

Research Article

A Cross-Modal Image and Text Retrieval Method Based on Efficient Feature Extraction and Interactive Learning CAE

Xiuye Yin ¹ and Liyong Chen²

¹*School of Computer Science and Technology, Zhoukou Normal University, Henan, Zhoukou 466001, China*

²*School of Network Engineering, Zhoukou Normal University, Henan, Zhoukou 466001, China*

Correspondence should be addressed to Xiuye Yin; 20111036@zkn.edu.cn

Received 10 November 2021; Revised 1 December 2021; Accepted 9 December 2021; Published 10 January 2022

Academic Editor: Le Sun

Copyright © 2022 Xiuye Yin and Liyong Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In view of the complexity of the multimodal environment and the existing shallow network structure that cannot achieve high-precision image and text retrieval, a cross-modal image and text retrieval method combining efficient feature extraction and interactive learning convolutional autoencoder (CAE) is proposed. First, the residual network convolution kernel is improved by incorporating two-dimensional principal component analysis (2DPCA) to extract image features and extracting text features through long short-term memory (LSTM) and word vectors to efficiently extract graphic features. Then, based on interactive learning CAE, cross-modal retrieval of images and text is realized. Among them, the image and text features are respectively input to the two input terminals of the dual-modal CAE, and the image-text relationship model is obtained through the interactive learning of the middle layer to realize the image-text retrieval. Finally, based on Flickr30K, MSCOCO, and Pascal VOC 2007 datasets, the proposed method is experimentally demonstrated. The results show that the proposed method can complete accurate image retrieval and text retrieval. Moreover, the mean average precision (MAP) has reached more than 0.3, the area of precision-recall rate (PR) curves are better than other comparison methods, and they are applicable.

1. Introduction

With the advancement of digitalization, more and more people use the Internet to obtain the information they need. How to make users accurately and quickly search for the information they need has become a hot issue [1]. In the era of mobile Internet, each of us is receiving massive amounts of information from the Internet, while at the same time generating massive amounts of multimedia information, that is, multimodal data [2]. The original form of cross-modal retrieval is similar to that of single-mode retrieval. With the growth of multimodal data, it is more difficult for users to retrieve the information they are interested in efficiently and accurately [3]. There are many retrieval methods so far, most of which are based on a single modality, such as searching for articles by text, searching for pictures by pictures, or multimodal search on the surface. In fact, it is in the form of search keywords to query and request the most matching content among many resources on the Internet.

In order to meet people's actual needs and provide better retrieval services, scholars are committed to the research on relevant methods and practice in the field of cross-modal retrieval. Therefore, the cross-modal retrieval method has a wide range of application scenarios and research significance. How to mine the effective information in these multimodal data is an important problem in the research field of cross-modal retrieval.

Researchers found a semantic gap between the low-level features of data and high-level semantics, and the data of different modalities are heterogeneous [4, 5]. It can be seen that the core of cross-modal retrieval research is to mine the associated information between different modal data. How to mine this associated information has become the key to the research of cross-modal retrieval technology.

In recent years, with the rapid development of deep learning technology, people have become more and more capable of solving more complex machine learning problems and have made great progress in analyzing and processing multimodal data [6]. Multimodal content analysis has broad

application prospects in various fields such as smart cities, smart homes, and smart transportation. Based on the breakthrough progress in the application research of deep learning in the monomodal field, it is applied to the theoretical research of cross-modal retrieval tasks, and technical practice is provided at the same time [7].

The current cross-modal retrieval system modelling mainly solves two problems: one is how to complete the unified mapping of different modal information features and the second is how to ensure the retrieval rate on the basis of improving the retrieval rate of retrieval models [8]. These two problems are interdependent. Due to the diversity and heterogeneity of different modal information, the feature extraction method and unified representation form of each modal become the key to solving the problem [9, 10]. In addition, the corpus with three modalities and above is less researched, and the corpus with two modalities is more common. In particular, the corpus with the modal alignment of images and text is more common.

2. Related Research

Because there is a huge heterogeneous gap in different modal data, how to effectively measure the content similarity of different modal data has become a major challenge [11]. Nowadays, many cross-modal retrieval methods have been proposed [12].

2.1. Real-Valued Cross-Modal Retrieval Method. Cross-modal retrieval methods based on real-valued representation can generally be divided into two categories: canonical correlation analysis (CCA) and deep learning [13]. CCA uses different modal data to form sample pairs, learns a projection matrix, and projects different modal data to a common latent subspace, and then in the subspace, measures the similarity between modal data [14]. Reference [15] proposed a new multilabel kernel canonical correlation analysis (ml-KCCA) method for cross-modal retrieval, which uses the high-level semantic information reflected in multilabel annotations to enhance the kernel CCA. Reference [16] proposed cross-media correlation learning with deep canonical correlation analysis (CMC-DCCA). It can better mine the complex correlation between cross-media data and achieve better cross-media retrieval performance. However, the performance of its feature extraction algorithm highly depends on the size of the sample set, and it is difficult to obtain training samples for noncooperative targets in actual situations. How to efficiently set the parameter range still needs further exploration.

The cross-modal retrieval method based on deep learning makes full use of the powerful feature extraction capabilities of deep learning models, learns the feature representation of different modal data, and then establishes semantic associations between modalities at a high level [17]. Reference [18] proposed a two-stage deep learning method for supervised cross-modal retrieval, extending the traditional norm-related analysis from 2 views to 3 views and conducting supervised learning in two stages. The evaluation

results on two publicly available datasets show that the proposed method has a better performance. However, there is still room for optimization for the detection accuracy of complex retrieval environments. At present, the dimensionality obtained by the representation learning model when automatically extracting features is relatively high. Particularly for the cross-modal retrieval model based on deep learning, the sample feature dimension obtained in the representation stage is usually not less than 4096, and the final feature dimension is still too high [19]. Reference [20] proposed an image retrieval method combining deep Boltzmann machine (DBM) and CNN to extract high-order semantic features of the image.

2.2. Cross-Modal Retrieval Method Based on Hash Transformation. The cross-modal retrieval method based on real-valued representation has the problems of time-consuming calculation and large demand space when facing large-scale data. Therefore, an information retrieval method based on hash transformation appears. This method is based on the paired sample pairs of different modal data, learns the corresponding hash transformation, maps the corresponding modal data features to the Hamming binary space, and then realizes faster cross-modal retrieval in this space [21]. The premise of hash transformation is that the hash codes of similar samples are also similar. Reference [22] proposed a method called DNDCMH. This algorithm uses binary vectors specifying the existence of specific facial attributes as input queries to retrieve relevant facial images from the database. Secondly, the dimension reduction methods such as principal component analysis (PCA) can reduce the feature dimension to a certain extent, but under the premise of maintaining the necessary retrieval accuracy, the dimension that can be reduced is quite limited and lacks efficient and reasonable retrieval mechanism that can adapt to large-scale image sets [23]. Reference [24] proposed a new self-supervised deep multimodal hashing (SSDMH) method. However, cross-modal retrieval still only realizes the matching of image content and subject words, ignoring a large amount of content-based, subtle, and important image information [25]. Reference [26] proposed a deep hashing method that can combine stacked convolutional autoencoders with hash learning and hierarchically map the input image to a low-dimensional space. Some additional relaxation constraints are added to the objective function to optimize the hash algorithm. Experimental results on ultra-high-dimensional image datasets show that the proposed method has good stability in cross-modal retrieval, but the detection timeliness needs to be optimized. However, various models have their specific adaptation targets, advantages, and limitations. How to combine the advantages of models and various algorithms in practical applications to construct a universal cross-modal retrieval model is one of the urgent problems to be solved in the current cross-modal retrieval research.

