Deep Discrete Cross-Modal Hashing for Cross-Media Retrieval

Fangming Zhong, Zhikui Chen, Geyong Min

Abstract

Cross-modal hashing has drawn increasing research interests in multimedia retrieval due to the explosive growth of multimedia big data. It is such a challenging topic due to the heterogeneity gap and high storage cost. However, most of the previous methods based on conventional linear projections and relaxation scheme fail to capture the nonlinear relationship among samples and suffers from large quantization loss, which result in an unsatisfactory performance of cross-modal retrieval. To address these issues, this paper is dedicated to learning discrete nonlinear hash functions by deep learning. A novel framework of cross-modal deep neural networks is proposed to learn binary codes directly. We formulate the similarity preserving in the framework, and also bit-independent as well as binary constraints are imposed on the hash codes. Specifically, we consider intra-modality similarity preserving at each hidden layer of the networks. Inter-modality similarity preserving is formulated by the output of each individual network. By so doing, the cross correlation can be encoded into the network training (i.e. hash functions learning) by back propagation algorithm. The final objective is solved by alternative optimization in an iterative fashion. Experimental results on four datasets i.e. NUS-WIDE, MIR Flickr, Pascal VOC, and LabelMe demonstrate the effectiveness of the proposed method, which is significantly superior to state-of-the-art cross-modal hashing approaches.

1. Introduction

Recently, multimedia retrieval has gained considerable attention, due to the explosive growth of multimedia data on the Internet, especially in the social network with social media big data being produced by users. As a significant component of multimedia retrieval, cross-modal retrieval has drawn increasing interests in many real life applications, such as visual search [1], image captioning [2], and machine translation [3]. However, the heterogeneity gap [4] among various modality data, such as images and texts, makes it difficult to perform Approximate Nearest Neighbor (ANN) search. In addition, the cross-modal retrieval in large-scale and high-dimensional datasets becomes quite challenging due to the high storage cost and computational complexity. To address these challenges, hashing has attracted increasing research interests due to its effectiveness in reducing storage cost and improving retrieval speed. The goal of hashing is to learn hash functions which can map high-dimensional data to a Hamming space in which data are represented by compact binary codes. Thus, the similarities in ANN search can be measured by Hamming distances, which can be obtained efficiently via bit-wise XOR operation in the expected Hamming space [5]. Hence, cross-modal hashing (CMH), which projects different modalities data into a common Hamming space enabling cross-modal retrieval is receiving more and more attentions. It is crucial that learning discriminative hash functions for each modality to boost the cross-modal retrieval.

There is a wide range of cross-modal hashing approaches proposed in recent years [5–15]. Cross-view hashing (CVH) [11] formulated the problem of learning hash functions as a generalized eigenvalue problem. Most recently, Collective Matrix Factorization Hashing (CMFH) [12] uses matrix factorization to learn the latent concepts from each modality which has achieved an impressive result on cross-modal retrieval. Inspired by CMFH, several extensions based on matrix factorization have been proposed to formulate the supervised label information, such as supervised matrix factorization hashing (SMFH [5], SMFCMH [13]), cluster-based joint matrix factorization (C-JMFH) [16], and Supervised CMFH (SCMH) [17] etc. In particular, SMFCMH [13] integrates the graph regularization into the collective non-negative matrix factorization. Furthermore, label information is used to refine the graph regularizer. With the supervised label information being taken in to account, SMFCMH learns more discriminative hash codes. Zhou et al.
[18] proposed Latent Semantic Sparse Hashing (LSSH), which learns the semantic concepts of images and text by sparse coding and matrix factorization respectively. The learned latent semantic features from images and text are then mapped to a common abstraction space in which the unified hash codes are generated by quantization. Although a number of efforts have been made on achieving impressive performance on cross-modal retrieval, there are still several limitations that remain to be exploited in cross-modal hashing.

One fundamental limitation of these methods is that the binary constraints of hash codes are relaxed to real values. In order to learn compact binary codes, the discrete constraints are imposed to most of the existing object functions. However, the object function with mixed-integer optimization results in an NP-hard optimization problem. To simplify the optimization involved in the binary hash functions learning, most of them discard the discrete constraints and then address the problem in a real-valued space with continuous solution. Then, the binary codes are obtained by quantizing the continuous solution. Unfortunately, using such a relaxation scheme will lead to large quantization loss. In such case, the accumulated quantization error, especially when learning long binary codes, is bound to degrade the discrimination capability of binary codes. To this end, it is crucial to learn binary codes directly and design discrete hash functions that can minimize the quantization error.

Another limitation of the recent methods is that they fail to formulate the nonlinear relationship of instances in each modality. Most of the existing cross-modal hashing methods [5,12,13,16,17] impose linear transformations as hash functions that project data from the original high-dimensional space into a lower dimensional Hamming space. Although impressive performances are achieved, however, in many applications, the data are linearly inseparable and thus the nonlinear manifold structure cannot be well captured by such simple linear projections. Therefore, to encourage the learned hash functions for encoding the nonlinear relationship of samples, it is an urgent required to design innovative nonlinear hash functions.

To tackle the nonlinear hashing challenge, several promising approaches introduced deep learning to cross-modal retrieval, [19–23]. However, few work explored the similarity preservation including intra-modality and inter-modality similarities in the learning phase of hash functions. Generally, the instances from different modalities while describing the same semantic object should share similar hash codes. In contrast, the samples representing different semantic objects are supposed to be pushed far away with dissimilar binary codes.

To address the above challenges, we propose a discrete cross-modal hashing based on deep neural network which is termed Deep Discrete Cross-Modal Hashing (DDCMH). We learn the compact binary codes directly by formulating it as a quantization optimization problem incorporated into the final object function. Owing to the successful application in feature extraction in computer vision of deep learning, it is reliable to learn high level nonlinear hash functions by using deep neural network. Therefore, we develop a new framework of cross-modal deep neural networks to seek multiple hierarchical nonlinear transformations to jointly learn compact binary codes and nonlinear hash functions. By so doing, the nonlinear relationship of instances in each modality would be well captured. Furthermore, we consider the similarity preservation in the learning of hash functions. We embed the intra-modality similarity preservation into every hidden layer of each modality. Additionally, the inter-modality preservation is formulated between the outputs of two deep networks at the top layers. Thus, the cross-modal correlation will be incorporated into the updating of deep neural networks by back propagation algorithm, which encourages the learned networks i.e. hash functions, to be more discriminative. Moreover, inspired by [24,25] that learn binary codes using classifiers, we integrate a linear classification with expected binary codes as input and label information as output. The framework of DDCMH is illustrated in Fig. 1. We employ the visual and text features extracted from images and texts as input of our method. In the testing phase, a query is also firstly transformed to visual or text representations followed by hash codes learning. Finally, cross-modal retrieval can be conducted based on the generated binary codes.

Motivated by nonlinear discrete hashing (NDH) [25], our proposed DDCMH is also optimized under four constraints: 1) quantization loss of learned binary codes and the output of cross-modal deep neural networks, 2) intra-modality similarity preservation in each hidden layer of each modality and inter-modality similarity preservation between the output layers of two networks, 3) independent bits in the learned binary codes, and 4) minimized the classification loss between expected binary codes and supervised label information. Comparing against the previous proposed methods, the main contributions of our work are summarized as follows:

- We develop a novel framework of cross-modal deep neural networks. Two deep neural networks are trained as hash functions to learn binary codes for image and text modality, respectively. Different from linear transformations, such cross-modal deep neural networks can well capture the nonlinear relationship in each modality.
- Binary codes of training data are learned directly without any relaxation. Furthermore, a quantization loss between deep neural networks and to-be-learned binary codes are imposed to minimize the quantization error. By so doing, we can jointly learn binary codes and hash functions with low quantization loss, which is especially significant for out-of-sample instances.
- Additionally, intra-modality and inter-modality similarity preservation is considered in the learning of nonlinear hash functions and compact binary codes. We introduced a graph regularization in each layer of the deep network to preserve intra-modality similarity. Similarly, a graph regularization based on semantic label information is imposed into the output layer of deep neural network for preserving inter-modality similarity. Moreover, the inter-modal similarity preservation will contribute to the whole network updating by back propagation algorithm. Hence, the learned binary codes will possess more discriminative power.

