Hybrid-attention based Feature-reconstructive Adversarial Hashing Networks for Cross-modal Retrieval

DOI : 10.36909/jer.ICCSCT.19467

Chen Li, Hua Wang

School of Information and Communication Engineering, Communication University of China, Bei Jing, China

* Corresponding author: lichengood@163.com, wanghua0514@cuc.edu.cn

ABSTRACT

With the massive growth of data of various modal types, people no longer use a single modal retrieval method, but a cross-modal retrieval method when performing retrieval tasks. Such methods often need to store data efficiently while maintaining the characteristics of fast query. Because the hashing learning method can represent the original high-dimensional data through a simple and compact binary hash code, which can greatly compress the data size and facilitate data storage and mutual retrieval, the cross-modal hashing retrieval has gradually become a hot topic in recent years. However, how to bridge the gap between modalities to improve the retrieval performance further is still a challenging problem. In order to solve this problem, we propose a Hybrid-attention based Feature-reconstructive Adversarial Hashing (HFAH) networks for cross-modal retrieval. First, a label semantic guidance module is introduced to guide the extraction process of text features and image features through the learning of labels, so as to fully maintain the semantic similarity between different modal data. Then, the hybrid-attention module is introduced to make the extracted data contain richer semantic information. Finally, the feature reconstruction network is used to make the relevant degree between similar cross-modal data pairs higher than that between dissimilar data pairs. Related experiments on two benchmark datasets confirm to us that HFAH performs better than several existing cross- modal retrieval methods.

Keywords: Hybrid attention; Feature reconstruction; Cross-modal retrieval; Deep hashing; Adversarial autoencoder.

INTRODUCTION

In recent years, due to the rapid development of computer networks and the rapid popularization of portable handheld devices, a large amount of multimedia data of various modalities, such as texts, images and audios, have entered people's lives, showing an explosive growth trend. The massive data contain rich information, have considerable economic value and bring new opportunities for technological progress and social development. In the face of more and more large-scale multimodal data, how to carry out cross-modal retrieval is still a challenging topic. On the one side, the data of different modalities show the characteristics of heterogeneity, and the similarity measurement of heterogeneous data is a issue that needs to be properly addressed in cross-modal retrieval. On the other side, because of a large quantity of data on the Internet and the high dimension of data representation, how to implement cross-modal retrieval accurately and efficiently has become an urgent problem (Peng et al.,2017) (Feng et al.,2021) (Liu et al.,2022).

The main challenge of cross-modal retrieval is "modality gap". Its content can be understood as the data representation of distinct modal types is incongruent and exists in diverse kinds of feature spaces. For this reason, measuring the similarity between them is very challenging. By analyzing the abundant correlation included in the cross-modal data, many methods have been put forward to solve the cross-modal retrieval task. For instance, the current mainstream method is to learn a common space in the middle based on the features of data of different modal forms, and get the similarity between them in a public space, which is called the common space learning method (Hotelling. H., 1935.) (Ngiam et al.,2009) (Liang et al.,2016) (Zhai et al.,2013) (Jiang et al.,2015) (Wang et al.,2012). At the same time, the cross-modal similarity measurement method (Clinchant et al.,2011) (Yi ta al.,2010) (Tong et al.,2005) is also proposed. This kind of method calculates the cross-media similarity directly by analyzing the given data relationship, rather than attaining an explicit public space. Among many methods for cross-modal retrieval, the cross-modal retrieval based on hashing learning (Gionis. A., 1999) (Shen et al., 2015) is a more common method. The core idea of this method can be summarized as the binary encoding mapping of similarity preserving, so the research of hash retrieval mainly focuses on the design of mapping and similarity preserving strategy (Tao. Y., 2017).

The cross-modal retrieval method usually directly maps the extracted modal features multiple times based on the full connected network, thus mapping the modal features to the corresponding dimensions of the hash code. In this process, the structural relationship in the original semantic space will be gradually destroyed which results in the loss of semantic feature information when mapping high-dimensional semantic features extracted from different modalities to low-dimensional features. In this case, the common mapping of visual features and text features will not be conducive to maintaining semantic integrity, thus affecting the subsequent retrieval performance.

For reason of solving the above-mentioned issues better, a hybrid-attention based feature-reconstructive adversarial hashing method is raised to deal with the cross-modal retrieval problems. The main contributions of this paper are concluded as below:

1) For the sake of In order to take full advantage of the multi-label information carried by the data, under the guidance of the label semantic information, the semantic relevance between different modal data is closer, and the semantic relevance can be better preserved.

2) In this paper, the hybrid-attention module is drew into the image feature learning stage to maintain the high-caliber semantic relevance of images. By this means, we can be provided with the more distinguished image features.

3) Each modal data reconstructs the features of its own modal data as well as the features of other modal data. Through adversarial learning, the correlation within the same modal

data is kept to the maximum in the feature extraction space and Hamming space, and the heterogeneous problem between different modal data can be solved to a great extent.

4) A large number of experiments on two datasets we selected indicate that the experimental effect of our proposed HFAH is obviously better than that of the more advanced cross-modal hash methods, whether it is traditional means or deep learning methods.

