Multi-Relational Deep Hashing for Cross-Modal Search

Xiao Liang, Erkun Yang, Yanhua Yang, and Cheng Deng, Senior Member, IEEE

Abstract—Deep cross-modal hashing retrieval has recently made significant progress. However, existing methods generally learn hash functions with pairwise or triplet supervisions, which involves learning the relevant information by splicing partial similarity between data pairs; notably, this approach only captures the data similarity locally and incompletely, resulting in sub-optimal retrieval performance. In this paper, we propose a novel Multi-Relational Deep Hashing (MRDH) approach, which can fully bridge the modality gap by comprehensively modeling the similarity relationship between data in different modalities. To more detail, to investigate the inter-modal relationships, we constrain the consistency of cross-modal pairwise similarities to maintain the semantic similarity across modalities. Moreover, to further capture complete similarity information, we design a new similarity metric, which we term cross-modal global similarity, by encouraging hash codes of similar data pairs from different modalities to approach a common center and hash codes for dissimilar pairs to converge to different centers. Adopting this approach enables our model to generate more discriminative hash codes. Extensive experiments on three benchmark datasets demonstrate the superiority of our method on cross-modal hashing retrieval.

Index Terms—Cross-modal retrieval, hash code learning, similarity learning, metric learning.

I. INTRODUCTION

WITH the development of social networks and the Internet, visual data (e.g., photos and videos) have experienced explosive growth, giving rise to renewed enthusiasm in the field of nearest-neighbor search. Cross-modal retrieval, one of the most popular approaches of this kind, aims to search semantically similar data points in one modality (e.g., image) by using a query from another modality (e.g., text) [1], [2]. Moreover, due to the low storage requirements and fast query speed, hashing methods, which maps high dimensional multi-modal data points into common Hamming space, endowing similar cross-modal data points with similar hash codes [3], [4]. However, the modality gap caused by the heterogeneous nature of the data obtained from different modalities still remains a challenge for cross-modal hashing.

Many deep cross-modal hashing retrieval methods have previously been proposed [5], [6], [7], [8], [9], [10]. These approaches extract features via convolution neural networks (CNNs), then learn hash codes simultaneously in an end-to-end training fashion, thereby generating accurate hash codes. However, most of these methods employ the similarity matrix directly as semantic constraints to generate hash codes, and learn the hash functions from pairwise or triplet data relationships.

Such approaches intrinsically give rise to the following issues: 1) Incomplete similarity relationship. As the query and retrieval are obtained from different modalities, the generated hash codes are expected to preserve the semantic similarities among the original features between modalities. However, the similarity matrix, as a very simple semantic constraint, cannot reflect the similarity distance between samples, and has limited power to model multi-modal similarity relationships, meaning that it can only learn incomplete similarity knowledge. 2) Insufficient coverage of data distribution. Pairwise/triplet similarity-based methods utilize only partial relationships between data pairs and then splice this local knowledge together, which may attenuate the discriminative ability of the generated hash codes. Therefore, the question of how to incorporate the global data similarity...
information and the semantic similarities between different modalities into the hash code learning procedure is of unprecedented importance.

To address the problem 1, in this paper, we propose a novel Multiple Relation Learning (MRL) framework, which not only utilizes the partial relationships between data pairs, but also employs relation as a manner to investigate the inter-modal relationships. Concretely, we aim to maintain the consistency of pairwise similarities among different samples for two different modalities. For example, as shown in Figure 1, we consider two samples $a$ and $b$, where $a_1$, $b_1$ are corresponding features of modality 1, while $a_2$, $b_2$ are corresponding features of modality 2. The traditional pairwise/triplet methods simply pull $a_1$, $a_2$ or $b_1$, $b_2$ closer and push $a_1$, $b_2$ or $b_1$, $a_2$ further away, without considering the relation information between them. We believe that the semantic difference between $a$ and $b$ should be maintained across modalities; that is, the distance $d(a_1, b_2)$ should be consistent with the distance $d(a_2, b_1)$. By applying this approach, the distance can be modeled as a similarity distribution between data pairs, and can accordingly be used as a metric to align the same samples in different modalities, meaning that the relationships between different samples can be maintained across different modalities. Compared with a single pairwise/triplet loss, our method can not only learn the relationship inter/intra-modalities, but also use the "relationship of relations" to constrain the alignment between modalities. This multiple relations method has stronger binding force and can learn more robust features.

However, the approach outlined above can only solve the problem of incomplete similarity relationship. In order to efficiently learn global similarity information and further narrow the gap between modalities (problem 2), inspired by TSDH [11], we propose a cross-modal global similarity (CGS) learning method. In more detail, the cross-modal global similarity measures the Hamming distances between hash codes and the adaptive cross-modal global similarity center, which is defined as a set of adaptive points in the Hamming space with a sufficient mutual distance. The goal of CGS learning is to encourage the generated hash codes to approach the corresponding hash center. Notably, TSDH proposes CSQ as a similarity measure by initializing some discrete centers and then bringing different classes of features closer to different centers. However, CSQ’s hash centers are fixed and equally spaced after initialization, which goes against intuition (the category centers of “eat” and “dog” should be closer than those of “tree”). Inspired by this, our CGS learns adaptive learnable similarity centers that can be adaptively updated as the network is trained. With a time complexity of only $O(nm)$ for $n$ data points and $m$ centers, cross-modal global similarity-based hashing can generate discriminative hash codes from the global data distribution, which enables it to overcome the limitations of hashing methods based on the pairwise/triplet similarity. The contributions of this work can be summarized as follows:

