High-order nonlocal Hashing for unsupervised cross-modal retrieval

Peng-Fei Zhang[1] · Yadan Luo[1] · Zi Huang[1] ⓘ · Xin-Shun Xu[2] · Jingkuan Song[3]

Abstract

In light of the ability to enable efficient storage and fast query for big data, hashing techniques for cross-modal search have aroused extensive attention. Despite the great success achieved, unsupervised cross-modal hashing still suffers from lacking reliable similarity supervision and struggles with handling the heterogeneity issue between different modalities. To cope with these, in this paper, we devise a new deep hashing model, termed as High-order Nonlocal Hashing (HNH) to facilitate cross-modal retrieval with the following advantages. First, different from existing methods that mainly leverage low-level local-view similarity as the guidance for hashing learning, we propose a high-order affinity measure that considers the multi-modal neighbourhood structures from a nonlocal perspective, thereby comprehensively capturing the similarity relationships between data items. Second, a common representation is introduced to correlate different modalities. By enforcing the modal-specific descriptors and the common representation to be aligned with each other, the proposed HNH significantly bridges the modality gap and maintains the intra-consistency. Third, an effective affinity preserving objective function is delicately designed to generate high-quality binary codes. Extensive experiments evidence the superiority of the proposed HNH in unsupervised cross-modal retrieval tasks over the state-of-the-art baselines.

Keywords Multimodal · Unsupervised Hashing · Cross-Modal search · Representation learning

1 Introduction

In recent years, the rapid development of the information society has brought a tremendous explosion of data in multiple media types, such as texts, images, audios and videos, which has consequently boosted the demands for effective data storage and similarity search

✉ Zi Huang
huang@itee.uq.edu.au

Extended author information available on the last page of the article.

techniques. To relieve that, great efforts have been made [1, 9, 32, 42, 51, 60], where one representative solution is hashing learning. Hashing techniques aim to project data from an arbitrary high dimensional space into a binary space where similarity relationships of the data in the original space are preserved. As data is represented by compact binary codes, the storage and searching costs would be reduced dramatically. In virtue of these major advantages, hashing based methods have attracted much attention and a great variety of excellent work has been presented [6, 14, 31, 34, 35, 39, 54, 58].

Most prior work mainly focuses on data retrieval within a single modality [11, 23, 25, 28, 43, 44, 53, 55], for instance, using texts to search similar texts and images for finding relevant images. However, in practice, it is more common to search data across modal borders, e.g., using several words or a sentence to search relevant videos or images. Handling such cross-modal retrieval tasks is challenging due to the existence of heterogeneity between different modalities. To cope with it, cross-modal hashing (CMH) has been investigated, which aims to learn effective hash codes for multi-modal data with the original intra- and inter-correlations well maintained [37, 40, 52, 56, 59, 61].

Depending on whether supervised information is involved, existing cross-modal hashing can be roughly categorized into supervised and unsupervised methods. Commonly, supervised methods [19, 20, 24, 45] utilize annotated labels or a pre-computed similarity matrix as a uniform guidance for all modalities to ensure the intra- and inter-modality consistencies during the binary learning procedure, thus obtaining precise binary codes and effective mapping functions. Unsupervised methodologies [7, 8, 36, 57] can only analyse raw features to uncover their intrinsic relations to guide hash learning. Therefore, the quality of the to-be-learnt binary codes and functions highly relies on how much useful information could be excavated from the original datasets. Generally, supervised hashing can achieve better performance as more information is available, while unsupervised counterparts are more practical in real applications, where data often comes with no tags and manually annotating large collections of data is very time-consuming and difficult.

Recently, with the strong power in extracting insightful non-linear features, deep learning has shown its superiority and been prevalent in many fields, including information retrieval, computer vision, etc. Combining deep learning and hashing techniques has become more and more ubiquitous and a large family of proposed methods [2–4, 13, 21, 47] have exhibited encouraging performance compared to conventional non-deep hashing methods. Thereinto, existing deep cross-modal hashing methods mainly focus on supervised scenes while unsupervised learning receives relatively less attention yet it is crucial in reality as we mentioned above. In this work, we concentrate on improving the performance of deep cross-modal hashing in unsupervised searching scenarios.

Technically, the key objectives of deep unsupervised cross-modal learning are to obtain reliable guide information, relieve the modal discrepancy and build effective connection between multimodal data. To achieve the goals, many existing methods construct affinity graphs by various approaches so as to capture the underlying similarity relationships, and then learn binary codes by approximating the affinity relationships [17, 38, 49].

However, there are still some problems that need to be further considered. First, current methods widely compute the pair-wise distance such as the Euclidean distance and the Cosine distance between samples to indicate their similarity relationships, which then is processed as the guidance for learning. Nevertheless, the aforementioned similarity measures only consider the pair-wise points independently, neglecting their relationships with other points (i.e., neighbours). Such a local-view affinity construction makes the gained similarity information insufficient for desired results. Second, conventionally unsupervised

methods treat each individual modality separately, constructing modality-specific affinity matrices. Nevertheless, it is rarely taken into consideration the diversity and complement benefits from different modalities. In practice, different modalities contain various information, thereby it is highly likely that the intrinsic information carried by multiple modalities could be complementary to each other. Intuitively, an interactive learning strategy through incorporating multi-modality information is expected to yield competitive results. Third, the similarity-preserving strategy for the learning purpose has not been studied adequately by previous methods, in result, limiting the learning performance.

To address the above issues, in this paper, we propose a novel unsupervised deep cross-modal method, called High-order Nonlocal Hashing (HNH) for cross-modal retrieval. An overview of our proposed method is illustrated in Figure 1 and the contributions are summarised as follows.

– The proposed HNH takes a deep exploration into the underlying semantic structure of multi-modal data, which elegantly distills similarity information in both the local and nonlocal views, and integrates multi-modal correlation information to construct a high-order unified similarity matrix as the learning supervision. In this way, it can significantly mitigate the heterogeneity problem and provide a more precise guidance for cross-modal learning.
– (Re-P2.4) To reduce modality discrepancy and improve performance, HNH keeps intra-consistency and inter-affinity both implicitly and explicitly. This is implemented by introducing a common representation for all modalities and reconstructing similarity relationships with the common representation and modal-specific embeddings.
– Extensive experiments demonstrate that the proposed algorithm outperforms baseline methods for cross-modal hashing retrieval.

The rest of the paper is organized as follows. In Section 2, we give a brief review of the related work. In Section 3, we elaborate the proposed work including the notation, framework and optimization scheme, followed by comprehensive experiments in Section 4. The conclusion is given in Section 5.
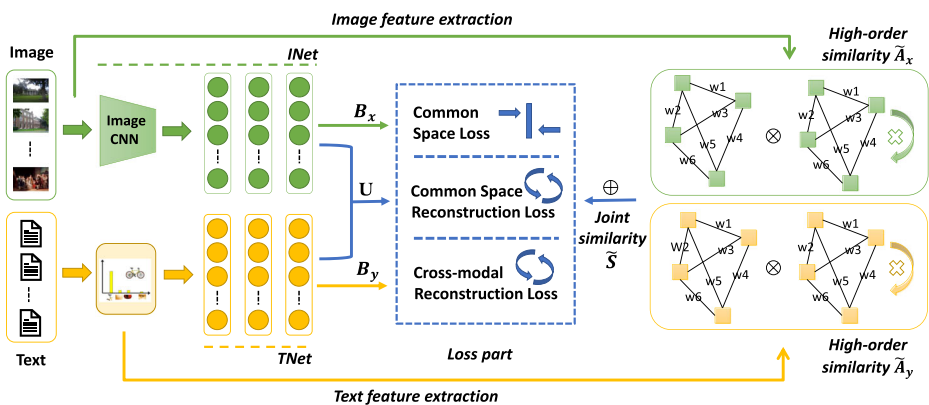

Figure 1 Illustration of the proposed HNH framework, which consists of three main parts: similarity matrix construction, feature learning and hash code learning

2 Related work

In the literature, existing cross-modal hashing methods can be divided into shallow (i.e., non-deep) hashing and deep hashing according to if deep neural networks are used in the learning procedure. In this section, we briefly review the related work of these two categories.

2.1 Shallow cross-modal Hashing

Early work in hashing learning usually learns binary codes and embedding functions by designing effective algorithms to deal with hand-craft features. Based on whether supervised information is utilized, existing shallow cross-modal hashing methods can be roughly categorized into unsupervised and supervised parts. Usually, the former generates hash codes or projection functions by exploiting the intra- and inter-modality similarity structures of the training data. For instance, Inter-Media Hashing (IMH) [36] learns consistent binary representations for multimedia data by exploring the intra-media and inter-media connections and consistencies. To eliminate the semantic gap and capture more semantic information, Latent Semantic Sparse Hashing (LSSH) [57] adopts the sparse coding and matrix factorization to deal with images and texts, respectively. Collective Matrix Factorization Hashing (CMFH) [7] learns unified hash codes of instances by the collective matrix factorization with a latent factor model for different views. Composite Correlation Quantization (CCQ) [26] introduces an isomorphic latent space, in which different modalities can be more effectively correlated.

