Deep Normalized Cross-Modal Hashing with Bi-Direction Relation Reasoning

Changchang Sun[1], **Hugo Latapie**[2], **Gaowen Liu**[2], **Yan Yan**[1]
[1]Department of Computer Science, Illinois Institute of Technology, USA
[2]Emerging Technologies and Incubation, Cisco Research, USA
csun39@hawk.iit.edu; hlatapie@cisco.com; gaoliu@cisco.com; yyan34@iit.edu

Abstract

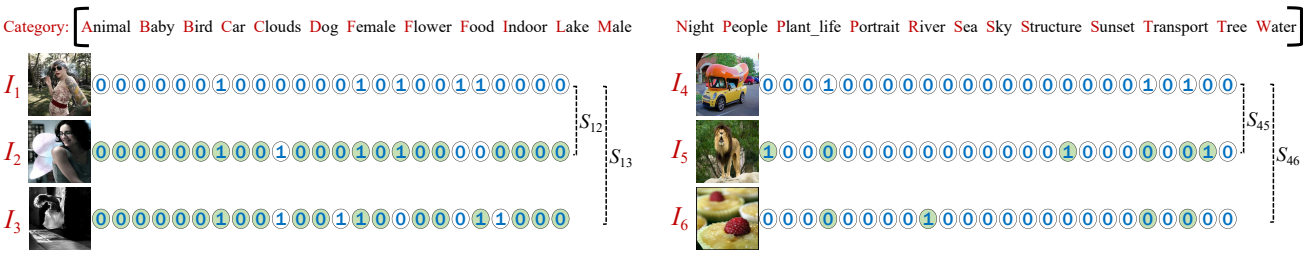
Due to the continuous growth of large-scale multi-modal data and increasing requirements for retrieval speed, deep cross-modal hashing has gained increasing attention recently. Most of existing studies take a similarity matrix as supervision to optimize their models, and the inner product between continuous surrogates of hash codes is utilized to depict the similarity in the Hamming space. However, all of them merely consider the relevant information to build the similarity matrix, ignoring the contribution of the irrelevant one, i.e., the categories that samples do not belong to. Therefore, they cannot effectively alleviate the effect of dissimilar samples. Moreover, due to the modality distribution difference, directly utilizing continuous surrogates of hash codes to calculate similarity may induce suboptimal retrieval performance. To tackle these issues, in this paper, we propose a novel deep normalized cross-modal hashing scheme with bi-direction relation reasoning, named Bi_NCMH. Specifically, we build the multi-level semantic similarity matrix by considering bi-direction relation, i.e., consistent and inconsistent relation. It hence can holistically characterize relations among instances. Besides, we execute feature normalization on continuous surrogates of hash codes to eliminate the deviation caused by modality gap, which further reduces the negative impact of binarization on retrieval performance. Extensive experiments on two cross-modal benchmark datasets demonstrate the superiority of our model over several state-of-the-art baselines.

1. Introduction

With the increasing prevalence of portable digital devices and the popularity of social media platforms, it has become the daily habit for most netizens to record and share massive amounts of data in various modalities. For example, users can upload images with textual descriptions on Instagram[1]. Accordingly, obtaining the relevant information via cross-modal retrieval has become a great demand of users when surfing the Internet. For instance, users may seek the desired images or videos by textual queries. Towards improving the retrieval efficiency, cross-modal hashing [2,5,11,20–23,25,30,31,42] that maps the multi-modal data into a unified Hamming space with similar binary hash codes for semantically similar data has become a promising topic recently.

