

CrossModal Image and Text Retrieval Algorithm Based on Deep Learning

Jiana Bi^{1,a*}

¹*School of Software and Big Data, Changzhou College of Information Technology, Changzhou213164, Jiangsu, China*

Multi-modal data contains rich and complementary information that provides a more comprehensive means of understanding and analyzing things deeply. Cross-modal image retrieval is related to the effective association between two different modal data, which can accurately retrieve a semantically matched image through the input text, or accurately find the corresponding text through the input image. In this study, the algorithm model is built based on deep learning, and the dual tower structure and multi-modal fusion structure are combined to give full play to the advantages of different structures. Image construction based on CNN is responsible for feature extraction of input images, and text construction based on Transformer is used to accurately capture semantic dependencies in text. The addition of a fusion layer to the twin tower structure can better mine the potential semantic associations between images and text. Two open data sets, MSCOCO and Flickr30k, were used to carry out experiments that demonstrated the superiority of the proposed model in cross-modal image retrieval task.

Keywords: deep learning; cross modal; image and text retrieval; twin tower structure; CNN; Transformer

1. INTRODUCTION

With the rapid development of information technology, there has been an explosive increase in the volume of generated data. In this environment, multi-modal data has become a key part of the data field, and is widely found in many fields such as social media, news information, e-commerce, medical imaging and intelligent security. Multi-modal data contains rich and complementary information, providing a more comprehensive perspective for in-depth analysis and understanding of things. Cross-modal retrieval requires the establishment of a common underlying space to capture and correlate data modes from different sources [1]. Cross-modal image retrieval is particularly challenging because images of similar content captured by different modes may have little common structure [2]. Cross-modal image retrieval is devoted to the effective correlation and retrieval between the

two different modal data, image and text. It can accurately retrieve the semantically matched image through the input text, or accurately find the corresponding text description according to the input image. In regard to search engines, cross-modal graphic retrieval technology enables users to query related pictures by entering text, which enriches the presentation form of search results and improves the efficiency and accuracy of the information acquired by users. Via image database management, users can easily find from the massive number of images the image that conforms to the specific text description, and efficiently manage and utilize image resources. In the field of intelligent aided design, users can input text description, quickly retrieve relevant image materials, obtain inspiration and reference for design work, and improve design efficiency and quality. The powerful feature learning ability and nonlinear modeling ability of deep learning provide a new opportunity and method for the development of cross-modal image retrieval. It can automatically learn the deep semantic features of

*Corresponding author. "Email: bijiana666@163.com

images and texts from a large amount of data, effectively address the limitations of traditional methods in dealing with complex semantic relations, and significantly improve the performance and effect of cross-modal image retrieval. The application of deep learning in cross-modal image retrieval can better reveal the potential semantic association between images and text, and achieve more accurate retrieval.

2. RELATED WORKS

Cross-modal retrieval aims to identify a common potential space [3]. Earlier studies used traditional machine learning to establish connections between text and images. For example, CCA is used to find linear correlations between text and image features, enabling cross-modal retrieval. Traditional methods are inadequate for dealing with complex semantic relations, and it is difficult to capture the deep semantic association between text and image effectively. With the emergence of deep learning, cross-modal text retrieval based on DNN has gradually become the mainstream. CNN has shown strong ability in image feature extraction. For example, AlexNet, VGGNet, ResNet and other models [4-6] can extract highly discriminative features from images, providing a visual feature basis for cross-modal image retrieval. In terms of text processing, RNN, its LSTM and GRU can effectively process text sequence information and capture semantic and contextual relationships [7]. For example, by fusing image features extracted by CNN with text features extracted by LSTM, a joint representation space is constructed to achieve similarity measurement and retrieval between images and texts.

The cross-modal matching model based on twin tower network has become an important direction. For example, the VSE++ model uses VGG convolutional network for image coding, GRU for text coding, and then uses triplet loss function optimization to map the picture and text to the shared subspace, and realizes retrieval based on the global similarity matching of the picture and text. However, because global features contain all information, and not all information is equally important in image retrieval, it is difficult to achieve a breakthrough in matching accuracy. Later, with the development of target detection models such as FasterRCNN and MaskRCNN, the similarity matching of target objects at the region level was used to accurately model the correspondence between image regions and sentence words or phrases, and the attention mechanism was used to estimate local similarity and accumulate global similarity, thus achieving a breakthrough in matching accuracy [8]. However, this kind of model requires complex process involving detection first and then coding, which is difficult to train end-to-end, and the mapping based on object region is simple, which limits the feature expression ability of the model.