The rest of this paper is organized as follows. The previous work on cross-modal hashing is reviewed and analyzed in Section 2. Section 3 presents the detailed design of our proposed method. In Section 4, extensive experimental details and results are described in comparison with the state-of-the-art methods on four benchmark datasets. Finally, this work is concluded in Section 5.

2. Related Work

Due to the efficiency of retrieval and low storage cost, hashing methods are widely investigated in both unimodal hashing and cross-modal hashing. In the last decade, a number of cross-modal hashing approaches have been proposed due to the increasing interest attracted by cross-modal retrieval in various applications. In this section, we will review and analyze the main differences of these methods from three aspects: 1) inter-modality and intra-modality similarity preserving, 2) using of relaxation scheme, and 3) nonlinear relationship capturing.
2.1. Inter-modality and Intra-modality Similarity Preserving

Inter-modality similarities are considered in many cross-modal hashing methods for better learning the common semantic relationship. Among them, a few methods are unsupervised [12,18], and they measure the inter-modality relationship by training the paired samples which describes the same object. For example, Ding et al. [12] firstly proposed collective matrix factorization to learn the latent concepts from different modalities. Although CMFH does not consider the label information, it has achieved an impressive performance on cross-modal retrieval that demonstrated the power of matrix factorization in latent structure learning. Nonetheless, CMFH does not preserve the local structure information within each individual modality i.e. intra-modality similarity, nor takes the inter-modality correlations into consideration. Liu et al. [26] proposed a collaborative hashing to learn compact binary codes by preserving both intra-view and inter-view relationships simultaneously. Different from collaborative hashing which uses linear projections as hash functions, our work aims to learn the nonlinear hash functions.

Others considered the label information to enhance the common semantic relationship [11,17,27]. Cross-view hashing (CVH) [11] is extended from spectral hashing [28], which formulated the hash functions learning as a generalized eigenvalue problem. With the label information taken in to account, SMFCMH [13] constructs a graph regularization to investigate the correlations across different modalities. Different from CMFH, it uses multiplicative iteration and a subsample method for efficient optimization. Thanks to the contribution of matrix factorization and supervised label information, SMFCMH has obtained impressive retrieval performance.

By integrating the label information, users can embed the must-link or must-not-link constraints into the hash functions learning. For instance, the samples from different modalities with the same label information should share similar representation in the Hamming space. Thus, the learned hash functions will be more powerful to extract the intrinsic semantics across different modalities, as well as to improve the cross-modal retrieval performance. However, a limitation of the above methods is that, the relationships among samples in the same modality are not considered, i.e. intra-modality similarity.

In order to preserve the intra-modality relationship in each individual modality, most of the existing methods formulated it as a graph regularization. In [29], IMH explores the intra-media consistency by the affinity relationship in each modality. Rafailidis et al. [16] proposed cluster-based joint matrix factorization hashing (C-JMFH) to generate cross-modal cluster representations for instances, which are incorporated into a joint matrix factorization later to measure the inter-modality and intra-modality similarities. In [30], Wang et al. proposed leaning bridging mapping for cross-modal hashing (LBMCH) to explore the semantic correspondence of distinct Hamming spaces which can characterize the discriminative local structure for each modality. IMH, C-JMFH, and LBMCH are unsupervised methods. There are also many supervised cross-modal hashing approaches that take the intra-modality similarity into consideration [5,31,32].

One common weakness of the aforementioned approaches is that they solved the objective in the real-valued space, and obtained the binary codes with simple quantization. In such case, the discrete constraints are relaxed to avoid addressing a mixed-integer optimization problem, which is generally NP-hard. However, such relaxation mode used here will decrease the discriminative capability and accuracy of learned binary codes due to the accumulated quantization errors. Therefore, discrete hashing without relaxation are gained more and more research interests.

2.2. Discrete Hashing

Considering the large quantization error can degrade the performance of binary codes, many discrete cross-modal hashing methods extended from unimodal discrete hashing have been presented. Discrete cross-modal hashing (DCH) was proposed by Xu et al. [23], in which the discriminant binary codes are directly learned without relaxation, and label information is used to enhance the discriminability of binary codes through linear classifiers. In [33], Ding et al. proposed a rank-order preserving hashing (RoPH) to explore the ranking information when learning hash functions. The hash codes are also learned directly without any relaxation trick by using the widely used kernel hash function, which is a linear hash function. We can observe that, most of the existing discrete hashing methods, such as DCH and RoPH, learn binary codes directly without relaxation, but they cannot capture the nonlinear relationship of samples by using simple linear hash functions. The same problem occurs in the methods mentioned in subsection 2.1.

2.3. Nonlinear Relationship Capturing

Extensive efforts have been devoted to migrate deep learning to deep hashing due to its powerful semantic representation
capability. For example, in [19], a deep multimodal hashing with orthogonal regularization (DMHOR) is presented, in which multimodal autoencoder (MAE) and cross-modality autoencoder (CAE) are used to preserve intra-modality and inter-modality correlation respectively by using reconstruction scheme. Our method differs DMHOR in that the deep network model and similarity preserving scheme used are different. More importantly, DMHOR removes the discrete constraint, which results in a suboptimal binary codes.

In recent years, several discrete cross-modal hashing based on nonlinear hash functions, such as deep learning and kernel function, have been investigated. Yan et al. [24] presented a supervised robust discrete multimodal hashing (SRDMH), where the label information is preserved in the final binary codes, and a flexible $\ell_{2,p}$ loss with nonlinear embedding is integrated to make it robust to noise. Unfortunately, the hash functions based on nonlinear kernel limits the scalability. Extend from canonical correlation analysis, a model based on deep neural network was proposed in [34] to extract the canonical correlation between two modalities, which also can capture the nonlinear feature within each modality. A novel deep sketch hashing (DSH) [35] based on convolutional neural network (CNN) was proposed to address sketch-based image retrieval, which can capture the cross-view similarities and the intrinsic semantic correlations. Our work differs from DSH in that we exploit the relationship between images and texts. In [36], Cao et al. proposed a collective deep quantization for efficient cross-modal retrieval (CDQ) to jointly learn deep hash functions and quantizers for both modalities. However, CDQ ignores the intra-modality correlation preservation. In [21], Jiang et al. proposed a deep cross-modal hashing (DCM) which use deep neural network as hash functions, one for each modality. However, DCM only preserve the intra-modality similarity while neglecting the inter-modality similarity preservation, which is critical for enhancing cross-modal retrieval performance. Pairwise relationship guided deep hashing (PRDH) [22] is the extension of DCMH, which exploits different pairwise constraints to enforce the hash codes from intra-modality and inter-modality preserving. Different from PRDH, we formulate the similarity preservation as a graph regularization. Furthermore, we consider intra-modality similarity preserving in each layer of the deep network, which can reduce the accumulated errors. It can be observed that, most of the previous proposed deep cross-modal hashing models fail to formulate the similarity preservation including both intra-modality and inter-modality similarities, which are significant for nearest neighbor search across different modalities. In [37], Zhuang et al. exploited the neural network to learn the hash functions, termed cross-media neural network hashing (CMNNH), which preserves both inter-modal pairwise correspondence and intra-modal discriminative capability. However, the manifold structure in each modality is ignored as well as discrete constraint.

There are also several discrete and nonlinear hashing methods [25,38–40] proposed for unimodal tasks, such as image retrieval. For example, the quantization based methods such as adaptive binary quantization [41] and structure sensitive hashing [42] have been proposed to generate compact binary codes with discrete constraints. But these methods cannot deal with multimodal data. Chen et al. [25] proposed a nonlinear discrete hashing (NDH) which combines the discrete optimization and nonlinear hash functions. However, NDH can only deal with unimodal data, and it cannot cope with the out-of-sample data. Motivated by the promising results delivered by unimodal NDH in learning nonlinear binary codes, we make further efforts to investigate the nonlinear discrete hashing in cross-modal scenario, and explore the cross-correlation preserving in the binary codes.

3. Proposed DDCMH

In this section, the details of the proposed DDCMH are presented. For cross-modal retrieval, we take image and text as an example. Thus, the goal of our method is to jointly learn unified binary codes of image and text modalities, as well as nonlinear hash functions for image and text, respectively. After that, out-of-sample extension is introduced for new query instances from both modalities which are not present in the training database.

