Computer Systems Science & Engineering
DOI:10.32604/csse.2021.014563
Article

Tech Science Press

Cross-Modal Hashing Retrieval Based on Deep Residual Network

Zhiyi Li[1,2,*], Xiaomian Xu[2], Du Zhang[1] and Peng Zhang[2]

[1]Faculty of Information Technology, Macau University of Science and Technology, Macau
[2]School of Economics and Management, South China Normal University, Guangzhou, 510006, China
[*]Corresponding Author: Zhiyi Li. Email: leeds@scnu.edu.cn
Received: 28 September 2020; Accepted: 06 November 2020

Abstract: In the era of big data rich in We Media, the single mode retrieval system has been unable to meet people's demand for information retrieval. This paper proposes a new solution to the problem of feature extraction and unified mapping of different modes: A Cross-Modal Hashing retrieval algorithm based on Deep Residual Network (CMHR-DRN). The model construction is divided into two stages: The first stage is the feature extraction of different modal data, including the use of Deep Residual Network (DRN) to extract the image features, using the method of combining TF-IDF with the full connection network to extract the text features, and the obtained image and text features used as the input of the second stage. In the second stage, the image and text features are mapped into Hash functions by supervised learning, and the image and text features are mapped to the common binary Hamming space. In the process of mapping, the distance measurement of the original distance measurement and the common feature space are kept unchanged as far as possible to improve the accuracy of Cross-Modal Retrieval. In training the model, adaptive moment estimation (Adam) is used to calculate the adaptive learning rate of each parameter, and the stochastic gradient descent (SGD) is calculated to obtain the minimum loss function. The whole training process is completed on Caffe deep learning framework. Experiments show that the proposed algorithm CMHR-DRN based on Deep Residual Network has better retrieval performance and stronger advantages than other Cross-Modal algorithms CMFH, CMDN and CMSSH.

Keywords: Deep residual network; cross-modal retrieval; hashing; cross-modal hashing retrieval based on deep residual network

1 Introduction

The development of network communication technology and network social media has brought massive network multimedia information, which makes the information retrieval work face great challenges. Most of the existing search engines are limited to the retrieval of single-modal data, that is, they realize the functions of text-to-text and image-to-image retrieval. This single-modal retrieval method cannot meet users' requirements of cross-modal retrieval, such as text search for images and text search for audio [1]. To solve this problem, scholars have carried out the research of cross-modal retrieval. Cross-modal retrieval

first performs feature extraction on the underlying information of different modalities, integrates the features of different modal information into feature vectors that are easy to calculate, and models the relationship between different modal information to achieve retrieval matching. How to establish an association representation between different modal data is the main challenge faced by cross-modal retrieval.

Methods based on subspace learning are often used in cross-modal association modeling. In this method, the data of different modalities are mapped into a common subspace, and the similarity is measured by calculating the spatial distance between the data of different modalities. In the early stage, statistical Correlation Analysis was mainly used, and the representative algorithm is Canonical Correlation Analysis (CCA). For example, Rasiwasla et al. [2] used CCA algorithm to map data of different modalities into a common subspace, and realized cross-modal retrieval by calculating the subspace distance. However, the model based on the typical correlation analysis method is a one-to-two-layer shallow model, which cannot learn the deep semantics of multimodal data well. Moreover, only the correlation between paired samples is learned, and sample class information, such as paired constraint information and class label information, is not utilized. In addition, the features learned by canonical correlation analysis are linear, which is not suitable for learning nonlinear features.

The proposal and development of deep learning in recent years can enable deep neural network to extract more fine-grained multi-modal data features and improve the performance of cross-modal retrieval. For example, the deep Autoencoder model [3,4] is built, and the deep neural network is used to conduct feature learning for image and text data respectively. In the process of learning different modal features, association constraints are carried out, and the correlation between image features learned by the Autoencoder and text features is established to achieve the unified representation of different modalities. Compared with CCA model, the structure of the Autoencoder is relatively simple, so the training speed of the model is faster, the time consumption is shorter and the performance is better. At the same time, literature [5,6] introduced the idea of generative confrontation into cross-modal retrieval, and learned better and more effective public subspace representation by using the deep generative confrontation networks. Deep learning promotes the performance of cross-modal retrieval, but when processing large-scale multimedia data, it will face the problems of insufficient storage space caused by excessive dimensionality, excessive calculation and reduced retrieval efficiency. In view of this, the literatures [7–9] applied the Hash method to Cross-Modal Retrieval. The Hash method uses the binary low-dimensional vector composed of “+1” and “–1” to represent and store data, and maps the multi-modal data to the public Hamming space of the binary hash code. Using the Hash method can greatly reduce storage space, reduce the amount of calculation, and improve the efficiency of Cross-Modal Retrieval. However, the existing Cross-Modal Hashing methods such as CMSSH [7] and CMFH [8] are shallow models and rely on traditional manual design features, which are not universal and costly. Inspired by deep learning, people combine deep learning with hash learning. However, as the depth of the model deepens, the accuracy will gradually saturate, and then rapidly degrade [10]. Experiments show that [11,12] adding more layers in a model of appropriate depth will lead to higher training error and reduce retrieval accuracy.

In order to solve these problems, this paper proposes a Cross-Modal Hash retrieval algorithm based on Deep Residual Network (CMHR-DRN) [10]. This algorithm uses the Deep Residual Network to extract the features of images and texts, and maps them uniformly to the public Hamming space, so as to realize the Cross-Modal Retrieval. At the same time, the problem of network convergence caused by the increase of the number of network layers can be reduced by using the Deep Residual Network. Compared with the shallow Hash cross-mode retrieval model, our model can achieve better retrieval effect and is suitable for large-scale data retrieval.

2 Literature Review

At present, a lot of research results have been achieved in the field of Cross-Modal Retrieval, and a large number of open algorithms have been proposed, among which the more typical ones include: the CCA algorithm in the associated learning stage, the Word2Vec algorithm in text object modeling, and the Deep Residual Network (DRN) algorithm, etc. [1]. From the perspective of cross-modal data association modeling strategies, early people used shared layer and subspace learning methods to achieve data association modeling. With the development of Hash methods, many Hash modeling methods have also appeared in recent years. This article will introduce the following three aspects:

2.1 Establish the Association between Image and Text Data Based on the Shared Layer

A typical representative is a content-based automatic interaction system model proposed by Amir et al. [13], which proposes a learning algorithm based on feature extraction for images, establishes a multimodal semantic retrieval system for video content, and uses a 0–1 loss function in the algorithm to estimate the degree to which the predicted value of the model differs from the true value. Subsequently, Zheng et al. [14] applied the Cross-Modal Retrieval technology between text and image to automatic annotation of biological images, and established an automatic annotation and retrieval model for cell migration. Jia et al. [15] proposed a model combining Markov Random Field and LDA. This model regards topic distribution as a shared layer of images and texts, and optimizes *corpus* that is not closely related to images and texts. Zhong et al. [16] used a Bagging method consisting of multiple SVMs to map images and text. The advantages of this method are simplicity, easy implementation of the algorithm, but the disadvantage is that the accuracy rate is low and there is room for greater accuracy improvement. However, due to the heterogeneity between different modal data, the Cross-Modal Retrieval model constructed by this strategy is difficult to fully and effectively learn the association between different modal data. LDA needs to be trained on a larger *corpus*, and the quality of the *corpus* seriously affects the effectiveness of text features. The Cross-Modal Retrieval based on the shared layer has not yet involved deep learning algorithms, and the association of cross-modal data is still in the stage of conception and exploration, the modeling effect is not ideal, and a standardized cross-modal retrieval model has not yet been formed.

2.2 Image-Text Retrieval Model Based on Subspace Learning Method

Methods based on subspace learning can be divided into statistical correlation analysis methods and deep learning methods.

(1) Statistical Correlation Analysis Method

Among the statistical association analysis methods, the CCA method, as a classic method, is widely used in computer vision, natural language processing and other fields, but it also has obvious shortcomings. To this end, Peng et al. [17] proposed a semi-supervised canonical correlation analysis method (Semi-CCA), which improves classification performance by introducing supervised information given in the form of paired constraints. Borges et al. [18] proposed the concept of Manifold Learning, which uses a non-linear feature learning method to reduce dimensionality through the technique of Local Linear Embedding. Akaho et al. [19] combined nuclear technology with CCA and proposed a nuclear CCA technology that can realize nonlinear feature learning. Rasiwasia et al. [2] mapped the underlying image features and the text topic distribution features obtained through deep learning methods into the same space, and established a semantic-based deep learning cross-modal retrieval model. After that, CCA developed slowly in the field of cross-modal retrieval, without effective improvement and development.