The rest of this paper is arranged as below: other related work on the cross-modal hash methods will be presented in Section 2; the HFAH raised in this paper and the corresponding learning algorithm and optimization process can be found in Section 3; the experimental results will be discovered in Section 4; Section 5 sums up this work.

RELATED WORK

With the continuous increase of data modality types in the scene, a lot of multimodal technique has practiced to show the correlation between multiple perspectives. The key issues of hash learning for cross modal retrieval is how to construct the potential correlation within multiple modalities and maintain the correlation between modalities. Generally, these techniques are

divided into two classes: Multi-Source Hashing (MSH) and Cross Modal Hashing (CMH). According to whether label information aided training is added, cross modal hash learning can also be separated into unsupervised hash learning and supervised hash learning (Jiang et al., 2017).

Unsupervised cross-modal hash method can only use the same event information of data. Latent Semantic Sparse Hashing (LSSH) (Zhou et al.,2014) uses sparse encoding and matrix decomposition for image data and text data, and then maps these constructed features to generate a unified hash code. Unsupervised Deep Imputed Hashing (UDIH) (Chen et al.,2020) is a two phases learning tactics with the help of enhanced data, the correlation graph is constructed to enhance the capability to maintain the semantic consistency and difference between text and image. In paper (Zhang et at.,2020), a multipath generation adversary hash method for unsupervised cross modal retrieval is proposed.

Compared with unsupervised hash learning method, supervised hash learning methods can use existing supervised information to reduce semantic differences, enhance the relevance of different modal data, and have more advantages than the cross-modal retrieval methods based on unsupervised hash learning in improving the accuracy of retrieval. Adaptive Label correlation based asymmEtric Cross-modal Hashing (ALECH) (Li et al.,2021) uses both the hash code and semantic labels to improve the hash capability and maintain the similarity. Asymmetric Correlation Quantization Hashing (ACQH) (Wang et al.,2020) uses pairwise semantic similarity preservation and point by point label regression to construct combined quantization to generate hash codes. Scalable Discriminative Discrete Hashing (SDDH) (Qin et al.,2021) introduces a composite learning framework for compact hash code learning.

Most of the above cross-modal retrieval means based on hash learning use superficial structures to realize the feature extraction, which cannot explain the complex non-linear relationship between text and image. In recent years, due to the deep learning, many cross modal retrieval applications use deep learning networks. Deep Cross Modal Hashing (DCMH) (Jiang et al., 2017) can be divided into the feature learning and hash learning in the unity framework to maintain the similarity between text data and image data through negative logarithmic likelihood loss. Hierarchical Semantic Structure Preserving Hashing (HSSPH) (Wang et al.,2022) is a deep networks cross modal hash method which directly uses label level information to learn and identify hash codes. Multi-task Consistency-Preserving Adversarial Hashing (CPAH) (Xie et al.,2020) designs a consistency refinement module (CR) and a multitask confrontation learning module (MA), and more effectively collect semantic consistency information between text data and image data. Pairwise Relationship Guided Deep Hashing (PRDH) (Yang et al.,2017) explores pairwise constraints between text data and image data. Self-Supervised Adversarial Hashing Networks (SSAH) (Li et al.,2018) integrates adversarial learning into the cross-modal hash retrieval task in a self-supervised way.

PROPOSED METHOD

Figure 1 is the overall frame of the hybrid-attention based feature-reconstructive adversarial hashing method put forward in this paper. The framework mainly includes four parts: label semantic guidance module, text data and image data feature extraction module and hash learning module. The label semantic guidance module aims to fully utilize the superiorities of multi-label data and retain more semantic information. Text data and image data modules are designed to generate features of key points that retain the original data. The hash learning module not only takes the data similarity relationship within the modalities into account, but also attaches importance to the data similarity between modalities, so that the generated target hash code maintains the similarity between the original data.

In the section that follows, this article will illustrate each module in more detail and explain the learning algorithms involved.

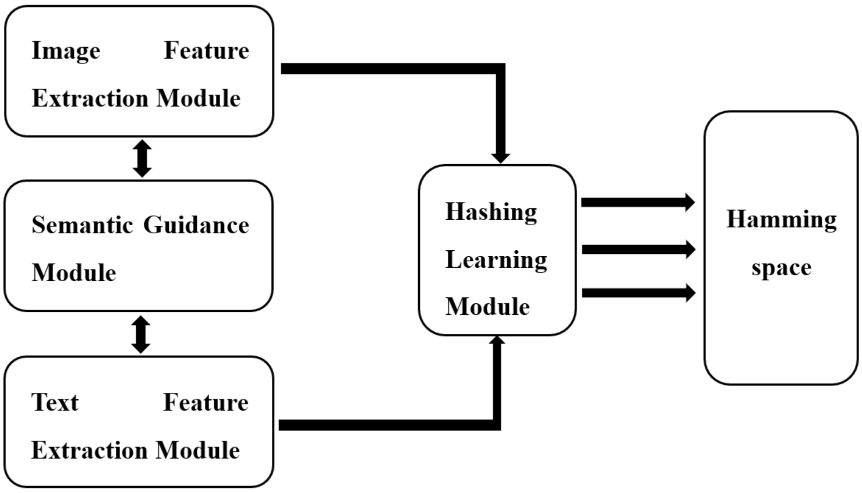
Image Feature Extraction Module
Semantic Guidance Module
Text Feature Extraction Module
Hashing Learning Module
Hamming space

Figure 1. HFAH Framework.