3.1. Problem Formulation

In this paper, matrices as represented by boldface uppercase letters, and vectors are denoted by boldface lowercase letters. Given the cross-modal data $X = [x_i]_{i=1}^{n}$ and $Y = [y_i]_{i=1}^{n}$, such as images and the associated text, $x_i$ denotes the visual feature of image $i$, and $y_i$ represents the corresponding text feature. Thus, we have $n$ samples from each modality, i.e. $X \in \mathbb{R}^{d_1 \times n}$, $Y \in \mathbb{R}^{d_2 \times n}$ where $d_1$ represents the dimension of image feature, and $d_2$ denotes the dimension of text descriptor (usually $d_1 \neq d_2$).

The goal of our work is to jointly learn unified binary codes $B$ of both modalities, and learn nonlinear hash functions for each modality. Thus, the original data can be transformed to a compact Hamming space, where the similarity of different modalities can be measured directly by Hamming distance. The hash function can be defined as:

$$f(i) : \mathbb{R}^s \mapsto \{-1, +1\}^k, t = 1, 2$$

(1)

where $k$ is the length of hash code. Here, we use $\{-1, +1\}$ to represent hash codes $B$, which can be easily transformed to binary codes via mean thresholding stated as follows:

$$H^t = \frac{1}{2}(B^t + 1), t = 1, 2$$

(2)

where $H^t \in \{0, 1\}^k \times n$ is the binary codes which can be stored directly in memory. For convenience, we define the hash codes as $\{-1, +1\}^k$ in the rest of this paper.

Here, we aim at learning nonlinear hash functions which can capture the nonlinear relationship of samples in each modality. As shown in Fig. 1, we learn two deep neural networks, one for each modality, as the hash functions. Taking image modal as an example, we assume that there are $M + 1$ layers in the deep neural network with $M$ hidden layers. The number of units in each hidden layer is denoted as $u_m$, where $1 \leq m \leq M$. Given an input data sample $x_i \in \mathbb{R}^{1 \leq i \leq n}$, the output of first hidden layer can be obtained by $h(1) = s(W(1)x_i + c(1))$, where $W(1) \in \mathbb{R}^{u(1) \times d_1}$ is the weight matrix, $c(1) \in \mathbb{R}^{u(1)}$ is the bias vector, and $s(\cdot)$ is an activation function such as tanh or sigmod function. Subsequently, the second hidden layer can be also computed as $h(2) = s(W(2)h(1) + c(2))$, where $W(2) \in \mathbb{R}^{u(2) \times u(1)}$ and $c(2) \in \mathbb{R}^{u(2)}$ are the parameters should be learned for the second layer. Analogously, we can achieve the representation of the $m$-th hidden layer by $h(m) = s(W(m)h(m-1) + c(m))$. Finally, we can have the representations at the top layer, i.e. output of the whole network $F_{st}(M) = s(W(M)h(M-1) + c(M))$, where $F_{st}(M)$ is the $i$-th column of $F(M)$ which corresponds to the dataset $X$. The binary codes are obtained by the sign of $F(M)$. Therefore, we can employ deep neural network as a nonlinear hash function stated as follows:

$$B = \text{sign}(F(M)) = \text{sign}(f(X, W_s, c_s))$$

$$s.t. B \in \{-1, +1\}^{k \times n}$$

(3)

where $\text{sign}$ is the sign function, $\text{sign}(u) = 1$ if $u > 0$ and -1 otherwise for all $u \in \mathbb{R}$.

For cross-modal hashing, we intend to learn hash functions for each modality. Thus, the hash functions for image modality $X$, and
text modality $Y$ can be written as:
\[
B^{(x)} = \text{sign}(F^{(M)}_x(x)) = \text{sign}(f(x)(X, W_x, c_x))
\] (4)
\[
B^{(y)} = \text{sign}(F^{(M)}_y(y)) = \text{sign}(f(y)(Y, W_y, c_y))
\] (5)
where $B^{(x)} \in \{-1, +1\}^{k \times n}$ and $B^{(y)} \in \{-1, +1\}^{k \times n}$ are the learned binary codes, $F^{(M)}_x$ and $F^{(M)}_y$ are the outputs of the top layers from deep neural networks for image and text, respectively. In this work, the aim is to jointly learn the unified binary codes and nonlinear hash functions. Therefore, we set the to-be-learned hash codes as the same, i.e., $B = B^{(x)} = B^{(y)}$. In the following of this paper, we mainly focus on investigating the unified binary codes from the image and text modalities.

In order to learn more discriminative binary codes, and to preserve inter-modality similarity, we incorporate supervised label information into the proposed framework. Thus we also have the label information of given cross-modal data. Let $C$ denote the label matrix $C \in \mathbb{N}^{c \times n}$, where $c$ is the total number of classes. Since the cross-modal data describe the same objects, they share the same label information. The $i$-th column of $C$ i.e., $C_i \in \{0, 1\}^c$ is the label of image $x_i$ and text $y_i$. We assume that the images and texts belong to at least one of the $c$ classes. If the $i$-th item belongs to the $j$-th class, $C_{ij} = 1$. Otherwise $C_{ij} = 0$.

### 3.2. Objective Function

In order to learn more discriminative binary codes and more flexible hash functions, the objective function consists of four main parts as stated in Eq. (6):
\[
\arg \min_B \mathcal{L}(C, B) + \mathcal{L}(B, X) + \mathcal{L}(B, Y) + \mathcal{L}(X, Y)
\] (6)
where $\mathcal{L}(C, B)$ represents the label consistency preserving term, $\mathcal{L}(B, X)$ and $\mathcal{L}(B, Y)$ are the loss information between original data and the to-be-learned unified binary codes, and $\mathcal{L}(X, Y)$ embeds the cross-modal correlation, which enables inter-modality similarity preserving.

#### 3.2.1. Label Consistency Preserving

To take advantage of the labels, and also motivated by the linear classification, we impose a classification error to ensure that the learned binary codes preserve the label information. Hence, given the label information $C \in \mathbb{N}^{c \times n}$, we aim at learning a linear classifier $P \in \mathbb{R}^{c \times n}$, which can classify the binary codes into different categories accurately. The classification error of linear classifier is defined as:
\[
\mathcal{L}(C, B) = \left\| C - PB \right\|_F^2
\]
\[
s.t. \quad B \in \{-1, +1\}^{k \times n}
\] (7)
where $\left\| \cdot \right\|_F^2$ represents the squared Frobenius norm. Here, the widely used least square loss function is employed to measure the error between predicted labels and supervised label information.

#### 3.2.2. Loss Information Modeling

In addition, modeling the loss information can facilitate the direct binary codes learning and nonlinear hash functions learning, which are the main goal of this paper. Different hash functions are learned for each modality. Thus, we model the loss information for both image and text modalities. Here, the loss information is composed of quantization loss and intra-modality similarity preserving term. As mentioned previously, we use the output of deep neural network as the binary codes. Thus, the quantization loss is computed as follows:
\[
\mathcal{L}(B, F^{(M)}_x) = \left\| B - F^{(M)}_x \right\|_F^2
\] (8)
\[
\mathcal{L}(B, F^{(M)}_y) = \left\| B - F^{(M)}_y \right\|_F^2
\] (9)
We use Laplacian Eigenmaps (LE) [43] to formulate the similarity preserving based on unipolarized weight graph which indicates neighbor relations of pairwise data. Hence, the objective function of intra-modality similarity preservation in the top layer of each modality can be stated as follows:
\[
\min_{F^{(M)}_x} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( F^{(M)}_x(:, i) - F^{(M)}_x(:, j) \right)^2 S^{(x)}_{ij}
\] (10)
\[
\min_{F^{(M)}_y} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( F^{(M)}_y(:, i) - F^{(M)}_y(:, j) \right)^2 S^{(y)}_{ij}
\] (11)
where $S^{(x)}$ and $S^{(y)}$ are the affinity matrices of the image modality and text modality, respectively. $S^{(y)}$ is defined as below:
\[
S^{(y)}_{ij} = \begin{cases} 
1, & \text{if } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i), \\
0, & \text{otherwise}.
\end{cases}
\] (12)
where $N_k(x_i)$ represents the $k$-nearest neighbors of $x_i$. Here, $k$ is a constant, and the neighborhood relationships are computed based on Euclidean distance. $S^{(y)}$ can be calculated similarly to $S^{(x)}$. Accordingly, Eq. (10) can be rewritten as the form of graph regularization stated as follows:
\[
\min_{F^{(M)}_x} {\text{tr}}(F^{(M)}_x L_x(F^{(M)}_x)^T)
\] (13)
where $L_x$ is the Laplacian matrix of $S^{(x)}$, $L_x = D_x - S^{(x)}$, where $D_x \in \mathbb{R}^{n \times n}$ is a diagonal matrix. The diagonal entries of $D_x$ are the column sum of $S^{(x)}$, i.e., $D_x(i, i) = \sum_j S^{(x)}_{ij}$. Similarly, we have the graph regularization of text modality as:
\[
\min_{F^{(M)}_y} {\text{tr}}(F^{(M)}_y L_y(F^{(M)}_y)^T)
\] (14)
where $L_y = D_y - S^{(y)}$, $D_y \in \mathbb{R}^{n \times n}$, $D_y(i, i) = \sum_j S^{(y)}_{ij}$, and $\text{tr} \cdot$ is the trace norm of a matrix.