Since 2010, Chandrika et al. [20] proposed LSI algorithm optimized for image retrieval, and built a multimodal latent semantic probability model, which effectively improves the accuracy of cross-modal

image retrieval. Lin et al. [21] proposed a new multimodal integration and extension model (MMIP) based on PLSA and found that it can effectively improve the recall rate of image cross-modal retrieval through experiments. Wang et al. [22] proposed a multi-modal subspace learning algorithm (JGRMSL) based on the regularization of related images for the learning problem of hidden space. Zhuang et al. [23] proposed a multi-modal retrieval system with supervised learning organization structure, which further improved the cross-modal retrieval image retrieval effect of supervised learning. Chen et al. [24] based on the pairing of recipes and food photos, discovered the details of food pictures through fine-grained search, and learned a joint space to locally capture the correspondence between the image and the recipe. Since the image learning is carried out at the regional level, and the recipe learning is carried out at the ingredient level, the model can generalize the recognition to invisible food categories. The advantage is to start with fine-grained, which enhances the generalization ability of recognition. The disadvantage is that the area captured image causes ambiguity and increases the error rate.

(2) Deep Learning Methods

The proposal of deep neural network brings new development opportunities for cross-modal retrieval. By building a multi-layer neural network, we can extract more fine-grained features from pictures and texts. Ngiam et al. [3] proposed a Bimodal Deep Autoencoder model based on joint representation. First, the vector method is used to learn and represent the characteristics of different modalities, and then the association models of different modalities are constructed for unified representation. Feng et al. [4] proposed a Corr-AE model (Correspondence Autoencoder) with association constraints. When learning each single modal feature, a unified constraint algorithm is used to make the learned features of various single modalities relevant united relationship.

Kim et al. [25] proposed the combination of DAE and CCA to achieve cross-modal information matching, which provides a good theoretical foundation and research direction for the combination of deep learning algorithms and statistical correlation algorithms. Subsequently, Verma et al. [26] proposed a new structured support vector machine (SVM), which built a unified framework for img2text and text2img, and proved the effectiveness of this method in cross-modal retrieval of images and text by training datasets within the network size. Compared with other algorithm models, this algorithm has achieved better results. It can not only be applied to two modalities of image and text, but also can be extended to retrieval of other modalities.

Wang et al. [27] proposed a distributed training platform (SINGA) that supports effective large-scale deep learning model training. In order to construct effective image modal and text modal mapping functions, unsupervised learning algorithm Stacked Auto-Encoder (SAE) and supervised learning algorithm deep convolutional neural network (DCNN), neural language model are used respectively (NLM) to learn the mapping function. Ding et al. [28] used LDA and BOW models as the feature expression methods of text and image resources, and then used the least square method to learn feature subspace projection functions to improve the accuracy of cross-modal information retrieval. Dutta et al. [29] proposed a unified framework that can handle all challenging scenarios from different sources without any modification. This method projects the data in different modes into a common semantic feature space, retains the semantic relationship given by the class name embedding (attribute), and finds and retrieves similar data. Its advantages are unified framework and easy implementation; its disadvantages are that compatibility and accuracy are difficult to choose from. Wu et al. [30] proposed a new convolutional neural network data representation method for representing different forms of data. And learn the CNN model of each modal data, map different modal data to a public space, and regularize the new representation in the public space through a cross-model correlation matrix. The advantage of this method is that the method is unified, and the learning problem is summarized as a minimization problem; the disadvantage is that the unified data format limits the versatility of the method.

In recent years, the rapid development of the generation of antagonistic networks has also attracted the attention of researchers. Generative adversarial networks have strong data distribution fitting capabilities, and are especially good at generating continuous data such as images, which helps to build more effective public subspace representations. Gu et al. [5] proposed GXN (generative cross-modal learning framework) generative cross-modal feature learning framework, using image-text and text-image two generative models, not only can learn high-level global abstract representation, It can also learn the local underlying feature representation. He et al. [6] proposed an unsupervised generative confrontation cross-modal retrieval model ACMR, which used the generative confrontation mechanism in cross-modal feature association. The generator is used to form text-image sample pairs with the same semantics in the common subspace. The discriminator judges which modal the samples of the sample pairs in the common subspace come from, and learns an effective common subspace representation through adversarial training.

2.3 Image-Text Retrieval Model Based on Hash Algorithm

As an efficient retrieval algorithm, the Hash algorithm has a long history of development. The Hash method was originally used for Approximate Nearest Neighbor search, that is, to map data to a Hash table through a Hash function, and use the index of the Hash table to achieve retrieval. With the expansion of data scale, Indyk et al. [31] proposed a Local Sensitive Hashing (LSH) method. This method is widely used in single-modal retrieval. Inspired by the LSH algorithm, people apply the hash algorithm to cross-modal retrieval, and have proposed many excellent algorithms. For example, the CMSSH (Cross Modal Similarity Sensitive Hashing) algorithm proposed by Bronstein et al. [7] establishes a supervised hash learning framework, which maps input data from any two spaces into a common Hamming space based on similar semantic hash (SSH) [32], and learns the similarity between different modes through the Boosting algorithm. Li et al. [33] proposed a hash algorithm based on subspace ordering, which learns two sets of linear subspaces together, one set for each modal data, and maintains maximum similarity between cross-modes according to the ordering order of features in different subspaces. The Collective Matrix Factorization Hashing algorithm (CMFH) proposed by Ding et al. [8] uses a latent factor model to learn a unified hash code from different modes of an instance through integrated matrix factorization, which not only supports Cross-view search, and can improve retrieval accuracy by combining multiple view information sources.

However, most of the above-mentioned hash algorithms adopt manual design features, which are costly and prone to errors. With the development of deep learning, people have found that deep neural networks have strong data fitting capabilities and perform well when processing large-scale images, texts and other data. So people combined deep learning and hashing methods, and proposed some deep hash cross-modal retrieval algorithms.

(1) Supervised Deep Hash

The supervised deep hash method learns the hash function through the semantic tags of the data set, which can achieve more accurate retrieval results. Jiang et al. [34] proposed Deep Cross-Modal Hashing (DCMH), which puts feature learning and hash code learning in the same end-to-end learning framework, and maintains the similarity between the learned hash functions and the original data pairs through paired labels. Li et al. [35], proposed the self-supervised adversarial hashing (SSAH) algorithm, which incorporates antagonistic learning into cross-modal hashing and uses two antagonistic networks to maximize the semantic correlation and consistency of representations between different modes. At the same time, a self-supervised semantic network is used to discover high-level semantic information in the form of multi-label annotations. Deng et al. [36] proposed a triple-based deep hashing (TDH) network based on triplet, who uses triplet tags to obtain more semantic association information between modalities; it also uses graph regularization to maintain the semantic similarity between the hash code and the original data.

(2) Unsupervised Deep Hash

The data set of the unsupervised deep hash method has no labels, and retrieval is realized by learning and mining the structural associations between the data, which is suitable for cross-modal retrieval of large-scale data. Su et al. [37] proposed a Deep Joint-Semantics Reconstructing Hashing (DJSRH) algorithm, constructed a joint semantic association matrix containing different modal primitive domain information, used to capture potential semantic association information between multimodal data, and trained the network to generate binary hash codes using the reconstruction framework. Wang et al. [38] proposed an Unsupervised Deep Cross-modal Hashing with Virtual Label Regression (UDCH-VLR) algorithm, which is a unified framework for deep hashing Function training, virtual label learning and regression. The unified hash code is learned by decomposing the collaborative matrix of the multi-modal deep representation, maintaining the multi-modal shared semantics, and integrating the virtual label learning into the objective function, and returning the learned virtual label to the hash code. Hoang et al. [39] proposed Deep Cross-modality Spectral Hashing (DCSH) algorithm. It first uses algorithms based on spectral hashing to learn single modal and binary cross-modal representations, and uses deep convolutional networks to map text and images into binary hash codes. Wu et al. [40] proposed depth generation cross-modal hash algorithm, which introduces a loss of circular consistency to learn paired hash functions in antagonism training without paired training samples, and generates a hash function through a network to reduce the loss of associated information.

It is worth noting that in deep learning, increasing the depth of the network can allow model fitting to express complex functions more effectively. However, when the number of layers of the network is increased to a certain number, the training accuracy of the network will drop significantly. In order to solve this problem, He et al. [10] proposed a Deep Residual Network. That is, adding an identity mapping layer after the network so that the system error will not increase due to the increase in the number of network layers. In recent years, literatures [41–43] used deep residual networks in large-scale image retrieval, which demonstrated the excellent performance of deep residual networks in processing large-scale image data.