- We design a novel Multiple Relation Learning (MRL) framework that narrows the modality gap between image and text modalities by maintaining the relational consistency between data from different modalities.
- We propose a novel adaptive global similarity metric, named Cross-modal Global Similarity (CGS), which effectively and globally models the similarity information of different samples in different modalities.
- Extensive experimental results on three benchmark datasets demonstrate that our proposed MRD achieves state-of-the-art performance in cross-modal image-text hashing retrieval methods.


text

II. RELATED WORK

Existing cross-modal hashing methods can be broadly divided into two distinct categories [12], [13], namely unsupervised and supervised approaches. The unsupervised methods use only co-occurrence information to learn hash functions for multi-modal data. For example, CVH [14] extends spectral hashing from uni-modal to multi-modal scenarios, while LSSH [15] jointly learns latent features from images and texts with sparse coding. For their part, the supervised methods exploit label information in order to learn more discriminative common representation. SCM [16] utilizes non-negative matrix factorization and a neighbor-preserving algorithm to maintain inter-modal and intra-modal semantic correlations. SMFH [17] considers both the label consistency across different modalities and the local geometric consistency in each modality based on collective matrix factorization. However, almost all these existing shallow hashing methods are based on hand-crafted features; as a result, the discriminative representation of instances may be limited, which will degrade the accuracy of the learned binary hash codes.

Models based on deep networks [5], [6], [7], [8], [9], [18], [19], [20], [21], [22] are highly regarded and can provide better access to more discriminative features compared to those with hand-crafted features, which boosts the deep cross-modal retrieval performance. The classic method, Cross-Modal Hamming Hashing (CMHH) [23], generates favorable hash codes for accurate retrieval by jointly optimizing a novel exponential focal loss and an exponential quantization loss in a Bayesian learning framework. Deep Cross-Modal Hashing (DCMH) [7] and Adaptive Label correlation based asymmetric Cross-modal Hashing (ALECH) [5] both learn hash codes by preserving the label similarity correlation through the similarity matrix, and focus on designing a more efficient pairwise-similarity loss function. ASCSH [6] formulates pairwise similarity and semantic labels to guide the hash code learning process. In addition, [24] utilizes a triplet ranking loss for similarity learning. Overall, these methods achieve satisfactory results; however, DCMH forces hash codes to maintain semantic relevance for similar data points through the direct use of the similarity matrix, and without paying adequate attention to the latent relationships among cross-modal data, while ALECH learns the relevant information by splicing partial similarity between data pairs, meaning that the data similarity is captured only locally and incompletely. In addition, these methods utilize pairwise or...
triplet data similarity, where the data relationships are captured from a local perspective.

Our proposed method belongs to the category of supervised cross-modal hashing (CMH), which leverages semantic supervision information to generate hashing codes for different modalities. Moreover, our proposed method focuses on the relational consistency between the data under different modalities, and employs the similarity metric based on global information to excavate data semantic information; thus, both modality and semantic similarities are well preserved, leading to optimal hash codes and better retrieval performance.

III. METHODOLOGY

In this work, we focus on improving the accuracy of retrieval between image and text modalities, which are the two modalities most commonly used in daily life. As shown in Fig. 2, the proposed MRDH model comprises three main types of components: two feature encoding networks for both modalities, a Multiple Relation Learning (MRL) framework, and a Cross-modal Global Similarity (CGS) quantization, all of which will be introduced in more detail below.

A. Problem Definition

Assume that we have \( n \) training entities (data points), each of which has two modalities of features. Without loss of generality, we use text–image datasets for illustration in this paper, which means that each training point has both a text modality and an image modality. We use \( X = \{ x_i \}_{i=1}^{n} \) to denote the image modality, where \( x_i \) can denote the handcrafted features or the raw pixels of image \( i \). Moreover, we use \( Y = \{ y_i \}_{i=1}^{n} \) to denote the text modality, where \( y_i \) is typically the text information related to the image \( i \).

Given the above training information, the goal of cross-modal hashing is to learn two hash functions for the two modalities, namely: \( h^{(\text{img})}(x) \in \{-1, 1\}^K \) for the image modality and \( h^{(\text{text})}(y) \in \{-1, 1\}^K \) for the text modality, where \( K \) is the length of binary code.

B. Multiple Relation Learning

Traditional cross-modal hashing methods directly employ similarity matrix as semantic constraints to generate hash codes. That is, if two samples belong to the same class, the similarity matrix is set to 1, while if the two samples do not belong to the same class, this value is set to 0. Notably, this naive method only utilizes partial relationships between data pairs and cannot fully explore the semantic information between data from different modalities. Considering that in the cross-modal setting, the relationships of similarity between different data points include both the semantic information between the data and the heterogeneous information between the modalities. So we use this similarity relationship as a manner to investigate the inter-modal relationships.

Let \( f(x_i; \theta_x) \in \mathbb{R}^c \) denote the learned image feature for point \( i \), which corresponds to the output of the CNN for the image modality. Furthermore, let \( g(y_j; \theta_y) \in \mathbb{R}^c \) denote the learned text feature for point \( j \), which corresponds to the output of the deep neural network for the text modality. Here, \( \theta_x \) is the network parameter of the CNN for the image modality, while \( \theta_y \) is the network parameter of the deep neural network for the text modality. We then calculate the similarity relationship between the data from different modalities. This can be measured by \( \Theta_{ij} = \frac{1}{2} F_{ij}^\top G_{si} \), where \( F \in \mathbb{R}^{c \times n} \) with \( F_{si} = f(x_i; \theta_x) \). \( G \in \mathbb{R}^{c \times n} \) with \( G_{sj} = g(y_j; \theta_y) \).