2.3. Other Cross-Modal Retrieval Methods. In addition to the above classical methods, there are some other methods. For example, Feng et al. [27] proposed an automatic encoder

(Corr-AE) model, which is characterized by using two autoencoder networks to encode image vectors and text vectors with each other to obtain two correlation loss terms for model training. Reference [28] proposed a retrieval method based on multimodal semantic autoencoder. This method uses an encoder decoder to learn projection and preserve feature and semantic information while ensuring embedding. The 2-way net model proposed in [29] also applies the idea of autoencoder, which is optimized in more detail than Corr-AE. Reference [30] proposed a graphic matching method based on semantic concepts and order (SCO), which is characterized by introducing a multilabel classification mechanism when retrieving images. Specifically, SCO performs a multilabel classification operation for each candidate image extracted by the target detection network so that each candidate image can not only carry entity category information but also add some attribute labels.

According to the above analysis, (1) in CCA method, the single-mode feature representation of different data is extracted first, and then associated learning is carried out. This two-stage method cannot ensure that the extracted single-mode feature is the effective representation required by associated learning. (2) In the deep learning method, most networks use shallow networks to model the association learning part, ignoring the high-level semantic association between modes. (3) In the deep hashing method, some information will be lost when it converts the modal representation to hash coding.

Therefore, effective feature extraction and feature association learning are key to improving the accuracy of cross-modal retrieval. In order to make better association learning between different modal data, a cross-modal image and text retrieval method combining efficient feature extraction and interactive learning convolutional autoencoder (CAE) is proposed in this paper. The innovations of the proposed method are as follows:

- (1) Image feature extraction: The new convolution kernel constructed by 2DPCA is integrated into the image feature extraction based on residual network, which avoids the complex operation of traditional PCA and reduces the dimension of image spatial features.
- (2) Cross-modal CAE architecture: Based on the traditional multimodal CAE architecture, a feature association module (i.e., joint public representation) is integrated to associate the representations of each mode to realize interactive learning, make the learned intermediate representation of each mode contain the association relationship between modes, and improve the accuracy of cross-modal retrieval.

3. Method Framework

3.1. Overall Framework. In order to make full use of the advantages of complementary information of multimodal data, in the training stage, the proposed method takes image data and text data as the input of the network at the same

time, carries out interactive learning of image and text features through multimodal CAE model, and generates the classification model of the retrieval system. In the test stage, the image or text features are input into the classification model for discrimination, and the retrieval results are obtained. The overall architecture of the proposed method is shown in Figure 1.

Among them, the image data use the residual network as the image feature extractor and introduce two-dimensional principal component analysis (2DPCA) to construct a new convolution kernel. The text data use word2vec and long short-term memory (LSTM) network as the text feature extractor. The network fusion layer is designed using cross-modal convolution CAE based on interactive learning, and the two modal data features are fused and sent to the next fully connected layer. In order to learn the nonlinear mapping from the image-text data feature space to the semantic label space and prevent overfitting, the Batch Norm layer and the ReLU layer are added to the fully connected layer. The output dimension of the final fully connected layer is consistent with the data dimension of the real label. The proposed method takes full advantage of the complementary information of different modal data for multimodal data-image data and text data.

3.2. Improved Image Feature Extraction of Convolution Kernel

3.2.1. Convolution Neural Network Is Used to Extract Image Features. For the extraction of image features, a very mainstream residual network, which is more suitable for image features, is selected. The network has five convolution stages, each of which has a corresponding pooling operation. After inputting a piece of image data, it is processed in layers of convolution, and the size of the output image feature map is $7 \times 7 \times 2048$, which can be processed according to the needs of subsequent machine learning tasks.

Image modal data have high dimensionality and rich content information. The selection of a deep convolutional neural network will extract effective visual monomodal representation features. Using W_x to simplify the model parameters of the entire embedded subnetwork, the feature output h_x of the image modal data after passing through this network is

$$h_x = f_x(X; W_x), \quad (1)$$

where X is the input image modal data.

3.2.2. Constructing a New Convolution Kernel by Introducing 2DPCA. PCA is a linear analysis method to extract the main features of data in high-dimensional space and transform it into low-dimensional vector space. 2DPCA directly utilizes the two-dimensional information of the image, avoiding the complicated calculations brought about by PCA's row and column vector conversion while retaining the spatial characteristics of the image. Assuming that there are M images $I = \{I_1, I_2, \dots, I_M\}$ of size $w \times h \times c$, the average image of the sample can be expressed as