3.1 Problem definition

This means is easily expanded to the case of multiple modal data, not just two, but we take text and image modalities as example in this experiment. Supposed that there are n groups of training data, and each group of training data contains data of two modalities. In this experiment, the text-image dataset is used for explanation, that is, each group of data contains data of both text and image modalities, and there is also corresponding label data. Let $X = \{x_i\}_{i=1}^{n}$is denoted as text modal data, x_i is original text data or extracted features. Let $Y = \{y_i\}_{i=1}^{n}$ is denoted as image data, y_i is original image pixel or manually extracted features. Let $L = \{l_i\}_{i=1}^{n}$ is denoted as label data, l_i=[l_{i1}，l_{i2} ，… ，l_{ic}] is the label corresponding to the i^{th} text/image, where ic is the quantity of label forms, l_{im}=1 indicates that the data belongs to category m, l_{im}=0 means it does not belong to category m. Such a set of data can also be denoted by $\{x_i，y_i，l_i\}$. At the same time, in the cross-modal similarity matrix S, s_{ij}=1 indicates that the text data x_i is semantically similar to the image data y_j, while s_{ij}=0 indicates that it is not similar. In a multi-label dataset, if one item in the label data is considered to be the same, then we can get s_{ij}=1, otherwise s_{ij}=0.

The target of the cross-modal hash method is to get a public Hamming space, so that data of distinct modalities are able to learn a unified hash code $B \in (-1，1)^K$ (K represents the bit), while preserving the similarity between the original data. More specifically, if the two samples are homologous, the hamming distance between the hash codes generated by them are supposed to be smaller, otherwise, the hamming distance between them should be larger.

For two given examples x and y, when we compute the similarity between two hash codes by using hamming distance to, it can be expressed as D(x ,y)=$\frac{1}{2}(k - x^T y)$, that is, we get the similarity of hash codes between two samples with inner product. Given the x_i and the y_j, the conditional probability of s_{ij} is:

$$p(s_{ij}|x_i, y_j) = \begin{cases} \sigma(\theta_{ij}) & s_{ij} = 1 \\ 1 - \sigma(\theta_{ij}) & s_{ij} = 0 \end{cases} \qquad (1)$$

Where θ_{ij}=$\frac{1}{2}{x_i}^T y_j$, generally $\sigma(\theta_{ij})$=$\frac{1}{1+e^{-\theta_{ij}}}$. If the inner product is large, we consider the similarity between the data to be high. Therefore, instead of getting the similarity between the two, we can calculate the inner product between the Hamming codes in the Hamming space.

Combined with formula (1), the large likelihood estimation of the negative pole of the two is:

$$\begin{aligned} L &= -logp(s_{ij}|x_i, y_j) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n}(s_{ij}\theta_{ij} - \log(1 + e^{\theta_{ij}})) \end{aligned} \qquad (2)$$

3.2 Label semantic guidance module

Supervision information plays a great role in learning semantic information that has similar relationship with the original data. In a multi-label dataset, different categories of data have different similarity relationships. If only one category is the same as similarity to make judgments, the advantages of multi-label datasets cannot be fully exploited, and the semantic information contained in them cannot be fully explored and utilized, thus reducing the accuracy of cross-modal retrieval. The aim at label network is to train the features and hash codes of labels, and use the generated features to help the network training and learning in the next stage.

In the label network, let $H_i^l = f(l_i; \theta_l)$ denotes the hash function generated by the label network. Here θ_l is the parameter of label network, f_i^l is the semantic feature of the ith point. At this time, the loss function of the label network is delimited as below:

$$\begin{aligned}\min_{\theta_l\,,\,B^l} L^l &= L_1^l + L_2^l + \alpha L_3^l + \beta L_4^l \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n}(s_{ij}\varphi_{ij}^l - \log\left(1 + e^{\varphi_{ij}^l}\right)) \\ &\quad -\sum_{i=1}^{n}\sum_{j=1}^{n}(s_{ij}\theta_{ij}^l - \log\left(1 + e^{\theta_{ij}^l}\right)) \\ &\quad +\alpha||H^l - B^l||_F^2 + \beta||L - L^l||_F^2 \end{aligned} \qquad (3)$$

In the above formula, $\varphi_{ij}^l = \frac{1}{2} {f_i^l}^T f_j^l$,$\theta_{ij}^l = \frac{1}{2} {H_i^l}^T H_j^l$,B^l is the binary hash code of the label generated by the label network, and B^l= sign (H^l). α and β are super-parameters. In the above loss function, L_1^l is the semantic feature loss of similar instances, L_2^l is the semantic hash loss of similar instances. L_3^l is the quantization loss of the hash code. L_4^l indicates the classification error between the initial label and the prediction label. In this process, a random gradient descent back propagation algorithm is used to optimize the objective function, then we can attain the label network parameters θ_l and B^l. When the label network is trained, the required semantic features and semantic hash codes can be obtained.

3.3 Text and image modal feature learning

After the labels in the multi-label dataset are trained and learned through the label network, the obtained semantic features and binary hash codes are transferred into the text and image feature extraction network as the supervision information for the learning and training of the text and image feature extraction network. Because the output of the label network greatly retains the similarity between the original data and contains rich semantic information, the training of the text and image feature extraction network will be more accurate.