Transformer has revolutionized the process of cross-modal image retrieval. It can better handle long-sequence data and capture global dependencies, and has excellent performance in cross-modal image retrieval. The advantages of BERT are extended to the cross-modal field, and the semantic alignment between images and texts is improved through

the joint encoding of images and texts [9]. At the same time, pre-training technology also provides strong support for cross-modal image retrieval. With pre-training on large-scale data sets, models can learn generic semantic representations, and fine-tuning on specific cross-modal retrieval tasks can significantly improve the model's generalizability. For example, OpenAI's CLIP model [10] is pre-trained on large-scale graphic pairs through comparative learning so that the model can understand the semantic association between images and texts, and achieve good results in a variety of cross-modal tasks. Zhang investigated deep learning models built on CNNs to automatically extract image features, thereby boosting the accuracy and robustness of classification tasks [11]; Ke addressed the poor robustness of neural network models against adversarial samples [12]. The above studies provide valuable references for the present research.

Although the relevant research at home and abroad has made remarkable progress, there are still some limitations. The processing of complex semantics and the understanding of fine-grained semantics need to be improved, and the retrieval accuracy is still not ideal for some texts and images with fuzzy semantics and polysemy. The training of the model requires a large amount of annotated data, and the cost of annotating good-quality graphic data is high, which limits the further optimization and expansion of the model. In addition, due to the high computational complexity, the real time and scalability of the model are challenged in practical applications, and it is difficult to meet some scenarios with high real-time requirements.

3. ALGORITHM AND MODEL

The algorithm model constructed in this study uses the combination of the two-tower structure and the multi-modal fusion structure to give full play to the advantages of different structures, thereby achieving effective semantic association and accurate retrieval between images and texts. The two-tower structure is a common architecture in cross-modal text retrieval, consisting of two branches: image tower and text tower. The image tower is responsible for the feature extraction of the input image, converting the image data into the feature vector that can represent its semantic information. The text tower processes the input text and extracts the semantic features of the text. The overall structure of the model is shown in Figure 1.

In this study, a fusion layer was added on the basis of the twin tower structure to fuse the feature vectors of the image tower and the text tower. By calculating the attention weight between the image feature and the text feature, the fusion ratio of the two modal features is dynamically adjusted so as to better capture the semantic association between the image and the text.

3.1 Image Feature Extraction Based on Improved CNN

Image feature extraction is a key step in cross-modal image retrieval, which aims to transform visual information into

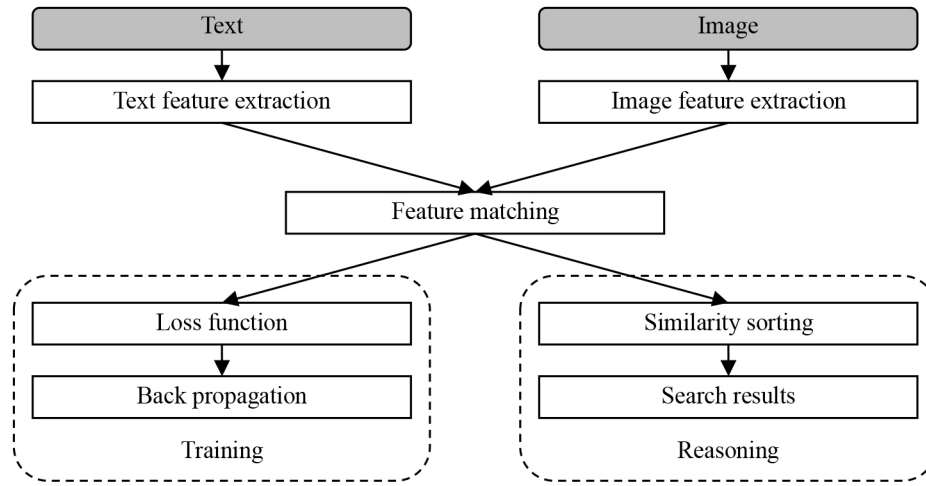


Figure 1 Overall architecture of the model.