A softmax layer can be adopted to process the calculated similarities, which then produces the following relationship distribution:

\[
p_i^1 = \frac{\exp(\Theta_{ij}/\tau_1)}{\sum_{k=1}^{K} \exp(\Theta_{ik}/\tau_1)},
\]

where \( \tau_1 \) is the temperature parameter. At the same time, we obtain another corresponding similarity relationship \( \Theta_{ij} = \frac{1}{2} F_{ij}^\top G_{si} \). The resulting relationship distribution is:

\[
p_i^2 = \frac{\exp(\Theta_{ij}/\tau_2)}{\sum_{k=1}^{K} \exp(\Theta_{ik}/\tau_2)},
\]

where \( \tau_2 \) is a different temperature parameter. We propose to push the relational consistency between \( p_i^1 \) and \( p_i^2 \) by minimizing the Kullback–Leibler divergence, which can be formulated as follows:

\[
\mathcal{L}_{\text{relation}} = D_{KL}(p_i^1 \| p_i^2)
= - \sum_{i=1}^{n} p_i^1 \log \frac{1}{p_i^1} + \sum_{i=1}^{n} p_i^1 \log \frac{1}{p_i^1}
= H(p_i^1, p_i^2) - H(p_i^1).
\]

In this way, we maintain the consistency of the similarity relationship between samples under different modalities, and also narrow the gap between modalities.

C. Cross-Modal Global Similarity Learning

Inspired by TSDH [11], we initialize a set of points \( \mathcal{C} = \{ c_1, c_2, \ldots, c_m \} \subset \{0, 1\}^K \) with a sufficient distance in the Hamming space as cross-modal global similarity centers for subsequent network learning, and propose to learn hash functions supervised by the similarity centers w.r.t. \( C \). The cross-modal global similarity will encourage similar multi-modal data pairs to be close to a common similarity center and dissimilar multi-modal data pairs to be distributed around different similarity centers, respectively. Through such central similarity learning, the global similarity information between data pairs can be preserved in \( h \), yielding high-quality hash codes. According to the concept of the hash center in CSQ [25], we assume that each cross-modal global similarity center should be more distant from the other centers than to the hash codes associated with it. In this way, the dissimilar pairs can be better separated, while similar pairs can be aggregated more cohesively. Based on these observations and intuitions, we define a set of points in the Hamming space as cross-modal
global similarity centers with the following properties:
\[
\frac{1}{T} \sum_{i=p}^{m} D_H(c_i, c_j) \geq \frac{K}{2},
\]
where \( D_H \) is the Hamming distance, \( m \) is the number of hash centers, and \( T \) is the number of combinations of different \( c_i \) and \( c_j \in C \).

Based on the above properties, we use Hadamard matrix \([25]\) to initialize the cross-modal global similarity centers, after which we associate the training data points \( X \) to the different modalities with their individual corresponding centers to compute the central similarity. For multi-label data, we first generate \( q \) cross-modal global similarity centers \( \{c_1, \ldots, c_q\} \) corresponding to the semantic labels \( \{l_1, \ldots, l_q\} \). Then, for the data belonging to two or more categories, we calculate the centroid of these centers, each of which corresponds to a single category.

Given the generated centers \( C = \{c_1, \ldots, c_q\} \) for training data \( X \) with \( q \) categories, we obtain the cross-modal global similarity centers \( C' = \{c'_1, c'_2, \ldots, c'_q\} \) for the multi-label data, where \( c'_i \) denotes the cross-modal global center of the data sample \( x_i \). We derive the central similarity learning objective by maximizing the logarithm posterior of the hash codes w.r.t. the semantic hash centers. Formally, the logarithm Maximum A Posterior estimation of the hash codes \( \mathcal{H} = [h_1, \ldots, h_N] \) for all the training data can be obtained by maximizing the following likelihood probability:
\[
\log P(\mathcal{H} | C') \propto \log P(C' | \mathcal{H}) P(\mathcal{H})
\]
\[
= \sum_{i} \log P(c'_i | h_i) P(h_i),
\]
where \( P(\mathcal{H}) \) is the prior distribution over the hash codes. \( P(C' | \mathcal{H}) \) is the likelihood function. \( P(c'_i | h_i) \) is the conditional probability of center \( c'_i \) given hash code \( h_i \). Here we choose to model \( P(C' | \mathcal{H}) \) as a Gibbs distribution. This is because: 1) In the conditional distribution of hash codes, the values of each bit are independent of each other given the values of other bits. The Gibbs distribution has assumed conditional independence, so it can be used to model the conditional probability distribution of a hash code. 2) Binary nature: Hash codes are represented by binary bits, and the value on each bit can only be 0 or 1. The Gibbs distribution is a probability distribution that applies to binary variables and can be used to model the conditional probability distribution of binary bits. 3) Gibbs distribution has good parameter learning and inference ability. Since \( c'_i \) and \( h_i \) are binary discrete variables, we typically model the dependence between the two using Binary Gibbs Distribution:
\[
P(c'_i | h_i) = \frac{1}{2} \exp (-\beta D_H(c'_i, h_i)),
\]
where \( \alpha \) and \( \beta \) are constants, and \( D_H \) measures the Hamming distance between a hash code and its hash center. By substituting \( \log P(c'_i | h_i) \) into the Maximum A Posterior estimation, we can obtain the optimization objective of the central similarity loss \( L_C \):
\[
L_C = \frac{1}{K} \sum_{i} \sum_{k \in K} [c'_{i,k} \log h_{i,k} + (1 - c'_{i,k}) \log (1 - h_{i,k})].
\]