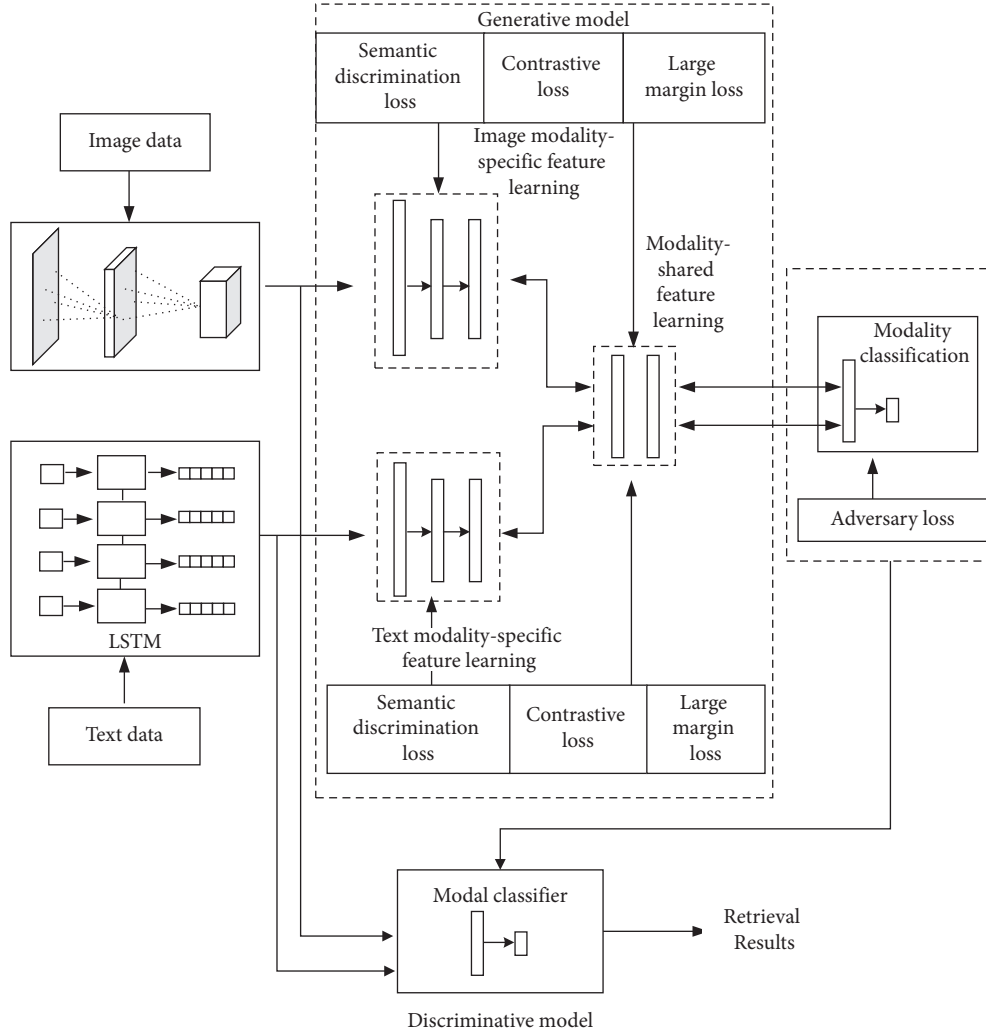


FIGURE 1: Overall framework of the proposed method.

$$\bar{I} = \frac{1}{M} \sum_{i=1}^M I_i. \quad (2)$$

The difference image between each sample and the average image is

$$Z(i) = I_i - \bar{I}. \quad (3)$$

The required covariance matrix is

$$C_{n \times n} = \frac{1}{M} \sum_{i=1}^M (I_i - \bar{I})^T (I_i - \bar{I}). \quad (4)$$

The optimal projection subspace $U = \{\eta_1, \eta_2, \dots, \eta_d\}$ can be constructed using the orthogonal eigenvectors corresponding to the first d eigenvalues of the covariance matrix. Mapping the original image to the projection space can obtain the feature image $T_i = Z_i U$ after dimensionality reduction. The flow of the 2DPCA algorithm is shown in Figure 2.

3.3. Text Feature Extraction. In the multimodal dataset used, the text modal data are mainly in the form of long text, so a reasonable representation that matches its characteristics is used for text feature extraction.

Short sentences: the text representation of short sentences is simpler than long sentences. It is represented by word vector (word2vec); that is, words are converted into vectors that can be accepted by machine learning tasks.

Long sentences: The representation of long sentences is more complicated because the words of the sentence are related to each other. The first or several words will affect the understanding of the following sentence, so the sentence's meaning should be grasped from the whole. In order to retain the previous information in the text, the LSTM network is used to first represent each word in the sentence by a word vector $Y = \{y_1, y_2, \dots, y_c\}$, and c represents the number of words in the sentence, so each sentence is represented as a 300-dimensional word vector sequence.

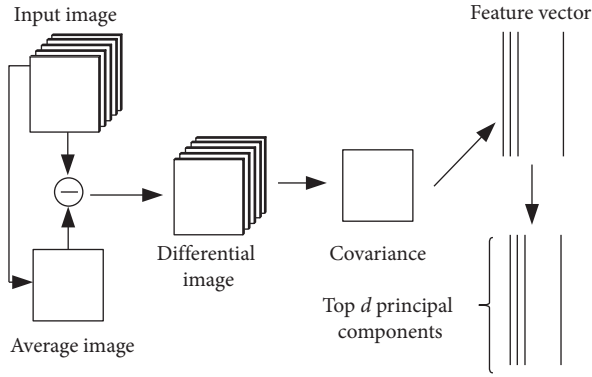


FIGURE 2: The flow of the 2DPCA algorithm.

4. Cross-Modal Convolutional Autoencoder

4.1. Classical Convolutional Autoencoder (CAE). An autoencoder (AE) is an unsupervised learning algorithm that makes the output close to the input by learning data representation. AE extracts data features through an encoder and then decodes the acquired features through a decoder to realize the reconstruction of input data. CAE is based on unsupervised AE, combining the convolution and pooling operations of CNN to convolve the encoder and decoder to achieve better feature extraction [31]. The single-layer CAE network model is shown in Figure 3. The coding part is composed of a convolutional layer and a maximum pooling layer.

Given M_{C1} feature maps $I = \{I_1, I_2, \dots, I_{C1}\}$, after convolution operation, a set of F_{C2} feature maps is obtained

$$g_n(i, j) = a \left(\sum_{u=-k}^k \sum_{v=-k}^k F_n^{(1)}(u, v) * I(i-u, j-v) + b_n^{(1)} \right), \quad (5)$$

where $g_n(i, j)$ is the activation value at pixel (i, j) in the activation map of the n -th channel and $a(\cdot)$ is a nonlinear activation function. The size of the filter is $F_{C2} = 2k + 1$. $F_n^{(1)}$ is the weight of the convolution filter in the encoding process, and the number of channels of each filter is the same as that of the input sample. $b_n^{(1)}$ is the offset of the encoder convolutional layer to the activation map of the n -th channel.

The convolutional layer of the convolutional encoding part outputs a feature map of size $(O_{C1} - F_{C2}/S_{C2} + 1)^2 \times M_{C2}$. After the maximum pooling operation, the final output of the encoding part is obtained. Among them, $O_{C1} = ((w - F_{C1} + S_{C1})/S_{C1}F_{P1})$ is the output feature map size of the convolution module $C1$.

The decoding process is the process of reconstructing the original image from the feature activation map. CAE is a fully convolutional network, so the decoding process is mainly realized through deconvolution operation. Considering that the size of the feature activation map obtained after encoding is smaller than the original image, the size information of the original image cannot be reconstructed only through the transposed convolution of the decoding process. Therefore, it is necessary to perform zero padding

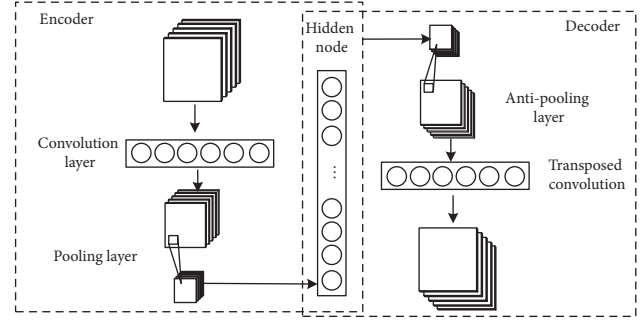


FIGURE 3: CAE model.

operation on the input feature map to decode later; a reconstructed image with the same size as the original image can be reconstructed. The convolution output of the encoding part is used as the input of the decoder and then convolved with the convolution filter $F^{(2)}$ to obtain the reconstructed image:

$$\tilde{I} = f(G * F_n^{(2)} + b_n^{(2)}), \quad (6)$$

where G is the set of feature maps obtained by encoding and $b_n^{(2)}$ is the offset of the activation map of the n -th channel corresponding to the decoder deconvolution layer.