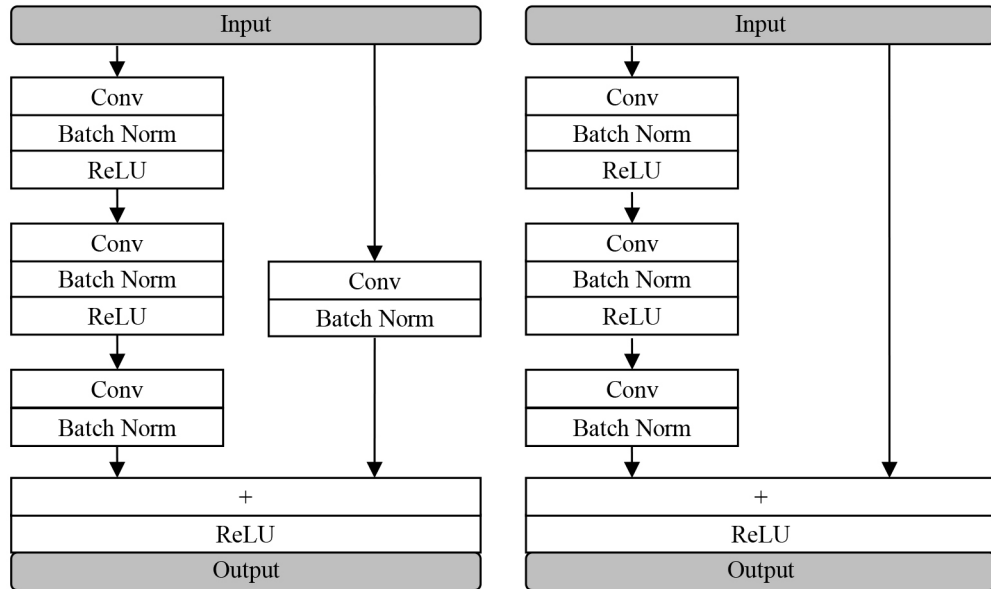


Figure 2 Conv Block and ID Block structure of ResNet-50.

feature vectors that can be understood and processed by computer. This study used an improved CNN to achieve this task. By optimizing and adjusting the classical CNN architecture, it can better extract the semantic features of images and improve the performance of cross-modal image retrieval.

3.1.1 ResNet-50 Model Structure

ResNet-50 was selected as the basic CNN architecture for image feature extraction. ResNet-50 is a deep residual network with 50 layers, which solves the problem of gradient disappearance and gradient explosion in deep networks by introducing residual blocks so that the network can learn deeper features [13]. The ResNet-50 consists of a Conv Block and an ID Block, as shown in Figure 2.

The Conv Block is used to change the dimension of the feature map, including the number of channels, height, and width. The Conv Block contains three convolution layers and under samples the main path, while the shortcut also requires convolution operations to match the dimensions of the main path output. Suppose the input feature graph is X , the three convolution layers on the main path are C_1 , C_2 , and C_3 , and the convolution layer on the shortcut path is C_s . The output of the main path is:

$$F(X) = C_3(ReLU(C_2(ReLU(C_1(x))))) \quad (1)$$

The output of the shortcut path is:

$$Y = C_s(X) \quad (2)$$

The output of the final convolutional block is:

$$Y = ReLU(F(X) + S(X)) \quad (3)$$

ID blocks are used to increase the depth of the network without changing the dimension of the feature map. It also contains three convolution layers, but there is no convolution operation on the shortcut path, and the input feature map is directly added to the output of the main path. The output of the ID Block main path is the same as that of the Conv Block, and the final constant output is:

$$Y = ReLU (F (X) + X) \quad (4)$$

The Conv Block is used to change the dimension of the feature graph, and there are convolution operations on the shortcut path to match the dimension of the main path output. The ID Block does not change the dimension of the feature map, and the identity mapping is directly carried out on the shortcut path. By using these two blocks alternately, ResNet-50 can build very deep network structures while avoiding gradient disappearance and gradient explosion problems, improving the performance of the model.