Therefore, the information loss of image modality including both quantization loss and intra-modality similarity preservation is defined as:
\[
\mathcal{L}(B, X) = \left\| B - F^{(M)}_x \right\|_F^2 + \alpha^{(x)} \text{tr}(F^{(M)}_x L_x(F^{(M)}_x)^T)
\] (15)
where $\alpha^{(x)}$ is the trade-off parameter. Similarly, the information loss for text modality can be stated as follows:
\[
\mathcal{L}(B, Y) = \left\| B - F^{(M)}_y \right\|_F^2 + \alpha^{(y)} \text{tr}(F^{(M)}_y L_y(F^{(M)}_y)^T)
\] (16)

Additionally, to reduce the accumulated error during the layer-wise training of deep neural work, we consider the local structure consistency preserving in each hidden layer. Thus, we have a new loss function as shown in Eq. (17).
\[
\mathcal{L}(B, X) = \mathcal{L}^{(M)}_k + \sum_{m=1}^{M} \alpha^{(m)} \tau^{(m)}
\]
\[
s.t. \quad B \in \{-1, +1\}^{k \times n}, \quad BB^T = nI_k
\] (17)
where
\[
\mathcal{L}^{(M)}_k = \left\| B - F^{(M)}_x \right\|_F^2 + \alpha^{(y)} \text{tr}(F^{(M)}_x L_x(F^{(M)}_x)^T)
\] (18)
represents the loss information on the top layer of image deep neural network, and
\[
\mathcal{L}^{(m)}_k = \text{tr}(F^{(M)}_x L_x(F^{(M)}_x)^T), \quad m = 1, 2, \ldots, M - 1
\] (19)
is the intra-modality similarity preserving term for each hidden layer. Concerning that the deep neural network might be interfered
by the direct pursuit of local consistency preserving at each hidden layer [25], we weak the impact of graph regularization during training by employing a hinge loss, which is defined as:

$$h(x) = \max(x, 0).$$

(20)

As shown in Eq. (17), we can control the regularization term to appear or not by setting a positive threshold $\tau^{(m)}$. In other words, the second term in Eq. (17) will equal to 0 when the regularization term $\mathcal{L}^{(m)} \leq \tau^{(m)}$. In this case, it will not impact the learning of the whole deep neural network. $\alpha^{(m)}$ is a trade-off parameter, which balances the effects of top layer and all hidden layers.

Analogously, the loss function of text modality can be defined as:

$$\mathcal{L}(\mathbf{Y}, \mathbf{Y}) = \mathcal{L}^{(M)} + \sum_{m=1}^{M} \frac{\alpha^{(m)}}{\alpha^{(m)}} h(\mathcal{R}^{(m)} - \tau^{(m)})$$

s.t. $\mathbf{B} \in \{-1, +1\}^{k \times n}$, $\mathbf{BB}^T = n \mathbf{I}_k$

(21)

where

$$\mathcal{L}^{(M)} = ||\mathbf{B} - \mathbf{F}^{(m)}||^2_F + \alpha^{(M)} \mathbf{tr} \left[ \mathbf{F}^{(M)} \mathbf{L} \mathbf{F}^{(M)^T} \right]$$

(22)

$$\mathcal{L}_{\alpha}^{(m)} = \mathbf{tr} \left[ \mathbf{F}^{(m)} \mathbf{L} \mathbf{F}^{(m)^T} \right]$$

(23)

3.2.3. Bit-independent Constraints

In order to make different bits of the binary codes be independent with each other, we enforce orthogonality constraints between the rows of $\mathbf{B}$, i.e. $\mathbf{BB}^T = n \mathbf{I}_k$, which can further least information redundancy. However, the bit-independent constraints will make hashing computationally intractable. Most of the existing approaches relax the independent constraint, which can lead to a poor performance of learning hash codes. In contrast, we take into consideration the independent constraint without resorting to such error-prone relaxations, by defining a real-valued matrix set

$$\Omega = \{ \mathbf{f} \in \mathbb{R}^{k \times n} | \mathbf{f} \mathbf{f}^T = n \mathbf{I}_k \}.$$ 

Thus, we can transform the independent constraint to minimizing the distance from matrix $\mathbf{B}$ to any matrix $\mathbf{f} \in \Omega$, which can be formulated as:

$$\mathcal{L}(\mathbf{B}, \mathbf{F}) = ||\mathbf{B} - \mathbf{F}||^2_F$$

(24)

3.2.4. Inter-modality Similarity Preserving

Up until now, we have not made any assumptions on the cross-modality correlation, due to the modality specific hash functions learning. Next, we consider the correlation of image and text modalities by modeling the inter-modality similarity preserving. Similar to intra-modality, the inter-modality similarity preserving can be formulated as:

$$\min_{\mathbf{f}_x, \mathbf{f}_y} \sum_{i=1}^{n} \sum_{j=1}^{n} ||\mathbf{F}^{(x)}(i, \cdot) - \mathbf{F}^{(y)}(j, \cdot)||^2_S^{(xy)}$$

(25)

where $S^{(xy)}$ denotes the affinity matrix across two modalities $\mathbf{X}$ and $\mathbf{Y}$. $S^{(xy)}$ is computed based on the label information as follows:

$$S^{(xy)}_{ij} = \begin{cases} 1, & \text{if } x_i \text{ and } y_j \text{ have the same label,} \\ 0, & \text{otherwise.} \end{cases}$$

(26)

Through algebraic calculation, the objective function defined in Eq. (25) can be reformulated as:

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}) = \mathbf{tr} \left[ \mathbf{F}^{(x)} \mathbf{D}^{(x)} \mathbf{F}^{(x)^T} \right] + \mathbf{tr} \left[ \mathbf{F}^{(y)} \mathbf{D}^{(y)} \mathbf{F}^{(y)^T} \right]$$

$$- \mathbf{tr} \left[ \mathbf{F}^{(x)} \mathbf{S}^{(xy)} \mathbf{F}^{(y)^T} \right] - \mathbf{tr} \left[ \mathbf{F}^{(y)} \mathbf{S}^{(xy)} \mathbf{F}^{(x)^T} \right]$$

(27)

where $\mathbf{D}^{(x)}$ is the diagonal matrix whose entries are $d^{(x)}_{ij} = \sum_j S^{(xy)}_{ij}$, $d^{(y)}_{ij} = d^{(x)}_{ij}$, and $S^{(xy)} = (S^{(xy)})^T$ are the affinity matrices defined as Eq. (26).