In comparison, supervised cross-modal hashing methods can additionally utilize the precise semantic information, e.g., semantic labels/tags. Various methods have been explored and also shown excellent performance, such as Cross View Hashing (CVH) [19], Semantics Preserving Hashing (SePH) [22], Discrete Cross-modal Hashing (DCH) [48] and Discrete Manifold-Embedded Cross-Modal Hashing (SDMCH) [27]. More specifically, CVH is the first work that considers hashing learning in the multiple view settings, which learns binary codes by minimizing the similarity-weighted Hamming distance. SePH learns hash representations by minimizing the KL-divergence of a probability distribution derived from the given semantic affinities. DCH introduces a classification framework with labels as supervision to directly learn discriminative hash codes without discarding the discrete constraints. SDMCH generates binary codes by preserving manifold correlations and leveraging various kinds of supervised information.

2.2 Deep cross-modal Hashing

By means of the prominent nonlinear modeling ability of deep networks, deep hashing outperforms non-deep methods in a landslide as it can learn high-level features from the original data. This learning pattern has become the mainstream in the hashing learning field and currently, more efforts are put in the supervised scenes. Representative supervised deep hashing methods include Deep Visual-Semantic Hashing (DVSH) [2], Deep Cross-Modal Hashing (DCMH) [16], Pairwise Relation Guided Deep Hashing (PRDH) [50] and Subspace Relation Learning for Cross-modal Hashing (SRLCH) [33]. DVSH takes into consideration the specific properties of different modalities and performs the joint visual-text representation learning by utilizing convolutional neural networks (CNN) and Long Short Term Memory (LSTM) to deal with different modalities. DCMH constructs an end-to-end deep discrete hash learning framework to enable the deep feature learning and hash learning

in one step. PRDH employs various kinds of pairwise constraints to capture the correlations across different modalities and imposes the decorrelation restriction to pursue more precise and discriminative binary codes. In order to learn more discriminative hash codes, SRLCH directly exploits the relation information of semantic labels by transforming class labels into a subspace.

On the other hand, solving unsupervised cross-modal problem is relatively tougher because there are only original features available. For instance, Deep Binary Reconstruction (DBRC) [12] proposes a binary reconstruction framework with the Adaptive Tanh (ATanh) activation, so that discrete hash codes can be directly learnt. Unsupervised Deep Cross-Modal Hashing (UDCMH) [46] performs the feature learning and binarization simultaneously under a novel Laplacian and discrete constraint based framework. Deep Joint-Semantics Reconstructing Hashing (DJSRH) [38] designs a joint-semantics affinity matrix constructing strategy to capture the underlying semantic relationships among data, and the binary codes can be obtained by reconstructing the built joint affinity.

3 Method

3.1 Notation and problem definition

In this paper, we use boldface uppercase letters, e.g., $\mathbf{U}$, to represent matrices and boldface lowercase letters, e.g., $\mathbf{u}$, to denote vectors. $\mathbf{U}_{i*}$ and $\mathbf{U}_{*j}$ represent the i-th row and the j-th column of $\mathbf{U}$, respectively. $\mathbf{U}_{ij}$ is the element at the position (i, j) of matrix $\mathbf{U}$. The transpose of $\mathbf{U}$ is indicated as $\mathbf{U}^T$ and $\mathbf{U}^{-1}$ denotes the inverse of the matrix $\mathbf{U}$. In addition, $\mathbf{I}_c$ indicates an identity matrix with dimensionality c, $\|\cdot\|_F$ denotes the Frobenius of a vector or matrix. The Hadamard matrix product $\otimes$ (i.e., element-wise product) between any two matrices, e.g., $\mathbf{U} = \{\mathbf{U}_{ij}\}_{i,j=1}^n$ and $\mathbf{V} = \{\mathbf{V}_{ij}\}_{i,j=1}^n$ is defined as:

$$\mathbf{U} \otimes \mathbf{V} = \left\{ \begin{array}{cccc} \mathbf{U}_{11} \cdot \mathbf{V}_{11} & \mathbf{U}_{12} \cdot \mathbf{V}_{12} & ... & \mathbf{U}_{1n} \cdot \mathbf{V}_{1n} \\ \mathbf{U}_{21} \cdot \mathbf{V}_{21} & \mathbf{U}_{22} \cdot \mathbf{Y}_{22} & ... & \mathbf{U}_{2n} \cdot \mathbf{V}_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{U}_{n1} \cdot \mathbf{V}_{n1} & \mathbf{U}_{n2} \cdot \mathbf{V}_{n2} & ... & \mathbf{U}_{nn} \cdot \mathbf{V}_{nn} \end{array} \right\} \tag{1}$$

$sign(\cdot)$ is an element-wise sign function which is defined as follows:

$$sign(\mathbf{u}) = \left\{ \begin{array}{ll} 1 & \mathbf{u} > 0 \\ -1 & \mathbf{u} \leq 0 \end{array} \right. . \tag{2}$$

Assume there is a multi-modal dataset with n image-text pair instances, indicated as $\mathcal{O} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{d_1 \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n] \in \mathbb{R}^{d_2 \times n}$ denote the d_1-dimensional image feature matrix and the d_2-dimensional text feature matrix (usually $d_1 \neq d_2$) with n training points, respectively. In experiments, we make an assumption that samples from both modalities in the training set are observed.

Given the training data and specific code length c, the purpose of our method is to learn modality-specific projecting functions $f(\mathbf{x}; \theta_{\mathbf{x}})$ and $g(\mathbf{y}; \theta_{\mathbf{y}})$ for images and texts, respectively, where $\theta_{\mathbf{x}}$ and $\theta_{\mathbf{y}}$ are network parameters, and gain the corresponding binary representations $\mathbf{B_x} \in \{-1, 1\}^{c \times n}$ and $\mathbf{B_y} \in \{-1, 1\}^{c \times n}$. In principle, $\mathbf{B_x}$ and $\mathbf{B_y}$ learnt by mapping functions should preserve well the similarity in the original multi-modal spaces.

3.2 Framework overview

As shown in Figure 1, our model is an end-to-end joint learning framework which mainly contains two subnetworks: INet and TNet, to perform image and text feature encoding, respectively. As the original image and text have specific characteristics, the INet and TNet are built on different components. More concretely, the INet is adapted from the pre-trained (on ImageNet) Alexnet [18] which is composed of five convolution layers and three fully-connected (fc) layers. We remove the last layer and add a fully-connected (fc) layer which contains c hidden units on top of the remaining layers. The TNet is a three-layer MultiLayer Perceptrons (MLP) [30].

Given the batch-input image-text pairs, at the very beginning of each iteration, we extract 4,096-dimensional vectors as the original high-level image features from the fc-7 layer of the pre-trained Alexnet, and use the original text features, e.g., BoW features, as the original text representations. Both of the features are then used to construct a unified similarity matrix for two modalities with the new designed strategy which will be progressively elaborated below. We take the original images and texts features from datasets as inputs fed into our network, which subsequently outputs the corresponding representative descriptors. To ensure the learnt representations can effectively preserve the original similarity relationships, we constantly optimize the whole network by minimizing the objective loss function defined in (11).

3.3 High-order nonlocal affinity matrix

Many previous work [10, 38, 49] has shown the feasibility of using features extracted from pre-trained deep architectures to construct affinity matrices as the learning guidance. In light of this, in each epoch of the training stage, we randomly select m ($m \ll n$) instance pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$ from two modalities as a batch input. We further extract their features $\{(\mathbf{x}_i', \mathbf{y}_i')\}_{i=1}^m$, where $\mathbf{x}_i' \in \mathbb{R}^{d_x \times 1}$ and $\mathbf{y}_i' \in \mathbb{R}^{d_y \times 1}$ from pre-trained deep neural networks and the original text dataset, respectively. Later, we calculate the pair-wise Cosine distance between these features to construct real-valued modality-specific affinity matrices $\mathbf{A_x}$ and $\mathbf{A_y} \in [0, 1]^{m \times m}$ for images and texts, respectively. They are defined as follows:

$$\mathbf{A_x}[i, j] = d_{cos}\left(\mathbf{x}_i', \mathbf{x}_j'\right) = \frac{\mathbf{x}_i'^T \mathbf{x}_j'}{\|\mathbf{x}_i'\|_2 \|\mathbf{x}_j'\|_2}, \tag{3}$$

$$\mathbf{A_y}[i, j] = d_{cos}\left(\mathbf{y}_i', \mathbf{y}_j'\right) = \frac{\mathbf{y}_i'^T \mathbf{y}_j'}{\|\mathbf{y}_i'\|_2 \|\mathbf{y}_j'\|_2}, \tag{4}$$

where $\mathbf{A_x}[i, j]$ and $\mathbf{A_y}[i, j]$ represent the (i, j)-th element of $\mathbf{A_x}$ and $\mathbf{A_y}$.