3.1.2 Multi-scale Feature Fusion

In order to better capture information at different scales in the image, this study extracts features at different levels of ResNet-50 and fuses these multi-scale features. Specifically, feature maps with different resolutions and semantic levels are extracted separately in the second, third, fourth, and fifth stages of ResNet-50 (each containing multiple residual blocks).

Feature extraction stage:

$$F2, F3, F4, F5 = ResNet - 50 (x) \quad (5)$$

Top-down fusion (F5→F4 as an example):

$$P4 = Conv_{1 \times 1} (F4, 256) + interp (Conv_{1 \times 1} (F5, 256), scale_factor = 2) \quad (6)$$

Layer by layer sampling fusion:

$$P3 = Conv_{1 \times 1} (F3, 256) + interp (P4, scale_factor = 2) \quad (7)$$

$$P2 = Conv_{1 \times 1} (F2, 256) + interp (P3, scale_factor = 2) \quad (8)$$

Horizontal connection (1×1 convolutional adjustment channel):

$$F_{adj} = W_{lat} \cdot F + b_{lat} \quad (9)$$

Where $W_{lat} \in \mathbb{R}^{C_{out} \times C_{in} \times 1 \times 1}$ and $b_{lat} \in \mathbb{R}^{C_{out}}$
Upsampling (bilinear interpolation):

$$F_{up} (x, y) = \sum_{i,j} F \left(\frac{x}{s} - i, \frac{y}{s} - j, \right) \cdot w (i, j) \quad (10)$$

Where, s is the upper sampling scale and $w (i, j)$ is the interpolation weight.

Feature fusion (add element by element):

$$P_l = F_{adj}^{(l)} + interp (P_{l+1}) \quad (11)$$

3.1.3 Attentional Mechanism

To make the model pay more attention to the parts of the image that are related to the text semantics, this study introduced the attention mechanism in ResNet-50. This involves adding an attention module after integrating multi-scale features [14]. Two mainstream attention mechanisms that are often cited in ResNet-50 are CAM and SAM. CAM is cited in this study, using the SE (Squeeze-and-Excitation) module. The mathematical model comprises three steps:

Step 1: Squeeze. Conduct global average pooling for each channel and generate channel descriptors:

$$z_c = F_{sq} (x_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_c (i, j) \quad (12)$$

Where $x_c \in \mathbb{R}^{H \times W}$ is the c channel of the input feature graph, and H and W are spatial dimensions.

Step 2: Excitation. Generate channel weights via two levels of fully connected layers (activated with ReLU in the middle):

$$s_c = F_{ex} (z, W) = Sigmoid (W_2 \cdot ReLU (W_1 \cdot z)) \quad (13)$$

Where, $W_1 \in \mathbb{R}^{(C/r) \times C}$ and $W_2 \in \mathbb{R}^{C \times (C/r)}$ are learnable parameters, and r is the compression ratio.

Step 3: Scale. Multiply the weights with the original feature map channel by channel.

$$x_c = s_c \cdot x_c \quad (14)$$

The purpose of the above formula is to enhance the characteristic response of important channels and suppress redundant channels.

3.2 Text Feature Extraction Based on Transformer

Before Transformer was introduced, RNN was commonly used to deal with long sequence generation. However, due to its recursive structure, the implicit state cannot be passed to the next moment until the information processing of the current moment is completed, and the RNN cannot capture the relationship between two-time steps when they are too far apart. Although the emergence of variants such as LSTM and GRU has alleviated the long-term dependence of RNN in information transmission, the issue has not been completely resolved. Transformer has been redesigned to address the weaknesses of RNN, to rectify the low efficiency and defects in information transmission, and to perform better than RNN in many tasks [15]. The encoder is a key component of the Transformer model and is responsible for converting the input sequence into a high-dimensional representation so that the model can understand and process text data. The Transformer encoder structure is shown in Figure 3.