In the text and image feature extraction network, let $H_i^x = f(x_i; \theta_x)$ denotes the hash function generated by the text network. Here θ_x is the parameter of text network which learns the text features and the hash codes, f_i^x is the semantic feature of the ith point, let $H_i^y = f(y_i; \theta_y)$ denotes the hash function generated by the image network. Here θ_y is the parameter of image network which learns the image features and the hash codes, f_i^y is the semantic feature of the ith point. The data of different modalities in the same sample can maintain semantic consistency, when the feature learning process of text and image is finished. If two image sample pairs are homologous, their correspondent features f_i^y and f_j^y also should be homologous. Therefore, with the conduct of the label semantic guidance module, the objective function of the feature extraction part of text data and image data is as follows:

$$\min_{f^*,\theta^*} L_f^{l,*} = -\sum_{i=1}^{n}\sum_{j=1}^{n}(s_{ij}\varphi_{ij}^{l,*} - log\left(1 + e^{\varphi_{ij}^{l,*}}\right)) \tag{4}$$

Where $\varphi_{ij}^{l,*} = \frac{1}{2} f_i^{l^T} f_j^*$, and we replace x or y with *, B^x is the binary hash code of the text network, B^y is the binary hash code of the image network, and B^x= sign (H^x), B^y= sign (H^y).

For the whole network of text modality and image modality, the objective functions are:

$$\min_{\theta^x,B^x} L^x = L_f^{l,x} + \beta L_c^x + \gamma L_r^x \tag{5}$$

$$\min_{\theta^y,B^y} L^y = L_f^{l,y} + \beta L_c^y + \gamma L_r^y \tag{6}$$

3.4 Hash learning module

Most of means based on generative adversarial mechanism only include one kind of modal discriminator, which is used to judge which modality the generated samples belong to, so it limits the accuracy of cross modal retrieval. In the hash learning module, feature reconstruction encoding is used to adversarial learning. Different from the general encoder, both text data features and image data features are reconstructed for each modality of data. The purpose of building the model in this way is: the data of another modality is reconstructed through the model to bring about the mining of the correlation of multimodal data in the consistent cross encoding model, but the learned features cannot well represent the data of the original modality. In order to make the learned features reflect both the cross-modal correlation and the characteristics of the original modal data, the encoding model and cross encoding model are combined to form a comprehensive modal encoding model.

In the training phase, the previously extracted image features are embedded in the public space, and the image embedding features are represented by B^y. The previously extracted text features are embedded in the public space, and the text embedding features are represented by B^x. Unlike the general auto-encoder, which only reconstructs the features of its own modality, this way not only reconstructs the features of its own modality, but also reconstructs the features of another corresponding modality. The image embedded features are reconstructed into image features f_y^y and text features f_x^y respectively. Text features are similar. Text embedding feature reconstruction generates text features f_x^x and image features f_y^x.For the reason of reducing the information loss of features during feature reconstruction process, the reconstruction loss function is defined:

$$L_r^x = ||f^x - f_x^x||_F^2 + ||f^x - f_y^x||_F^2 \tag{7}$$

$$L_r^y = ||f^y - f_y^y||_F^2 + ||f^y - f_x^y||_F^2 \tag{8}$$

In order to keep the discriminant relationship within the modality when the text data and image data are embedded in the common space, that is, to maintain the similarity relationship between the initial data when the text data and image data are embedded in the public space, the loss function is defined:

$$L_c^x = \big||H^x - B^x|\big|_F^2 + \Big||B^{x'} - B^x|\Big|_F^2 \tag{9}$$

$$L_c^y = \big||H^y - B^y|\big|_F^2 + \big||B^{y\prime} - B^y|\big|_F^2 \tag{10}$$

3.5 Adversarial learning

Under the supervision of label semantic guidance module, feature extraction of text and image data contains more abundant semantic information. However, the distributions of features extracted from different modal data are often different, and excessive differences will lead to the final retrieval results. For the sake of getting more uniform hash codes, we hope that the hash codes generated from data containing similar semantic information are as same as possible. In order to alleviate the difference between different modal data, this experiment adopts the way of adversarial learning to learn the public Hamming space, and try to solve the difference between modalities.

For the reconstructed text features and image features, two discriminators are selected to distinguish them: text modal discriminator and image modal discriminator. The text feature reconstructed from the original text feature and image is input to the text modality discriminator. If the discriminant is the original text feature reconstruction, the output is "1". If the discriminant is the image feature reconstruction, the output is "0". Regard f_x^x as real text features and f_x^y as false text features. Then taking them as the input of D1, train D1 to judge whether they are true or false, and define the adversarial loss function on D1. The input and output of image modal discriminator are similar to that of text modal discriminator. Regard f_y^y as real image features and f_y^x as false image features. Then taking them as the input of D2, train D2 to judge whether they are true or false, and define the adversarial loss function on D2. Therefore, the objective function can be describe as below:

$$\underset{\theta_{D1}}{min} L_{adv}^1 = -\frac{1}{n}\sum_{m=1}^{n}\left(logD_1\left(f_x^x, \theta_{D_1}\right) + \log\Big(1 - D_1\big(f_x^y, \theta_{D1}\big)\Big)\right) \tag{11}$$

$$\underset{\theta_{D2}}{min} L_{adv}^2 = -\frac{1}{n}\sum_{m=1}^{n}\left(logD_2\left(f_y^y, \theta_{D_2}\right) + \log\Big(1 - D_2\big(f_y^x, \theta_{D2}\big)\Big)\right) \tag{12}$$