After constructing the above affinity matrices, previous methods directly utilize them to supervise the hash code and function learning, where two potential problems exist. On the one hand, the above similarity construction approach exploits the relation of every two points in a local view, which implies that only the points themselves are involved when computing their similarity. As their relationships with other points are neglected, sufficient semantic similarity information can be hardly captured. For example, suppose we have three points a, b and c, and the distance between a and b is the same as that of a and c. According to the above similarity measure, two pairs have the same similarity. However, it may be not accurate because their neighbourhood relationships have not be considered. If a and c have more common similar neighbours than a and b, the relationship between a and b

would be closer compared to that of the other pair. This kind of high-order nonlocal relationships can not be observed when only the local-view pair-wise distance between samples independently is calculated.

On the other hand, the above similarity matrices are constructed purely based on one modality, i.e., a similarity matrix for one modality is built only based on the information within this modality. In the case that the amount of information within one modality is limited, extracting sufficient semantic information relying on one modality for learning is not always feasible. Besides, learning binary representations and functions by approximating these independent similarity matrices separately suffers from the discrepancy between different modalities, leading to inferior performance. In reality, the diverse and complementary information carried by different modalities can boost each other. Therefore, in cross-modal learning, merging information from various modalities can be an effective way to capture rich semantics and consequently gain a high-quality similarity matrix.

Considering this, we first re-weight the original similarities by exploiting pair-wise relationships with the holistic view, which is defined as follows:

$$\begin{aligned}
\widetilde{\mathbf{A}}_* &= \mathbf{A}_* \otimes \Psi_*, \\
\Psi_* &= \mathbf{A}_*^T \mathbf{A}_*, \\
s.t. \quad &* \in \{\mathbf{x}, \mathbf{y}\},
\end{aligned} \tag{5}$$

where the symbol represents the Hadamard matrix product (i.e., element-wise product). Since each row or column of $\mathbf{A}_*, * \in \{\mathbf{x}, \mathbf{y}\}$ indicates the relation between one point and all the other points, $\mathbf{A}_*^T \mathbf{A}_*$ can reflect how two samples are similar to each other based on their neighbours. It is common that if two points share more neighbours with larger similarity values, they should be closer than other points, and vice versa. For instance, as we can see in Figure 2, points a and b have 3 common neighbours {a,b,d}, and a and c also have 4 common neighbours {a,c,e,f}, according to the above method, the nonlocal similarity for (a,b) is $1 \times 0.2 + 0.2 \times 1 + 0.1 \times 0.1 = 0.41$, and for (a,c) is $1 \times 0.2 + 0.2 \times 1 + 0.1 \times 0.1 + 0.3 \times 0.5 = 0.56 > 0.41$, so a and c are more similar than a and b, which can not be observed via the traditional local-view similarity matrix construction. In a word ,our similarity computation can help capture more reliable and precise relationships between two points by considering both local and nonlocal information. In addition, we further regulate $\widetilde{\mathbf{A}}_* = k_* \cdot \widetilde{\mathbf{A}}_* - 1$ to give a flexible quantization area for the following quantization procedure.

After getting two relatively high-order nonlocal similarity matrices $\widetilde{\mathbf{A}}_{\mathbf{x}}$ and $\widetilde{\mathbf{A}}_{\mathbf{y}}$, we set out to solve the second problem. As we mentioned above, the information in one modality
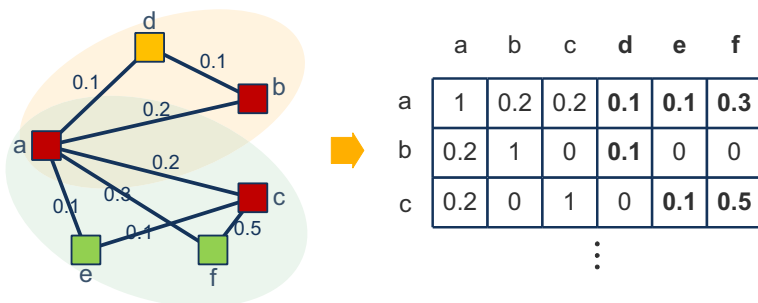

Figure 2 An example of neighbour relationships. Point a and b have the same Cosine similarity as point a and c, but a and c have more common neighbours with larger similarity values than a and b, so a is more similar to c

can complement and reinforce others, so we integrate the gained $\widetilde{\mathbf{A}}_{\mathbf{x}}$ and $\widetilde{\mathbf{A}}_{\mathbf{y}}$ together to get a new joint similarity matrix:

$$\widetilde{\mathbf{S}} = \gamma \widetilde{\mathbf{A}}_{\mathbf{x}} + (1 - \gamma) \widetilde{\mathbf{A}}_{\mathbf{y}},$$
$$s.t. \ \gamma \in [0, 1], \tag{6}$$

where γ is the balance parameter.

Unlike previous methods, our similarity matrix constructing strategy does not limit itself in a local view that only considers two points independently, but also captures the nonlocal similarity structure. Moreover, the bimodal similarity is merged together in an elegant way, such that sufficient information for data from different modalities can be obtained, making the following hash learning algorithm more efficient.

3.4 Objective function

After constructing the similarity matrix for both modalities, we turn to perform binary code and mapping function learning. We treat the similarity matrix $\widetilde{\mathbf{S}}$ as the supervisory signal to train the network. Specifically, we denote features extracted from the last hidden layer of INet as $f(\mathbf{X}; \theta_{\mathbf{x}})$, and text representations generated from TNet are $g(\mathbf{Y}; \theta_{\mathbf{y}})$. We then binarize them to obtain the binary representations $\mathbf{B}_{\mathbf{x}}$ and $\mathbf{B}_{\mathbf{y}}$ via $sign(\cdot)$ functions. They are further normalised, however for simplicity this process is not represented in the following objective functions and the optimization procedure. To eliminate the modality gap, we first introduce a common representation $\mathbf{U} \in \mathbb{R}^{c \times m}$ for both modalities and minimise the distance between it and the binarized representations $\mathbf{B}_{\mathbf{x}}$ and $\mathbf{B}_{\mathbf{y}}$, respectively. Based on this, the common space loss is formulated as:

$$\mathcal{J}_1 = \min_{\mathbf{U}, \theta_{\mathbf{x}}, \theta_{\mathbf{y}}} \|\mathbf{U} - \mathbf{B}_{\mathbf{x}}\|_F^2 + \|\mathbf{U} - \mathbf{B}_{\mathbf{y}}\|_F^2,$$
$$s.t. \ \mathbf{B}_{\mathbf{x}} = sign\left(f\left(\mathbf{X}; \theta_{\mathbf{x}}\right)\right) \in \{-1, 1\}^{c \times m},$$
$$\mathbf{B}_{\mathbf{y}} = sign\left(g\left(\mathbf{Y}; \theta_{\mathbf{y}}\right)\right) \in \{-1, 1\}^{c \times m}. \tag{7}$$

However, in the above formulation, adopting the $sign(\cdot)$ function as the activation function will inevitably lead to the intractable back-propagate gradient problem. To handle it, the $tanh(\cdot)$ function is applied to approximate $sign(\cdot)$ as:

$$\mathcal{J}_1 = \min_{\mathbf{U}, \theta_{\mathbf{x}}, \theta_{\mathbf{y}}} \|\mathbf{U} - \mathbf{B}_{\mathbf{x}}\|_F^2 + \|\mathbf{U} - \mathbf{B}_{\mathbf{y}}\|_F^2,$$
$$s.t. \ \mathbf{B}_{\mathbf{x}} = tanh\left(f\left(\mathbf{X}; \theta_{\mathbf{x}}\right)\right) \in [-1, 1]^{c \times m},$$
$$\mathbf{B}_{\mathbf{y}} = tanh\left(g\left(\mathbf{Y}; \theta_{\mathbf{y}}\right)\right) \in [-1, 1]^{c \times m}. \tag{8}$$

The objective function suggests that the binary representations for image and text modality would be consistent as much as possible. In light of this, the gap between two modalities can be bridged. Moreover, the common representation $\mathbf{U}$ can also act as the representation of each modality as $\mathbf{B}_{\mathbf{x}}$ and $\mathbf{B}_{\mathbf{y}}$ do, so that it can be further utilized to keep the intra-consistency as the following formulation defines:

$$\mathcal{J}_2 = \min_{\mathbf{U}, \theta_{\mathbf{x}}, \theta_{\mathbf{y}}} \underbrace{\|\widetilde{\mathbf{S}} - \mathbf{U}^T \mathbf{B}_{\mathbf{x}}\|_F^2}_{\text{image modality}} + \underbrace{\|\widetilde{\mathbf{S}} - \mathbf{U}^T \mathbf{B}_{\mathbf{y}}\|_F^2}_{\text{text modality}},$$
$$s.t. \ \mathbf{B}_{\mathbf{x}} = tanh\left(f\left(\mathbf{X}; \theta_{\mathbf{x}}\right)\right) \in [-1, 1]^{c \times m},$$
$$\mathbf{B}_{\mathbf{y}} = tanh\left(g\left(\mathbf{Y}; \theta_{\mathbf{y}}\right)\right) \in [-1, 1]^{c \times m}. \tag{9}$$

The two terms in (9) represent the common space reconstruction loss for image and text modalities, respectively. At the same time, as the common representation $\mathbf{U}$ is consistent to each modality representation, the cross-modal similarity can also be preserved implicitly, which further builds the connection between different modalities.