Transformer uses Self-Attention (SA), and the query vector is not given, but obtained by linear mapping of the input vector, which means that the model will determine the

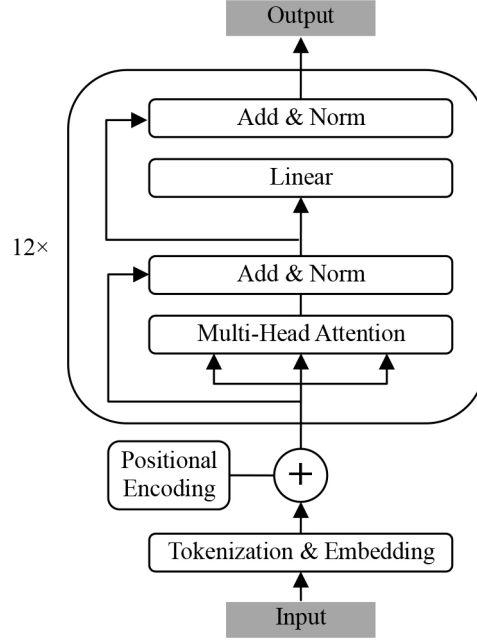


Figure 3 Transformer encoder.

most important information based on the input vector itself after receiving it. The input vector is represented as:

$$H = (h_1, h_2, \dots, h_N)^T \quad (15)$$

Firstly, it is mapped to a query (Q), key (K), and value (V) with the formula:

$$Q = HW^Q = (q_1, q_2, \dots, q_N) \quad (16)$$

$$K = HW^K = (k_1, k_2, \dots, k_N) \quad (17)$$

$$V = HW^V = (v_1, v_2, \dots, v_N) \quad (18)$$

Next, using the dot product of the query and key, the attention score is calculated for each position and the weighted results are summed:

$$c_i = \sum_{j=1}^N \text{softmax}(q_i \bullet k_j) \bullet v_j \quad (19)$$

In order to speed up the calculation and improve efficiency, the attention output vector of all positions can be calculated at once through the matrix:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (20)$$

Where, $Q \in \mathbb{R}^{N \times d_k}$ is a matrix composed of query vectors, $K \in \mathbb{R}^{N \times d_k}$ is a matrix composed of key vectors, $V \in \mathbb{R}^{N \times d_v}$ is a matrix composed of value vectors, N represents quantity, d_k is the dimension of query and key vectors, d_v is the dimension of value vectors, and $\sqrt{d_k}$ is a reduced term introduced to avoid excessive dot product.

Multi-Head Attention (MHA) can analyze input information from different perspectives to enhance the expressiveness

of the model. More comprehensive features can be extracted. The combination formula of MHA is:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \quad (21)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (22)$$

Where, W^O is also a linear change matrix that can be learned, and its function is to map the self-attention obtained by splicing back to the original dimension of the input token after all the heads have been spliced. Because Transformer uses a multi-head self-attention mechanism, it means that:

$$Q = K = V = H \quad (23)$$

4. EXPERIMENT AND ANALYSIS

4.1 Data Set Selection and Preprocessing

Two public datasets, MSCOCO and Flickr30k, were selected for this study. The MSCOCO dataset is a large-scale sentence-based image description dataset with scene understanding as its goal and image content extracted mainly from complex daily scenes. The Flickr30k dataset is collected from Flickr, with images and texts of high quality and standardized annotation, providing reliable data support for cross-modal image retrieval research. The two datasets were thoroughly cleaned to remove noisy data, mislabeling and duplicate data. Then, data annotation was carried out. Although MSCOCO and Flickr30k datasets already have certain annotation information, further detailed annotation was carried out to better meet the experimental requirements of this study. Finally, after cleaning and labeling, the MSCOCO

Table 1 Hardware and software environment for the experiment

Class	Name	Configuration
Hardware environment	GPU	NVIDIA Tesla V100
	GPU's VRAM	32GB
	CPU	Intel Xeon Platinum 8280
	CPU's main frequency	2.7GHz
	CPU's memory	128GB
	CUDA core count	5120
Software environment	OS	Ubuntu 18.04
	Programming language	Python 3.7
	Deep learning framework	PyTorch 1.7.1
	Numerical calculation	NumPy 1.2
	Data processing and analysis	Pandas 1.3.3
	Data visualization	Matplotlib 3.8