3.6 Optimization

The overall objective function can be written as:

$$L_{gen} = L_x + L_y + L_l \tag{13}$$

$$L_{adv} = L_{adv}^1 + L_{adv}^2 \tag{14}$$

The training of text feature selection network and image feature selection network is conducted in the way of adversarial learning. The optimization objectives of the generating model and the discriminant model are opposite: the goal of generating the model is to generate samples that make the discriminant model unable to identify the modal category, while the goal of the discriminant model is to learn how to determine the modal category accurately to which the sample belongs. They conduct iterative training in a adversarial way, so it can be seen as a minimum maximum game problem:

$$(B, \theta^{x,y,l}) = \underset{B,\theta^{x,y,l}}{argmin}\, L_{gen}(B, \theta^{x,y,l}) - L_{adv}(\hat{\theta}_{adv}) \tag{15}$$

$$\theta_{adv} = \underset{\theta_{adv}}{argmax}\, L_{gen}\left(\hat{B},\ \hat{\theta}^{x,y,l}\right) - L_{adv}(\theta_{adv}) \tag{16}$$

The algorithm flow of the hashing learning is shown in Table 1:

Table 1. The algorithm flow of the hashing learning.

```
Algorithm 1 The detailed learning algorithm of HFAH
Input: Text set X; Image set Y; Label set L; Length of hash code k;
Output: Optimal Binary codes B;
          Parameters $\theta_x$, $\theta_y$, $\theta_l$, $\theta_{D_1}$ and $\theta_{D_2}$;
Procedure:
Initialization: Initialize the parameters:$\theta_x, \theta_y, \theta_l, \theta_{D_1}, \theta_{D_2}$;
                      the hyperparameters: $\alpha, \beta, \gamma$;
repeat
    for t iteration do
    Update the parameters $\theta_l$ by using Back Propagation algorithm:
        $\theta_l \leftarrow \theta_l - \mu \cdot \nabla_{\theta_l} \frac{1}{n}(L_{gen} - L_{adv})$
    Update the parameters $\theta_x$ and $\theta_y$ by using Back Propagation algorithm:
        $\theta_* \leftarrow \theta_* - \mu \cdot \nabla_{\theta_*} \frac{1}{n}\left(L_{gen} - L_{adv}\right)$ , where x and y are replaced by *
    Update the parameters $\theta_{D_1}$ and $\theta_{D_2}$ by using Back Propagation algorithm:
        $\theta_{D_*} \leftarrow \theta_{D_*} - \mu \cdot \nabla_{\theta_{D_*}} \frac{1}{n}(L_{gen} - L_{adv})$ , where x and y are replaced by *
    end for
    Update the binary codes B by B = sign ( $H^l + H^x + H^y$ )
until convergence
```

EXPERIENT

4.1 Datasets

The **MIR-Flickr 25K** dataset (Huiskes et al.,2008) contains a total of 25000 instances, which are captured according to social photography website Flickr. Every example contains an image and a matching text, and is labeled with one of the 24 Categories of labels.

The **NUS-WIDE** dataset (Chua et al.,2009) includes 269648 images and their related labels, which are divided into 81 category concepts that have been marked for search evaluation. Each image has an average of 2 to 5 label statements, including 5018 independent labels. After deleting the data without any labels or label related information, a subset of 195834 image-text pairs of 21 of the most common category concepts was selected as the dataset of this experiment.

4.2 Assessment and Baseline

Evaluation: Among the retrieval protocols used to evaluate the performance of cross-modal retrieval tasks, Hamming Ranking and hash lookup are two classic ones. The Hamming Ranking method sorts the Hamming distance between the queried data and the retrieved results from the smallest to the largest, and then returns the highest ranked results of the specified items. The hash lookup method refers to returning the search results within the specified hamming radius, where the hamming radius value range is 0 to the hash code length. In this experiment, two more commonly used performance evaluation indicators were selected: the mean accuracy rate (MAP) was used to measure the accuracy of Hamming sorting based on Hamming distance, and the precision recall (PR) curve was used to measure the performance of hash learning.

Baseline: This experiment compares the proposed HFAH method with the experimental results of six other cross modal retrieval methods, including CVH (Kumar et al.,2011), SePH (Lin et al.,2015), DCMH (Jiang et al., 2017), PRDH (Yang et al.,2017), SSAH (Li et al.,2018), and SAAH (Li et al.,2022). Most of the comparison results are from SAAH.

4.3 Implementation Details

In the label semantic guidance module, because all label data are binary data, only feedforward neural network is needed to collect the semantic information in the label. The label network contains four connection layers. The first three full connection parts are used to extract feature. The number of neurons in the three full connection layers are c (number of labels) ,4096 and 512, respectively. The last layer generated the semantic hash code of the label. The number of neurons in this layer is k (the number of hash codes), and take tanh as the activation function. The label semantic features output from the label network will be used as supervision information to collect the features extraction of cross-modal data.