4.2. Cross-Modal CAE Based on Interactive Learning. Different from the existing multimodal CAE models [32, 33], while learning the representations of different modes, respectively, this method generates some association between the representations of each mode through a feature association module (i.e., joint public representation) after the hidden layer, to realize interactive learning. Therefore, the intermediate representation of each mode contains the correlation between modes, which helps to improve the accuracy of cross-modal retrieval. The proposed dual-mode interactive learning CAE architecture is shown in Figure 4.

The input text and image data are, respectively, passed through the convolution layer and the pooling layer to obtain the data representation. Then, through an intermediate interaction layer, the feature representation of text and image data is interactively learned to obtain a new joint public representation feature data. The original input can be obtained by deconvolution of the feature data [34–36].

In order to train the dual-mode interactive learning CAE, it is necessary to construct the objective function in the training stage. In classical CAE training, the objective function is usually to minimize the reconstruction error. However, in the dual-mode interactive learning CAE model, the interactive learning between multimodal features is integrated to improve the accuracy of model retrieval. Therefore, the objective function needs to include the goal of maximizing the correlation between the two modal features in the hidden layer.

The given input is $z_i = \{x_i; y_i\}$, where z_i is the associated representation of the input views x_i and y_i . Self-reconstruction loss and cross-reconstruction loss are defined as

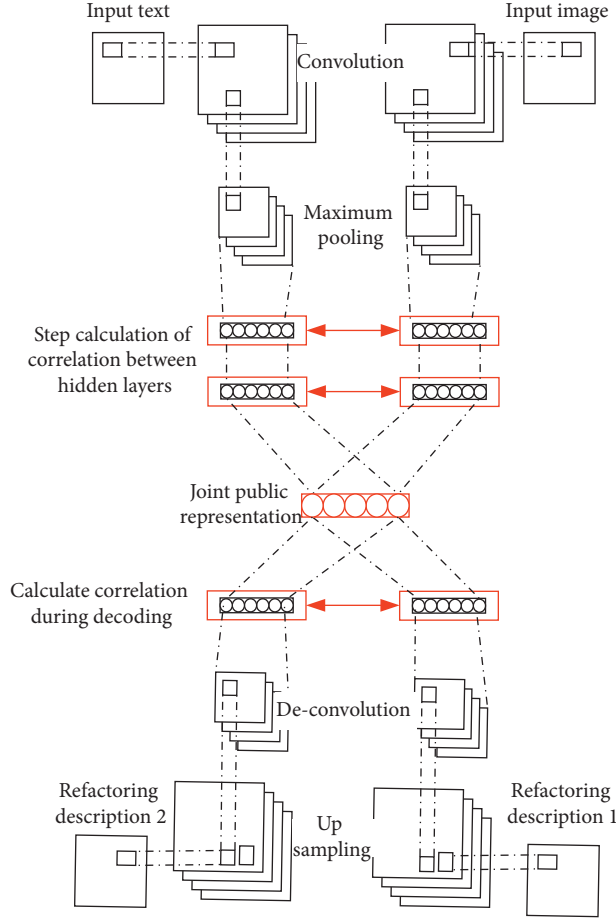


FIGURE 4: Dual-mode interactive learning CAE architecture.

$$\begin{aligned}
 L_1 &= \sum_{i=0}^N L(z_i, g(h(z_i))), \\
 L_2 &= \sum_{i=0}^N L(z_i, g(h(x_i))), \\
 L_3 &= \sum_{i=0}^N L(z_i, g(h(y_i))), \\
 L_4 &= \sum_{k=0}^K \sum_{i=0}^N L(h(x_i)^k, h(y_i)^k), \\
 L_5 &= \sum_{i=0}^N L(g(h(x_i)), g(h(y_i))),
 \end{aligned} \tag{7}$$

where g , h are the nonlinearity generally regarded as ReLU, $g(h(x_i^k))$ and $g(h(y_i^k))$ are the representations of the k^{th} intermediate hidden layer ($K=2$), and L is the error function. In the loss L_2 and L_3 (for cross reconstruction), the 0 vector is used instead of another view to calculate x_i and y_i .

Finally, in order to enhance the interaction between the two modal features, the objective function of correlation loss is expressed as follows:

$$\begin{aligned}
 L_6 &= \lambda \text{corr}(h(X), h(Y)), \\
 L_7 &= \sum_{k=0}^K \lambda_k \text{corr}(h(X)^k, h(Y)^k),
 \end{aligned} \tag{8}$$

where $h(X)$ and $h(Y)$ are the projections of the combined model (the projections of the joint public representation in Figure 4). X and Y are the representation of two modal features. λ_k is the relative regularization hyperparameter used for each k^{th} intermediate encoding step (similarly using λ in the decoding stage). In the encoding process, a convolution layer and two intermediate layers ($K=2$) are used. For decoding, the deconvolution layer and an intermediate layer ($K=1$) are used for reconstruction. λ affects the complexity of model training. When it is too small, the model is easy to overfit. When the value is large, it is easy to cause underfitting. Considering the search results on each

dataset, $\lambda_1 = 0.004$ and $\lambda_2 = 0.05$ in item L_7 and $\lambda = 0.02$ in item L_6 are uniformly set here.

The correlation between the two views $h(X)$ and $h(Y)$ is

$$\text{corr}(h(X), h(Y)) = \frac{\sum_{i=1}^n (h(x_i) - \overline{h(X)})(h(y_i) - \overline{h(Y)})}{\sqrt{\sum_{i=1}^n (h(x_i) - \overline{h(X)})^2 \sum_{i=1}^n (h(y_i) - \overline{h(Y)})^2}}, \quad (9)$$

where $\overline{h(X)}$ and $\overline{h(Y)}$ are the mean vectors of the hidden representations of the two views. $h(x_i)$ and $h(y_i)$ are hidden layer representations of a single modal view.

Integrate all objective functions to build a total objective function, which is expressed as follows:

$$L(\theta) = \sum_{i=1}^5 L_i - \sum_{j=6}^7 L_j, \quad (10)$$

where θ is the model parameter. The above formula minimizes self-reconstruction and cross-reconstruction and maximizes the association between views.

5. Experiment and Analysis

5.1. Experimental Dataset. In order to verify the performance of the proposed method, the effectiveness of the method is verified on three commonly used real cross-modal graphic retrieval datasets: Flickr30K dataset, MSCOCO dataset, and Pascal VOC 2007 dataset.