Table 2 Performance metrics of this research model on different datasets and tasks

Data set	Task	Accuracy	Recall	F1
MSCOCO	TBIR	0.82	0.78	0.80
MSCOCO	IBTR	0.80	0.76	0.78
Flickr30k	TBIR	0.78	0.74	0.76
Flickr30k	IBTR	0.76	0.72	0.74

Table 3 Comparison of this research model with VSE++ and SCAN models

Model	MSCOCO dataset (mAP)	Flickr30k dataset (mAP)
This research model	0.85	0.82
VSE++	0.78	0.75
SCAN	0.82	0.79

and Flickr30k datasets were divided into 80% data as the training set, 10% data as the verification set, and 10% data as the test set.

4.2 Experimental Hardware and Software Environment

Based on the existing experimental conditions in the laboratory and the requirements of study, the hardware and software environments were built, as shown in Table 1.

4.3 Experimental Results and Analysis

After completing the training and testing of the algorithm model, the experimental results were statistically analyzed in detail. The experimental results of text-based image retrieval (TBIR) and image-based text retrieval (IBTR) tasks of this research model on MSCOCO and Flickr30k datasets are shown in Table 2.

As can be seen from Table 2, the performance of the proposed model on the MSCOCO dataset is slightly better than that of the Flickr30k dataset, indicating that the proposed model performs well on different datasets and can effectively implement cross-modal image retrieval. In order to more intuitively compare the performance differences between this study model and other related models, the mean accuracy (mAP) of this study model and the two classical models

VSE++ and SCAN on the same data set is shown in Table 3.

As can be seen from the comparison results in Table 3, the mean accuracy of the proposed model on MSCOCO and Flickr30k datasets is higher than that of the VSE++ and SCAN models. On the MSCOCO dataset, the mAP of the model in this study reaches 0.85, which is 0.07 higher than that of the VSE++ model and 0.03 higher than that of the SCAN model. On the Flickr30k dataset, the mAP of the model in this study is 0.82, which is 0.07 higher than that of the VSE++ model and 0.03 higher than that of the SCAN model. This fully proves the superiority of this model in cross-modal image retrieval task, as it can more accurately retrieve the semantic results matching the queried text or image.

4.4 Analysis of Influencing Factors of Model Performance

To further improve the retrieval ability of this research model, two key factors affecting the performance of the model were systematically analyzed, providing a strong basis for the optimization and improvement of the model.

(1) Dataset size has a significant impact on model performance. In the experiment, by gradually increasing the amount of training data, we observe the change of model indexes on the MSCOCO and Flickr30k datasets. The results show that the performance of the model tends to improve with the increase of dataset size. When the amount of training

data is small, it is difficult to fully capture the complex semantic association between images and texts due to the limited sample information learned, resulting in low retrieval performance. For example, on the MSCOCO dataset, when training with 10% of the data, the mean accuracy is only 0.65. When the amount of training data increases to 80%, the mean accuracy increases to 0.85. This is because a greater amount of training data can provide the model with richer semantic information, enabling it to learn more mapping relationships under different scenes, themes and semantic expressions, thus enhancing the generalizability of the model and its capacity to understand complex semantics. However, when the size of the dataset increases to a certain extent, the performance improvement gradually decreases, indicating that after a certain amount of data increase, the improvement effect of continuing to increase the data on the model performance is limited, indicating that at this time, the model has learned the main semantic patterns in the data, and the marginal benefit obtained by further increasing the data decreases.

(2) Model structure is one of the key factors determining model performance. The model constructed in this study combines the two-tower structure with the multi-modal fusion structure, and includes an attention mechanism in the design. In order to analyze the effect of the model structure on the performance, a series of comparative experiments were carried out where the multimodal fusion structure and attention mechanism were removed respectively, and the performance of the model was observed. The results show that after removing the multimodal fusion structure, the mean precision of the model decreases by 0.08 in the MSCOCO dataset and 0.07 in the Flickr30k dataset. This shows that the multi-modal fusion structure can effectively facilitate the interaction and fusion between image and text features, enhance the semantic association between the two-modal data, and improve the retrieval performance of the model. After removing the attention mechanism, the average accuracy of the model in the two data sets also decreased by about 0.05, indicating that the attention mechanism can make the model pay more attention to the semantically relevant parts of the image and text and dynamically adjust the weight of features, thereby improving the accuracy of similarity calculation, which plays an important role in enhancing the overall performance of the model.