Next, we explicitly preserve inter-similarity by defining the following function:

$$\mathcal{J}_3 = \min_{\theta_\mathbf{x}, \theta_\mathbf{y}} \|\widetilde{\mathbf{S}} - \mathbf{B}_\mathbf{x}^T \mathbf{B}_\mathbf{y}\|_F^2,$$
$$s.t. \ \mathbf{B}_\mathbf{x} = tanh\left(f\left(\mathbf{X}; \theta_\mathbf{x}\right)\right) \in [-1, 1]^{c \times m},$$
$$\mathbf{B}_\mathbf{y} = tanh\left(g\left(\mathbf{Y}; \theta_\mathbf{y}\right)\right) \in [-1, 1]^{c \times m}. \tag{10}$$

Combining (8), (9) and (10) together, the final objective function is formed as:

$$\mathcal{J} = \min_{\mathbf{U}, \theta_\mathbf{x}, \theta_\mathbf{y}} \alpha \mathcal{J}_1 + \beta \mathcal{J}_2 + \lambda \mathcal{J}_3$$
$$= \alpha(\|\mathbf{U} - \mathbf{B}_\mathbf{x}\|_F^2 + \|\mathbf{U} - \mathbf{B}_\mathbf{y}\|_F^2)$$
$$+ \beta(\underbrace{\|\widetilde{\mathbf{S}} - \mathbf{U}^T \mathbf{B}_\mathbf{x}\|_F^2}_{\text{image modality}} + \underbrace{\|\widetilde{\mathbf{S}} - \mathbf{U}^T \mathbf{B}_\mathbf{y}\|_F^2}_{\text{text modality}})$$
$$+ \lambda \underbrace{\|\widetilde{\mathbf{S}} - \mathbf{B}_\mathbf{x}^T \mathbf{B}_\mathbf{y}\|_F^2}_{\text{corss-modality}},$$
$$s.t. \ \mathbf{B}_\mathbf{x} = tanh\left(f\left(\mathbf{X}; \theta_\mathbf{x}\right)\right) \in [-1, 1]^{c \times m},$$
$$\mathbf{B}_\mathbf{y} = tanh\left(g\left(\mathbf{Y}; \theta_\mathbf{y}\right)\right) \in [-1, 1]^{c \times m}, \tag{11}$$

where α, β, λ are the trade-off parameters to balance each loss item.

There are several advantages of the proposed scheme. First, we construct a high-order nonlocal similarity matrix by deeply and reliably exploiting the underlying semantic affinity among data, including the local-view and nonlocal-view similarity relationships. On this basis, we further leverage useful information from different modality as much as possible. In light of this, the matrix based similarity-preserving hash learning can produce high-quality hash codes and functions. Second, we introduce a common representation for different modalities so as to effectively relieve the modal discrepancy. By approximating the high-order similarity matrix with the binary representations of different modalities and the common representation, the intra- and inter-modality consistency is preserved both explicitly and implicitly, which further helps improve performance.

3.5 Optimization algorithm

An alternating optimization scheme is designed to solve the problem of (11) iteratively, which is summarized in **Algorithm 1**. In the following paragraphs, we elaborate the solving algorithm.

Step 1: Updating U with variables $\theta_\mathbf{x}$ and $\theta_\mathbf{y}$ fixed.

We first rewrite (11) as:

$$\mathcal{J} = \min_{\mathbf{U}} \alpha(\|\mathbf{U} - \mathbf{B}_\mathbf{x}\|_F^2 + \|\mathbf{U} - \mathbf{B}_\mathbf{y}\|_F^2)$$
$$+ \beta(\|\widetilde{\mathbf{S}} - \mathbf{U}^T \mathbf{B}_\mathbf{x}\|_F^2 + \|\widetilde{\mathbf{S}} - \mathbf{U}^T \mathbf{B}_\mathbf{y}\|_F^2). \tag{12}$$

For brevity, we expand each item and remove the irrelevant items. Thereafter, (12) is reformulated as follows:

$$\min_{\mathbf{U}} -2Tr(\mathbf{U}(\alpha\mathbf{B_x}^T + \alpha\mathbf{B_y}^T + \beta\widetilde{\mathbf{S}}\mathbf{B_x}^T + \beta\widetilde{\mathbf{S}}\mathbf{B_y}^T))$$
$$+2\alpha\|\mathbf{U}^T\mathbf{U}\|_F^2 + \beta(\|\mathbf{U}^T\mathbf{B_x}\|_F^2 + \|\mathbf{U}^T\mathbf{B_y}\|_F^2). \tag{13}$$

By setting the derivation of (13) w.r.t. $\mathbf{U}$ to zero, we obtain:

$$\mathbf{U} = (2\mathbf{I}_c + \frac{\beta}{\alpha}\mathbf{B_x}\mathbf{B_x}^T + \frac{\beta}{\alpha}\mathbf{B_y}\mathbf{B_y}^\mathsf{T})^{-1}[(\mathbf{B_x} + \mathbf{B_y})(\mathbf{I}_c + \frac{\beta}{\alpha}\widetilde{\mathbf{S}})]. \tag{14}$$

Step 2: Updating $\theta_\mathbf{x}$ with other variables fixed.

When other variables are fixed, (11) can be reformulated as follows:

$$\mathcal{J} = \min_{\theta_\mathbf{x}} \alpha\|\mathbf{U} - \mathbf{B_x}\|_F^2 + \beta\|\widetilde{\mathbf{S}} - \mathbf{U}^T\mathbf{B_x}\|_F^2 + \lambda\|\widetilde{\mathbf{S}} - \mathbf{B_x}^T\mathbf{B_y}\|_F^2,$$
$$s.t. \ \mathbf{B_x} = tanh\,(f\,(\mathbf{X};\theta_\mathbf{x})) \in [-1, 1]^{c \times m}. \tag{15}$$

We can update the parameter $\theta_\mathbf{x}$ by using mini-batch stochastic back-propagation.

$$\theta_\mathbf{x} \quad \longleftarrow \quad \theta_\mathbf{x} - \mu_\mathbf{x}\frac{\partial\mathcal{J}}{\partial\theta_\mathbf{x}} \tag{16}$$

where $\mu_\mathbf{x}$ is the learning rate.

Step 3: Updating $\theta_\mathbf{y}$ with other variables fixed.

When other variables are fixed, (11) can be reformulated as follows:

$$\mathcal{J} = \min_{\theta_\mathbf{y}} \alpha\|\mathbf{U} - \mathbf{B_y}\|_F^2 + \beta\|\widetilde{\mathbf{S}} - \mathbf{U}^T\mathbf{B_y}\|_F^2 + \lambda\|\widetilde{\mathbf{S}} - \mathbf{B_x}^T\mathbf{B_y}\|_F^2,$$
$$s.t. \ \mathbf{B_y} = tanh\left(f\left(\mathbf{Y};\theta_\mathbf{y}\right)\right) \in [-1, 1]^{c \times m}. \tag{17}$$

Similarly, we can solve the above equation with Stochastic Gradient Descent (SGD) by mini-batch back-propagation.

$$\theta_\mathbf{y} \quad \longleftarrow \quad \theta_\mathbf{y} - \mu_\mathbf{y}\frac{\partial\mathcal{J}}{\partial\theta_\mathbf{y}} \tag{18}$$

where $\mu_\mathbf{y}$ is the learning rate.

Algorithm 1 High-order nonlocal Hashing for cross-modal retrieval.