Inspired by this, this paper proposes a CMHR-DRN model. Aiming at the "Degradation problem" that the accuracy of image feature extraction quickly reaches saturation and tends to decline, we use the Deep Residual Network (DRN) model to adopt the method of identity mapping at the increased level, so as to better control the error of training results. Experiments show that CMHR-DRN model can improve the efficiency of unsupervised learning and the accuracy of image feature extraction.

3 Construction of Cross-Modal Hash Retrieval Model Based on Deep Residual Network

3.1 Symbolization of a Model

In order to solve the Cross-Modal Retrieval problem of image-text, we need to first mathematically describe the problem.

Firstly, it is assumed that dataset $\boldsymbol{D}$ contains two different modalities, and the eigenvector of image modality is represented as $\boldsymbol{X}$, which describes the image data in the dataset. The text modal feature vector is represented as $\boldsymbol{T}$, describing the text data corresponding to the image $\boldsymbol{X}$ in the dataset. Image data v_i^x and text data v_j^t constitute a cross-modal dataset $\boldsymbol{D}$, which can be represented by mathematical symbols as follows:

$$X = \left\{v_i^x\right\}_{i=1}^n, \quad T = \left\{v_j^t\right\}_{j=1}^n, \quad D = \left\{v_i^x, v_j^t\right\}_{i=1,\, j=1}^n \tag{1}$$

Then, the similarity between the image and the text data is described. If the image v_i^X has similarity to the text v_j^t, then define $S_{i,j} = 1$; otherwise $S_{i,j} = 0$. Cross-modal similarity matrix S is constructed.

$$S = \{S_{i,j}\}_{i=1,j=1}^{n} \tag{2}$$

$$\text{where} \quad S_{i,j} = \begin{cases} 1 & v_i^x = v_j^t \\ 0 & v_i^x \neq v_j^t \end{cases}$$

Since the retrieval belongs to the scope of supervised learning, the similarity of data can be understood as the similarity of semantic information. Commonly used semantic information is in the form of tag information. Therefore, according to the description of the similarity matrix, the similarity value of the image v_i^x and the text t v_j^t having the same type of label is set to 1; otherwise the similarity value is 0.

After the data description of the image and the text is set, the data features of different modalities are extracted. For image and text features, different depth neural networks are used to operate the two separately. The image is extracted by using the Deep Residual Network. The text is processed by the TF-IDF feature word weight calculation method and then connected to the neural network. Since feature extraction is performed by using different neural networks, different modal features are obtained. Therefore, it is necessary to map features of different modalities into a common feature space for modeling.

Cao et al. [44] have proved that constructing a Hash function can effectively improve the retrieval performance, and by mapping different features into the binary Hamming space, it can realize Cross-Modal Retrieval between large-scale images and texts. Combined with the feature representation of the image and text in this paper, the common feature space of different modalities is Hamming space, so the mapping function that needs to be determined next is a Hash function

Then for the image modal X, assume that the image Hash function it looks for is formula (3) and the text modal T is formula (4), where d is the length of the binary code.

$$h^X(X) \in \{0,1\}^d \tag{3}$$

$$h^T(T) \in \{0,1\}^d \tag{4}$$

In this way, the image is binary coded into the formula (5) by the mapping of the image Hash function, and the corresponding text code is the formula (6).

$$b_i^x = h^X\left(v_i^x\right) \tag{5}$$

$$b_j^t = h^T\left(v_j^t\right) \tag{6}$$

With the common feature space, we also need to keep the distance measure of the original distance measure and the common feature space unchanged during the mapping using the Hash function. That is, the measure of the distance between the original data and the distance between the binary codes remains unchanged. As the supervised information, the cross-modality similarity matrix S obviously plays a role in the process of maintaining the distance. If there is in the original data $S_{i,j} = 1$, the desired Hamming distance between the binary codes b_i^X and b_j^T can be as small as possible, and vice versa. In addition, it is necessary to ensure that the mapping using the Hash function is a mapping of the common feature space that maintains a constant distance, which is also the key to improving the accuracy of Cross-Modal Retrieval.

After mathematical modeling the Cross-Modal Retrieval problem, the research is carried out from two aspects in two stages. In the first stage, different deep neural networks are used to learn the features of image

data and text data to extract the eigenvalues of different modal data. In the second stage, supervised information is used to learn the Hash function to map the common space while keeping the distance metric unchanged.

3.2 Construction of Cross-Modal Retrieval Model

In this paper, the image-text cross-mode retrieval model as shown in Fig. 1 is adopted. The construction process is mainly divided into two stages: The first stage is feature extraction, including image and text feature extraction. As for image features, this paper adopts the Deep Residual Network ResNet in the convolutional neural network model to solve the problem that accuracy will not decrease with the deepening of the network. For text features, because of their sparsity, TF-IDF features are used as input, and then DCNN is connected to the deep network. The second stage is the Hash coding part. By keeping the invariant distance mapping, the image and text features can finally be encoded on Hamming Cube, so as to carry out rapid retrieval on Hamming space.

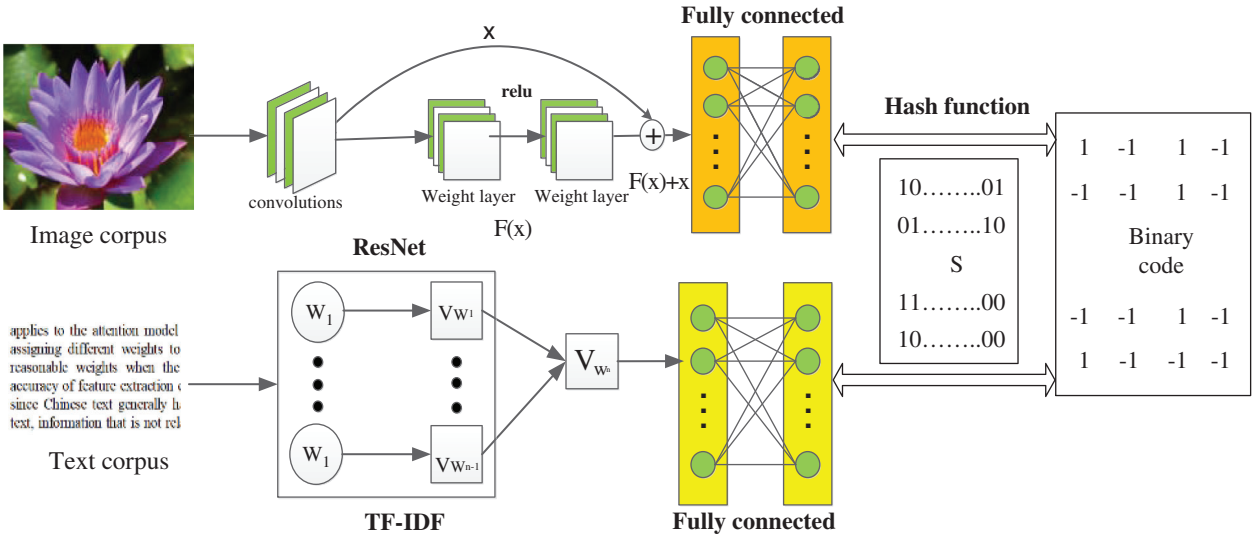

Figure 1: Image-text cross-modal retrieval model

(1) The First Stage: Model Construction of Feature Extraction

The deep learning network structure consists of two parts: one is oriented to image modal data, and the other is oriented to text modal data. The ResNet-50 model proposed by He is adopted for the deep network model of image modality [10], and the method of identity mapping is adopted at the increased level when constructing the deep network model, so as to control the error of training results. Although the residual network has many variations, in fact, at the application level, the effect of different network depths on the accuracy is not very different, but the effect on the training time is greater. Therefore, this paper adopts the ResNet-50 model, which consumes less training time.

There are two basic modules in the Deep Residual Network. One is identity block, whose input and output data dimensions are consistent. Therefore, it supports multiple continuous connections, whose function is to reduce training time consumption. The other is the convolution block, where the data dimensions of input and output are inconsistent and multiple consecutive conjunctions are not supported. Its function is to change the dimension of the feature vector and reduce the dimension.

The whole residual network model is formed by concatenating multiple residual units. In the input part of the network, the size of the image in different data sets is different, so this paper adjusts the image to make its size uniform. After passing through the network layer, the two layers are finally connected to the fully connected network. The goal is to further compress the data down to a specific dimension.