Text feature extracts text features based on multi-scale fusion module. Because the original text features (the feature vectors processed by the BoW method) inputted to the network are bag of words, in order to capture more information about the text itself, the data are processed from different scales of receptive fields, and the processed data are spliced to obtain new text data features. After the full connection layer, more semantic text features can be extracted.

The image network extracts image features based on the hybrid-attention network, connects the attention network with the CNN network, and can extract image features with more important information. These high-quality feature vectors can achieve good results in the hash learning task. Extracting effective image modal sample features is conducive to generating high-quality hash codes. CNN-F network model is used as the prototype to design our image modal network framework. For the purpose of extracting high-quality image features, we have made some adjustments based on the CNN model to better adapt to the hash learning task. That is, all layers and full connection layers in the original CNN model are retained and attention networks are added between the convolution layer and full connection layer.

In the feature reconstruction network, for different modal data, this paper constructs two groups of adversarial encoders. Each group includes two encoders and a discriminator. The data features of each modality obtained from the feature extraction network are reconstructed separately to further reduce the gap between the data of each modality.

4.4 Experimental result

Table 2 shows the average accuracy results of HFAH and other selected cross modal retrieval methods based on hash learning in the two datasets MIR-Flickr25K and NUS-WIDE. The tasks involved include: image retrieval text and text retrieval image. Among the selected methods, CVH and SePH are shallow hash learning methods, while DCMH, PRDH, SSAH and SAAH are deep cross-modal hash learning methods. For the methods to be compared, the results provided in paper (Li et al.,2018), and paper (Li et al.,2022) are selected.

MPA assessment: This may be because in the learning process, the network proposed in this paper can promote the learning of semantic relevance within the same modality and between different modalities more effectively, which means that HFAH method can learn more discriminatory representations. Therefore, HFAH can more accurately highlight the use of the correlation between data, in order to obtain better retrieval performance.

It can be seen from Table 2 that the HFAH method proposed in this paper has good performance results both on MIRFlickr25K dataset and NUS-WIDE dataset. With the development of the binary hash length, the effect of all methods is basically on the rise, which shows that increasing the hash code length can contain more data information. It can improve the cross modal retrieval accuracy, but the occupied encoding space will also increase. It can be seen from the comparison with other selected baseline methods that compared with shallow hash methods CVH and SePH, HFAH has a large improvement in MAP, achieving an improvement of more than 10%. Compared with other deep hash methods, HFAH has also improved to varying degrees. This improvement is attributed to our new hash learning method, especially the use of semantic label information and the maintenance of similar relationships within the same modal data and between different modal data, which shows that the network can more fully mine the potential semantic information associations.

P-R curve evaluation: Figure 2 and Figure 3 show the comparison of P-R curves on MIR Flickr25K and NUS-WIDE datasets between HFAH method proposed in this paper and other selected baseline methods. It can be seen from the figure that the P-R curves of HFAH on different data sets are basically above the curves of the selected comparison method, which shows that HFAH has more accurate retrieval performance than other comparison methods.

Table 2. MAP Assessment Results.

Task	Method	MIR-Flickr 25K			NUS-WIDE		
		16bit	32bit	64bit	16bit	32bit	64bit
Image retrieval text	CVH	0.557	0.554	0.664	0.400	0.392	0.386
	SePH	0.657	0.660	0.661	0.478	0.487	0.489
	DCMH	0.735	0.737	0.750	0.478	0.486	0.488
	PRDH	0.722	0.740	0.755	0.593	0.633	0.624
	SSAH	0.782	0.790	0.800	0.602	0.622	0.639
	SAAH	0.792	0.796	0.815	0.628	0.646	0.656
	HFAH	0.798	0.814	0.823	0.643	0.648	0.659
Text retrieval image	CVH	0.557	0.554	0.554	0.372	0.366	0.363
	SePH	0.648	0.652	0.654	0.449	0.454	0.458
	DCMH	0.763	0.764	0.775	0.638	0.651	0.657
	PRDH	0.755	0.764	0.777	0.594	0.610	0.602
	SSAH	0.791	0.795	0.803	0.612	0.637	0.640
	SAAH	0.795	0.803	0.806	0.651	0.663	0.659
	HFAH	0.798	0.810	0.819	0.648	0.655	0.661

Ablation experiment results: In order to verify the impact of different modules in the network on the performance of the experiment results, three variants are designed as the verification baseline for HFAH: HFAH-3 means that the feature reconstruction network is deleted from the original network architecture of HFAH, and the new network model is learned by iteration; HFAH-2 means that the hybrid attention model in the image feature extraction module is deleted on the basis of HFAH3 network; HFAH-1 means that on the basis of HFAH-2 network, the multi-scale feature fusion model of text feature extraction module is deleted, and the full connected network is directly used to build the text feature extraction network. From Table 2, we can draw the following conclusions: each key point we proposed, including the introduction of semantic guidance module, mixed attention module and feature reconstruction confrontation learning network, has a positive effect on improving retrieval performance. The contribution of the feature reconstruction adversarial learning network to improving the retrieval performance is more obvious, which verifies the effectiveness of the proposed hybrid- attention based feature reconstruction adversarial hashing for cross-modal hashing method.