- (1) Flickr30K: The Flickr30K dataset contains 31,783 images, and the English description of the images is 158,915 sentences. That is, each image corresponds to 5 sentences with different description sentences. The sentence descriptions of these images are obtained through manual annotation. The Flickr30K dataset is divided into three parts: 1000 images and corresponding descriptions as the verification dataset, 1000 images and corresponding descriptions as the test dataset, and the remaining part as the training dataset.
- (2) MSCOCO: The MSCOCO dataset contains 123287 images, and each image also corresponds to 5 different description sentences. This dataset is divided into four parts, including 82783 images as the training dataset, 5000 images as the verification dataset, 5000 as the test dataset, and 30504 images as the reserved dataset.
- (3) Pascal VOC 2007: The Pascal VOC 2007 dataset contains 5011 image-annotation pairs for training and 4952 image-annotation pairs for testing, all from the Flickr website. Each sample pair is labeled as one of 20 semantic categories. This dataset is randomly divided into three subsets: training set, test set, and validation set, which contain 800, 100, and 100 samples, respectively.

The experimental running environment is a PC configured with Intel Core i7-7700 CPU and Nvidia GTX1070Ti 8G video memory GPU. The deep learning framework used is PyTorch, and the development language is Python.

5.2. Performance Index and Comparison Method. The evaluation indexes commonly used in the cross-modal retrieval field are selected to compare and analyse the proposed methods: the mean average precision (MAP) and the precision-recall (PR) curve. Among them, MAP can effectively evaluate the experimental results through the positions of positive samples and negative samples in the search results. AP represents the average accuracy of each specific search, calculated as follows:

$$\text{AP} = \frac{\sum_{k=1}^n (P(k) \times \varphi(k))}{N}, \quad (11)$$

where N represents the total number of search results that belong to the same semantic category as the query. n is the number of all results returned by the search. k is the position index in the search result sequence. $P(k)$ is the accuracy of the first k search. $\varphi(k)$ indicates whether the k th search result and the query have the same semantic category (the same value is 1, and the value is 0 if they are different).

The value of MAP is the average of AP values corresponding to multiple searches:

$$\text{MAP} = \frac{\sum_{q=1}^Q \text{AP}(q)}{Q}, \quad (12)$$

where Q represents the total number of searches.

Use MAP@R to indicate that given a query, sort the top R results with the highest similarity according to the similarity. The accuracy of these R results was averaged:

$$\text{MAP@R} = \frac{\sum_{k=1}^R P(k) \times \varphi(k)}{N}. \quad (13)$$

The PR curve is the curve of the accuracy rate changing with the recall rate, which is used as the performance evaluation index in cross-modal retrieval.


In the experiment, the three selected datasets have two modes: image and text. This model is compared with the reference model on two retrieval tasks, namely, retrieving text with images and retrieving images with text. For example, when retrieving images based on text, the proposed method selects each text in the test set to retrieve all images in the test set and finally obtains the retrieval result.

In order to verify the effectiveness of the proposed method, it is compared with two classical methods: CCA and deep hashing method. The corresponding research is a multilabel kernel canonical correlation analysis (ml-KCCA) method proposed in [15] and a cross-modal hashing retrieval method (DNDCMH) proposed in [22]. In addition, in order to highlight the effectiveness of the interactive learning CAE model proposed in this paper, it is compared with other methods based on the CAE model, such as the text retrieval method based on multimodal semantic automatic encoder (SCAE) proposed in [28].

5.3. Cross-Modal Retrieval Example

5.3.1. Image-Text Retrieval Analysis. The image-text retrieval results obtained by the proposed method and [22] retrieval method are shown in Table 1. It is the text retrieval

TABLE 1: Comparison of text retrieval.

| Retrieving images | Methods | Text retrieval results (top 5) |
|---|---------------------------------|---|
|  | Reference [22] (Flickr30K) | <ol style="list-style-type: none"> 1. A man wearing a black sweater cook food in a pan while standing in a cluttered kitchen. 2. A man cooking food on the stove. 3. A man is cooking on a stove in a kitchen, using wooden utensil. A cook is posing for a camera while cooking. Man with a white T-shirt and black rimmed glasses cooking a pot of food on the stove. |
| | The proposed method (Flickr30K) | <ol style="list-style-type: none"> 1. A man preparing food in his kitchen. 2. A man wearing a black sweater cook food in a pan while standing in a cluttered kitchen. 3. A man cooking food on the stove. 4. A man is cooking on a stove in a kitchen, using wooden utensil. A man stirring a pot of liquid in this kitchen. |

result of the image on the Flickr30K test set. The text in bold is the correct recall text, and the text without bold is the wrong recall text.

It can be seen from Table 1 that the proposed method has better retrieval results in terms of recall index. Specifically, in the text retrieval task, the proposed method uses image search to find the correct text sorting more advanced. These visually presented phenomena more intuitively illustrate the effectiveness of the proposed method. In [22], DNDCMH is used to achieve text retrieval. Due to the lack of image feature extraction effect, the correct text is less.

5.3.2. Text-Image Retrieval Analysis. In order to compare the performance of the proposed method and the comparison method [15, 22, 28], in text-image retrieval, the ‘car’ is used as the query text to retrieve the image on the Pascal VOC 2007 dataset. The top 5 images retrieved by various methods are shown in Figure 5.

It can be seen from Figure 5 that compared to other comparison methods, the text retrieval results of the proposed method are more reasonable. Since the proposed method uses word2vec and LSTM network for text feature extraction, the extraction effect is better. Therefore, the retrieval images obtained through the CAE network of interactive learning are more accurate.

5.4. Performance Comparison. In order to demonstrate the retrieval performance of the proposed method in the three datasets, it is compared with the methods in [15, 28] and [22]. The MAP values of the first 50 results of the four methods are shown in Table 2.

It can be seen from Table 2 that, in the two retrieval tasks of retrieving images by text and retrieving text by images, the proposed method has significantly improved MAP on these three datasets compared with other comparison methods. Since the Pascal VOC 2007 dataset has the largest magnitude, the proposed method has the most significant improvement on Pascal VOC 2007. On Flickr30K, MSCOCO, and Pascal VOC 2007, three cross-modal graphic retrieval domain datasets, the average MAP on the two retrieval tasks of the proposed method are 0.359, 0.334, and 0.309, respectively.

Compared with [15], it increased by 58.85%, 44.59%, and 58.46%; compared with [28], by 14.14%, 9.57%, and 10.69%; and compared with [22], by 16.56%, 12.46%, and 24.10%.

In addition, with different methods on the Flickr30K dataset, the PR curves for two different retrieval tasks of image retrieval and text retrieval are shown in Figure 6. The ordinate represents the precision, and the abscissa represents the recall. Similarly, the PR curves of two different retrieval tasks on MSCOCO and Pascal VOC 2007 datasets with different methods are shown in Figures 7 and 8, respectively.