In this paper, the deep network structure is set by experiments. From the perspective of function space, single-layer neural network has simple nonlinear mapping characteristics, and its function is similar to that of vector base in mathematics, while multi-layer neural network combination is similar to that of using base to construct function space. From a statistical point of view, neural network is often considered as a function of approximation. The single-layer network structure makes the data present a single probability distribution, while the multi-layer superposition makes the probability form a mixture. Such a hybrid model makes the neural network present Universal Approximation Theorem [45]. Therefore, by deepening the hierarchical structure of the neural network and increasing the capacity of the model, more complex functions can be approximated. In fact, the process of image feature extraction using deep learning is a mapping process, that is, assuming that the parameter set learned through training for the deep network of image modality is θ_x, then the overall output result can be expressed as a function of the input image set $\boldsymbol{X}$: $f(X;\theta_x)$.

For the deep network model of the text part, we use the TF-IDF feature to represent each text data V_j^T and extract it as a text feature to the input of the deep network.

Since the TF-IDF is a statistical feature, in order to obtain similar features corresponding to the image data from the text, we attempt to access it into the fully connected network to re-map the TF-IDF feature results. The text deep network has three fully connected layers, and the specific parameters of each layer of data are shown in Tab. 1. In the first two layers of fully connected neural networks, ReLU function is the activation function, and in the last layer of output, it is the identity function.

Table 1: Parameter setting table of deep network

Layer	Number of layers
Full connection 1	TF-IDF input length
Full connection 2	4096
Full connection 3	Hash code length c

In this way, through the above text network, this paper assumes that the parameter learned in the network is θ_t, and the overall text output result can be expressed as a function $g(T;\theta_x)$of the input text set $\boldsymbol{T}$.

(2) The Second Stage: Hash Coding

The features extracted through the above feature learning process will be used as the input of the Hash coding function. Then, $\boldsymbol{F}$ is used to represent the image features acquired by data element t v_i^X learned from ResNet, and the corresponding $\boldsymbol{G}$ represents the text features obtained by data element V_j^T from the text deep network.

$$F : f(X;\theta_x) = \left\{ f\left(v_i^x ; \theta_x \right) \in R^c \right\} \tag{7}$$

$$G : g(T;\theta_x) = \left\{ g\left(v_{rj}^t ; \theta_t \right) \in R^c \right\} \tag{8}$$

where θ_x and θ_t represent parameter information learned in two different deep networks respectively.

In the feature extraction stage, this paper will discuss how to warp the results of the two Feed-Forward Neural Networks into Hamming space, how to realize the likelihood mapping process of the whole Hash coding decomposition and projection to the same space, and how to use the Hash method to maintain the constant distance for quantization.

Firstly, the features extracted in the first stage are likelihood mapped. From the perspective of machine learning, Discriminative Model can be used to solve this problem. Specifically, the spatial distance between image features $\boldsymbol{F}$ and text features $\boldsymbol{G}$ should be as close as possible in the case of the same category, while the distance should be as far as possible in the case of different categories. Therefore, it is necessary to ensure the accuracy of classification by distance. Thus, the similarity matrix $\boldsymbol{S}$, as the distance information of the original data, can be used as the discriminant criterion of the two feature distances. By optimizing the discriminant model, the same type of data can be clustered together as far as possible, that is, to maximize the logarithmic likelihood function:

$$L(S|F,G) = -\log_p(S|F,G) \tag{9}$$

The above probability distribution is divided to obtain the following formula:

$$p\left(S_{ij}|F^X{}_i, G^T{}_j\right) = \begin{cases} \sigma\left(\Theta_{ij}\right) & S_{ij} = 1 \\ 1 - \sigma\left(\Theta_{ij}\right) S_{ij} = 0 \end{cases} \tag{10}$$

where $\sigma\left(\theta_{ij}\right) = \frac{1}{1+e^{-\theta_{ij}}}$ is exponential function, $\theta_{ij} = \frac{1}{2}F_i^X G_j^T$ is the product between the i-th image feature and the j-th text feature, which are cross-modal variables that fuse different modal data. S_{ij} is similarity matrix of raw data.

Combine the piecewise function, there is a likelihood function L_1, which is expressed as follows:

$$L_1 = -\sum_{i,j=1}^{n}\left[S_{ij}\log(\sigma(\Theta_{ij})) + (1-S_{ij})\log(1-\sigma(\Theta_{ij}))\right] = -\sum_{i,j=1}^{n}(S_{ij}\Theta_{ij} - \log(1+e^{\Theta_{ij}})) \tag{11}$$

This allows the text data and image data to satisfy the likelihood principle by calculating and maximizing the value of L_1. Next, a Hash-quantization process of distance invariance is performed on the real-valued variable. In order to unify the modeling and facilitate the subsequent gradient descent learning, the Hash map result of the image is described in the form of a matrix, that is $B^{(X)}: \left\{b_i^x\right\}_{i=1}^n$. The matrix of the Hash map result of the text is $B^{(T)}: \left\{b_j^t\right\}_{j=1}^n$. Obviously the Hash mapping function should be after two network output results, as shown in Fig. 2. Then use the common symbolization function as a Hash mapping function, get $B^{(X)} = sign(F)$, $B^{(T)} = sign(G)$. The quantization process can be represented by a minimized loss function L_2, that is:

$$L_2 = \left|\left|B^{(X)} - F\right|\right|_F^2 + \left|\left|B^{(T)} - G\right|\right|_F^2 \tag{12}$$

where $||*||_F^2$ is the square of the matrix norm. To maintain the balance of the data, in the quantization process of the Hash function, it is necessary to make the final Chinese text encoding result and the image encoding result as close as possible, thus introducing a regular term.

$$R = B^{(X)} - B^{(T)} \tag{13}$$

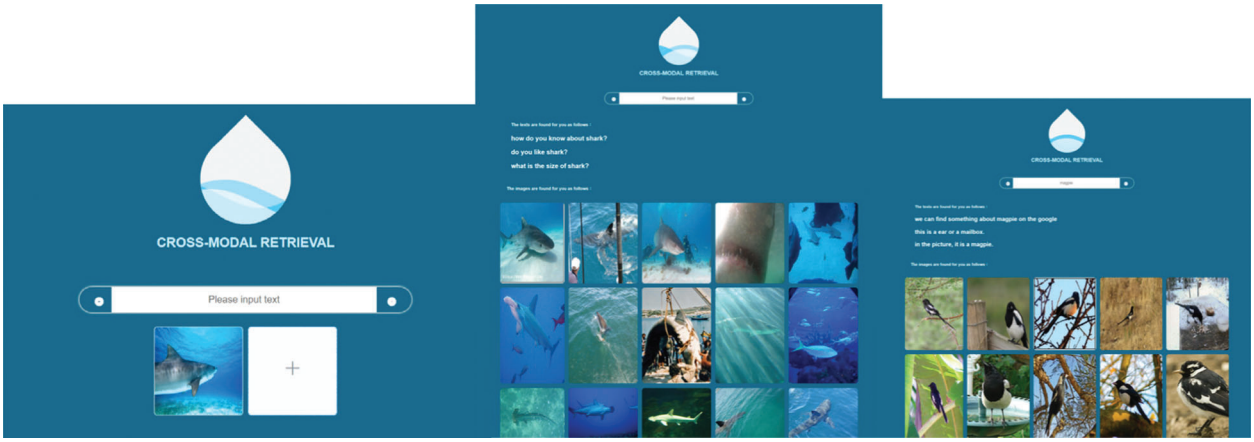

Figure 2: Image retrieval of text, images and text retrieval of images, text

Finally, through L_1, L_2 and expressions above, the Hash-encoded learning objective function in the deep network can be defined as:

$$\min_{B^{(X)},B^{(T)},\theta_x,\theta_t} J = -\sum_{i,\ j=1}^{n}\left(S_{ij}\theta_{ij} - \log\left(1+e^{\Theta_{ij}}\right)\right) + \alpha\left(\left||B^{(X)}-F\right||_F^2 + \left||B^{(T)}-G\right||_F^2\right) + \beta\left(B^{(X)}-B^{(T)}\right) \tag{14}$$

where α and β are Lagrange Multipliers $\alpha \geq 0$. In practical applications, $\mathrm{B} = \mathrm{B}^{(\mathbf{X})} = \mathrm{B}^{(\mathbf{T})}$is directly set, which simplifies the objective function of Hash coding. For optimization training, the above objective function is simplified to:

$$\min_{B,\theta_x,\theta_t} J = -\sum_{i,\ j=1}^{n}\left(S_{ij}\theta_{ij} - \log\left(1+e^{\Theta_{ij}}\right)\right) + \alpha\left(||B-F||_F^2 + ||B-G||_F^2\right) \quad s.t.\ \mathcal{B} \in \{-1,+1\}^{c\times n} \tag{15}$$

3.3 Model Parameter Learning

After the objective function is obtained, the parameters in the model are then learned. Since the parameters to be optimized in the model are θ_x, θ_y and B, the cross-learning strategy is used for parameter learning. The specific model parameter optimization process is divided into three steps.