Input: The training data $\mathcal{O} = (\mathbf{X}, \mathbf{Y})$, mini-batch size m, hash code length c, iteration time t, balance parameters $\alpha, \beta, \gamma, \lambda$.
Output: Parameters of network $\theta_\mathbf{x}$ and $\theta_\mathbf{y}$.
Procedure:
Randomly initialize the network parameters $\theta_\mathbf{x}$ and $\theta_\mathbf{y}$;
Repeat:
1. Randomly select m image-text pairs from the dataset to construct a mini-batch;
2. Extract image features and text features to construct the similarity matrix $\widetilde{\mathbf{S}}$;
3. Calculate the loss of the objective function in (11);
4. Update $\mathbf{U}$ by using (14);
5. Update the parameter $\theta_\mathbf{x}$ by backpropagation;
6. Update the parameter $\theta_\mathbf{y}$ via backpropagation;
Until convergent.
Return: $\theta_\mathbf{x}$ and $\theta_\mathbf{y}$.

3.6 Extensions

The out-of-sample extensions can be easily reached. Specifically, for an unseen image instance $\mathbf{x}_o \in \mathbb{R}^{d_1 \times 1}$ or an unseen text $\mathbf{y}_o \in \mathbb{R}^{d_2 \times 1}$, we can get its hash code as:

$$\begin{aligned} \mathbf{b}_{\mathbf{x}_o} &= sign\left(f\left(\mathbf{x}_o; \theta_{\mathbf{x}}\right)\right), \\ \mathbf{b}_{\mathbf{y}_o} &= sign\left(g\left(\mathbf{y}_o; \theta_{\mathbf{y}}\right)\right). \end{aligned} \tag{19}$$

Besides, The proposed method can also be easily extended to the cases with more modalities by adding a new subnetwork for each new modality and slightly modifying the objective function (11) as

$$\begin{aligned} \mathcal{J} &= \min_{\mathbf{U}, \theta_{\mathbf{x}}, \theta_{\mathbf{y}}} \alpha \mathcal{J}_1 + \beta \mathcal{J}_2 + \lambda \mathcal{J}_3 \\ &= \alpha \sum_{i \in W} \|\mathbf{U} - \mathbf{B}_i\|_F^2 + \beta \sum_{i \in W} \|\widetilde{\mathbf{S}} - \mathbf{U}^T \mathbf{B}_i\|_F^2 \\ &+ \lambda \sum_{i \in W} \sum_{j \in W} \|\widetilde{\mathbf{S}} - \mathbf{B}_i^T \mathbf{B}_j\|_F^2, \\ s.t. \ \ &\mathbf{B}_i = tanh\left(f_i\left(\mathbf{X}_i; \theta_{\mathbf{x}}\right)\right) \in [-1, 1]^{c \times m}, \\ &W = \{X, Y, ...\}, \end{aligned} \tag{20}$$

where f_i represents the projecting functions for the modality $i \in W = \{X, Y, ...\}$. The higher order similarity matrix can be obtained by slightly adjusting our high-order similarity constructions (6) as:

$$\begin{aligned} \widetilde{\mathbf{S}} &= \sum_{i \in |W|} \gamma_i \widetilde{\mathbf{A}}_i, \\ s.t. \ \ &\sum_{i \in |W|} \gamma_i = 1, W = \{X, Y, ...\}, \end{aligned} \tag{21}$$

where $|W|$ is the number of modalities.

The optimization problem of (20) can be easily solved by **Algorithm 1**.

3.7 Computational complexity analysis

In this section, we analyse the computational complexity of the proposed HNH. In each iteration of the training stage, we first construct the unified matrix $\widetilde{\mathbf{S}}$, which costs $O(nm^2)$. Then, it takes $O(n(mc + c^2))$ to update the intermediate representation $\mathbf{U}$. Totally, The computational cost of our iterative process is $O(n(m^3 + mc + c^2)t)$ $(m, c, t \ll n)$, which is linear to the size of datasets. The competitive computational efficiency shows the practicality and effectiveness of our proposed algorithm when dealing with large-scale real-world cross-modal searching tasks.

4 Experiments

To evaluate the effectiveness of our proposed HNH, we conduct extensive experiments on two widely-used benchmark datasets for cross-modal retrieval, i.e., MIRFlickr-25K [15], NUS-WIDE [5] and Wiki [29].

4.1 Datasets

MIRFlickr-25K [15] is a multi-label dataset which consists of 25,000 images with corresponding textual tags from 24 unique labels. All of instance pairs are crawled from Flickr website. The dataset also provides a 1,386-dimensional feature vector as the descriptor for each image and a 100-dimensional BoW SIFT feature derived from PCA on the binary tagging vector to represent the corresponding textual content.

NUS-WIDE [5] is a real-world multi-modal image dataset which is composed of 269,648 images and it corresponding textual tags, collecting from Flickr. In our experiments, we select a subset with 186,577 images-tags instance pairs that belong to 10 most commonly used concepts from the original data as the final experimental dataset. The dataset provides a 1,000-dimensional BoW feature for each text.

Wiki [29] is a single label dataset with 2,866 coupled image-text instances. All these pairs are labeled with one of ten semantic classes. The dataset has been divided into 2,173 training pairs and 693 testing pairs. We summarize the statistics of two dataset in Table 1.

4.2 Baselines and evaluation metric

We compare HNH with nine state-of-the-art unsupervised baselines, including shallow models : CVH [19], IMH [36], LCMH [61], CMFH [7], LSSH [57], RFDH [41], and deep learning based methods : DBRC [12], UDCMH [46], DJSRH [38].

We choose the widely-used Mean Average Precision (MAP) which can well reflect both ranking information and precision, to evaluate the performance of all methods.

More specifically, for a set of queries $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_p]$, the Mean Average Precision (MAP) is defined as follows:

$$MAP = \frac{1}{p} \sum_{i=1}^{p} \underbrace{\left(\frac{1}{N_g} \sum_{r=1}^{n} P_i(r)\zeta_i(r) \right)}_{\text{Average Precision (AP)}}, \tag{22}$$

where p is the size of the query set $\mathbf{Q}$, N_g is the number of ground-truth neighbors of the query $\mathbf{q}_i$ in the database, n is the number of entities in the database, $P_i(r)$ denotes the precision of the top r retrieved entities, and $\zeta_i(r) = 1$ if the r-th retrieved entity is a ground-truth neighbour and $\zeta_i(r) = 0$, otherwise. The ground-truth neighbors are defined as those sharing at least one semantic label.

Table 1 Statistics of two Benchmark datasets

Dataset	MIRFlickr-25K	NUS-WIDE	Wiki
Database	25,000	186,577	2,866
Training	5,000	5,000	2,173
Testing	2,000	2,000	693
Labels	24	10	10

We also adopt the top-N precision curves as a performance evaluation criterion to further evaluate the performance of our method and all compared methods.

4.3 Implementation details

We implement the proposed HNH via Pytorch on a workstation (CPU: Intel XEON E5-2650 v3 @ 2.60GHz, RAM: 128GB, GPU: NVIDIA 1080Ti). With regard to the parameters of HNH, we set the batch size as 32, momentum as 0.9, weight decay as 0.0005, learning rates are 0.0001 and 0.01 for ImgNet and TxtNet, respectively. Other balance parameters are selected by a validation procedure. We set $\gamma = 0.9$, $\alpha = 40$, $\beta = 1$, $\lambda = 1$, $k_* = 2 (* \in \{x, y\})$, For MIRFlickr-25K, $\gamma = 0.6$, $\alpha = 40$, $\beta = 1$, $\lambda = 1$, $k_* = 2$ for NUS-WIDE, and $\gamma = 0.8$, $\alpha = 40$, $\beta = 0.3$, $\lambda = 0.01$, $k_x = 2$, $k_y = 0.2$ for Wiki. The iteration time is fixed as 40, 80 and 200 for MIRFlickr-25K, NUS-WIDE and Wiki, respectively. In the evaluation of mAP, we set the number of retrieved points as 50. All experiments are repeated several times with random data partitions, and the averaged results are reported.

For fair comparison, following previous work [38, 46], on MIRFlickr-25K and NUS-WIDE, 5,000 samples are randomly chosen for training, and selected 2,000 samples as the testing samples. All data of Wiki are used in experiments. For all compared non-deep methods, the nonlinear image features from the pre-trained Alexnet and the BoW features are used as the original features.

4.4 Results and discussions

4.4.1 MAP results

We compare the proposed HNH with all baselines in the "Image-to-Text" and "Text-to-Image" search tasks with code length varying from 32 bits to 128 bits and summarize the MAP values in Table 2. From the results, we can have the following observations.