Existing cross-modal hashing techniques can be broadly categorized into unsupervised methods [9,10,15,24,28,38,42,43] and supervised ones [4,7,16,17,29,35,36,41]. The former focus on exploring the semantic affinities of training data to learn the hash projection function, yet neglecting the importance of semantic labels, resulting in the inferior performance. To bridge this gap, the latter mainly incorporate the similarity matrix built by semantic labels of instances to supervise the hash code learning. In this way, the similarity between instances can be better retained by the learned hash codes. Despite the promising performance of supervised cross-modal hashing, there are still several critical shortcomings. 1) Most of supervised methods [4,16,17,29,35,41] simply treat two instances similar as long as they share a common relevant category, and dissimilar otherwise. As each instance may belong to multiple categories, this naive binary similarity assessment may not precisely reflect the complex relations between two instances. For example, in Fig. 1a, compared with image I_3, image I_2 shares more common relevant and irrelevant categories with I_1. Therefore, the similarity between images I_1 and I_2 should be higher than that between I_1 and I_3. 2) Although some pioneering approaches [7,36] have considered the multiple categories in defining the similarity between instances, they merely focus on how similar two instances are, while thoroughly overlooking the dissimilarity between them. The dissimilar information in fact delivers pivotal cues regarding the complex relation between instances. As shown in Fig. 1b, both image I_5 and I_6 have no shared category with image I_4. However, I_5 has more non-overlapped

[1]https://www.instagram.com/.



(a) Consistent direction.

(b) Inconsistent direction.

Figure 1. Illustration of bi-direction relation reasoning according to multi-labels, where 1 and 0 separately represent the relevant and irrelevant category. As for the consistent direction, overlapped information between two instances is considered to measure their similarity score. Apparently, I_2 is more similar to I_1 compared to I_3, i.e., $S_{12} > S_{13} > 0$. As to the inconsistent direction, non-overlapped information is concerned. Obviously, I_5 is more dissimilar to I_4 compared to I_6, i.e., $S_{45} < S_{46} < 0$.

category information with I_4 compared to I_6, and hence more dissimilar to I_4. And 3) existing deep hashing methods commonly treat the last layer output of the deep neural network as the hash representation of each instance, and directly adopt the inner product between them as the similarity between two instances. Nevertheless, due to the distribution difference between heterogeneous modalities, this kind of measurement may hurt the model performance. Specifically, if the magnitudes of hash representations vary greatly between different modalities, the similarity will be determined by the one with larger magnitude. Thereby, the similarity determination of hash codes is adversely affected, causing poor retrieval performance ultimately.

To address the aforementioned issues, we propose a novel deep normalized cross-modal hashing scheme with bi-direction relation reasoning (Bi_NCMH), as shown in Fig. 2. In particular, on the one hand, to take full advantage of multi-labels, we design a bi-direction relation modeling method to construct the multi-level semantic similarity matrix, with values ranging from -1 to 1. Among them, the positive value represents the similarity degree of two similar instances sharing at least one relevant category, while the negative one depicts the dissimilarity extent of two samples without category shared. In this way, the semantic similarity can be expressed more precisely. On the other hand, we introduce the normalization operation to bridge the modality distribution gap and compress hash representations range from -1 to 1, while modules of them are 1. By this means, we can further reduce the binarization loss. In addition, we consider three loss indicators, namely inter-modal, intra-modal and regularization loss, to respectively constrain the inter-modal, intra-modal, and relaxation similarity.

The key contributions of this work are three-fold:

- We design a novel bi-direction relation reasoning scheme to capture the complex multi-level semantic similarity relying on multi-labels. Moreover, under this supervision, hash codes could preferably maintain original similarity relations.
- To the best of our knowledge, this is the first attempt to execute the feature normalization on the hash representation. It can effectively reduce the modality distribution gap and binarization penalization.
- Extensive experiments on two multimodal benchmark datasets demonstrate the superiority of our model over several state-of-the-art methods.

2. Related Work

2.1. Unsupervised Cross-modal Hashing

Unsupervised cross-modal hashing methods [10, 13, 15, 28, 42], similar to conventional subspace learning methods, generally aim to learn a feature projection function by constructing correlations between different modalities. In particular, Ding *et al*. [9] employed the collective matrix factorization technology to learn cross-view hash functions. To capture the high-level latent semantic information and bridge the semantic gap, Zhou *et al*. [42] proposed a novel latent semantic sparse hashing by combining the sparse coding and matrix factorization. This model can well capture the salient information of images and latent concepts of text. Moreover, as the quantization quality is very essential for improving retrieval performance, Irie *et al*. [15] presented an alternating co-quantization scheme. It alternately seeks the binary quantizer for each modality to minimize quantization errors. Different from previous approaches, to preserve the fusion similarity among different modalities, Liu *et al*. [19] put forward fusion similarity hashing by constructing an undirected asymmetric graph and explicitly embedding it into a common Hamming space. However, the performance of all above models are far from satisfactory since they thoroughly overlook the label information.