5. CONCLUSIONS

The results of this study theoretically enrich the research content in the field of cross-modal image retrieval, and provide a new idea and method for the application of deep learning in cross-modal retrieval. Moreover, it has important value in practical application, and can provide effective technical support for information retrieval and analysis in many fields such as medicine and security, improve the work efficiency and decision-making accuracy of related fields, and promote the intelligent development of these fields. Future in-depth research will be carried out to improve the model structure and algorithm, and learn the mapping relationship between images and texts by generating adversarial methods.

The application scenarios will be expanded, the cross-modal image retrieval technology will be applied to more fields, and new solutions will be proposed for information management and retrieval in these fields. Also, future work will explore a range of modal data fusion methods and determine ways to achieve efficient and accurate retrieval between graphic data in different languages. The focus will be on the interpretability and security of the model, making the decision-making process of the model more transparent and explainable, and ensuring the data security and privacy protection of the model in the application process.

REFERENCES

1. Mallea M, Nanculef R, Araya M. Intramodal consistency in triplet-based cross-modal learning for image retrieval[J]. *Machine Learning*, 2025, 114(4): 110-110.
2. Breznik E, Wetzler E, Lindblad J, et al. Cross-modality sub-image retrieval using contrastive multimodal image representations[J]. *Scientific Reports*, 2024, 14(1): 18798-18798.
3. Maleki D, Rahnamayan S, Tizhoosh R H. A self-supervised framework for cross-modal search in histopathology archives using scale harmonization[J]. *Scientific reports*, 2024, 14(1): 9724-9724.
4. Asnake W N, Ayalew M A, Engda A A. Detection of oral squamous cell carcinoma cancer using AlexNet on histopathological images[J]. *Discover Applied Sciences*, 2025, 7(3): 155-155.
5. Abdullah K, Asfandiyar K, Muneeb U, et al. A computational classification method of breast cancer images using the VG-Net model [J]. *Frontiers in Computational Neuroscience*, 2022, 161001803-1001803.
6. Anila S V, Nagarajan G, Perarasi T. Low-light image enhancement using retinex based an extended ResNet model[J]. *Multimedia Tools and Applications*, 2024, (prepublish): 1-16.
7. Chantangphol P, Sakdejayont T, Chalothorn T. Enhancing spoken term detection with deep acoustic word embeddings and cross-modal matching techniques[J]. *International Journal of Speech Technology*, 2024, 27(4): 1-12.
8. Choudhury S, Saini D, Banerjee B. Boosting cross-modal retrieval in remote sensing via a novel unified attention networks.[J]. *Neural networks: The Official Journal of the International Neural Network Society*, 2024, 180106718.
9. Viktor A, Dmitrij Š. Text Augmentation Using BERT for Image Captioning[J]. *Applied Sciences*, 2020, 10(17): 5978-5978.
10. Shukor M, Thome N, Cord M. Vision and Structured-Language Pretraining for Cross-Modal Food Retrieval[J]. *Computer Vision and Image Understanding*, 2024, 247104071-104071.
11. Zhang X. Deep Learning Model for Image Classification in Machine Vision[J]. *Engineering Intelligent Systems*, 2025, 33(4): 405-415.
12. Ke H. Algorithm for Detection and Defense of Neural Network Technology Based on Neural Network and Multimedia[J]. *Engineering Intelligent Systems*, 2025, 33(4): 377-384.
13. Gu, Y., Wang, M., Gan, J. et al. Text semantic-guided adaptive feature aggregation for image-text retrieval. *Multimedia Systems* 31, 300 (2025).
14. Chen, R., Qiang, B., Yang, X. et al. Cross-modal quaternion relation mining and gap bridging for image-text retrieval. *J Intell Inf Syst* 64, 715–734 (2026).
15. Liu, D., Wang, P., Zhao, Z. et al. Automated text annotation: a new paradigm for generalizable text-to-image person retrieval. *Appl Intell* 55, 593 (2025).