3.3.1 Fix θ_t and B, Optimize θ_x

When parameters θ_t and B are set as fixed values, only θ_x is a variable at this time, and the parameters obtained as image deep network learning can be learned through Back Propagation. Back Propagation is mainly performed by gradient descent algorithm for optimization learning. To avoid the time-consuming Gradient calculation of the entire data set by Gradient Descent method and the problem of falling into local minima in the process of Gradient Descent, this paper refers to the Mini-Batch Gradient Descent (MBGD) for parameter learning [46]. A portion of the sample is used to update each parameter as it is updated, and the gradient is calculated using a mini-batch data calculation gradient instead of all data. On the overall model, the convergence effect of the results of multiple stochastic gradients is only slightly lower than the original gradient, but in exchange for a huge improvement in performance. For the objective function, we first calculate the gradient under small batch data:

$$\frac{\partial J}{\partial F_i} = \frac{1}{2}\sum_{j=1}^{n}\left(\sigma\left(\Theta_{ij}\right)G_j - S_{ij}G_j\right) + 2\alpha(F_i - B_i) \tag{16}$$

Then calculate the target gradient $\frac{\partial J}{\partial \theta_x}$ by chain rule.

3.3.2 Fix θ_x and B, Optimize θ_t

When the parameter B and θ_x are fixed values, θ_t is a variable, and as a parameter obtained by text deep network learning, the same reason can also be obtained by using the Back Propagation learning by the stochastic gradient descent. The gradient of the small-lot data is obtained as follows:

$$\frac{\partial J}{\partial G_j} = \frac{1}{2}\sum_{i=1}^{n}\left(\sigma\left(\Theta_{ij}\right)F_i - S_{ij}F_i\right) + 2\alpha\left(G_j - B_j\right) \tag{17}$$

Similarly, the target gradient $\frac{\partial J}{\partial \theta_t}$ can be obtained from the above formula using the chain rule to update the parameters.

3.3.3 Fix θ_t and θ_x, Optimize B

When θ_t and θ_x are fixed, the network structure is fixed, so the objective function can be written as:

$$\max_B tr\left(B^T(\alpha(F+G))\right) = tr\left(B^T V\right) = \sum_{i,\ j} B_{ij}V_{ij}$$

$$s.t.\ \mathcal{B} \in \{-1, +1\}^{c \times n} \tag{18}$$

Among them, $V = \alpha(F + G)$ it is easy to find that the binary encoding B_{ij} has the same symbol as V_{ij}, so there is:

$$B = sign(V) = sign(\alpha(F + G)) \tag{19}$$

3.3.4 Expansion of New Samples

For the model, in addition to training samples, there are also test samples and verification samples, namely new samples. The model trained in this paper can be directly used for binary coding of new samples. Therefore, after training the model, a new sample point $\left(\mathrm{x}_q, \mathrm{t}_q\right)$ was given to represent the new image and text data. The binary code of the image x_q can be expressed as:

$$b_q^{(x)} = h^{(x)}\left(x_q\right) = sign\left(f\left(x_q; \theta_x\right)\right) \tag{20}$$

Similarly, the binary encoding for text t_q can be expressed as:

$$b_q^{(t)} = h^{(t)}\left(t_q\right) = sign\left(g\left(t_q; \theta_t\right)\right) \tag{21}$$

The model thus constructed has good scalability for newly added data.

3.4 Model Algorithm and Retrieval Algorithm

3.4.1 Model Algorithm

Algorithm 1: Cross-Modal Hash Retrieval algorithm based on Deep Residual Network (CMHR-DRN).

Input: Define the image dataset as $X = \left\{ v_i^x \right\}_{i=1}^{n}$.

The input size of the small batch is min _*batch*.

The number of model trainings is *epochs*.

Text data set $T = \left\{ v^t_i \right\}_{i=1}^{n}$.

Cross-modal similarity matrix $S = \{S_{i,j}\}_{i=1,j=1}^{n}$.

Output: Image depth network model parameter set θ_x.

Text depth network model parameter set θ_t.

Binary Hash coded value B.

Random initialization: random initialization image network parameters θ_x, text network parameters θ_t, small batch input size mini_batch iteration number is Iteration, the number of model iterations can be expressed as epochs, and Adam parameters are set to the default value.

The pseudo code is as follows:

$\theta_x = image\_drn\_parameters \qquad \theta_t = text\_fc\_parameters$

$min\_batch = 64, Sample\_size = N$

$iteration = \dfrac{N}{min\_batch}, epochs = 500, \ Adam\_\theta_a = default \ \ value$

Repeat the training model:

Step 1: Randomly sample the image sample set X. After learning through the residual network, the Feed Forward output is obtained as F_i, the gradient of the back propagation *gradient_x* is calculated. And then the parameters θ_x are updated according to the gradient. The pseudo code is as follows:

For (int i = 1; i<iteration; i++) {

$V_i^x = Random\_sampling\_image\ (x_i)$;

$F_i = f\left(v_i^x; \theta_x\right)$;

$gradient\_x = \dfrac{\partial J}{\partial F_i}$;

$\theta_x = \theta_x + gradient\_x$;

}

Step 2: Pre-processing the text sample set T to extract the TF-IDF feature and Vectorization to obtain $V^{tfidf}{}_j$. Randomly sample the whole data set to obtain V^t_j. And also calculate the output G_j through the neural network to calculate the gradient of the back propagation gradient_t. Parameter θ_t update based on the gradient. The pseudo code is as follows:

For (int i = 1; i<iteration; i++){

$V_j^{tiidf} = tiidf\left(t_j\right)$;

$v_j^t = Random\_sampling\_text\left(v_j^{\ tiidf}\right)$;

$G_j = g\left(v_j^t, \theta_t\right)$;

$gradient\_t = \dfrac{\partial J}{\partial G_j}$;

$\theta_t = \theta_t + gradient\_t$;

}

Step 3: When the parameters θ_t and θ_x are fixed, the entire network structure has been fixed

$B^{(X)} = sign\left(f\left(v_i^x, \theta_x\right)\right)$,

$B^{(T)} = sign\left(f\left(v_j^t, \theta_t\right)\right)$,

Step 4: Optimize $B = B^{(X)} = B^{(T)}$ to match the inputs of the two networks.

Step 5: Until the number of training reaches 500, the result is obtained.

3.4.2 Text-to-Image

Algorithm 2: Text-to-Image algorithm.

Input: text t_1

Output: a series of images $X_1 = \left\{x_i\right\}_{i=1}^{k}$ and the probability of each image $P_{X_1} = \left\{p\left(x_i\right)\right\}_{i=1}^{k}$.

Step 1: Using TF-IDF algorithm to extract feature vectors by text feature extraction. Then use binary function to convert to binary code $b^x_1 = h^X\left(v^x_1\right)$.

Step 2: All images in the image library are extracted by feature extraction. Use the convolutional neural network to obtain the feature vector $X = \left\{v^x_i\right\}_{i=1}^{m}$. Then convert it into binary coding $h^X(X) \in \{0,1\}^d$ by using the Hash function.

Step 3: Calculate the similarity $S = \left\{S_{i,1}\right\}_{i=1}^{m}$ of the binary encoding of the input text and the binary encoding of all images in the image library.

Step 4: Sort all similarities from high to low, and take the image $X_1 = \left\{x_i\right\}_{i=1}^{k}$ corresponding to the binary encoding of the similarity top k.

3.4.3 Image-to-Text

Algorithm 3: Image-to-Text algorithm.

Input: an image x_1.

Output: a series of texts $T_1 = \left\{t_j\right\}_{j=1}^{k}$ and the probability of each text $P_{T_1} = \left\{p\left(t_j\right)\right\}_{j=1}^{k}$.

Step 1: Using convolutional neural network to perform feature extraction on the input image x_1 to obtain a feature vector $v^x{}_1$, and then converting into a binary code $b^x_1 = h^X\left(v^x_1\right)$ by using a Hash function.

Step 2: All the texts in the text library are extracted by the feature extraction using the TF-IDF algorithm to obtain the feature vector $T = \left\{v^t_j\right\}_{j=1}^{n}$, and then converted into binary encoding $h^T(T) \in \{0,1\}^d$ by using the Hash function.