– On MIRFlickr-25K and NUS-WIDE, HNH outperforms all of the compared methods with various binary code length, demonstrating its superiority. In particular, our method achieves 3% - 5.6% and 2.3% - 9.3% improvements in the "Image-to-Text" task on MIRFlickr-25K and NUS-WIDE, respectively, while in the "Text-to-Image" task, HNH achieves the average improvement of 4.3% and 3.6% over the best results of the compared baselines on two datasets, respectively. It is also noticed that our performance on Wiki are not so promising as those on others, where a potential reason there are fewer training samples and larger noises in the dataset. Nevertheless, our method still achieve competitive performance compared to the baseline methods.
– most of methods achieve better performance with relatively long bit length, which means that utilizing longer hash codes can incorporate more useful information.
– In many cases, deep models outperform shallow models with deep features which well reflects that an end-to-end structure can facilitate the training of a model through back-propagate.

Particularly, the DJSRH method also devotes to construct a high-order similarity matrix by treating the original affinity matrix as the high-level feature matrix, so we further compare it with our proposed method in terms of MAP with different number of top returned points. The results are plotted in Figure 3. We can find that our method has

Table 2 The mAP results of all methods on three datasets

Task	Method	MIRFlickr-25K			NUS-WIDE			Wiki		
		32 bits	64 bits	128 bits	32 bits	64 bits	128 bits	32 bits	64 bits	128 bits
$I \rightarrow T$	CVH	0.599	0.596	0.598	0.362	0.406	0.390	0.162	0.153	0.149
	IMH	0.601	0.592	0.579	0.473	0.476	0.459	0.203	0.204	0.195
	LCMH	0.569	0.585	0.593	0.361	0.389	0.383	0.124	0.134	0.149
	CMFH	0.624	0.625	0.627	0.459	0.465	0.467	0.253	0.259	0.263
	LSSH	0.599	0.602	0.614	0.489	0.507	0.507	0.208	0.199	0.195
	DBRC	0.619	0.620	0.621	0.459	0.447	0.447	0.265	0.269	0.288
	RDFH	0.636	0.641	0.652	0.492	0.494	0.508	0.246	0.244	0.243
	UDCMH	0.698	0.714	0.717	0.519	0.524	0.558	0.318	0.329	0.346
	DJSRH	0.843	0.862	0.876	0.773	0.798	0.817	0.403	0.412	0.434
	HNH	**0.883**	**0.895**	**0.902**	**0.802**	**0.816**	**0.847**	**0.600**	**0.582**	**0.512**
$T \rightarrow I$	CVH	0.583	0.576	0.576	0.384	0.442	0.432	0.235	0.171	0.154
	IMH	0.595	0.589	0.580	0.483	0.472	0.462	0.478	0.453	0.456
	LCMH	0.569	0.582	0.582	0.387	0.408	0.419	0.142	0.154	0.157
	CMFH	0.662	0.676	0.685	0.577	0.614	0.645	0.601	0.616	0.622
	LSSH	0.659	0.659	0.672	0.617	0.642	0.663	0.593	0.593	0.595
	DBRC	0.626	0.626	0.628	0.459	0.468	0.473	0.588	0.598	0.599
	RDFH	0.693	0.698	0.702	0.641	0.658	0.680	0.596	0.603	0.610
	UDCMH	0.704	0.718	0.733	0.653	0.695	0.716	0.633	0.645	**0.658**
	DJSRH	0.822	0.835	0.847	0.744	0.771	0.789	**0.635**	**0.646**	**0.658**
	HNH	**0.854**	**0.868**	**0.878**	**0.776**	**0.796**	**0.802**	0.573	0.600	0.648

($I \rightarrow T$ means the search task of Image-to-Text, and vice versa). The best mAPs for each category are shown in boldface

a better performance than DJSRH in all cases, which verifies that our similarity constructing scheme can better capture the unified intrinsic manifold structure for different modalities.
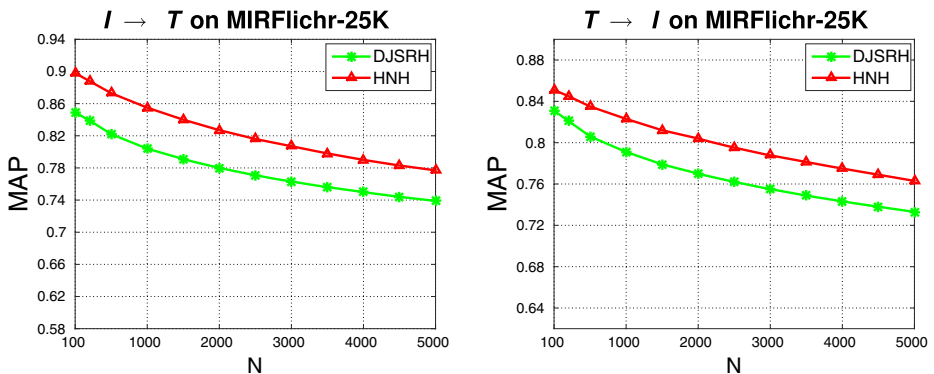

Figure 3 The MAP of DJSRH and HNH with different number of top returned points with 128-bit on MIRFlickr-25K

4.4.2 Top-N precision curves

The top-N precision curves of the cases with 128 bits on three datasets are plotted in Figure 4. From these results, we have the following observations. First, HNH maintains the best overall performance among those traditional learning methods and deep neural network based methods as it is in mAP results. In addition, our method can even achieve better performance with more retrieved points (i.e., 5000) than other methods with less retrieved points in NUS-WIDE dataset, which is very important in retrieval tasks. All these further verify the superiority of our learning strategy on unsupervised cross-modal retrieval tasks.

4.4.3 Ablation study

The proposed HNH tries to improve the retrieval performance mainly by constructing the high-order affinity matrix and introducing the common representation. To validate the effectiveness of two strategies, we investigate two variants of HNH, i.e., 'HNH-1', 'HNH-2'. Detailedly, we use 'HNH-1' to represent the variant which does not consider nonlocal-view relationships in the similarity matrix construction, only adopting low-level affinity matrix $\mathbf{S} = \gamma \mathbf{A_x} + (1 - \gamma)\mathbf{A_y}$ as the supervision. We remove the the common representation $\mathbf{U}$ from the proposed HNH to constitute 'HNH-2' which final objective function is $\mathcal{J} = \alpha\|\widetilde{\mathbf{S}} - \mathbf{B_x}^T \mathbf{B_x}\|_F^2 + \beta\|\widetilde{\mathbf{S}} - \mathbf{B_y}^T \mathbf{B_y}\|_F^2 + \lambda\|\widetilde{\mathbf{S}} - B_x^T \mathbf{B_y}\|_F^2$. We conduct experiments on MIRFlickr-25K dataset and measure the corresponding performance with mAP. The results are reported in Table 3. It can be seen that without one of the components, the performance of the proposed method decreases, which means both components make contributions to improve retrieval accuracy. In addition, these variants all perform better than all of the compared methods in previous experiments, demonstrating the superiority of our method.
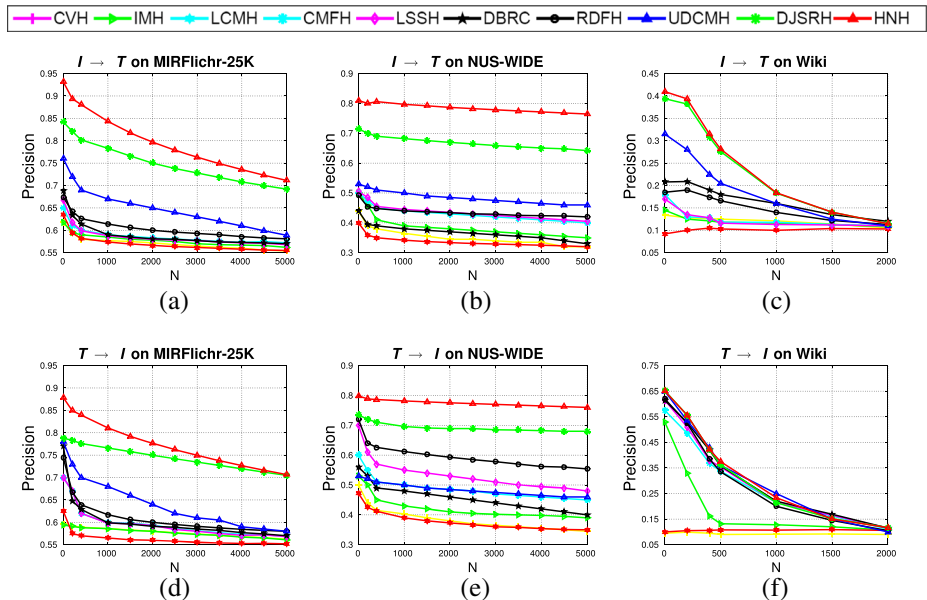