Subsequently, adopting a similar approach to DHN \([26]\), we use the bi-modal Laplacian prior for quantization; this is defined as \( L_Q = \sum_{i \neq j} ||2h_i - 1||_1 \), where \( 1 \in \) the cross-modal global similarity centers with the following properties:
\[
\frac{1}{T} \sum_{i=p}^{m} D_H(c_i, c_j) \geq \frac{K}{2},
\]
where \( D_H \) is the Hamming distance, \( m \) is the number of hash centers, and \( T \) is the number of combinations of different \( c_i \) and \( c_j \in C \).

Based on the above properties, we use Hadamard matrix \([25]\) to initialize the cross-modal global similarity centers, after which we associate the training data points \( X \) to the different modalities with their individual corresponding centers to compute the central similarity. For multi-label data, we first generate \( q \) cross-modal global similarity centers \( \{c_1, \ldots, c_q\} \) corresponding to the semantic labels \( \{l_1, \ldots, l_q\} \). Then, for the data belonging to two or more categories, we calculate the centroid of these centers, each of which corresponds to a single category.

Given the generated centers \( C = \{c_1, \ldots, c_q\} \) for training data \( X \) with \( q \) categories, we obtain the cross-modal global similarity centers \( C' = \{c'_1, c'_2, \ldots, c'_q\} \) for the multi-label data, where \( c'_i \) denotes the cross-modal global center of the data sample \( x_i \). We derive the central similarity learning objective by maximizing the logarithm posterior of the hash codes w.r.t. the semantic hash centers. Formally, the logarithm Maximum A Posterior estimation of the hash codes \( \mathcal{H} = [h_1, \ldots, h_N] \) for all the training data can be obtained by maximizing the following likelihood probability:
\[
\log P(\mathcal{H} | C') \propto \log P(C' | \mathcal{H}) P(\mathcal{H})
\]
\[
= \sum_{i} \log P(c'_i | h_i) P(h_i),
\]
where \( P(\mathcal{H}) \) is the prior distribution over the hash codes. \( P(C' | \mathcal{H}) \) is the likelihood function. \( P(c'_i | h_i) \) is the conditional probability of center \( c'_i \) given hash code \( h_i \). Here we choose to model \( P(C' | \mathcal{H}) \) as a Gibbs distribution. This is because: 1) In the conditional distribution of hash codes, the values of each bit are independent of each other given the values of other bits. The Gibbs distribution has assumed conditional independence, so it can be used to model the conditional probability distribution of a hash code. 2) Binary nature: Hash codes are represented by binary bits, and the value on each bit can only be 0 or 1. The Gibbs distribution is a probability distribution that applies to binary variables and can be used to model the conditional probability distribution of binary bits. 3) Gibbs distribution has good parameter learning and inference ability. Since \( c'_i \) and \( h_i \) are binary discrete variables, we typically model the dependence between the two using Binary Gibbs Distribution:
\[
P(c'_i | h_i) = \frac{1}{2} \exp (-\beta D_H(c'_i, h_i)),
\]
where \( \alpha \) and \( \beta \) are constants, and \( D_H \) measures the Hamming distance between a hash code and its hash center. By substituting \( \log P(c'_i | h_i) \) into the Maximum A Posterior estimation, we can obtain the optimization objective of the central similarity loss \( L_C \):
\[
L_C = \frac{1}{K} \sum_{i} \sum_{k \in K} [c'_{i,k} \log h_{i,k} + (1 - c'_{i,k}) \log (1 - h_{i,k})].
\]

Subsequently, adopting a similar approach to DHN \([26]\), we use the bi-modal Laplacian prior for quantization; this is defined as \( L_Q = \sum_{i \neq j} ||2h_i - 1||_1 \), where \( 1 \in \)
Algorithm 1 The Learning Algorithm

Input: Image set $X$, text set $Y$;

Output: Parameters $\theta_x$ and $\theta_y$ in the hash layer and binary codes $B$

Initialization

Initialize neural network parameters $\theta_x$ and $\theta_y$, mini-batch size $N_x = N_y = 128$, and iteration number $t_x = n/N_x$, $t_y = n/N_y$.

repeat

for iter = 1, 2, ..., $t_x$ do

Randomly sample $N_x$ points from $X$.

For each sampled point $x_i$ in the mini-batch, calculate $F_{si} = f(x_i; \theta_x)$ by forward propagation.

Calculate the derivative according to Eq. 9. Update the parameter $\theta_x$ by using back propagation.

end for

for iter = 1, 2, ..., $t_y$ do

Randomly sample $N_y$ points from $Y$.

Calculate $G_{ij} = g(y_i; \theta_y)$.

Calculate the derivative according to Eq. 9. Update the parameter $\theta_y$ by using back propagation.

end for

Learn $B$ according to Eq. 10. until a fixed number of iterations.