2.2. Supervised Cross-modal Hashing

Supervised cross-modal hashing methods [16–18, 26, 39, 40] work on leveraging semantic labels as the supervision to guide hash codes learning and boost the retrieval performance. Thereinto, Zhang *et al*. [36] proposed a novel supervised multimodal hashing method, named semantic correla-

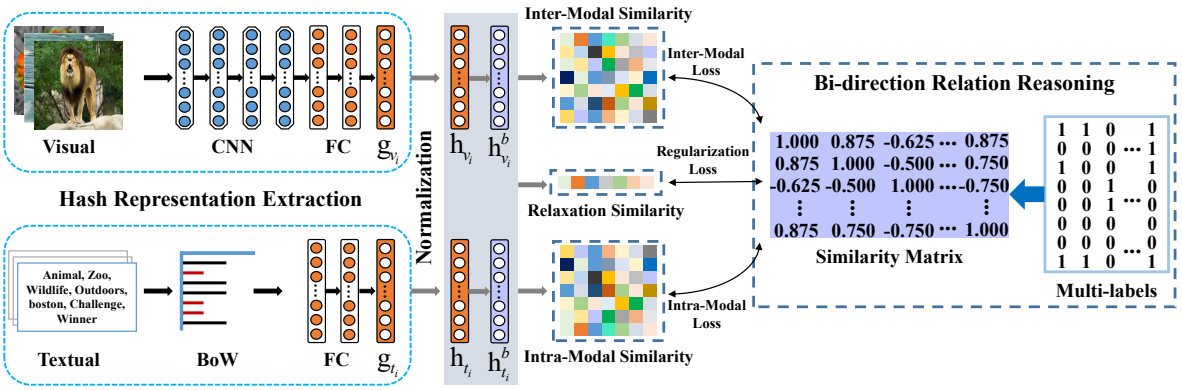


Figure 2. Pipeline of our proposed cross-modal hashing scheme. Under the supervision of a pre-defined bi-direction multi-level similarity matrix, we simultaneously consider the inter-modal, intra-modal, and regularization loss to learn the optimal hash codes. In this way, the inter-modal, intra-modal, and relaxation similarity could be well preserved.

tion maximization (SCM), where semantic labels are seamlessly integrated into the hash learning procedure. In addition, given semantic affinities of training data, Lin *et al*. [18] formulated a semantic-preserving hashing paradigm. They first transformed semantic affinities into a probability distribution, and then approximated it with to-be-learnt hash codes via minimizing the KL-divergence. Furthermore, to achieve discriminative binary codes while retaining discrete constraints, Xu *et al*. [34] introduced discrete cross-modal hashing (DCH) by formulating a linear classification framework. It is worth noting that above methods mainly rely on hand-crafted features, where feature extraction and hash codes learning procedures are separate.

Recently, deep hashing has attracted more and more attention due to its strong representation ability. For example, Jiang *et al*. [16] established an end-to-end cross-modal hashing (DCMH) framework with deep neural networks, to perform the feature learning from scratch. Besides, to exploit the hierarchical correlation among labels, Sun *et al*. [29] introduced a new supervised hierarchical cross-modal hashing method to unify the hierarchical discriminative learning and regularized cross-modal hashing. Although compelling successes have been achieved by these supervised methods, they excessively focus on the relevant information to build the modality relation, throughly ignoring the contribution of irrelevant one. Furthermore, reducing the binarization loss is also an important issue to be considered. To this end, we propose a novel normalized cross-modal hashing with bi-direction relation reasoning.