3.2.5. Overall Objective Function

Consisting of label consistency preserving, deep discrete hash functions learning, intra-modality similarity preserving, bit-independent constraints, and inter-modality similarity preservation, the overall objective function of our proposed DDCMH can be defined as follows:

$$\arg \min_{\mathbf{F}, \mathbf{B}, \mathbf{C}} \mathcal{L} = \mathcal{L}(\mathbf{C}, \mathbf{B}) + \lambda_1 \mathcal{L}(\mathbf{B}, \mathbf{F})$$

$$+ \lambda_2 (\mathcal{L}(\mathbf{X}, \mathbf{Y}) + \mathcal{L}(\mathbf{B}, \mathbf{Y}))$$

$$+ \lambda_3 \mathcal{R}_{\alpha}(\cdot)$$

s.t. $\mathbf{B} \in \{-1, +1\}^{k \times n}$

(28)

where $\lambda_1, \lambda_2, \lambda_3$, and $\lambda_4$, are trade-off parameters of the corresponding terms, and $\mathcal{R}(\cdot)$ denotes the regularization term to avoid overfitting defined as follows:

$$\mathcal{R}(\cdot) = \sum_{m=1}^{M} \left( ||\mathbf{W}^{(m)}||^2_F + ||\mathbf{C}^{(m)}||^2_F + ||\mathbf{W}^{(m)}||^2_F + ||\mathbf{C}^{(m)}||^2_F + ||\mathbf{P}||^2_F \right)$$

(29)

3.3. Optimization Algorithm

Since the optimization problem in Eq. (28) is non-convex with matrix variables $\mathbf{P}, \mathbf{B}, \mathbf{W}^{(m)}, \mathbf{C}^{(m)}$ ($m = 1, 2, \cdots, M$), and $\mathbf{F}$, it is intractable to be directly minimized. Fortunately, it is convex with respect to any one of the five variables in the case that the others are fixed [5]. Therefore, we propose an alternative optimization algorithm in an iterative manner to address the optimization problem until convergence. The detailed optimization steps are listed as follows:

P-step: Fix other variables but $\mathbf{P}$, then the objective function shown in Eq. (28) can be simplified as:

$$\arg \min_{\mathbf{P}} \mathcal{L} = ||\mathbf{C} - \mathbf{P}^{(x)}||^2_F + \lambda_4 ||\mathbf{P}||^2_F$$

(30)

Let $\frac{\partial \mathcal{L}}{\partial \mathbf{P}} = 0$, we can have the closed-form solution stated as follows:

$$\mathbf{P} = (\mathbf{BB}^T + \lambda_4 \mathbf{I})^{-1} \mathbf{BC}$$

(31)

F$_x$-step: The objective of optimizing $\mathbf{F}^{(m)}_x$ is learning the image deep neural network as a hash function. It can be transformed to the following sub-problem:

$$\arg \min_{\mathbf{W}^{(m)}, \mathbf{C}^{(m)}} \mathcal{L} = \lambda_2 \mathcal{L}^{(M)} + \sum_{m=1}^{M} \alpha^{(m)} h(\mathcal{R}^{(m)} - \tau^{(m)})$$

$$+ \lambda_4 \mathcal{R}_{\alpha}(\cdot)$$

(32)

where the final objective is to learn the parameters $\mathbf{W}^{(m)}$ and $\mathbf{C}^{(m)}$ at the top layer are calculated as follows:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(m)}} = \lambda_2 \frac{\partial \mathcal{L}^{(M)}}{\partial \mathbf{W}^{(m)}} + \lambda_3 \frac{\partial \mathcal{L}^{(X, Y)}}{\partial \mathbf{W}^{(m)}} + 2 \lambda_4 \mathbf{W}^{(m)}$$

(33)

$$\frac{\partial \mathcal{L}}{\partial \mathbf{C}^{(m)}} = \lambda_2 \frac{\partial \math{L}^{(M)}}{\partial \mathbf{C}^{(m)}} + \lambda_3 \frac{\partial \mathcal{L}^{(X, Y)}}{\partial \mathbf{C}^{(m)}} + 2 \lambda_4 \mathbf{C}^{(m)}$$

(34)

While for each hidden layer $m (m = 1, 2, \cdots, M - 1)$, the gradients with respect to $\mathbf{W}^{(m)}$ and $\mathbf{C}^{(m)}$ are computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(m)}} = \lambda_2 \left( \frac{\partial \mathcal{L}^{(M)}}{\partial \mathbf{W}^{(m)}} + \sum_{t=m}^{M-1} \alpha^{(t)} h(\mathcal{R}^{(t)} - \tau^{(t)}) \frac{\partial \mathcal{L}^{(t)}}{\partial \mathbf{W}^{(m)}} \right)$$

(35)
\[ + \lambda_3 \frac{\partial L(X, Y)}{\partial W^{(m)}} + 2\lambda_4 W^{(m)}_x \]  
(35)

where \( h'(x) \) is the derivative function of hinge loss \( h(x) \), which equals 0 at non-differentiable point \( x = 0 \). Since the network is computed by forward propagation, we have to discuss different cases for computing \( \frac{\partial c}{\partial c^{(m)}} \) and \( \frac{\partial c}{\partial c^{(m)}} \). The \( F_x^{(m)} \) is computed firstly by forward propagation using the following formula:

\[ F_x^{(m)} = s(W_x^{(m)}F_{x}^{(m-1)} + c_x^{(m)}) \]  
(37)

Hence, the derivations of \( F_x^{(m)} \) with respect to \( W_x^{(m)} \) and \( c_x^{(m)} \) are stated as follows:

\[ \frac{\partial F_x^{(m)}}{\partial W_x^{(m)}} = s'(W_x^{(m)}F_{x}^{(m-1)} + c_x^{(m)}) \]  
(38)

\[ \frac{\partial F_x^{(m)}}{\partial c_x^{(m)}} = s'(W_x^{(m)}F_{x}^{(m-1)} + c_x^{(m)}) \]  
(39)

Thus, for \( 1 \leq m \leq l \leq M \), we have:

\[ \frac{\partial c_x^{(m)}}{\partial c_x^{(m)}} = \frac{\partial c_x^{(m)}}{\partial F_x^{(m-1)}} (F_x^{(m-1)})^T \]  
(40)

While for \( 1 \leq m < l \leq M \), the gradient of \( c_x^{(l)} \) with respect to \( c_x^{(m)} \) is computed by:

\[ \frac{\partial c_x^{(m)}}{\partial c_x^{(m)}} = 2F_x^{(m)}L_x \odot s'(Z_x^{(m)}) \]  
(41)

where \( Z_x^{(m)} \) is the total weighted sum of inputs in \( m \)-th layer, and \( s'(\cdot) \) is derivative function of active function. Otherwise,

\[ \frac{\partial c_x^{(m)}}{\partial c_x^{(m)}} = 2\alpha_x^{(M)}F_x^{(m)}L_x + F_x^{(m)} - B \odot s'(Z_x^{(M)}) \]  
(42)

when \( m = l = M \).

If \( 1 \leq m < l \leq M \), \( \frac{\partial c_x^{(m)}}{\partial c_x^{(m)}} \) can be computed as:

\[ \frac{\partial c_x^{(m)}}{\partial c_x^{(m)}} = (W_x^{(m+1)})^T \frac{\partial c_x^{(m)}}{\partial c_x^{(m)}} \odot s'(Z_x^{(m)}) \]  
(43)

Additionally, we have a inter-modality similarity preserving term at the top layer to connect cross-modalities. The gradients of \( \mathcal{L}(X, Y) \) with respect to \( W_x^{(m)} \) and \( c_x^{(m)} \) at the top layer of image modality are defined as:

\[ \frac{\partial \mathcal{L}(X, Y)}{\partial W_x^{(m)}} = \frac{\partial \mathcal{L}(X, Y)}{\partial c_x^{(m)}} (F_x^{(m-1)})^T \]  
(44)

\[ \frac{\partial \mathcal{L}(X, Y)}{\partial c_x^{(m)}} = 2(F_x^{(m)}D_{xy} - F_y^{(M)}S_{xy}) \odot s'(Z_x^{(M)}) \]  
(45)

For the hidden layer \( m = 1, 2, \ldots, M - 1 \), the gradient of \( \mathcal{L}(X, Y) \) with respect to \( c_x^{(m)} \) is computed as:

\[ \frac{\partial \mathcal{L}(X, Y)}{\partial c_x^{(m)}} = (W_x^{(m+1)})^T \frac{\partial \mathcal{L}(X, Y)}{\partial c_x^{(m)}} \odot s'(Z_x^{(m)}) \]  
(46)

Similarly, we have

\[ \frac{\partial \mathcal{L}(X, Y)}{\partial W_x^{(m)}} = \frac{\partial \mathcal{L}(X, Y)}{\partial c_x^{(m)}} (F_x^{(m-1)})^T \]  
(47)

From Eq. (46) and Eq. (47), we can observe that, the cross-modal correlation are embedded into the learning phase of image modality. That is to say, our proposed method can learn discriminative codes carrying the cross-modal relationship.

Finally, we can update the network of image modality by:

\[ W_x^{(m)} = W_x^{(m)} - \eta \frac{\partial \mathcal{L}}{\partial W_x^{(m)}} \]  
(48)

\[ c_x^{(m)} = c_x^{(m)} - \eta \frac{\partial \mathcal{L}}{\partial c_x^{(m)}} \]  
(49)

where \( \eta \) denotes the learning rate.