$\mathbb{R}^K$ is an all-one vector. As $L_Q$ is a non-smooth function, which makes it difficult to calculate its derivative, we adopt the smooth function log cosh [27] to replace it. So $|x| \approx \log \cosh x$. The quantization loss $L_Q$ then becomes

$$L_Q = \sum_i \sum_k \log \cosh ([2h_{i,k} - 1] - 1). (7)$$

We then have the following cross-modal global similarity loss:

$$L_{CG} = \lambda L_C + \mu L_Q. (8)$$

Therefore, the overall objective function of the hash learning is written as follows:

$$L_T = \lambda L_C + \mu L_Q + \eta L_{\text{relation}}. (9)$$

where $\lambda$, $\mu$, and $\eta$ are hyper-parameters. We determine the specific values of $\mu$ and $\eta$ through parameter analysis experiments. In addition, compared with other optimization methods (such as random search or Bayesian optimization), Grid Search is more feasible in the case of limited computational resources when the number of hyper-parameters is relatively small and the value range of each hyper-parameter is limited. It can traverse the parameter space through a pre-determined discrete grid, without additional sampling and iteration processes. Since the $\lambda$ is a parameter with a small value range [25], usually within 0–0.1, we obtained the value of $\lambda$ as 0.05 through grid search.

D. Optimization

Let $L(T_k; \theta_i)$ $\in \mathbb{R}^c$ denote the learned cross-modal global similarity center for point $k$, which corresponds to the output of the deep neural network for global similarity learning, and $\theta_i$ is the parameter of the deep neural network. We adopt an alternating learning strategy to learn $\theta_x$, $\theta_y$, $\theta_l$, and $B$. Each time, we optimize one parameter with the other parameters remaining fixed. An overview of the whole alternating learning algorithm is provided in Algorithm 1. When $\theta_x$, $\theta_y$, and $B$ are fixed, we learn the CNN parameter $\theta_l$ of the image modality by using a back propagation (BP) algorithm. In line with most existing deep learning methods [28], we utilize stochastic gradient descent (SGD) to learn $\theta_l$ with the BP algorithm. In each iteration, we sample a mini-batch of points from the training set, then execute our learning algorithm based on the sampled data.

In more detail, for each point $x_i$, we first compute the gradient $\frac{dL}{\theta_l}$. Then we can compute $\frac{dL}{\theta_l}$ using the chain rule, with which BP can be used to update the parameter $\theta_l$. Similarly, when $\theta_x$, $\theta_y$, and $B$ are fixed, we also learn the neural network parameter $\theta_l$ of the text modality by using SGD with a BP algorithm, then update the parameter $\theta_l$. When $\theta_x$, $\theta_y$, and $B$ are fixed, we learn the neural network parameter $\theta_l$ of the global similarity learning, then update the parameter $\theta_l$. When $\theta_x$, $\theta_y$ and $\theta_l$ are fixed, it can be easily determined that the binary code $B_{ij}$ should keep the same sign as $V_{ij} = \gamma(F_{ij} + G_{ij})$. Therefore, we have:

$$B = \text{sign}(V) = \text{sign}(\gamma(F + G)). (10)$$

where $\gamma$ is a hyper-parameter, we set it to 1 in accordance with the previous convention [7]. In this way, the binary code of each sample can be derived.

E. Superiority Analysis

In this section, we prove the superiority of CGS through theoretical analysis of the optimization process. In general, we believe that the hash codes should be optimal for joint learning classifiers. A linear classifier $f^x(\cdot)$ is commonly used to model the relationship between hash codes (e.g., $b^x_i$) and semantic labels (e.g., $l^x_i$); in this context, it can be defined as follows:

$$f(b^x_i) = W^x b^x_i. (11)$$

where $W^x = [w^x_1, w^x_2, \ldots, w^x_o]$ is the parameter for the classifier $f^x(\cdot)$, and $w^x_j$ is the weight vector for the $d$-th class. The following loss is typically used to optimize the classifier:

$$L^x_c(b^x) = -\log \left[ l^x_i + \frac{1}{1 + \exp(-f(b^x_i))} \right]$$

$$= -\sum_{d=1}^o \left[ l^x_i d \log \left( 1 + \frac{1}{1 + \exp(-w^x_{jd} b^x) \right]} + (1 - l^x_i d \log \left( 1 + \frac{1}{1 + \exp(-w^x_{jd} b^x) \right] \right). (12)$$

Through the above optimization, the hash codes can be brought closer to the weight vectors of their corresponding
classes and pushed away from the weight vectors of other classes. In our method, cross-modal global similarity centers behave similarly to the class weight vectors. We note that the class weight vectors contain continuous values that may have different scales, while the cross-modal global similarity centers are all binary hash codes. Using Eq. 12 will cause the learned hash code to focus on hash dimensions with larger weights (i.e., those that are more important); however, this may lead to inconsistent weights of different categories of hash dimensions, which will significantly hamper both the accurate calculation of Hamming distances and the search performance. In addition, the uneven distribution of class weight vectors will cause the learned hash codes to be concentrated in certain specific classes, thus preventing the model from being extended to other classes. Therefore, the optimization used in Eq. 12 may not be optimal for hash code learning. To address this issue, our method explicitly learns the binary cross-modal global similarity centers $C^r = \{c^r_1, c^r_2, \ldots, c^r_N\}$ and directly minimizes the distances between the hash codes and their cross-modal global similarity centers, enabling it to generate more consistent hash codes with better discrimination.

IV. EXPERIMENTS

In this section, we evaluate the proposed MRDH by implementing the image–text retrieval on three benchmark datasets. Several existing state-of-the-art methods are adopted as baselines to demonstrate the effectiveness of the proposed approach.

