonawane, B. (2023). Online meeting summarization based on text and image processing. In International Conference on ICT for Sustainable Development, Goa, India, pp. 193-205. https://doi.org/10.1007/978-981-99-5652-4_19

[11] Lee, J., Herskovitz, J., Peng, Y.H., Guo, A. (2022). ImageExplorer: Multi-layered touch exploration to encourage skepticism towards imperfect AI-generated image captions. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, pp. 1-15. https://doi.org/10.1145/3491102.3501966

[12] Zhou, R., Jiang, C., Xu, Q. (2021). A survey on generative adversarial network-based text-to-image synthesis. Neurocomputing, 451: 316-336. https://doi.org/10.1016/j.neucom.2021.04.069

[13] Zhao, S., Zhang, T., Hu, M., Chang, W., You, F. (2022). AP-BERT: Enhanced pre-trained model through average pooling. Applied Intelligence, 52(14): 15929-15937. https://doi.org/10.1007/s10489-022-03190-3

[14] Chen, H., Ming, T., Liu, S., Gao, C. (2019). Semantic summarization of reconstructed abstract meaning representation graph structure based on integer linear pragramming. Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology, 41(7): 1674-1681. https://doi.org/10.11999/JEIT180720

[15] Wu, B., Liang, X., Zhang, S., Xu, R. (2022). Advances and applications in graph neural network. Jisuanji Xuebao/Chinese Journal of Computers, 45(1): 35-68.

[16] Zhu, J., Xiang, L., Zhou, Y., Zhang, J., Zong, C. (2021). Graph-based multimodal ranking models for multimodal summarization. Transactions on Asian and Low-Resource Language Information Processing, 20(4): 1-21. https://doi.org/10.1145/3445794

[17] Yu, G., He, R., Liu, Y., Dang, J. (2017). Context based model for temporal Twitter summarization. Ruan Jian Xue Bao/Journal of Software, 28(10): 2654-2673. https://doi.org/10.13328/j.cnki.jos.005146

[18] Lee, L.S., Chen, S.C., Ho, Y., Chen, J.F., Li, M.H., Li, T.H. (2004). An initial prototype system for Chinese spoken document understanding and organization for indexing/browsing and retrieval applications. In 2004 International Symposium on Chinese Spoken Language Processing, Hong Kong, China, pp. 329-332. https://doi.org/10.1109/CHINSL.2004.1409653

[19] Li, Y., Ma, S., Jiang, H., Liu, Z., Hu, C., Li, X. (2018). An approach for storytelling by correlating events from social networks. Jisuanji Yanjiu yu Fazhan/Computer Research and Development, 55(9): 1972-1986. https://doi.org/10.7544/issn1000-1239.2018.20180155

[20] Song, J.Q., Lyu, M., Hwang, J.N., Cai, M. (2003). PVCAIS: A personal videoconference archive indexing system. In 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698), Baltimore, MD, USA, pp. II-673. https://doi.org/10.1109/ICME.2003.1221706