3. Problem Formulation

Suppose that we have N multi-labeled instances $\mathcal{E} = \{e_i\}_{i=1}^{N}$, where e_i refers to the i-th instance. Each instance is comprised of an image, a text, and a label vector, i.e., $e_i = (\mathbf{v}_i, \mathbf{t}_i, \mathbf{y}_i), i \in \{1, 2, \ldots, N\}$, where $\mathbf{y}_i \in \{0, 1\}^K$ and K denotes the number of categories. In particular, if instance e_i is labeled with the k-th category, the k-th element of $\mathbf{y_i}$ is 1; otherwise is 0. Besides, according to label vectors, we construct a $N \times N$ pair-wise similarity matrix $\mathbf{S}$, with values ranging from -1 to 1, by bi-direction relation reasoning.

In this work, we aim to devise a deep normalized cross-modal hashing learning scheme, with $\mathbf{S}$ as the supervision information, to learn the hash projection function. It could map the visual and textual data of the input instance into L-bit hash codes, namely, $\mathbf{b}_{v_i} \in \{-1, 1\}^L$ and $\mathbf{b}_{t_i} \in \{-1, 1\}^L$, $i \in \{1, 2, \ldots, N\}$, and well maintain their similarities. In light of this, we can conduct the cross-modal retrieval via measuring the Hamming distance, i.e., $dis_H(\mathbf{b}_{v_i}, \mathbf{b}_{t_j}) = \frac{1}{2}(L - \mathbf{b}_{v_i}^T \mathbf{b}_{t_j})$.

4. The Proposed Model

As Fig. 2 illustrates, our proposed cross-modal hashing scheme comprising four components: 1) a hash representation extraction module; 2) a hash representation normalization module; 3) a bi-direction relation reasoning module; and 4) the model training and optimization. In what follows, we will introduce each module in detail.

4.1. Hash Representation Extraction

Inspired by the huge success of deep representation learning, we adopt deep networks to extract powerful image and text hash representations, which are continuous surrogates of image and text hash codes. Regarding the visual modality, we exploit the classical deep convolution neural network (CNN), such as CNN-F [3] and VGG19 [27], to extract image features. In this work, both CNN-F and VGG19 are considered, and analyses of them can be found in the experiment section. Specifically, we choose the output of the last layer as the image hash representation. As for the textual modality, we first construct a word bag by filtering words that appear below the specific word frequency[2]. And then we obtain text representations based on the bag-

[2]In our work, the word frequency is set to 20.

of-words (BoW) model. Afterwards, we employ a fully-connected neural network to transform them into text hash representations.

Formally, the above process can be summarized as follows,

$$\begin{cases} \mathbf{g}_{v_i} = f^v(\mathbf{v}_i; \mathbf{\Theta}_v), \\ \mathbf{g}_{t_i} = f^t(\mathbf{t}_i; \mathbf{\Theta}_t), \end{cases} \tag{1}$$

where f^v and f^t respectively refer to the image encoder and text encoder with parameters $\mathbf{\Theta}_v$ and $\mathbf{\Theta}_t$ to be learned; $\mathbf{g}_{v_i}$ and $\mathbf{g}_{t_i}$ denote the image hash representation and text hash representation of the i-th instance, separately.

4.2. Hash Representation Normalization

Having obtained hash representations for visual and textual modalities, most existing studies [16, 17] directly adopt the inner product to measure the similarity between them. However, due to the modality gap and distribution difference, this kind of measurement may lead to deviation. Specifically, if magnitudes of these multi-modal hash representations vary greatly, the similarity will be determined by the one with the larger magnitude. This may dramatically affect the similarity calculation between target hash codes, therefore causing poor retrieval performance. To mitigate this issue, we execute normalization on the learnt hash representations to compress them ranging from -1 to 1, while keeping the modules of them with 1. By this means, hash representations of two modalities play the same role, which effectively alleviate the influence of the modality with higher magnitude. Particularly, the normalization operation can be formally represented as,

$$\begin{cases} \mathbf{h}_{v_i} = \dfrac{\mathbf{g}_{v_i}}{\left\|\mathbf{g}_{v_i}\right\|_2}, \\ \mathbf{h}_{t_j} = \dfrac{\mathbf{g}_{t_j}}{\left\|\mathbf{g}_{t_j}\right\|_2}. \end{cases} \tag{2}$$

Afterwards, we feed the normalized hash representations to the sign function, and hence obtain hash codes for visual and textual modalities (i.e., $\mathbf{b}_{v_i} = sgn(\mathbf{h}_{v_i})$ and $\mathbf{b}_{t_i} = sgn(\mathbf{h}_{t_i})$), respectively.