A. Datasets

1) MS COCO: MS COCO [29] is a public dataset consisting of 123,287 multi-labeled images of 80 classes, and is commonly used for several tasks, including image recognition, image text description, and segmentation. Each image depicts a complex regular scene and is annotated with five descriptive sentences. In total, 117,218 data points are used in our experiment, with 10,000 data points for the training set and 5,000 for the query set; the remaining data points serve as the retrieval set. Each text is represented as a 2,000-dimension bag-of-words vector.

2) MIRFLICKR-25K: MIRFLICKR-25K [30] contains 25,000 data points collected from FLICKR. A total of, 20,015 data points are selected in our experiment (10,000 for training, 2,000 for query, and the remainder for retrieval). The text for each point is represented as a 1,386-dimensional bag-of-words vector.

3) NUS-WIDE: NUS-WIDE [31] contains about 269,648 web images with 81 ground truth concepts. After pruning the data without any label or tag information, a subset of 188,321 data points belonging to the 21 most frequent concepts are selected; ultimately, 10,500 data points are used for training and 2,100 for query, while the remainder serve as the retrieval set.

B. Implementation Details

We implement our MRDH based on the open-source deep toolbox PyTorch. All experiments are run on a server with one NVIDIA TITAN X GPU. In the training procedure, both multi-modal representation learning and hash learning are optimized in an alternating way. The batch size is 64, and the total number of epochs is 80. The initial learning rate is set to 0.0001 and then decreased by one-fifth every 30 epochs. In the testing procedure, we use only the modality-common representations of image and text modalities to construct the binary codes.

For the image modality, the input images are first resized to $224 \times 224 \times 3$. We use two CNN architectures as the extractor $E^V$ to extract the image representation: ResNet50 [32] and VGG16 [33], which both consist of five convolutional blocks and three fully connected layers. The output of their $f_C$ layer serves as the image representation. For the text modality, the text representation extraction $E^T$ is constructed by two $1 \times 1$ convolutional layers. After extracting the representations of the image and text, we design two consistency refining modules, i.e., $C^R^V$ and $C^R^T$, which consist of two different fully-connected layers, both of which have 512 neural nodes. For the Cross-modal Global Similarity learning module, we frame a MLP $E^C$, which is made up of fully connected layers. For hash learning, two hash layers $f_c^V$ and $f_c^T$ are designed for hash codes generation, which are two fully-connected layers with $K$ neural nodes.

Regarding the activation functions used in MRDH, sigmoid [34] activation is used to output the predicted label, tanh [35] activation is used to output the hash codes, and the rest of the layers are uniformly activated by the ReLU [36] function.

C. Evaluation and Baselines

1) Evaluation: The Hamming ranking and hash lookup are two classical retrieval protocols used to evaluate

<table>
<thead>
<tr>
<th>TASK</th>
<th>Method</th>
<th>MIRFLICKR-25K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16 bits</td>
</tr>
<tr>
<td>Image Query v.s.</td>
<td>LSSH</td>
<td>0.557</td>
</tr>
<tr>
<td></td>
<td>SePH</td>
<td>0.681</td>
</tr>
<tr>
<td></td>
<td>DCMH</td>
<td>0.733</td>
</tr>
<tr>
<td></td>
<td>ALECH</td>
<td>0.730</td>
</tr>
<tr>
<td></td>
<td>ASCSH</td>
<td>0.751</td>
</tr>
<tr>
<td></td>
<td>FDDH</td>
<td>0.810</td>
</tr>
<tr>
<td></td>
<td>GCDH</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>0.837</td>
</tr>
<tr>
<td>Text Query v.s.</td>
<td>LSSH</td>
<td>0.557</td>
</tr>
<tr>
<td></td>
<td>SePH</td>
<td>0.622</td>
</tr>
<tr>
<td></td>
<td>DCMH</td>
<td>0.749</td>
</tr>
<tr>
<td></td>
<td>ALECH</td>
<td>0.795</td>
</tr>
<tr>
<td></td>
<td>ASCSH</td>
<td>0.805</td>
</tr>
<tr>
<td></td>
<td>FDDH</td>
<td>0.791</td>
</tr>
<tr>
<td></td>
<td>GCDH</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>0.818</td>
</tr>
</tbody>
</table>

| Image Database | LSSH | 0.557 | 0.559 | 0.561 |
|                | SePH | 0.622 | 0.632 | 0.633 |
|                | DCMH | 0.749 | 0.764 | 0.766 |
|                | ALECH | 0.795 | 0.803 | 0.800 |
|                | ASCSH | 0.805 | 0.829 | 0.836 |
|                | FDDH | 0.791 | 0.812 | 0.825 |
|                | GCDH | 0.815 | 0.822 | 0.832 |
|                | Ours | 0.818 | 0.831 | 0.839 |
the performance of a cross-modal retrieval task. In our experiments, we use two evaluation criteria: mean average precision (MAP) [37], which is used to measure the accuracy of the Hamming distances; the precision-recall (PR) curve, which is used to measure the accuracy of the hash lookup protocol.

2) Baselines: We compare our proposed MRDH with seven state-of-the-art methods, including two shallow-structure-based methods (SePH [38], LSSH [15]), and several deep-structure-based methods (DCMH [7], ALECH [5], ASCSH [5], FDDH [8], GCDH [9]). To facilitate fair comparison, we utilize both ResNet50 [32] and VGG-19 [33], both of which have been pretrained on the ImageNet dataset [39], to extract deep features for all shallow-structure-based baselines. For the image modality, we initialize the first seven layers of ImgNet with the ResNet50 network pre-trained on the ImageNet dataset. For the text modality, TxtNet is randomly initialized. The learning rate is chosen from $10^{-4}$ to $10^{-8}$. Subsequently, we report the average results of the 10 runs.