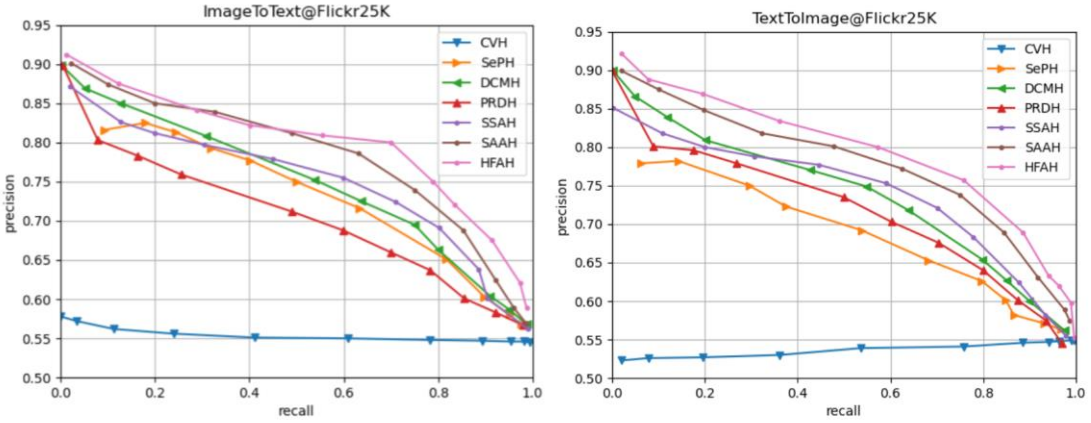

Figure 2. PR Curves of Various Methods on MIR Flickr 25K.

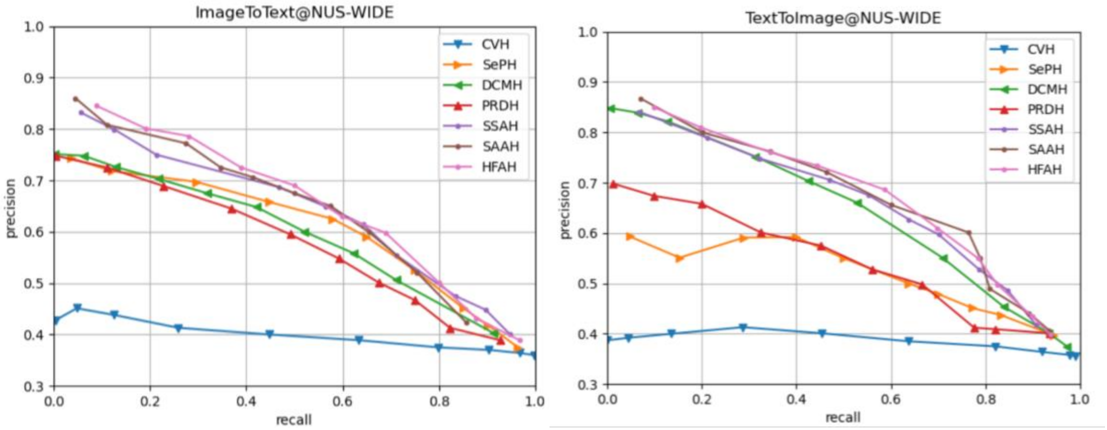

Figure 3. PR Curves of Various Methods on MIR Flickr 25K.

Table 2. 64bit MPA on MIR Flickr 25K Data Set of Ablation Experiment.

Task	Method	16bit	32bit	64bit
Image retrieval text	HFAH	0.7981	0.8143	0.8228
	HFAH-1	0.7704	0.7778	0.7935
	HFAH-2	0.7719	0.7915	0.7967
	HFAH-3	0.7776	0.7915	0.8041
Text retrieval image	HFAH	0.7976	0.8103	0.8189
	HFAH-1	0.7654	0.7767	0.7827
	HFAH-2	0.7679	0.7801	0.7847
	HFAH-3	0.7760	0.7854	0.7869

CONCLUSION

In this paper, we propose a new cross-modal retrieval method based on hash learning, that is, hybrid-attention based feature reconstruction adversarial hash (HFAH) method. This method can solve the cross-modal retrieval task more effectively. The proposed method framework mainly includes four parts: label semantic guidance module, text data and image data feature extraction module and hashing learning module. The label semantic guidance module maximizes the semantic guidance of category labels by making full use of multiple labels carried by data. The feature extraction module of text data and image data is mainly used to learn powerful feature representation of data. Through the introduction of hybrid-attention and multi-scale fusion module, more semantic information is given to the extracted features. The features reconstructed in the hashing learning module are trained through adversarial learning to maximize the semantic relevance between and within data modalities, so that the original similar sample data still maintains the similar relationship when mapped to the Hamming space. In order to improve the accuracy of our proposed method, we conducted experiments on two benchmark datasets, and compared with several representative advanced cross media retrieval methods, HFAH has achieved relatively leading retrieval performance.

REFERENCES

Peng, Y., Huang, X. & Zhao, Y. 2017. An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges. IEEE Transactions on Circuits and Systems for Video Technology. 99.