Step 3: Calculate the similarity between the binary encoding of the input image and the binary encoding of all text in the text library:

$$S = \left\{S_{1,j}\right\}_{j=1}^{n}, \quad S_{1,j} = \begin{cases} 1 & v^x_1 = v^t{}_j \\ 0 & v^x_1 \neq v^t_j \end{cases};$$

Step 4: Sort all similarities from high to low, and take the similarity of the text $T_1 = \left\{t_j\right\}_{j=1}^{k}$ corresponding to the binary value of top k.

3.4.4 Rendering Effect of Cross-Modal Retrieval System Design

In order to achieve better results, the program was modified to support dual mode retrieval. The effect is shown in Fig. 2. Click the link to upload the picture, as shown in Fig. 2 (left). Upload the picture to be retrieved. Click the link to retrieve the text and image that meet the requirements, as shown in Fig. 2 (middle). Similarly, enter "Magpie" in the text box to retrieve a matching image and text, as shown in Fig. 2 (right).

4 Experimental Designs and Results Analysis

4.1 Experimental Operating Environment

According to the above model, the Caffe framework was used to build a Cross-Modal Retrieval system, and multiple text-image datasets were used for experiments. The experiment adopts assembled personal workstation, the operating system is Ubuntu 16.01-LTS, and the programming language is Python 3.6.

4.2 Data Set and Its Feature Extraction

To validate the algorithm's effectiveness, we used three common cross-modal data sets, including Wikipedia, MIR FLICKR-25K and NUS-WIDE.

The Wikipedia dataset [2] is a recommended article crawled from Wikipedia, containing 10 different kinds of 2,866 encyclopedias. The image, as a further description of the textual information, constitutes a pairwise relationship. Each image in the original data set is composed of 128-dimensional SIFT feature vectors, and the text extracts corresponding TF-IDF features.

MIR KLICKR-25K [47] contains 25,000 images collected from the image social networking site Flickr. Each image corresponds to several different text word tags. We selected data containing at least 20 text labels as experimental data. Among them, the text was represented as a 1386-dimensional word bag vector. Each image used a 512-dimensional SIFT feature vector. Each data point was marked with multiple text labels.

The NUS-WIDE [48] dataset contains 260,648 site images, some of which have text labels. It is a multi-label dataset created by the Media Search Lab of the National University of Singapore with 81 text labels. The original image dataset extracts low-level features such as color histograms, wavelet textures, and word vectors described by SIFT.

In terms of data partitioning, we divided the training set and test set by 8:2 for each type of data set. At the same time, cross-validation was used to carry out three experiments on the training set and the test set at different positions. Finally the average value was obtained.

4.3 Performance Evaluation Index

The search evaluation indicators mainly used in this experiment included MAP, PR curve and F-Measure.

MAP (Mean Average Precision) refers to the mean of the average precision of multiple query results [49]. For retrieval, you first need to calculate the average precision of the results (AP). In the retrieval task, the precision and recall rate are single-valued metrics based on the results of the entire document list, but the order in which the sorted documents are returned must also be considered. Assuming a recall rate of r and a precision $p(r)$ as a function of the recall rate, the AP is:

$$\text{AveP} = \int_0^1 P(r)dr \tag{22}$$

The above equation shows that the average precision is the area under the PR curve. For the results obtained by a finite number of retrieval tasks, the average precision can be described as discrete:

$$\mathrm{AveP} = \sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{p(k)}\ \Delta \mathrm{r(k)} \tag{23}$$

where k is the rank in the retrieved document sequence and n is the number of retrieved documents. p(k) is the precision of the cutoff value k in the list. r(k) is the recall change of the items $k-1$ to k.

MAP, as the mean of the average accuracy of multiple query results, can be expressed as:

$$\mathrm{MAP} = \frac{\sum_{\mathrm{q}=1}^{\mathrm{Q}} \mathrm{AveP(q)}}{\mathrm{Q}} \tag{24}$$

where Q is the retrieval times, MAP index can comprehensively calculate the average value of multiple retrieval results, and it is a single-value index of all documents. Therefore, it has good global characteristics, so it is quite appropriate as an evaluation index of the retrieval system.

In addition to use the MAP value to measure the average accuracy of the search, there is also a PR [50] (Precision-Recall) curve, which is used to evaluate the accuracy and recall rate of the search model.

F-Measure is the weighted harmonic mean of the precision rate P and the recall rate R, which is calculated as follows:

$$F = \frac{(\alpha^2+1)P*R}{\alpha^2(P+R)} \tag{25}$$

Evaluation of a search model mainly depends on whether the precision rate and the recall rate are optimal at the same time, that is, the closer the PR curve is to the coordinates (1,1), the better the retrieval effect [51].

4.4 Benchmark Model and Parameter Settings

We adopted CMFH, CMSSH and CMDN as the benchmark comparison model of this experiment, and compared the advantages and disadvantages of the cross-modal model, retrieval algorithm and other algorithms used in this paper. As a cross-modal algorithm for applying metric learning, CMSSH linearly maps data from different spaces to a unified space, ensuring the similarity of data distances before and after mapping. CMFH decomposes the data into different modalities to obtain the mapping of implicit variables. This paper only considers CMFH under unsupervised conditions. In the deep learning method, CMDN is used as a similar comparison of the algorithm.

In the first part of the model construction, the Deep Residual Network Resnet-50 is used. In the feature extraction of the image, there is no need to adjust the original parameters too much. This section uses the default parameter settings. In the final output stage, two layers of fully connected networks are connected, and the first layer is set to $512 \times 3 \times 4$ dimensions, and the second layer is 512 dimensions. On the text network, this experiment is connected to a three-layer fully connected network. The first layer of the first two layers is the length of the TF-IDF vector. The output of the second layer is 4096. And the third layer is the identity mapping. This ensures that the probability distribution of the data will not be completely changed. The Hash function of the second part set Hash codes of different lengths, which are 16 bits, 32 bits, and 64 bits. In the calculation of stochastic gradient descent, the Adam [52] (Adaptive Moment Estimation) method is used. As an adaptive moment estimation algorithm, Adam has good optimization performance.

4.5 Experimental Results and Analysis

Tab. 2 lists the average accuracy MAP retrieved by the different types of cross-modal algorithms on the Wikipedia data set. In the task of image-to-text, in the case of using three different length Hash codes, the

MAP value of the algorithm used in this experiment is higher than the two algorithms of CMFH and CMSSH, but there is still a certain gap compared with the CMDN algorithm. In the task of text-to-image, when the Hash code length reaches 32 bit and 64 bit, the MAP value of the algorithm used in this experiment is higher than the MAP values of the other three algorithms.

Table 2: Comparison of map values with other algorithms on Wikipedia datasets

Wikipedia dataset	Algorithms	Hash code length		
		16 bits	32 bits	64 bits
Image-to-text	CMFH	0.2447	0.2536	0.2652
	CMSSH	0.1886	0.1749	0.1702
	CMDN	0.3591	0.3630	0.3922
	CMHR-DRN	0.2707	0.2816	0.2914
Text-to-image	CMFH	0.6116	0.6298	0.6398
	CMSSH	0.1802	0.1768	0.1918
	CMDN	0.3107	0.325	0.3590
	CMHR-DRN	0.5459	0.6626	0.7258

Tab. 3 lists the average accuracy of different types of cross-modal algorithms retrieved on the MIRKLICKR-25K dataset. In the two tasks, the MAP value of the algorithm in this paper is higher than the MAP values of the other three algorithms.

Table 3: Comparison of map values with other algorithms on the mirklickr-25K dataset

MIRKLICKR-25K	Algorithms	Hash code length		
		16 bits	32 bits	64 bits
Image-to-text	CMFH	0.6155	0.6221	0.6299
	CMSSH	0.5890	0.6069	0.5790
	CMDN	0.6420	0.6342	0.7062
	CMHR-DRN	**0.6927**	**0.7143**	**0.7201**
Text-to-image	CMFH	0.6424	0.6563	0.6649
	CMSSH	0.5997	0.5688	0.5835
	CMDN	0.6321	0.6533	0.6873
	CMHR-DRN	**0.7504**	**0.7574**	**0.7704**

Tab. 4 lists the average accuracy of different types of cross-modal algorithms retrieved on the NUS-WIDE dataset. In the task of image-to-text, the MAP value of the algorithm in this paper is higher than the MAP value of the other three algorithms. In the task of text-to-image, when the Hash code length is 16 bits, the MAP value of the algorithm in this experiment is higher than the MAP values of the other three algorithms. When the Hash code length is 32 bit and 64 bit, the MAP value of the algorithm in this experiment is higher than the MAP values of the two algorithms CMSSH and CMDN, which is slightly lower than the MAP value of the CMFH algorithm.