D. Performance

1) Hamming Ranking: Table II, Table IV, and Table VI report the MAP results for both our MRDH and the other comparison methods with ResNet50 features on three popular datasets (MIRFLICKR-25K, NUS-WIDE and MS COCO) in cross-modal retrieval. Compared with the shallow baselines of SePH [38], and LSSH [15], our MRDH achieves more than a 10% absolute increase in terms of MAP for Image $\rightarrow$ Text/Text $\rightarrow$ Image on the MIRFLICKR-25K dataset. Moreover, when comparing our MRDH with the deep-learning-based methods (DCMH [7], ALECH [5], ASCSH [5], FDDH [8], GCDH [9]), we run the source code provided by the authors. Here, it can be seen that MRDH can achieve more than a 5% increase in MAP. For another two datasets NUS-WIDE and MS COCO which have more instances and more complex content, and are thus more challenging, MRDH always achieves superior performance when measured against
Fig. 3. The Precision-recall curves of different methods for the “image-query-text” and “text-query-image” on three datasets respectively (Code length 64).

The MAP Scores of Two Retrieval Tasks on the NUS-WIDE Dataset With Different Hash Code Lengths. The Baselines Are Based on ResNet50 Features and the Best Accuracy Is Shown in Boldface

<table>
<thead>
<tr>
<th>TASK</th>
<th>Method</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16 bits</td>
</tr>
<tr>
<td>Image Query</td>
<td>LSSH</td>
<td>0.434</td>
</tr>
<tr>
<td>v.s.</td>
<td>SePH</td>
<td>0.592</td>
</tr>
<tr>
<td>Text Database</td>
<td>DCMH</td>
<td>0.526</td>
</tr>
<tr>
<td></td>
<td>ALECH</td>
<td>0.650</td>
</tr>
<tr>
<td></td>
<td>ASCSH</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>FDDH</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td>GCDH</td>
<td>0.720</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td><strong>0.841</strong></td>
</tr>
<tr>
<td>Text Query</td>
<td>LSSH</td>
<td>0.399</td>
</tr>
<tr>
<td>v.s.</td>
<td>SePH</td>
<td>0.561</td>
</tr>
<tr>
<td>Image Database</td>
<td>DCMH</td>
<td>0.577</td>
</tr>
<tr>
<td></td>
<td>ALECH</td>
<td>0.725</td>
</tr>
<tr>
<td></td>
<td>ASCSH</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>FDDH</td>
<td>0.799</td>
</tr>
<tr>
<td></td>
<td>GCDH</td>
<td>0.736</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td><strong>0.802</strong></td>
</tr>
</tbody>
</table>

The MAP Scores of Two Retrieval Tasks on the NUS-WIDE Dataset With Different Hash Code Lengths. The Baselines Are Based on VGG16 Features and the Best Accuracy Is Shown in Boldface

<table>
<thead>
<tr>
<th>TASK</th>
<th>Method</th>
<th>NUS-WIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16 bits</td>
</tr>
<tr>
<td>Image Query</td>
<td>LSSH</td>
<td>0.481</td>
</tr>
<tr>
<td>v.s.</td>
<td>SePH</td>
<td>0.608</td>
</tr>
<tr>
<td>Text Database</td>
<td>DCMH</td>
<td>0.523</td>
</tr>
<tr>
<td></td>
<td>ALECH</td>
<td>0.645</td>
</tr>
<tr>
<td></td>
<td>ASCSH</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>FDDH</td>
<td>0.837</td>
</tr>
<tr>
<td></td>
<td>GCDH</td>
<td>0.722</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td><strong>0.845</strong></td>
</tr>
<tr>
<td>Text Query</td>
<td>LSSH</td>
<td>0.445</td>
</tr>
<tr>
<td>v.s.</td>
<td>SePH</td>
<td>0.536</td>
</tr>
<tr>
<td>Image Database</td>
<td>DCMH</td>
<td>0.566</td>
</tr>
<tr>
<td></td>
<td>ALECH</td>
<td>0.650</td>
</tr>
<tr>
<td></td>
<td>ASCSH</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>FDDH</td>
<td>0.810</td>
</tr>
<tr>
<td></td>
<td>GCDH</td>
<td>0.738</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td><strong>0.811</strong></td>
</tr>
</tbody>
</table>

the comparison methods. This may be because, during the learning process, the proposed relational global similarity hashing network can more effectively facilitate the learning of semantic relevance between different modalities, which means that more discriminative representations can be learned using our MRDH. As a result, MRDH can more accurately capture correlations between modalities.

We further verify our MRDH using VGG-19 features [33] that have been pre-trained on the ImageNet dataset. Table I, Table III, and Table V show the MAP results on three
Fig. 4. MAP varies with the parameters \( \eta \), \( \mu \), and the temperature parameters \( \tau_t \), \( \tau_s \) for different bit lengths on the MS COCO.

2) Hash Lookup: When considering the lookup protocol, we compute the PR curve for the returned points given any Hamming radius. The PR curve can be obtained by varying the Hamming radius from 0 to K with a step-size of 1. Fig 3 illustrate the precision-recall curves on three datasets with 64-bit hash codes. In this way, it can be seen that our MRDH significantly outperforms all of its state-of-the-art competitors.