Feng, X., Hu, Z. & Liu, C. 2021. Survey of Research Progress on Cross-modal Retrieval. Computer Science. 48(8) 11.

Liu, Y., Guo, Y., Fang, J., Fan, J., Hao, Y. & Liu, J. 2022. A Survey of Research on Deep Learning Image-Text Cross-Modal Retrieval. Journal of Frontiers of Computer Science and Technology. 16(3) 489-511.

Hotelling, H. 1935. Relations Between Two Sets of Variates. Biometrika. 28 321-377.

Ngiam, J., Khosla, A., Kim, M., Nam, J. & Ng, A.Y. 2009. Multimodal Deep Learning. International Conference on Machine Learning. DBLP.

Liang, J., Li, Z., Cao, D., He, R. & Wang, J. 2016. Self-Paced Cross-Modal Subspace Matching. ACM SIGIR FORUM 569-578. ACM.

Zhai, X., Peng, Y. & Xiao, J. 2013. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. National Conference on Artificial Intelligence. AAAI Press.

Jiang, X., Fei, W., Xi, L., Zhou, Z. & Zhuang, Y. 2015. Deep Compositional Cross-modal Learning to Rank via Local-Global Alignment. the 23rd ACM international conference. ACM.

Wang, S., Zhang, D., Yan, L. & Quan, P. 2012. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE.

Clinchant, S., Ah-pine, J. & Csurka, G. 2011. Semantic combination of textual and visual information in multimedia retrieval. ICMR '11: Proceedings of the 1st ACM International Conference on Multimedia Retrieval. April 2011 1-8.

Yi, Y., Fei, W., Dong, X., Zhuang, Y. & Chia, L.T. 2010. Cross-media retrieval using query dependent search methods. Pattern Recognition. 43(8) 2927-2936.

Tong, H., He, J., Li, M., Zhang, C. & Ma, W. 2005. Graph based multi-modality learning.

Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005. ACM.

Gionis, A. 1999. Similarity Search in High Dimensions via Hashing. Proc of VLDB Conference. Morgan Kaufmann Publishers Inc.

Shen, F., Shen, C., Liu, W. & Shen, H.T. 2015. Supervised Discrete Hashing.2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Tao, Y. 2017. Research on Cross-media Retrieval Based on Hash Method. Dalian University of Technology.

Jiang, Q. & Li, W. J. 2017. Deep Cross-Modal Hashing. IEEE Computer Society. 3270-3278.

Zhou, J., Ding, G. & Guo, Y. 2014. Latent semantic sparse hashing for cross-modal similarity search general terms. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 415–424.

Chen, D., Cheng, M., Min, C. & Jing, L.2020. Unsupervised Deep Imputed Hashing for Partial Cross-modal Retrieval. 2020 International Joint Conference on Neural Networks (IJCNN).

Zhang, J. & Peng, Y. 2020. Multi-Pathway Generative Adversarial Hashing for Unsupervised Cross-Modal Retrieval. IEEE Transactions on Multimedia. 22(1) 174-187.

Li, H., Zhang, C., Jia, X., Gao, Y. & Chen, C. 2021. Adaptive Label Correlation Based Asymmetric Discrete Hashing for Cross-modal Retrieval. IEEE Transactions on Knowledge and Data Engineering.

Wang, L. Zareapoor, M., Yang, J. & Zheng, Z. 2022. Asymmetric Correlation Quantization Hashing for Cross-modal Retrieval. IEEE Transactions on Multimedia. 24 3665 – 3678.

Qin, J., Fei, L., Zhu, J., Wen, J. & Wu, S. 2021. Scalable Discriminative Discrete Hashing For Large-Scale Cross-Modal Retrieval. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Wang, D., Zhang, C., Wang, Q., Tian, Y., He, L.& Zhao, L. 2022. Hierarchical Semantic Structure Preserving Hashing for Cross-Modal Retrieval. IEEE Transactions on Multimedia.

Xie, D., Deng, C., Li, C., Liu, X. & Tao, D. 2020. Multi-Task Consistency-Preserving Adversarial Hashing for Cross-Modal Retrieval. IEEE Transactions on Image Processing. 29 3626 - 3637

Yang, E., Deng, C., Liu, W., Liu, X., Tao, D. & Gao, X. 2017. Pairwise Relationship Guided Deep Hashing for Cross-Modal Retrieval. AAAI 1618-1625.

Li, C., Deng, C., Li, N. Liu, W., Gao, X. & Tao, D. 2018. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Huiskes, M. J. & Lew, M. S. 2008. The MIR Flickr Retrieval Evaluation. In ACM: CIVR 39–43.

Chua, T., Tang, J., Hong, R., Li, H., Luo, Z. & Zheng, Y. 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In ACM: CIVR pp 48.

Kumar, S. & Udupa, R. 2011. Learning Hash Functions for Cross-View Similarity Search.

International Joint Conference on Artificial Intelligence. AAAI Press.

Lin, Z., Ding, G., Hu, M. & Wang, J. 2015. Semantics-preserving hashing for cross-view retrieval. Computer Vision & Pattern Recognition. IEEE.

Li, M., Li, Q., Ma, Y. & Yang, D. 2022. Semantic-guided autoencoder adversarial hashing for large-scale cross-modal retrieval. Complex & Intelligent Systems 8(2) 1603-1617.