**F**<sub>2</sub> - step: To learn the deep neural network for text modality, we have the similar procedure of that for image modality. **Algorithm 1** summarizes the training procedure of cross-modal deep neural network for image modality. Since the deep neural network training of image modality and text modality is similar, we can train the network for text modality using **Algorithm 1** with different inputs.

**B**-step: Here, we learn the unified binary codes directly by solving the reformulated optimization stated as follows:

\[ \arg \min_{B} \mathcal{L} = \left\| C - P^T B \right\|^2_F + \lambda_1 \left\| B - \Phi \right\|^2_F + \lambda_2 \left\| B - \Phi \right\|^2_F + \lambda_3 \left( \left\| B - F_x^{(M)} \right\|^2_F + \left\| B - F_y^{(M)} \right\|^2_F \right) \]  
(50)

Due to the binary constraints \( B \in \{-1, +1\}^{k \times n} \), it is challenging to address the optimization in Eq. (50). Fortunately, we can have a closed-form solution to a single row of \( B \) when fixing the other rows. Consequently, \( B \) can be optimized by using discrete cyclic coordinate descent (DCC) algorithm [44].

**Φ**-step: In order to obtain the optimal \( \Phi \), we have

\[ \arg \min_{\Phi} \mathcal{L} = \lambda_1 \left\| B - \Phi \right\|^2_F \]  
(51)

The solution of this problem can be easily achieved with the aid of the singular value decomposition (SVD) of \( B \).

The overall algorithm of our DDCMH is summarized in **Algorithm 2**.
Algorithm 2: Discrete Deep Cross-Modal Hashing

**Input:** Training data \( X \) and \( Y \), label \( C \), the length of hash codes \( k \), iterative number \( T_2 \), number of layers \( M \), network structure \((s_x^{(m)}, s_y^{(m)})_{m=1}^{M}\), parameters \( \lambda_1, \lambda_2, \lambda_3, \lambda_4, \alpha_x^{(m)}, \alpha_y^{(m)}, \gamma_y^{(m)} \), \( m=1 \).

**Output:** \((W_x^{(m)}, c_x^{(m)})_{m=1}^{M}, (W_y^{(m)}, c_y^{(m)})_{m=1}^{M}, B\).

Initialize \( B, \Phi, (W_x^{(m)}, c_x^{(m)}, W_y^{(m)}, c_y^{(m)})_{m=1}^{M}\).

Calculate \( S^{(y)}, D^{(y)}, L_x, L_y \).

for \( r = 1, 2, ..., T_2 \) do

**P-step:** update \( P \):

**F-step:** update \((W_x^{(m)}, c_x^{(m)})_{m=1}^{M}\);

**\( \Phi \)-step:** update \( \Phi \);

**B-step:** update \( B \);

Compute \( L \) with Eq. (28);

if convergence then
  return;
end

end

3.4. Out-of-Sample Extension

For the new coming query data, we firstly transform them to binary codes. In particular, given a query sample from image modality \( X \), we can generate the corresponding binary code by using the learned deep neural network through forward propagation as follows:

\[
b_q^{(x)} = \text{sign}(f^{(x)}(X_q; W_x, c_x))
\]

(52)

Similarly, given a text query sample \( y_q \), the binary code is computed as:

\[
b_q^{(y)} = \text{sign}(f^{(y)}(y_q; W_y, c_y))
\]

(53)

Therefore, our proposed DDCMH can easily perform cross-modal retrieval for out-of-sample instances.

3.5. Complexity Analysis

In this subsection, we discuss the computational complexity of the proposed DDCMH. The complexity of computing \( S^{(y)}, D^{(y)}, L_x, \) and \( L_y \) is \( O(dn^2) \), where \( d = \max(d_1, d_2) \). According to **Algorithm 2**, the computational complexity of DDCMH is dominated by the iterative optimization. **P-step** occupies \( O(nk^2 + k^3) \) in each iteration. **F-step** costs \( O(nW_y + T_1(nW_x + kn^2 + U_yn^2)) \), where \( U_y = \sum_{m=1}^{M-1} u_y^{(m)} \) is the sum of units in each hidden layer, \( W_y = u_y^{(1)} \times d_1 + u_y^{(2)} \times u_y^{(1)} + \cdots + k \times u_y^{(M-1)} \). \( W_y \) and \( Y_y \) are the number of parameters in image and text networks, respectively. Similarly, the complexity of **\( \Phi \)-step** scales \( O(nW_y + T_1(nW_x + kn^2 + U_yn^2)) \), where \( U_y = \sum_{m=1}^{M-1} u_y^{(m)} \). Due to the SVD of \( B \) in updating \( \Phi \), it leads to a cost of \( O(\text{kn}^2) \). For the discrete optimization of \( B \), the time complexity is \( O(T_2(nk^2 + nk)) \), where \( T_2 \) is a small number (10 in our experiments). Hence, the overall time complexity of training is \( O(nW_x + T_2(kn^2 + nk^2 + k^2 + T_3(nk^2 + nk^2 + nW_y + T_1(nW_x + kn^2 + U_xn^2 + nW_y + kn^2 + U_yn^2))) \). It is worth noting that the training of cross-modal networks is a "one-time" cost. Once the hash functions are learned, the computational cost to generate the binary codes of one image is \( O(W_x) \), and \( O(W_y) \) for text.

4. Experiments and Results Analysis

In this section, experiments are conducted on four benchmark datasets to validate the effectiveness of the proposed DDCMH. In order to evaluate the performance of cross-modal retrieval, two cross-modal retrieval tasks are designed, i.e., text to image (**Task 1**) and image to text (**Task 2**). Text to image utilizes text as a query to search relevant images, and image to text uses image as a query to search relevant text. Here, an image and a text are considered to be relevant if they share at least one common semantic label.

4.1. Experimental Settings

4.1.1. Datasets

**NUS-WIDE** [45] is composed of 269,648 images crawled from Flickr, together with the associated raw tags from 81 semantic concepts. Following [27], we select the top 10 largest categories consisting of 186577 labeled image-text pairs. Moreover, we prune the 186577 image-text pairs by further selecting the sample with nonzero feature vector, which results in a new dataset containing 181365 image-text pairs. Each image is represented by a 500-dimensional Bag-of-Visual-Words (BoVW) SIFT histogram and its corresponding text is represented by 1000-dimensional tag occurrence feature vectors. In the new experimental dataset, we randomly select 5000 image-text pairs for training, and 1000 image-text pairs are randomly sampled as the query set.

**MIR Flickr** [46] consists of 25000 images collected from Flickr associated with tags. These images are belong to at least one of the 24 semantic classes. We select experimental data as the way in [47], which leads to a dataset with 16738 instances. We randomly select 5% (836) for testing, and 5000 instances for training. Here, the images are represented by 150-dimensional edge histogram (EH), and associated tags are described by binary tagging vectors. For convenience, PCA is applied to reduce the dimensionality of text features, resulting in a 500-dimensional feature representation.

**Pascal VOC** [48] dataset contains 9963 testing image-tag pairs, which can be classified into 20 categories. Since several image-tag pairs are multi-labeled, we select the pairs with only one label as the way in [49] resulting in 2808 training and 2841 testing image-tag pairs. The image modality is represented by 512-dimensional GIST features [50], and the representations of text modality are 399-dimensional word frequency features.

**LabelMe** [51] dataset contains 2688 outdoor scenes from eight different classes. Each image is annotated with tags which are from a 781 words dictionary. We discard the words that occur in less than 3 times, resulting in 366 unique words. Thus, the representation of text modality is a 366-dimensional word frequency. In terms of image modality, we employ LabelMe Tollebox to generate a 512-dimensional GIST [50] feature. In addition, we delete the samples without tags, which results in a dataset with 2688 image-text pairs. In our experiments, 75% of the data are selected for training, and the rest are chose as the query set for testing.

4.1.2. Baselines

In order to evaluate the effectiveness of our DDCMH, we employ six state-of-the-art methods, i.e., Cross-View Hashing (CVH) [11], Collective Matrix Factorization Hashing (CMFH) [12], Latent Semantic Sparse Hashing (LSSH) [18], Supervised Matrix Factorization Hashing (SMFH) [5], Deep Canonical Correlation Analysis (DCCA) [34], and Cross-Media Neural Network Hashing (CMNNH) [37] as baselines to compare with DDCMH. Among these methods, CMFH, LSSH, and DCCA are unsupervised, while CVH, SMFH, CMNNH and our DDCMH are supervised. The hash functions of CVH, CMFH, LSSH, and SMFH are linear projections and the optimizations of them are conducted without discrete constraints. In
terms of DCCA and CMNNH, the proposed DDCMH is quite similar to them because of the deep neural networks that are used for nonlinear hashing. In the experiments, all the parameters in these competitors are set based on the original papers.