4.3. Bi-direction Relation Reasoning

The goal of cross-modal hashing is to learn the hash function that maps original features into binary hash codes, while preserving the cross-modal similarity. By using the similarity matrix to guide the model training, relations between samples can be well maintained in the Hamming space. However, most of existing supervised methods merely consider the relevant information to build the similarity matrix, ignoring the contribution of irrelevant one, i.e., the categories that samples do not belong to. In fact, multi-level complex relations can be reasoned by fully exploring all available label information, whether belong to or do not belong to.

To this end, we introduce the bi-direction relation reasoning module to construct the similarity matrix. Concretely, to better depict the complex multi-level relations between two similar instances that share at least one category, we adopt the consistent direction reasoning, i.e., the overlapped information is utilized for similarity estimation. Moreover, if two instances do not share any category, we select another direction reasoning to infer their scores. Note that in the similarity matrix, the positive score denotes the similarity degree of two instances in the consistent direction, while the negative one represents the dissimilarity extent in the inconsistent direction.

Consistent Direction: Let $\mathbf{S}_{ij}$ denote the similarity score between instances e_i and e_j, whose label vectors are $\mathbf{y}_i$ and $\mathbf{y}_j$. As analyzed before, categories that they both belong to and do not belong to play the same role when estimating the similarity. In light of this, we resort to the XOR operation to obtain the overlapped label information of two instances. Thereby, the label information is maximumlly utilized to model the similarity relation between two similar instances. Besides, the quotient operation is adopted to constrain the similarity score ranging from 0 to 1. Formally, we summarize the above process as follows,

$$\mathbf{S}_{ij} = \frac{K - (\mathbf{y}_i \oplus \mathbf{y}_j)}{K}, \tag{3}$$

where $\oplus$ is the XOR operation and K is the number of categories. Apparently, the maximum value of $\mathbf{y}_i \oplus \mathbf{y}_j$ is K, and $\mathbf{S}_{ij}$ hence is non-negative, representing the similarity degree between instances e_i and e_j.

Inconsistent Direction: As the irrelevant information plays a critical role in estimating the dissimilarity score, we assume that the more non-overlapped label information two instances have, the more dissimilar they are. Inspired by this, we define the score as follows,

$$\mathbf{S}_{pq} = -\frac{\mathbf{y}_p \oplus \mathbf{y}_q}{K}, \tag{4}$$

where $\mathbf{S}_{pq}$ ranges from -1 to 0, representing the dissimilar extent between instances e_p and e_q.

4.4. Model Training and Optimization

Preserving similarity among instances, i.e., generating similar binary hash codes for semantically similar data, is the major concern of cross-modal hashing [12, 26, 29, 32, 33, 37]. However, as hashing is essentially a discrete learning problem, we thus adopt the inner product between normalized hash representations to model the similarity between two instances in the training phase. Accordingly, we construct three similarity matrices $\mathbf{C}^*$, $\mathbf{C}^v$, and $\mathbf{C}^t$ as follows,

$$\begin{cases} \mathbf{C}_{ij}^{*} = \mathbf{h}_{v_i}^{T}\mathbf{h}_{t_j}, \\ \mathbf{C}_{ij}^{v} = \mathbf{h}_{v_i}^{T}\mathbf{h}_{v_j}, \\ \mathbf{C}_{ij}^{t} = \mathbf{h}_{t_i}^{T}\mathbf{h}_{t_j}, \end{cases} \tag{5}$$

where $\mathbf{C}_{ij}^{*}$ denotes the inter-modal similarity between $\mathbf{v}_i$ and $\mathbf{t}_j$. Likewise, $\mathbf{C}_{ij}^{v}$ and $\mathbf{C}_{ij}^{t}$ respectively refer to the intra-modal similarity between $\mathbf{v}_i$ and $\mathbf{v}_j$, as well as $\mathbf{t}_i$ and $\mathbf{t}_j$. In the training phase, by narrowing the gap between the similarity matrix $\mathbf{S}$ and each of these three matrices, similarity between data can be well maintained in the Hamming space.