It can be seen from Figure 6 that whether it is image retrieval text or text retrieval image, the area of the PR curve of the proposed method is larger than other comparison methods. Because it adopts the cross-modal retrieval method of image and text interactive CAE and incorporates 2DCPA into the feature extraction, the accuracy of retrieval is improved. Reference [15] proposed a ml-KCCA method to achieve cross-mode retrieval, but the retrieval performance is low due to poor feature extraction. Reference [28] combined low-level features and high-level semantic information to learn feature representation. Although it solves the problem of feature representation, due to the lack of feature interaction, the retrieval accuracy for complex environments still needs to be improved. Reference [22] used the DNDCMH method to complete cross-modal retrieval. However, this method has poor universality, so the retrieval performance is inferior to the proposed method.

It can be seen from Figure 7 that the retrieval performance of the proposed method is better than other comparison methods in the two retrieval tasks of image retrieval text and text retrieval image. When the recall is 0.2, the accuracy of each method reaches the maximum, and the recall increases and decreases continuously. Since the MSCOCO dataset has relatively few samples, the area composed of PR curves of different methods has increased compared to the Flickr30K dataset.

It can be seen from Figure 8 that, like the first two datasets, the retrieval performance of the proposed method on the Pascal VOC 2007 dataset is better than other comparison methods. The proposed method uses the residual network to extract image features and introduces 2DPCA to construct a new convolution kernel. At the same time, using

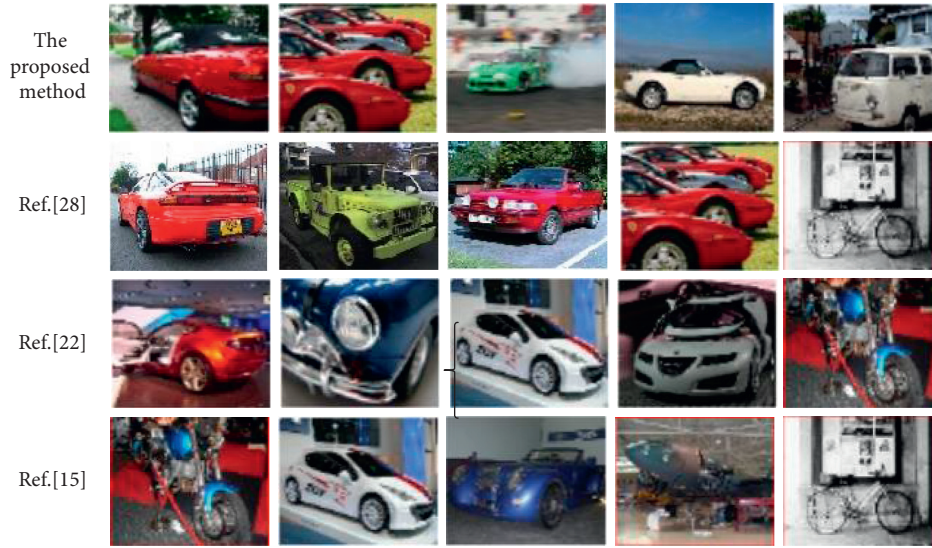


FIGURE 5: An example of image retrieval with text “car.”

TABLE 2: MAP ($R = 50$) values of different methods on three datasets.

| Datasets | Methods | MAP values | | |
|-----------------|---------------------|-------------|------------|---------|
| | | Image query | Text query | Average |
| Flickr30K | Reference [15] | 0.215 | 0.237 | 0.226 |
| | Reference [28] | 0.304 | 0.312 | 0.328 |
| | Reference [22] | 0.281 | 0.335 | 0.308 |
| | The proposed method | 0.338 | 0.379 | 0.359 |
| MSCOCO | Reference [15] | 0.198 | 0.264 | 0.231 |
| | Reference [28] | 0.293 | 0.319 | 0.301 |
| | Reference [22] | 0.275 | 0.318 | 0.297 |
| | The proposed method | 0.324 | 0.343 | 0.334 |
| Pascal VOC 2007 | Reference [15] | 0.192 | 0.198 | 0.195 |
| | Reference [28] | 0.279 | 0.295 | 0.262 |
| | Reference [22] | 0.251 | 0.247 | 0.249 |
| | The proposed method | 0.306 | 0.311 | 0.309 |

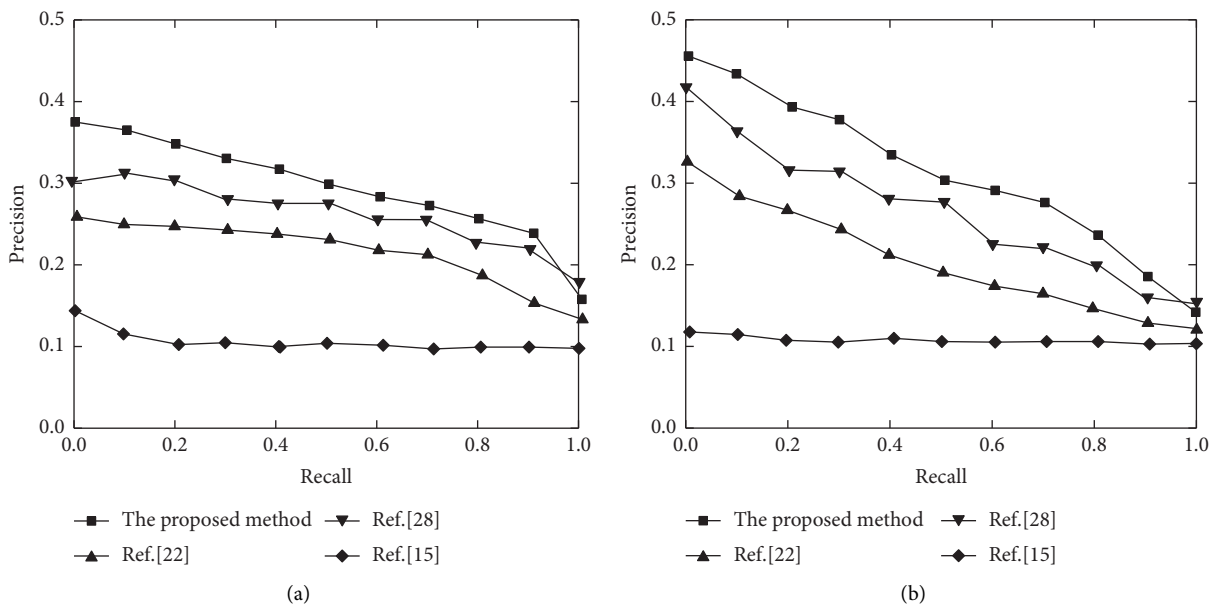


FIGURE 6: PR curves on Flickr30K datasets. (a) Retrieving text with images. (b) Retrieving images with text.