Table 4: Comparison of map values with other algorithms on the nus-wide dataset

NUS-WIDE	Algorithms	Hash code length		
		16 bits	32 bits	64 bits
Image-to-text	CMFH	0.5532	0.5620	0.5699
	CMSSH	0.4823	0.4833	0.4731
	CMDN	0.3241	0.3916	0.3985
	CMHR-DRN	**0.6249**	**0.6355**	**0.6438**
Text-to-image	CMFH	0.6521	**0.6877**	**0.7092**
	CMSSH	0.4080	0.3927	0.3822
	CMDN	0.3667	0.3571	0.3766
	CMHR-DRN	**0.6791**	0.6829	0.6906

Take MIRFLICKR-25K as an example to draw the PR curve of image-to-text and text-to-image, as shown in Figs. 3 and 4.

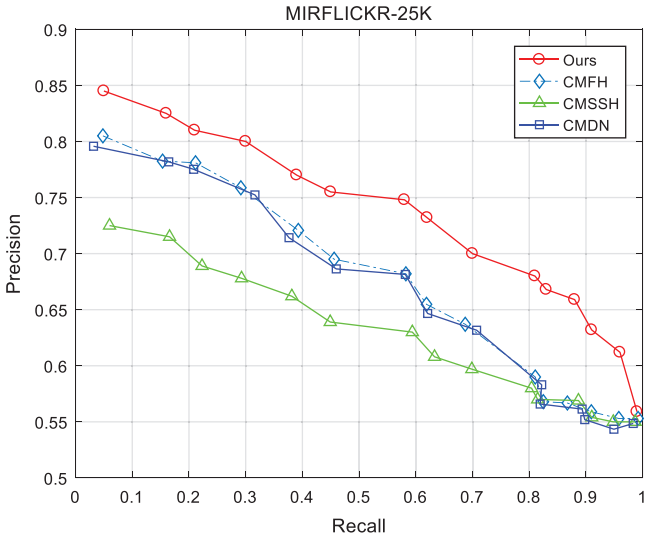

Figure 3: The PR curve of image-to-text

The experimental results show that the CMHR-DRN algorithm is better than the other three Cross-Modal Retrieval algorithms in the MIRFLICKR-25K dataset when it is used in Cross-Modal Retrieval. It improves the retrieval performance. In the task of text-to-image, when using the 16-bit Hash code in the Wikipedia dataset, it is better than the other three cross-model retrieval algorithms. In the task of image-to-text in the NUS-WIDE dataset, its performance is better than other three cross-modal search algorithms, which improves the search performance. It can be seen that the cross-modal Hash retrieval algorithm based on Deep Residual Network (CMHR-DRN) proposed in this experiment has stronger advantages in Cross-Modal Retrieval tasks than other types of algorithms CMSSH and CMDN. Compared with the same type CMFH, this algorithm also has improved retrieval performance.

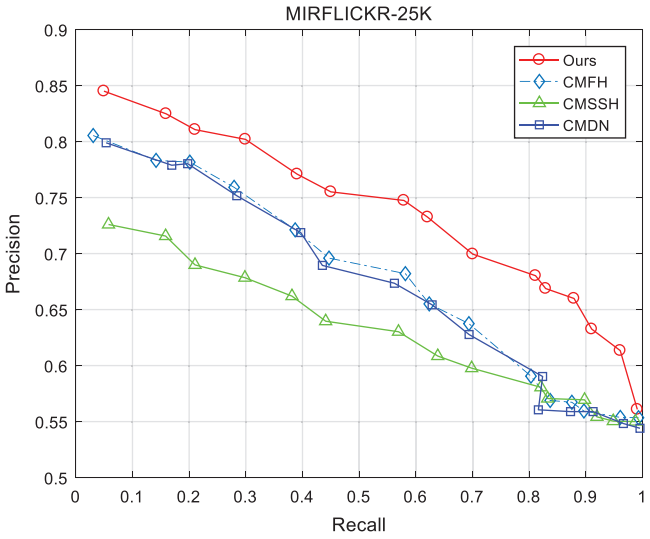

Figure 4: The PR curve of text-to-image

5 Conclusion

This paper proposes a cross-modal hash retrieval algorithm CMHR-DRN based on a deep residual network, and solves the objective hash function through a stochastic gradient descent algorithm. In the optimization process, the Adam method is used as the main gradient descent optimization algorithm, which is suitable for most Non-convex optimization, thus learning a very effective hash function. The two-stage model training method can help the corresponding model to be fully trained in each stage, reduce the difficulty of training, and help the model to converge faster. Experiments have proved that the proposed algorithm has a strong advantage over other types and similar cross-modal hashing algorithms in image-text cross-modal retrieval tasks, and the retrieval performance has also been improved. In the image retrieval, the MAP value of CMHR-DRN is higher than the other three algorithms; in the retrieval image, the CMHR-DRN algorithm has also achieved better results. On the PR curve, we can clearly find that the value of the BEP (Break Even Point) point (P = R) of the PR curve of the CMHR-DRN algorithm is higher than other algorithms, indicating the overall performance of the CMHR-DRN algorithm better.

The essence of cross-modal retrieval is to extract key and useful information from the data features of various modules to the greatest extent, and then unify these information data into the same feature space to compare and sort similarities to achieve Search function. The algorithm proposed in this paper has two focuses: one is to focus on the feature extraction of input data. For the large amount of data in image data, the content information is mainly reflected in the combination of data spatial distribution and data elements; for text data, it reflects the problem of keywords and arrangement structure contained in text, which are different from each other. Therefore, the algorithm in this paper uses the most advanced and classic methods in the field of image feature extraction and text understanding to do feature extraction; the second is data matching. Classic hashing methods can be used for matching, and the similarity between the data can be obtained from the pairwise comparison of the data, which helps to sort the data retrieval results. Sorting according to similarity can reflect the relevance of retrieval.

The algorithm proposed in this paper also has excellent accuracy and recognition, and has great advantages compared to other algorithms. The core is that the algorithm provided in this paper makes an efficient and accurate feature extraction for the modal information data. Without the basic clear and valuable feature information, it is difficult to provide accurate information for the next phase of algorithm

matching, and it is also difficult to achieve excellent results in cross-modal retrieval. The key of the algorithm is to use the depth residual network to extract information from the pictures. The depth network plays an important role in practicability and accuracy. By combining the advantages of the two algorithms, the accuracy and applicability of the entire algorithm framework is provided.

For the next step of research, there are two aspects that need to be studied in depth:

1. With the rapid development of deep learning technology, the deep network framework is continuously optimized, but the deep network framework has not yet been used to achieve cross-modal retrieval tasks between speech and image, and between speech and text. Therefore, the next need to consider using a deep network framework and designing corresponding algorithms to extract voice features, realize the unified mapping of voice-image-text modal features, and finally realize multi-modal retrieval.
2. Due to the increasing number of deep network layers, the size of training samples is also growing rapidly. The data format and content of multimedia information are more complex and diverse, including pictures, voice, text, video, etc. How to achieve efficient generalization of data conversion and provide more real-time solutions? It will be an important impetus to improve the cross-modal retrieval, enhance the cross-modal retrieval ability, and promote the application of cross-modal retrieval.

Acknowledgement: This paper would like to thank all the authors cited in the reference for their contributions to this field.

Funding Statement: This paper is supported by the National Office for Philosophy and Social Sciences Project "Research on Cross-Modal Retrieval Model and Feature Extraction Based on Representation Learning" (No. 17BTQ062). The paper comprises the research results of the National Office for Philosophy and Social Sciences Project. The total amount of the project is RMB 200000. Z. Y. Li is the person in charge of the project. In this paper, he is mainly responsible for the formulation of the topic, proposing the research ideas, improving the algorithm, and revising the manuscript; X. M. Xu is also an important member of the research group of the national Social Science Fund project, which is supported by the fund. In the paper, she is responsible for data collection and model training and drafting of the first draft. Professor Zhang is responsible for reviewing the manuscript. P. Zhang has participated in the literature search and review revision.

Conflicts of Interest: All authors of this paper declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Y. Li, Z. Q. Huang and X. M. Xu, "A review of the cross-modal retrieval model and feature extraction based on representation learning," *Journal of the China Society for Scientific and Technical Information*, vol. 37, no. 4, pp. 422–435, 2018.

[2] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. Lanckriet *et al.,* "A new approach to cross-modal multimedia retrieval," *in Proc. of the 18th ACM Int. Conf. on Multimedia*, Firenze, Italy, pp. 251–260, 2010.

[3] J. Q. Nglam, A. Khosla, M. Y. Kim, J. H. Nam and A. Y. Ng, "Multimodal deep learning," in *Proc. of the 28th Int. Conf. on Machine Learning*, Bellevue, Washington, USA, pp. 689–696, 2011.