Figure 4 The top-N precision curves of various models with 128-bit on three datasets

Table 3 The ablation comparison among different variants of HNH on MIRFlickr-25K

Task	Method	32 bits	64 bits	128 bits
$I \rightarrow T$	HNH	**0.883**	**0.895**	**0.902**
	HNH-1	0.873	0.876	0.892
	HNH-2	0.869	0.88	0.889
$T \rightarrow I$	HNH	**0.854**	**0.868**	**0.878**
	HNH-1	0.831	0.855	0.867
	HNH-2	0.836	0.845	0.868

The best mAPs for each category are shown in boldface

4.4.4 Parameter sensitivity analysis

There are several parameters that may influence the performance of the proposed HNH. To analyse this, we conduct experiments on MIRFlickr-25K. The MAP curves of the case with 128 bits by varying α, β, λ and γ are plotted in Figure 5. Specifically, the mAP value increases with α increasing from 1 to 20, then the performance remains stable. The mAP
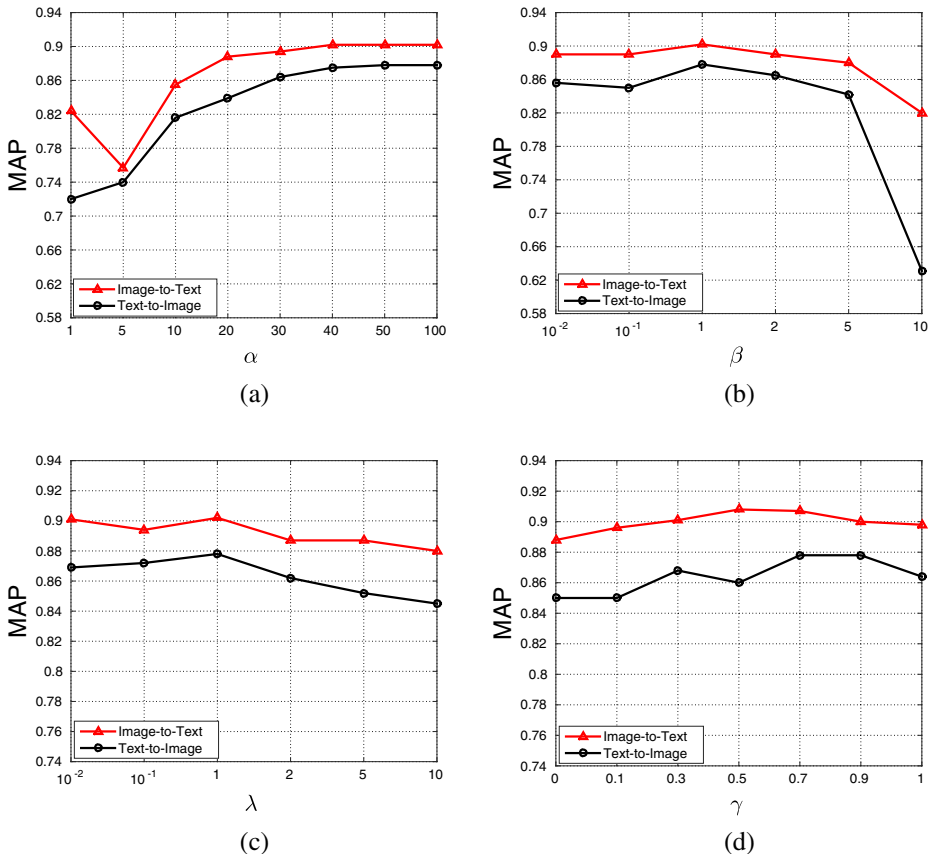

Figure 5 Sensitivity analysis of parameter α, β, λ and γ with 128-bit on MIRFlickr-25K
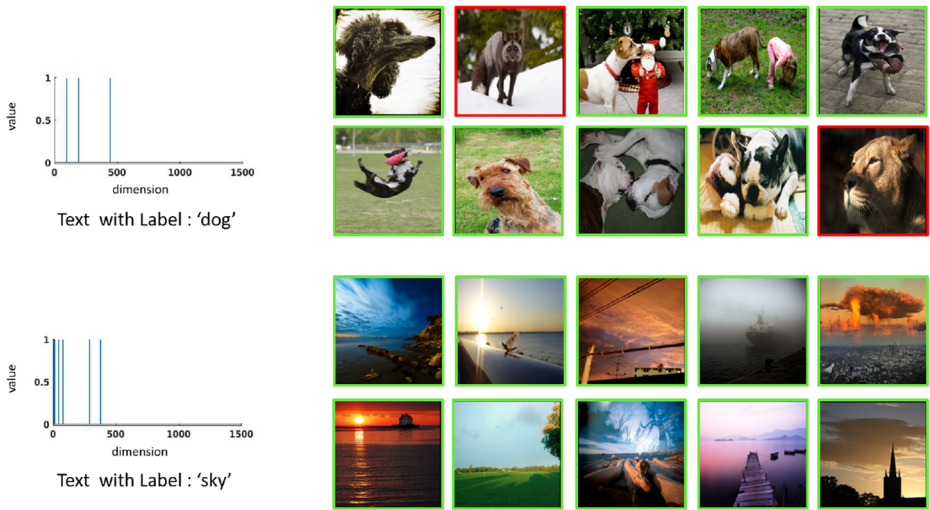
Figure 6 Example retrieval results

value decreases rapidly when β is larger than 5. As for λ, our method achieves promising results when λ ranges from [0.01,1], then it declines. γ controls the weight of the two similarity matrices of two modalities, we can see that our method can achieve the best overall performance when γ is larger than 0.6. In addition, when $\gamma = 1$ and $\gamma = 0$ which means we only use $\widetilde{\mathbf{A}}_{\mathbf{x}}$ and $\widetilde{\mathbf{A}}_{\mathbf{y}}$, respectively, the performance is inferior compared to our integrated method but it is still better than other compared methods, which demonstrates the effectiveness of the proposed HNH and shows that integrating information form different modalities during similarity construction is valid.

4.4.5 Visualization

In Figure 6, we present examples of top 10 retrieval results on MIRFlickr-25K. The correct searched data is marked in green while the incorrect results are in red. From the results, we can see that given queries, the proposed HNH can return relevant data that shares the same label with high probability. Despite a few irrelevant data returned (e.g., 'lion' for the first query), these returned irrelevant data appears visually similar to the correct retrieved data.

5 Conclusion

In this paper, we propose a novel deep unsupervised method, i.e., High-order Nonlocal Hashing (HNH) for cross-modal retrieval. We focus on capturing high-order supervisory signals to facilitate searching across various modalities. For this purpose, apart from capturing local similarity information, the relation between samples in an overall point of view is also taken into consideration. On this basis, we further leverage the complementary advantages of multi-modal data to build a unified similarity matrix which potentially carries comprehensive semantics. In addition, a common space representation is introduced to establish connections between different modalities, which helps to eliminate the modal

discrepancy. Finally, by preserving similarity among image, text, and common space representations, we can learn high-quality binary codes and effective hash functions. Extensive experiments with excellent results demonstrate the superiority of the proposed HNH in cross-modal retrieval tasks.

References

1. Andoni, A., Razenshteyn, I.: Optimal data-dependent hashing for approximate near neighbors. In: Proceedings of Annual Symposium on Foundations of Computer Science, pp. 793–801 (2015)
2. Cao, Y., Long, M., Wang, J., Yang, Q., Yu, P.S.: Deep visual-semantic hashing for cross-modal retrieval. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1445–1454 (2016)
3. Chaidaroon, S., Ebesu, T., Fang, Y.: Deep semantic text hashing with weak supervision. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 1109–1112 (2018)
4. Chaidaroon, S., Fang, Y.: Variational deep semantic hashing for text documents. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 75–84 (2017)
5. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: A real-world Web image database from national university of singapore. In: Proceedings of ACM International Conference on Image and Video Retrieval, p. 48 (2009)
6. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of Annual Symposium on Computational Geometry, pp. 253–262 (2004)
7. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2075–2082 (2014)
8. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of ACM International Conference on Multimedia, pp. 7–16 (2014)
9. Gionis, A., Indyk, P., Motwani, R., et al.: Similarity search in high dimensions via hashing. In: Proceedings of International Conference on Very Large Data Bases, pp. 518–529 (1999)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
11. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2916–2929 (2013)
12. Hu, D., Nie, F., Li, X.: Deep binary reconstruction for cross-modal hashing. IEEE Trans. Multimed. **21**(4), 973–985 (2018)
13. Hu, P., Zhen, L., Peng, D., Liu, P.: Scalable deep multimodal learning for cross-modal retrieval. In: Proceedings of ACM SIGIR International conference on Research and Development in Information Retrieval, pp. 635–644 (2019)
14. Huang, F., Zhang, L., Yang, Y., Zhou, X.: Probability weighted compact feature for domain adaptive retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 9582–9591 (2020)
15. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: Proceedings of ACM International Conference on Multimedia Information Retrieval, pp. 39–43 (2008)
16. Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3270–3278 (2017)
17. Kang, W.C., Li, W.J., Zhou, Z.H.: Column sampling based discrete supervised hashing. In: Proceedings of AAAI Conference on Artificial Intelligence, pp. 1230–1236 (2016)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
19. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 1360–1365 (2011)