4.4.1 Inter-modal Similarity Constraint

To ensure the cross-modal retrieval performance, the most important thing is to eliminate the semantic gap between different modalities. In light of this, the inter-modal similarity matrix $\mathbf{C}^{*}$ should be consistent with the ground truth similarity $\mathbf{S}$. Therefore, to effectively capture correlations across different modalities, we regularize the difference between the ground truth similarity $\mathbf{S}$ and the inter-modal similarity $\mathbf{C}^{*}$, formulating the inter-modal loss Ψ_1 as follows,

$$\Psi_1 = \sum_{i,j=1}^{N} \left\| \mathbf{S}_{ij} - \mathbf{C}_{ij}^{*} \right\|_F^2, \tag{6}$$

where $\mathbf{C}_{ij}^{*}$ denotes the inter-modal similarity score between $\mathbf{v_i}$ and $\mathbf{t_j}$.

4.4.2 Intra-modal Similarity Constraint

Obviously, the similarity retention of unimodal data is the premise of maintaining the cross-modal similarity. Particularly, only when similarities of images and texts themselves are respectively maintained, cross-modal retrieval performance can be guaranteed. Therefore, we define the intra-modal loss Ψ_2 in the same way as follows,

$$\Psi_2 = \sum_{i,j=1}^{N} \left(\left\| \mathbf{S}_{ij} - \mathbf{C}_{ij}^{v} \right\|_F^2 + \left\| \mathbf{S}_{ij} - \mathbf{C}_{ij}^{t} \right\|_F^2 \right), \tag{7}$$

where $\mathbf{C}_{ij}^{v}$ and $\mathbf{C}_{ij}^{t}$ represent intra-modal similarity scores between instance e_i and e_j, regarding the visual modality and textual modality, respectively.

4.4.3 Relaxation Similarity Constraint

Apart from the inter- and intra-modal similarity constraint, we further regularize binarization differences between normalized hash representations and hash codes of two modalities, so as to derive optimal continuous surrogates of hash codes. Specifically, we first normalize hash code vectors of visual and textual modality, and then directly utilize the inner product of normalized hash representation and hash codes to estimate the similarity between them. In this way, we could obtain $1 \times N$ self-supervised similarity matrices for the visual and textual modality, i.e., $\mathbf{C}^{bv}$ and $\mathbf{C}^{bt}$. The details of them are as follows,

$$\begin{cases} \mathbf{C}_{1i}^{bv} = \mathbf{h}_{v_i}^{T}\mathbf{h}_{v_i}^{b}, \\ \mathbf{C}_{1j}^{bt} = \mathbf{h}_{t_j}^{T}\mathbf{h}_{t_j}^{b}, \end{cases} \tag{8}$$

where $\mathbf{h}_{v_i}^{b}$ and $\mathbf{h}_{t_j}^{b}$ separately represent normalized image and text hash codes of instance e_i and e_j, derived via the following formulations,

$$\begin{cases} \mathbf{h}_{v_i}^{b} = \dfrac{\mathbf{b}_{v_i}}{\left\| \mathbf{b}_{v_i} \right\|_2}, \\ \mathbf{h}_{t_j}^{b} = \dfrac{\mathbf{b}_{t_j}}{\left\| \mathbf{b}_{t_j} \right\|_2}, \end{cases} \tag{9}$$

where $\mathbf{b}_{v_i} = sgn(\mathbf{h}_{v_i})$ and $\mathbf{b}_{t_j} = sgn(\mathbf{h}_{t_j})$.

For each image and text, the ground truth similarity between its normalized hash representation and hash codes is 1. Therefore, we define following regularization loss Ψ_3,

$$\Psi_3 = \sum_{i,j=1}^{N} \left(\left\| \mathbf{1} - \mathbf{C}_{1i}^{bv} \right\|_F^2 + \left\| \mathbf{1} - \mathbf{C}_{1j}^{bt} \right\|_F^2 \right), \tag{10}$$

where $\mathbf{1} \in \{1\}^{1\times N}$ is a matrix whose elements are 1.