4.1.3. Evaluation Protocols

In this paper, we use mean Average Precision (mAP) as the metric to evaluate the performance of cross-modal retrieval. The mAPs over the top 50 retrieved instances are reported. In addition, we evaluate the precision and recall through the precision-scope and precision-recall curves, which can reveal the performance of cross-modal retrieval significantly. Moreover, we also present the cross-modal retrieval precision within Hamming radius 2 (PH2) based on hash lookup table to show the effectiveness of the learned hash functions.

4.1.4. Implementation details

In the following experiments, several parameters in our DDCMH should be determined. The structure of cross-modal deep neural networks for four datasets are shown in Table 1. We use hyperbolic tangent (tanh) as the active function. $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ are selected from $1e-5$, $1e-4$, $\ldots$, $1$. $\alpha_y^{(m)}$, $\tau_x^{(m)}$, $M^{y}_{m-1} = \frac{1}{y-1}$ and $\alpha_y^{(m)}$, $\tau_y^{(m)}$, $M^{x}_{m-1} = \frac{1}{y-1}$ are set according to the input data. The learning rate $\eta = 0.1$, $T_1$ is set to 5, and $T_2$ is set to 100. In the construction of intra-modality similarity, the value of $k$-nearest neighbors is set to 10. We empirically tune the parameters and report the best performance. Although we mainly concentrate on the performance of our method with short hash code, we still investigate the performance with different hash codes length $k$, which varies in 8, 16, 32, and 64 bits.

4.2. Experimental Results

The mAP scores of all the methods on four datasets are shown in Table 2. From this table, we can see that the mAP scores of Task 1 are generally higher than that of Task 2. The reason is that the text feature can better represent the semantic information of object than the image feature which only describes the low level visual information. However, our DDCMH achieves different result on the MIR Flickr dataset. The mAP scores of Task 1 and Task 2 on this dataset are virtually same. It demonstrates the superior performance of the proposed DDCMH on learning unified binary codes. The results also verify the aforementioned assumption that paired images and texts which describe the same object should share similar binary codes. We own this to the high level semantic information extracted by the cross-modal deep neural networks. Therefore, our DDCMH has the capability to learn virtually same binary codes for the semantically correlated images and texts, which is crucial for cross-modal retrieval.

In addition, it can easily be observed that the proposed DDCMH significantly outperforms the other state-of-the-art approaches. Specifically, DDCMH yields a substantial increase in performance of both Task 1 and Task 2 on NUS-WIDE and MIR Flickr when the hash code length is 8 bits. In the NUS-WIDE dataset, DDCMH has achieved superior performance to the best of the counterparts with performance gains of 21.3%, and 14.5% on Task 1 and Task 2, respectively. Similarly, significant performance gains of 19.3%, and 27.4% on Task 1 and Task 2, respectively, are achieved on the MIR Flickr dataset. For Pascal VOC dataset, DDCMH outperforms the second best LSSH by 31.2% on Task 1. In addition, we observe that the mAP scores of our DDCMH when hash code length equals to 8 bits are even higher than that of other counterparts when their hash code lengths are 32 or 64 bits. It demonstrates the effectiveness of our method on learning more compact and discriminative binary codes.

Furthermore, the superior performance of DDCMH achieved under the hash code length of 8 bits enables a quite dramatic reduction in storage cost. This can be attributed to the nonlinear hash functions learned by cross-modal deep neural networks which can capture the nonlinear relationship among samples. Additionally, learning unified hash codes directly with discrete optimization can eliminate the quantization loss, which can further improve the accuracy of binary codes. Moreover, compared with SMFH, we can find that the bit-independent constraints and incorporation of the inter-modality and intra-modality similarities preserving term is beneficial to learn more discriminative hash functions for each modality, which is crucial to out-of-sample data.

Compared against the deep learning based cross-modal retrieval methods DCCA and CMNNH, our DDCMH also obtains superior performance on NUS-WIDE, MIR Flickr, and Pascal VOC. It indicates the effectiveness of our method for considering discrete optimization and similarity preservation. Generally, it can also be observed that CMNNH and our DDCMH outperform DCCA due to the relationship preservation with labels rather than pair-wise constraints that DCCA uses.

Table 1
Structure of cross-modal deep neural networks.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>M</th>
<th>Text Network</th>
<th>Image Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUS-WIDE</td>
<td>4</td>
<td>800-256-128-8</td>
<td>450-200-128-8</td>
</tr>
<tr>
<td>MIR Flickr</td>
<td>5</td>
<td>500-256-128-64-8</td>
<td>400-256-128-64-8</td>
</tr>
<tr>
<td>Pascal VOC</td>
<td>3</td>
<td>350-8</td>
<td>500-8</td>
</tr>
<tr>
<td>LabelMe</td>
<td>3</td>
<td>350-320-8</td>
<td>500-480-8</td>
</tr>
</tbody>
</table>

Table 2
MAP results on NUS-WIDE, MIR Flickr, Pascal VOC, and LabelMe dataset with varied hash code lengths.

<table>
<thead>
<tr>
<th>Task</th>
<th>Method</th>
<th>NUS-WIDE</th>
<th>MIR Flickr</th>
<th>Pascal VOC</th>
<th>LabelMe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8 bits</td>
<td>16 bits</td>
<td>32 bits</td>
<td>64 bits</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVH</td>
<td>0.4688</td>
<td>0.4496</td>
<td>0.4335</td>
<td>0.4510</td>
<td>0.6436</td>
</tr>
<tr>
<td>CMFH</td>
<td>0.4744</td>
<td>0.4748</td>
<td>0.4833</td>
<td>0.4918</td>
<td>0.7035</td>
</tr>
<tr>
<td>DCCA</td>
<td>0.5138</td>
<td>0.4918</td>
<td>0.4954</td>
<td>0.4846</td>
<td>0.6527</td>
</tr>
<tr>
<td>CMNNH</td>
<td>0.5285</td>
<td>0.5546</td>
<td>0.5999</td>
<td>0.5816</td>
<td>0.6868</td>
</tr>
<tr>
<td>DDCMH</td>
<td>0.6655</td>
<td>0.7174</td>
<td>0.7464</td>
<td>0.6922</td>
<td>0.8471</td>
</tr>
<tr>
<td>CVH</td>
<td>0.4620</td>
<td>0.4508</td>
<td>0.4423</td>
<td>0.4390</td>
<td>0.6368</td>
</tr>
<tr>
<td>CMFH</td>
<td>0.4908</td>
<td>0.4830</td>
<td>0.4966</td>
<td>0.4901</td>
<td>0.6332</td>
</tr>
<tr>
<td>DCCA</td>
<td>0.5130</td>
<td>0.4991</td>
<td>0.5021</td>
<td>0.4985</td>
<td>0.6176</td>
</tr>
<tr>
<td>CMNNH</td>
<td>0.5285</td>
<td>0.5546</td>
<td>0.5999</td>
<td>0.5816</td>
<td>0.6868</td>
</tr>
<tr>
<td>DDCMH</td>
<td>0.6655</td>
<td>0.7174</td>
<td>0.7464</td>
<td>0.6922</td>
<td>0.8471</td>
</tr>
<tr>
<td>CVH</td>
<td>0.4620</td>
<td>0.4508</td>
<td>0.4423</td>
<td>0.4390</td>
<td>0.6368</td>
</tr>
<tr>
<td>CMFH</td>
<td>0.4908</td>
<td>0.4830</td>
<td>0.4966</td>
<td>0.4901</td>
<td>0.6332</td>
</tr>
<tr>
<td>DCCA</td>
<td>0.5130</td>
<td>0.4991</td>
<td>0.5021</td>
<td>0.4985</td>
<td>0.6176</td>
</tr>
<tr>
<td>CMNNH</td>
<td>0.5285</td>
<td>0.5546</td>
<td>0.5999</td>
<td>0.5816</td>
<td>0.6868</td>
</tr>
<tr>
<td>DDCMH</td>
<td>0.6655</td>
<td>0.7174</td>
<td>0.7464</td>
<td>0.6922</td>
<td>0.8471</td>
</tr>
</tbody>
</table>
In terms of label information, we can see that the supervised methods SMFH, CMNNH, and DDCMH are generally superior to the unsupervised CMFH and LSSH. The reason is that exploring label information enhances the discriminative capability of hash functions. In contrast, CVH is the worst performing supervised method which neglects the extraction of common semantic information.