20. Li, C.X., Chen, Z.D., Zhang, P.F., Luo, X., Nie, L., Zhang, W., Xu, X.S.: Scratch: a scalable discrete matrix factorization hashing for cross-modal retrieval. In: Proceedings of ACM International Conference on Multimedia, pp. 1–9 (2018)

21. Li, W.J., Wang, S., Kang, W.C.: Feature learning based deep supervised hashing with pairwise labels. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 1711–1717 (2016)

22. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-preserving hashing for cross-view retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3864–3872 (2015)

23. Liu, J., Zhang, L.: Optimal projection guided transfer hashing for image retrieval. In: Proceedings of AAAI Conference on Artificial Intelligence, vol. 33, pp. 8754–8761 (2019)

24. Liu, W., Wang, J., Ji, R., Jiang, Y.G., Chang, S.F.: Supervised hashing with kernels. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2074–2081 (2012)

25. Liu, W., Wang, J., Kumar, S., Chang, S.F.: Hashing with graphs. In: Proceedings of International Conference on Machine Learning, pp. 1–8 (2011)

26. Long, M., Cao, Y., Wang, J., Yu, P.S.: Composite correlation quantization for efficient multimodal retrieval. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 579–588 (2016)

27. Luo, X., Yin, X.Y., Nie, L., Song, X., Wang, Y., Xu, X.S.: Sdmch: Supervised discrete manifold-embedded cross-modal hashing. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 2518–2524 (2018)

28. Luo, Y., Yang, Y., Shen, F., Huang, Z., Zhou, P., Shen, H.T.: Robust discrete code modeling for supervised hashing. Pattern Recogn. **75**, 128–135 (2018)

29. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of ACM International Conference on Multimedia, pp. 251–260 (2010)

30. Rumelhart, D.E., Hinton, G.E., McClelland, J.L., et al.: A general framework for parallel distributed processing. Parallel Distributed Processing: Explorations in the Microstructure of Cognition **1**(26), 45–76 (1986)

31. Shen, F., Xu, Y., Liu, L., Yang, Y., Huang, Z., Shen, H.T.: Unsupervised deep hashing with similarity-adaptive and discrete optimization. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 3034–3044 (2018)

32. Shen, H.T., Jiang, S., Tan, K.L., Huang, Z., Zhou, X.: Speed up interactive image retrieval. VLDB J. **18**(1), 329–343 (2009)

33. Shen, H.T., Liu, L., Yang, Y., Xu, X., Huang, Z., Shen, F., Hong, R.: Exploiting subspace relation in semantic labels for cross-modal hashing. IEEE Trans. Knowl. Data Eng. (2020)

34. Song, J., Yang, Y., Huang, Z., Shen, H.T., Luo, J.: Effective multiple feature hashing for large-scale near-duplicate video retrieval. IEEE Trans. Multimed. **15**(8), 1997–2008 (2013)

35. Song, J., Yang, Y., Li, X., Huang, Z., Yang, Y.: Robust hashing with local models for approximate similarity search. IEEE Trans. Cybern. **44**(7), 1225–1236 (2014)

36. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 785–796 (2013)

37. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Proceedings of Advances in Neural Information Processing Systems, pp. 2222–2230 (2012)

38. Su, S., Zhong, Z., Zhang, C.: Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of IEEE International Conference on Computer Vision, pp. 3027–3035 (2019)

39. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of ACM International Conference on Multimedia, pp. 154–162 (2017)

40. Wang, D., Cui, P., Ou, M., Zhu, W.: Deep multimodal hashing with orthogonal regularization. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 2291–2297 (2015)

41. Wang, D., Wang, Q., Gao, X.: Robust and flexible discrete hashing for cross-modal similarity search. IEEE Trans. Circ. Syst. Video Technol. **28**(10), 2703–2715 (2017)

42. Wang, Z., Zhang, Z., Luo, Y., Huang, Z.: Deep collaborative discrete hashing with semantic-invariant structure. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 905–908 (2019)

43. Wang, Z., Zhang, Z., Luo, Y., Huang, Z., Shen, H.T.: Deep collaborative discrete hashing with semantic-invariant structure construction. IEEE Trans. Multimed. (2020)

44. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1753–1760 (2009)

45. Wu, B., Yang, Q., Zheng, W.S., Wang, Y., Wang, J.: Quantized correlation hashing for fast cross-modal search. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 3946–3952 (2015)

46. Wu, G., Lin, Z., Han, J., Liu, L., Ding, G., Zhang, B., Shen, J.: Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 2854–2860 (2018)

47. Xu, R., Li, C., Yan, J., Deng, C., Liu, X.: Graph convolutional network hashing for cross-modal retrieval. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 10–16 (2019)

48. Xu, X., Shen, F., Yang, Y., Shen, H.T., Li, X.: Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Trans. Image Process. **26**(5), 2494–2507 (2017)

49. Yang, E., Deng, C., Liu, T., Liu, W., Tao, D.: Semantic structure-based unsupervised deep hashing. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 1064–1070 (2018)

50. Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., Gao, X.: Pairwise relationship guided deep hashing for cross-modal retrieval. In: Proceedings of AAAI Conference on Artificial Intelligence, pp. 1618–1625 (2017)

51. Yang, Y., Luo, Y., Chen, W., Shen, F., Shao, J., Shen, H.T.: Zero-shot hashing via transferring supervised knowledge. In: Proceedings of ACM International Conference on Multimedia, pp. 1286–1295 (2016)

52. Zhang, D., Li, W.J.: Large-scale supervised multimodal hashing with semantic correlation maximization. In: Proceedings of AAAI Conference on Artificial Intelligence, pp. 2177–2183 (2014)

53. Zhang, D., Wang, J., Cai, D., Lu, J.: Self-taught hashing for fast similarity search. In: Proceedings of ACM SIGIR International conference on Research and Development in Information Retrieval, pp. 18–25 (2010)

54. Zhang, P., Zhang, W., Li, W.J., Guo, M.: Supervised hashing with latent factor models. In: Proceedings of ACM SIGIR International conference on Research and Development in Information Retrieval, pp. 173–182 (2014)

55. Zhang, Z., Xie, G.S., Li, Y., Li, S., Huang, Z.: Sadih: Semantic-aware discrete hashing. In: Proceedings of AAAI Conference on Artificial Intelligence, pp. 5853–5860 (2019)

56. Zhen, Y., Yeung, D.Y.: Co-regularized hashing for multimodal data. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1376–1384 (2012)

57. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 415–424 (2014)

58. Zhou, X., Shen, F., Liu, L., Liu, W., Nie, L., Yang, Y., Shen, H.T.: Graph convolutional network hashing. IEEE Trans. Cybern. 1–13 (2018)

59. Zhu, L., Huang, Z., Liu, X., He, X., Sun, J., Zhou, X.: Discrete multimodal hashing with canonical views for robust mobile landmark search. IEEE Trans. Multimed. **19**(9), 2066–2079 (2017)

60. Zhu, X., Huang, Z., Cheng, H., Cui, J., Shen, H.T.: Sparse hashing for fast multimedia search. IEEE Trans. Image Process. **31**(2), 1–24 (2013)

61. Zhu, X., Huang, Z., Shen, H.T., Zhao, X.: Linear cross-modal hashing for efficient multimedia search. In: Proceedings of ACM International Conference on Multimedia, pp. 143–152 (2013)

Affiliations

Peng-Fei Zhang[1] · Yadan Luo[1] · Zi Huang[1] (iD) · Xin-Shun Xu[2] · Jingkuan Song[3]

Peng-Fei Zhang
mima.zpf@gmail.com

Yadan Luo
lyadanluol@gmail.com

Xin-Shun Xu
xuxinshun@sdu.edu.cn

Jingkuan Song
jingkuan.song@gmail.com

[1]   School of Information Technology, Electrical Engineering, University of Queensland,
       Brisbane, Australia

[2]   School of Software, Shandong University, Jinan, China

[3]   School of Computer Science, Engineering, University of Electronic Science and
       Technology of China, Chengdu, China