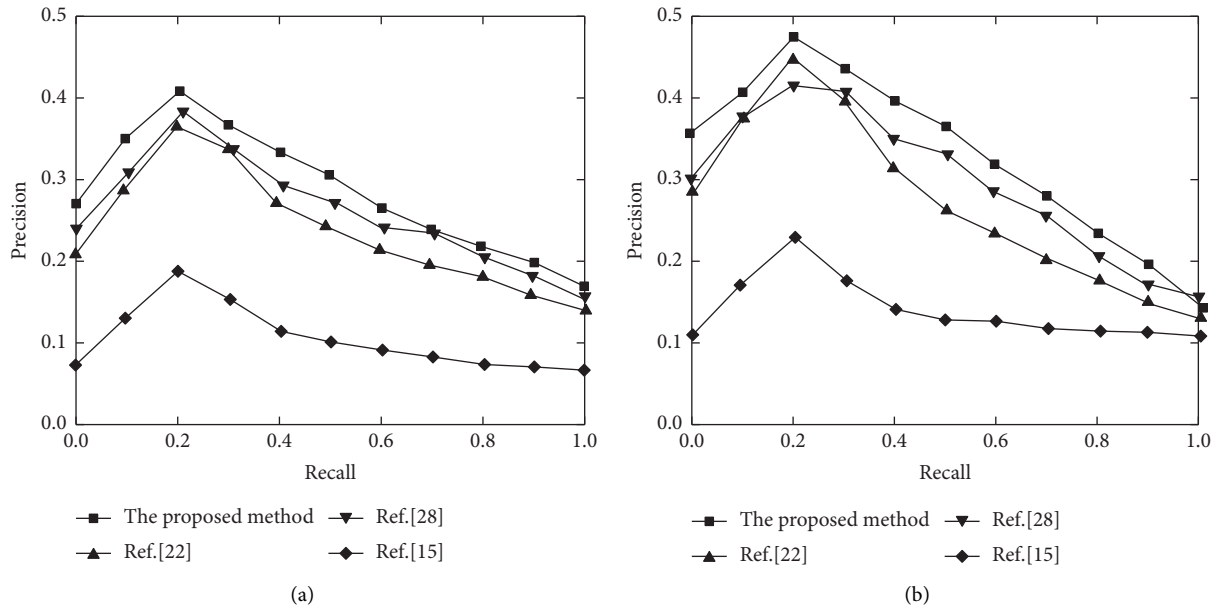


FIGURE 7: PR curves on the MSCOCO datasets. (a) Retrieving text with images. (b) Retrieving images with text.

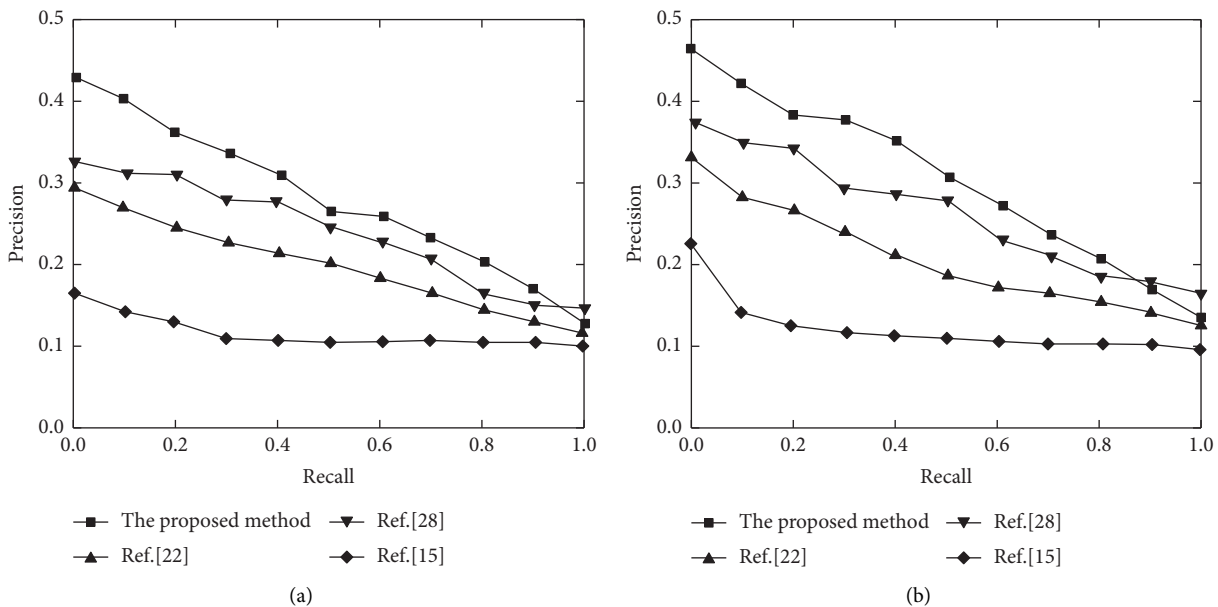


FIGURE 8: PR curves on Pascal VOC 2007 datasets. (a) Retrieving text with images. (b) Retrieving images with text.

word2vec and LSTM network for text feature extraction, feature extraction is more efficient. It is better than [15] using existing label information and [22] using specific images. In addition, [28] used the semantic CAE method to learn multimodal mapping and projected multimodal data into low dimensional space to retain feature and semantic information and improve retrieval accuracy. However, the proposed method uses the CAE model with interactive learning, and the fusion effect of image and text feature learning is better, so the retrieval performance is more ideal.

In summary, it can be seen from the PR curves on different datasets that the proposed method shows the best

results under different recall. This proves that the deep interactive learning method constructed by it is effective.

6. Conclusion

Cross-modal retrieval technology meets people's more diverse retrieval needs and solves the problems of heterogeneous gap and semantic gap between different modal data. However, the retrieval accuracy still needs to be improved. For this reason, a cross-modal image retrieval method combining efficient feature extraction and interactive learning CAE is proposed. The residual network convolution

kernel is improved by incorporating 2DPCA to extract image features, and text features are extracted through LSTM and word vectors to obtain image and text features. After that, the two features are input into the cross-modal CAE of interactive learning, and through the interactive learning of the middle layer, the image-text retrieval is realized. In addition, the proposed method is experimentally demonstrated based on the Flickr30K, MSCOCO, and Pascal VOC 2007 datasets. The results show that the proposed method can complete accurate image retrieval and text retrieval. Moreover, the average MAP on the two retrieval tasks is 0.359, 0.334, and 0.309, which are higher than other comparison methods. The same is true for the area formed by the PR curve.

At present, the method proposed in this paper is only suitable for cross-modal retrieval between text and image, but there are many types of multimodal data on the network. Next, more data of different media types such as audio and video will be expanded to meet people's broader retrieval needs.

Data Availability

The data included in this paper are available without any restriction.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61402350, 61103143, U1404620, and U1404622), the Key Scientific and Technological Project of Henan Province (182102310034, 172102310124, and 212102210400), and the Key Research Projects of Henan Provincial Department of Education (20A520046).