[4] F. Feng, X. Wang and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. of the 22nd ACM Int. Conf. on Multimedia*, Orlando, FL, USA, pp. 7–16, 2014.

[5] J. Gu, J. Cai, S. Joty, L. Niu and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, USA, pp. 7181–7189, 2018.

[6] L. He, X. Xu, H. M. Lu, Y. Yang, F. M. Shen *et al.,* "Unsupervised cross-modal retrieval through adversarial learning," in *Proc. of the 2017 IEEE Int. Conf. on Multimedia and Expo (ICME)*, Hong Kong, China, pp. 1153–1158, 2017.

[7] M. M. Bronstein, A. M. Bronstein, F. Michel and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 3594–3601, 2010.

[8] G. Ding, Y. Guo and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 2083–2090, 2014.

[9] X. Xu, F. Shen, Y. Yang, H. T. Shen and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2494–2507, 2017.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.

[11] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 5353–5360, 2015.

[12] R. K. Srivastava, K. Greff and J. Schmidhuber, *Highway Networks*. 2015. [Online]. Available: https://arxiv.org/abs/1505.00387.

[13] A. Amir, S. Basu, G. Iyengar, C. Y. Lin, M. Naphade *et al.,* "A multi-modal system for the retrieval of semantic video events," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 216–236, 2004.

[14] C. X. Zheng, A. Long, Y. Volkov, A. Davies, D. Kelleher *et al.,* "A cross-modal system for cell migration image annotation and retrieval," in *Proc. of the 2007 Int. Joint Conf. on Neural Networks*, Orlando, FL, USA, pp. 1738–1743, 2007.

[15] Y. Q. Jia, M. Salzmann and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. of the 2011 Int. Conf. on Computer Vision*, Barcelona, Spain, pp. 2407–2414, 2011.

[16] F. Zhong, G. Wang, Z. Chen, F. Xia and G. Min, "Cross-modal retrieval for CPSS data," *IEEE Access*, vol. 8, pp. 16689–16701, 2020.

[17] Y. Peng and D. Q. Zhang, "Semi-supervised canonical correlation analysis algorithm," *Journal of Software*, vol. 19, no. 11, pp. 2822–2832, 2008.

[18] V. R. P. Borges, "Visualizing multidimensional data based on Laplacian eigen maps projection," in *Proc. of the 2014 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, San Diego, CA, USA, pp. 1654–1659, 2014.

[19] S. Akaho, "A kernel method for canonical correlation analysis," *Computer Science*, vol. 40, no. 2, pp. 263–269, 2007.

[20] P. Chandrika and C. V. Jawahar, "Multi modal semantic indexing for image retrieval," in *Proc. of the 9th ACM Int. Conf. on Image and Video Retrieval*, Xi'an, China, pp. 342–349, 2010.

[21] W. X. Lin, T. Lu and F. Su, "A novel multi-modal integration and propagation model for cross-media information retrieval," in *Advances in Multimedia Modeling—18th Int. Conf.*, Klagenfurt, Austria, pp. 740–749, 2012.

[22] K. Y. Wang, W. Wang, R. He, L. Wang and T. Tan, "Multi-modal subspace learning with joint graph regularization for cross-modal retrieval," in *Proc. of the 2013 2nd IAPR Asian Conf. on Pattern Recognition*, Naha, Japan, pp. 236–240, 2013.

[23] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. of the Twenty-Seventh AAAI Conf. on Artificial Intelligence*, Bellevue, Washington, USA, pp. 1070–1076, 2013.

[24] J. J. Chen, L. Pang and C. W. Ngo, "Cross-modal recipe retrieval with stacked attention model," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29457–29473, 2018.

[25] J. S. Kim, J. Y. Sim and C. S. Kim, "Multiscale saliency detection using random walk with restart," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 2, pp. 198–210, 2014.

[26] Y. Verma and C. V. Jawahar, "A support vector approach for cross-modal search of images and texts," *Computer Vision and Image Understanding*, vol. 154, pp. 48–63, 2017.

[27] W. Wang, X. Yang, B. C. Ooi, D. X. Zhang and Y. T. Zhuang, "Effective deep learning-based multi-modal retrieval," *VLDB Journal*, vol. 25, no. 1, pp. 79–101, 2016.

[28] H. Ding and W. Lu, "A study on correlation-based cross-modal information retrieval," *Data Analysis and Knowledge Discovery*, vol. 32, no. 1, pp. 17–23, 2016.

[29] T. Dutta and S. Biswas, "Cross-modal retrieval in challenging scenarios using attributes," *Pattern Recognition Letters*, vol. 125, pp. 618–624, 2019.

[30] Y. Wu, L. Wang, F. Cui, H. Zhai, B. Dong *et al.,* "Cross-model convolutional neural network for multiple modality data representation," *Neural Computing and Applications*, vol. 30, no. 8, pp. 2343–2353, 2018.

[31] M. Datar, N. Immorlica, P. Indyk and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. of the Twentieth Annual Sym. on Computational Geometry (SCG'04)*, Brooklyn, New York, USA, pp. 253–262, 2014.

[32] G. Shakhnarovich, "Learning task-specific similarity," Ph.D. dissertation, Department of Electrical Engineering and Computer Science, MIT University, Cambridge, MA, USA, 2006.

[33] K. Li, G. J. Qi, J. Ye and K. A. Hua, "Linear subspace ranking hashing for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1825–1838, 2017.

[34] Q. Y. Jiang and W. J. Li, "Deep cross-modal hashing," in *Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 3270–3278, 2017.

[35] C. Li, C. Deng, N. Li, W. Liu, X. Gao *et al.,* "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 4242–4251, 2018.

[36] C. Deng, Z. Chen, X. Liu, X. Gao and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018.

[37] S. Su, Z. Zhong and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proc. of the 2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, South Korea, pp. 3027–3035, 2019.

[38] T. Wang, L. Zhu, Z. Y. Cheng, J. Li and Z. Gao, "Unsupervised deep cross-modal hashing with virtual label regression," *Neurocomputing*, vol. 386, pp. 84–96, 2020.

[39] T. Hoang, T. Do, T. V. Nguyen and N. Cheung, "Unsupervised deep cross-modality spectral hashing," *IEEE Transactions on Image Processing*, vol. 29, pp. 8391–8406, 2020.

[40] L. Wu, Y. Wang and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1602–1612, 2019.

[41] S. Conjeti, A. G. Roy, A. Katouzian and N. Navab, "Hashing with residual networks for image retrieval," in *Proc. of the Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Quebec City, Quebec, Canada, pp. 541–549, 2017.

[42] L. Yang, M. Zhuang, H. Ming, Y. F. Zhang and H. Li, "Deep attention residual hashing," *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, vol. 101, no. 3, pp. 654–657, 2018.

[43] B. H. Qiang, P. L. Wang, S. P. Guo, Z. Xu, W. Xie *et al.,* "Large-scale multi-label image retrieval using residual network with hash layer," in *Proc. of the 2019 Eleventh Int. Conf. on Advanced Computational Intelligence (ICACI)*, Guilin, China, pp. 262–267, 2019.

[44] Y. Cao, M. Long, J. Wang, Q. Yang and P. S. Su, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, pp. 1445–1454, 2016.

[45] J. Heaton, I. Goodfellow, Y. Bengio and A. Courville, "Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, no. 1–2, pp. 305–307, 2017.

[46] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*. Berlin, Germany, pp. 421–436, 2012.

[47] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. of the 1st ACM SIGMM Int. Conf. on Multimedia Information Retrieval*, Vancouver, British Columbia, Canada, pp. 39–43, 2008.

[48] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo *et al.,* "Nus-wide: A real-world web image database from National University of Singapore," in *Proc. of the 8th ACM Int. Conf. on Image and Video Retrieval*, Santorini Island, Greece, pp. 1–9, 2009.

[49] C. D. Manning, P. Raghavan and H. Schutze, "Introduction to information retrieval," *Journal of the American Society for Information Science & Technology*, vol. 43, no. 3, pp. 824–825, 2008.

[50] E. Minkov and W. W. Cohen, "Adaptive graph walk-based similarity measures for parsed text," *Natural Language Engineering*, vol. 20, no. 3, pp. 361–397, 2014.

[51] X. H. Zhang, "Biomimetic principle and methods of objects recognition and classification in complex scenes," Ph.D. dissertation. Jilin University, Jilin, China, 2012.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of the 3rd Int. Conf. for Learning Representations*, San Diego, CA, USA, pp. 1–15, 2015.