E. Ablation Study

In this part, we conduct rich ablation experiments to analyze the respective contributions of the Multiple Relation Learning (MRL), and Cross-modal Global Similarity (CGS) module. More specifically, for our proposed MRDH, we implement two variants: (1) GSLH, which is the variant with the multiple relation learning module removed; (2) RSLH, which is the variant with the cross-modal global similarity module removed. As shown in Table VII, Table VIII, and Table IX, MRDH (equipped with the MRL and CGS modules) achieves more promising performance than GSLH and RSLH. We can also make the following observations: 1) CGS alone has a significantly stronger effect on the network than MRL. This is because MRL focuses on the relationship between modalities, but cannot effectively exploit the relationship between categories within the same modality, while CGS models both relationships. 2) The MAP values for MS COCO are generally inferior to those for MIRFLICKR-25K. This is partly due to the fact that MS COCO is a fine-grained dataset, in which most categories lack diversity, meaning that the learned hash codes are less discriminative. 3) The MAP values are positively related to code lengths, which indicates that the discriminative ability of hash codes improves as the code length increases.

F. Parameter Analysis

To further evaluate the MRDH, two parameters \( \mu \), \( \eta \) and two temperature parameters \( \tau_t \), \( \tau_s \) of our objective function...
Fig. 5. Qualitative results of text retrieval given image queries on MIRFLICKR-25K and MS COCO datasets. For each image query we show the top-5 ranked texts.

Fig. 6. Qualitative results of image retrieval given text queries on MIRFLICKR-25K. For each text query, we show the top-5 ranked images, ranking from left to right. We outline the true matches in green boxes and false matches in red boxes. In the examples we show, our model retrieves the ground truth image in the top-5 list. Note that other results are also reasonable outputs.

are analyzed on the MS COCO. We present the results in Fig. 4. First, we set $\mu$ and vary the value of $\eta$. It can be seen from Fig. 4(a) that our MRDH obtains the best MAP value when $\eta$ is set to 1. Next, we set $\eta = 1$ and vary the value of $\mu$. Fig. 4(b) shows the results. It can be seen that the MRDH method obtains the best MAP value when $\mu$ is set to
1. The temperature parameters are also tested by this method of rotation tuning. Hence, we use the parameter values $\mu = \eta = 1$, temperature parameter values $\tau_f = 0.05$, $\tau_s = 0.1$ in all our experiments.

G. Visualization

Fig. 5. and Fig. 6 present the visualizations of our method on the cross-modal retrieval task. The feature used is the ResNet feature, and the hash code length is 64. Fig. 5. shows qualitative descriptive search results using image query on MIRFLICKR-25K and MS COCO. Below each image, we show the top five retrieval sentences in descending order of confidence. The sentences in green are the true matches, while the descriptions in red are the false matches. Note that some general descriptions are also reasonable outputs. Fig. 6. shows the qualitative image search results obtained using text query. The results are sorted from left to right in order of their confidence. The images marked with green boxes are the true matches, while those outlined in red are the false matches. Cross-modal retrieval utilizes the labels of data pairs to determine whether the two are correct matches. Although some sample pairs have inconsistent labels, the actual semantic correlation is great. In practical application, these false matches are also reasonable outputs. In addition, some labels may be missing some key information, resulting in a semantic gap between the labels and the samples. However, we calculate the confidence by the semantic similarity between the sample pairs, that is why incorrect matches may have higher confidence than correct matches.

V. Conclusion

In this work, we proposed a novel deep hashing approach, dubbed Multi-Relational Deep Hashing (MRDH) for cross-modal retrieval. The proposed MRDH can fully bridge the modality gap by comprehensively modeling the similarity relationship between data in different modalities. Specifically, to investigate the inter-modal relationships, we constrain the consistency of cross-modal pairwise similarities to maintain the semantic similarity across modalities (MLR). Moreover, to further capture complete similarity information, we design a new similarity metric, dubbed cross-modal global similarity (CGS), by encouraging hash codes of similar data pairs from different modalities to approach a common center and hash codes for dissimilar pairs to converge to different centers. Adopting this approach enables our model to generate more discriminative hash codes. Extensive experiments on three benchmark datasets demonstrate the superiority of our method on cross-modal hashing retrieval.

With the emergence of large-scale visual-language pre-trained models (VLPs), fine-tuning VLPs has become a new paradigm for cross-modal retrieval task, which can achieve better performance than ours. However, our approach mainly focuses on constructing comprehensive relational constraints to capture complete multi-modal similarity information without relying on large models. In the future, we will combine our method with large-scale VLPs to achieve better performance.

REFERENCES


Xiao Liang received the B.E. and M.S. degrees from Xidian University, China, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree with the School of Electronic Engineering. His research interests include computer vision and machine learning.

Erkun Yang received the B.E., M.S., and Ph.D. degrees from the School of Electronic Engineering, Xidian University. He is currently an Associate Professor with the School of Electronic Engineering, Xidian University. He is the author or coauthor of more than 40 scientific articles at top venues. His research interests include computer vision and machine learning.

Yanhua Yang received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi’an, China, in 2004, 2007, and 2017, respectively. She is currently an Associate Professor with the School of Computer Science and Technology, Xidian University. Her main research interests include complex action recognition and event detection.

Cheng Deng (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi’an, China. He is currently a Full Professor with the School of Electronic Engineering, Xidian University. He is the author or coauthor of more than 200 scientific articles at top venues. His research interests include multimodal representation learning, multimodal retrieval, and visual generation.