References

- [1] L. Ma, H. Li, F. Meng, Q. Wu, and K. N. Ngan, "Global and local semantics-preserving based deep hashing for cross-modal retrieval," *Neurocomputing*, vol. 312, no. 10, pp. 49–62, 2018.
- [2] Y. Wu, S. Wang, and Q. Huang, "Online fast adaptive low-rank similarity learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1310–1322, 2020.
- [3] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6699–6711, 2018.
- [4] J. Zhang and Y. Peng, "Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 174–187, 2020.
- [5] Y. Tao, K. Xiangwei, Y. Lianshan, T. Wenjing, and T. Qi, "Efficient discrete supervised hashing for large-scale cross-modal retrieval," *Neurocomputing*, vol. 385, no. 04, pp. 358–367, 2020.
- [6] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 29, no. 02, pp. 3626–3637, 2020.
- [7] X. Yuan, H. Zhong, Z. Chen, F. Zhong, and Y. Hu, "Multimedia feature mapping and correlation learning for cross-modal retrieval," *International Journal of Grid and High Performance Computing*, vol. 10, no. 3, pp. 29–45, 2018.
- [8] P. Hu, D. Peng, X. Wang, and Y. Xiang, "Multimodal adversarial network for cross-modal retrieval," *Knowledge-Based Systems*, vol. 180, pp. 38–50, 2019.
- [9] H. Chen, G. Ding, Z. Lin et al., "ACMNet: adaptive confidence matching network for human behavior analysis via cross-modal retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 1, pp. 1–21, 2020.
- [10] C. Ushasi, B. Biplab, B. Avik, and D. Mihai, "A deep learning based model for cross-modal retrieval in remote sensing," *Pattern Recognition Letters*, vol. 131, no. 03, pp. 456–462, 2020.
- [11] G. Guanghai, L. Jiangtao, L. Zhuoyi, H. Wenhua, and Z. Yao, "Joint learning based deep supervised hashing for large-scale image retrieval," *Neurocomputing*, vol. 385, no. 93, pp. 348–357, 2020.
- [12] Y. Fang, H. Zhang, and Y. Ren, "Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing," *Knowledge-Based Systems*, vol. 171, no. 05, pp. 69–80, 2019.
- [13] C. Jing-Jing, P. Lei and N. Chong-Wah, "Cross-modal recipe retrieval with stacked attention model," *Multimedia Tools and Applications*, vol. 77, no. 22, Article ID 29457, 2018.
- [14] A. Jiang, H. Li, Y. Li, and M. Wang, "Learning discriminative representations for semantical crossmodal retrieval," *Multimedia Systems*, vol. 24, no. 1, pp. 111–121, 2018.
- [15] Y. Jia, L. Bai, S. Liu, P. Wang, J. Guo, and Y. Xie, "Semantically-enhanced kernel canonical correlation analysis: a multi-label cross-modal retrieval," *Multimedia Tools and Applications*, vol. 78, no. 10, Article ID 13169, 2019.
- [16] S. Wang and Z. Shi, "Cross-media semantic retrieval with deep canonical correlation analysis," *Journal of University of Science and Technology of China*, vol. 48, no. 4, pp. 322–330, 2018.
- [17] Y. Liu, K. Cai, C. Liu, and F. Zheng, "Csrncva: a model of cross-media semantic retrieval based on neural computing of visual and auditory sensations," *Neural Network World*, vol. 28, no. 4, pp. 305–323, 2018.
- [18] J. Shao, Z. Zhao, and F. Su, "Two-stage deep learning for supervised cross-modal retrieval," *Multimedia Tools and Applications*, vol. 78, no. 12, Article ID 16615, 2019.
- [19] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, "Category-based deep CCA for fine-grained venue discovery from multimodal data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 1250–1258, 2019.
- [20] Q. Wu, "Image retrieval method based on deep learning semantic feature extraction and regularization softmax[J]," *Multimedia Tools and Applications*, vol. 79, no. 13, pp. 9419–9433, 2020.
- [21] Z. Xie, L. Li, X. Zhong, Y. He, and L. Zhong, "Enhancing multimodal deep representation learning by fixed model reuse," *Multimedia Tools and Applications*, vol. 78, no. 21, Article ID 30769, 2019.
- [22] F. Taherkhani, V. Talreja, M. C. Valenti, and N. M. Nasrabadi, "Error-corrected margin-based deep cross-modal hashing for

- facial image retrieval,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 3, pp. 279–293, 2020.
- [23] Z. Xi, L. Hanjiang, and F. Jiashi, “Attention-aware deep adversarial hashing for cross-modal retrieval,” in *Proceedings of the European Conference on Computer Vision*, pp. 614–629, Munich, Germany, September 2018.
- [24] G. Wu, J. Han, Z. Lin, G. Ding, B. Zhang, and Q. Ni, “Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning,” *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9868–9877, 2019.
- [25] M. Zhang, H. Zhang, J. Li, Y. Fang, L. Wang, and F. Shang, “Multi-modal graph regularization based class center discriminant analysis for cross modal retrieval,” *Multimedia Tools and Applications*, vol. 78, no. 19, Article ID 28285, 2019.
- [26] M. Zareapoor, J. Yang, D. K. Jain, P. Shamsolmoali, N. Jain, and S. Kant, “Deep semantic preserving hashing for large scale image retrieval,” *Multimedia Tools and Applications*, vol. 78, no. 17, Article ID 23831, 2019.
- [27] F. Feng, X. Wang, and R. Li, “Cross-modal retrieval with correspondence autoencoder,” in *Proceedings of the ACM International Conference on Multimedia*, pp. 7–16, New York, NY, USA, November 2014.
- [28] Y. Wu, S. Wang, and Q. Huang, “Multi-modal semantic autoencoder for cross-modal retrieval,” *Neurocomputing*, vol. 331, no. 28, pp. 165–175, 2019.
- [29] A. Eisenschtat and L. Wolf, “Linking image and text with 2-way nets,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.
- [30] Y. Huang, Q. Wu, and L. Wang, “Learning Semantic Concepts and Order for Image and Sentence matching,” 2017, <https://arxiv.org/abs/1712.02036>.
- [31] C. Zhou, L. M. Po, W. Y. F. Yuen et al., “Angular deep supervised hashing for image retrieval,” *IEEE Access*, vol. 7, no. 99, Article ID 127521, 2019.
- [32] G. Wu, J. Han, Y. Guo et al., “Unsupervised deep video hashing via balanced code for large-scale video retrieval,” *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1993–2007, 2018.
- [33] S. Basavaraju and A. Sur, “Multiple instance learning based deep CNN for image memorability prediction,” *Multimedia Tools and Applications*, vol. 78, no. 24, Article ID 35511, 2019.
- [34] Y. Li, L. Wan, T. Fu, and W. Hu, “Piecewise supervised deep hashing for image retrieval,” *Multimedia Tools and Applications*, vol. 78, no. 17, Article ID 24431, 2019.
- [35] Y. Gu and J. Yang, “Densely-connected multi-magnification hashing for histopathological image retrieval,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1683–169, 2019.
- [36] H. Hu, K. Wang, C. Lv, J. Wu, and Z. Yang, “Semi-supervised metric learning-based anchor graph hashing for large-scale image retrieval,” *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 739–754, 2018.