From Table 2, we further observe that the mAP scores of all the methods are fluctuant with the variation of hash code length. In terms of CMFH, LSSH, SMFH, and CMNNH the mAP scores increase gradually as the hash code length varies from 8 bits to 64 bits. Whereas the results of DCCA for both Task 1 and Task 2 decrease as the hash code length increasing. While CVH presents different trends on four datasets which discloses its sensitivity to datasets. Different from the above methods, DDCMH does not show a obvious trend on any dataset. The reason is that we only perform the parameter tuning in the experiments when the hash code length is 8 bits. In the other experiments when hash code lengths are 16, 32, and 64 bits, we use the same networks and parameters which are tuned under the code length equaling 8 bits. However, we can still observe an increasing trend implicitly, such as on Pascal VOC and NUS-WIDE. Therefore, we conjecture that better performance would be obtained, if we carefully tune the parameters for our method when hash code lengths are 16, 32, and 64 bits, respectively. Fortunately, the performance achieved with 8 bits can be significantly superior to that of the competitors with longer hash code length.

As shown in Table 2, DDCMH does not always obtain the best results with different hash code lengths and different tasks. It is worth noting that, our DDCMH is still superior to all the counterparts but LSSH and DCCA when hash code length is 8 bits. The mAP of DDCMH (0.2845) is comparable to DCCA (0.2899) for Task 2 on Pascal VOC at 8 bits. Although LSSH obtains the highest mAPs in Task 2 on LabelMe dataset, the overall performances of it on all four datasets fluctuate dramatically. Overall, the proposed DDCMH shows the superior performance compared against the baselines.

The precision-recall curves of different methods on the four datasets are plotted in Fig. 2. As shown in this figure, the performance keeps consistent with the results of mAP scores reported in Table 2, which demonstrates that our DDCMH consistently outperforms the competitors. In addition, we can observe that the deep neural networks based cross-modal hashing CMNNH and our DDCMH outperform others on NUS-WIDE and MIR Flickr. However, CMNNH performs inferior on Pascal VOC and LabelMe. Fig. 3 shows
Fig. 4. PH2 with 8 bits hash codes on MIR Flickr dataset.

Fig. 5. The validation of intra-modality similarity preservation on MIR Flickr dataset.

Fig. 6. An example of cross-modal retrieval Task 1 i.e. text to image on the LabelMe dataset. Top 10 images retrieved with text query (first column) are presented in the third column. The second column shows the corresponding image to the text query.

Fig. 7. Performance variation with respect to parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ for both Text to Image and Image to Text tasks on the NUS-WIDE dataset when the length of hash code is 8 bits. (Best viewed in color).

the precision-scope curves of various methods at top 1000 on the four datasets. From this figure, we can observe that, DDCMH achieves superior performance consistently. Specifically, the precisions for top 100 retrieval are significantly higher than other approaches.

We also report the results of retrieval precisions within Hamming radius 2 (PH2) on MIR Flickr dataset when hash code length is 8 bits. As shown in Fig. 4, for both tasks, the proposed method outperforms the others, which demonstrates the effectiveness of our method. It also reveals that more relevant samples are dis-
tributed around the query sample under the Hamming space, which is essential to cross-modal retrieval. In addition, the deep neural networks based methods CMNNH and our DDCMH usually outperform others.

In order to evaluate the effect of intra-modality similarity preservation in our method, we conduct additional experiments without the consideration of intra-modality similarity preservation terms. The mAP scores of our method with and without intra-modality similarity preservation for Task 1 and Task 2 on MIR Flickr are plotted in Fig. 5. DDCMH is our method including intra-modality similarity preservation, while DDCMH-Intra excluding it. Obviously, DDCMH outperforms DDCMH-Intra in both tasks on four datasets. From this figure, we can observe that the intra-modality similarity preservation has been preserved and plays an important role in our method.

An example of Task 1 on the LabelMe dataset is shown in Fig. 6. The top 10 images retrieved by our DDCMH are presented according to the similarity descending from left to right.

Generally speaking, owing to the integration of nonlinear hash functions and discrete optimization, the proposed DDCMH consistently outperforms state-of-the-art approaches. In addition, the proposed DDCMH can achieve satisfactory cross-modal retrieval performances with more compact (8 bits) and discriminative binary codes. This advantage can contribute to a more efficient storage and effective retrieval for large-scale multimedia retrieval.

4.3. Parameter Sensitivity Analysis

In this section, we also evaluate the robustness of our method to parameters by analyzing the parameter sensitivity. We empirically set $\lambda_1 = 1 e^{-3}$, $\lambda_2 = 0.1$, $\lambda_3 = 1 e^{-4}$, and $\lambda_4 = 1 e^{-5}$. Then, we conduct experiments to investigate the performance with respect one parameter by fixing the others on the NUS-WIDE dataset. The results are shown in Fig. 7. We can observe that all the hyperparameters are insensitive, and the values of them can be set in a reasonable wide range.

4.4. Convergence

Since the alternative optimization in an iterative manner is used to optimize the objective function, we further investigate the convergence of our algorithm. We conduct experiments on four datasets to record the objective function value in each iteration, resulting to the convergence curves in Fig. 8. In addition, we also investigate the convergence of training deep neural network. As illustrated in Fig. 8, our algorithm can often converge within a reasonable number of iterations which further demonstrates that our DDCMH is timing effective for cross-modal retrieval.

Additional experiments are conducted on MIR Flickr dataset to reflect the computational complexity of the proposed DDCMH. Fig. 9 depicts the running time of each iteration in Algorithm 2 with varied training size. All the running times are recorded using Matlab on a laptop with an Intel i7-7500 CPU @ 2.70 GHz and 16 GB of RAM.

5. Conclusion

In this paper, we have proposed a discrete deep cross-modal hashing termed DDCMH for performing cross-media retrieval. Different from most of the existing methods, DDCMH investigates the nonlinear relationship among samples by using cross-modal deep neural networks. Furthermore, discrete optimization is employed to learn binary codes directly. In order to facilitate cross-modal retrieval performance, we imposed the intra-modality similarity preserving into each hidden layer of the two individual networks, and formulated inter-modality similarity preserving at the top layer these two networks. Extensive experiments have been carried out on four datasets to validate the effectiveness of DDCMH. The results shown that, the proposed DDCMH presents an advantage i.e. achieving satisfactory cross-modal retrieval performance with more compact (8 bits) and discriminative binary codes, which can contribute to an efficient storage and effective retrieval for large-scale multimedia retrieval. In the future, we intend to extend DDCMH to multimodal scenario, and to reduce the complexity imposed by the consideration of intra-modality and inter-modality similarity preservation.

Acknowledgment

This work was supported in part by the Nature Science Foundation of China [grant numbers 61672123]; the National Key Research and Development Program of China [grant numbers 2016YFD0800300]; and the Science and Technology Planning Key Project of Guangdong Province [grant numbers 2015B010110006].

References

Fangming Zhong received the B.S. and M.S. degree in Software Engineering from Dalian University of Technology in 2012 and 2014, respectively. He is now pursuing his Ph.D. at School of Software Technology in Dalian University of Technology, Dalian, China. His research interests include multimodal learning, cross-modal retrieval, and subspace learning.

Zhikui Chen received his Ph.D. degree in Digital Signal Processing and M.S. degree in Mechanics from Chongqing University, China, in 1998 and 1993, respectively. He obtained his B.S. degree in the Department of Mathematics and Computer Science from Chongqing Normal University, China. Zhikui Chen is working as a full professor at Dalian University of Technology, China. He is leading the Institute of Ubiquitous Network and Computing of Dalian University of Technology. His research interests are big data processing, mobile cloud computing, ubiquitous network and its computing. He is a senior member of IEEE.

Geyong Min is a Professor of High-Performance Computing and Networking with the Department of Mathematics and Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, U.K. He received the Ph.D. degree in Computing Science from the University of Glasgow, Glasgow, U.K., in 2003, and the B.Sc. degree in Computer Science from the Huazhong University of Science and Technology, Wuhan, China, in 1995. His research interests include next-generation Internet, wireless communications, multimedia systems, information security, high-performance computing, ubiquitous computing, modeling, and performance engineering.