In conclusion, we devise the objective function Ψ consisting of above three loss components and reach the final objective function as follows,

$$\min_{\mathbf{\Theta}_v, \mathbf{\Theta}_t} \Psi = \alpha\Psi_1 + \beta\Psi_2 + \gamma\Psi_3, \tag{11}$$

where α, β, and γ are balancing parameters, $\mathbf{\Theta}_v$ and $\mathbf{\Theta}_t$ respectively refer to parameters of the image encoder and text encoder. Once the model has been trained, we can directly use f^v and f^t, combining with the element-wise sign function $sgn(\cdot)$, to generate hash codes and fulfill the cross-modal retrieval task. Moreover, it is worth noting that although we assume that both modalities of each instance are observed in the training phase, our scheme can also be easily extended to handle other scenarios, where some training instances miss certain modality.

5. Experiment

5.1. Datasets

For evaluation, we adopted two widely used cross-modal datasets: MIRFLICKR-25K [14] and NUS-WIDE [6], where images are assigned to multiple category labels.

MIRFLICKR-25K. This dataset includes $25,000$ images with the fixed size of $224 \times 224 \times 3$, which are originally collected from the Flickr website[3]. And each image is

[3]http://www.flickr.com/.

manually annotated with several textual tags and at least one of the 24 labels. In our experiments, we merely utilized images that are associated with at least 20 textual tags. Therefore, there are $20,015$ images retained. Afterwards, we split these images into two subsets: query and gallery. Specifically, $2,000$ images are randomly selected as the query subset, and the remaining ones are set as gallery set. To learn the hash function, $10,000$ images are randomly chosen from the gallery subset as training data. Moreover, to reduce too noisy tags, we removed tags that appear below 20 from those retained images. We hence obtained $1,386$ unique tags, constituting a word bag. Based on the BoW strategy, the textual modality of each instance is represented by a $1,386$-d vector.

NUS-WIDE. It is a large-scale social image dataset including $269,648$ images associated with $5,018$ unique tags, where the image size is $224 \times 224 \times 3$. Moreover, each image is manually annotated by a predefined set of 81 labels. In our work, we retained $195,834$ images that are associated with at least one of the 21 most frequent labels. we formed a query set of $2,100$ images, while the training set and gallery set of $10,500$ and $193,734$ images, respectively. And we removed those tags that appear below 20 to construct the word bag and obtained $1,000$ unique tags. In this way, the textual modality of each instance is represented by a $1,000$-d vector.

5.2. Experimental Settings

Evaluation Protocols. In this work, we evaluated our proposed model on two classic cross-modal retrieval tasks: querying the image database with given textual vectors ("Text→Image") and querying the text database with given image examples ("Image→Text"). For each cross-modal retrieval task, we adopted two widely utilized performance metrics, i.e., Hamming ranking and hash lookup, to compare the retrieval performance of our method with other state-of-the-art methods. In particular, mean average precision (MAP) [34], a representative method to measure the accuracy of Hamming ranking, is adopted in our work. Meanwhile, the precision-recall (P-R) curve is utilized to measure the accuracy of hash lookup protocol. Notably, to be consistent with baseline methods, two instances are considered to be similar if and only if they share at least one label in the testing phase.

Baselines. To justify the effectiveness of our proposed Bi_NCMH, we chose five state-of-the-art methods as baselines, including four supervised methods: SCM [36], DCH [34], DCMH [16], and SSAH [17], and one unsupervised one: CCA [13]. As SCM presents two learning models, i.e., orthogonal projection and sequential one, we respectively denoted them by SCM-Or and SCM-Se. Among these baselines, CCA, SCM-Or, SCM-Se, and DCH are shallow learning methods, namely they rely on hand-crafted image features. Meanwhile, we resorted to CNN-F and VGG19 networks as image encoders for deep learning based methods. For fairness, we also separately extracted image features from CNN-F and VGG19 networks that are pre-trained on the Imagenet [8], for shallow learning approaches. Note that the source codes and involved parameters of above baselines are kindly provided by corresponding authors, we tried our best to tune the models and reported their best performance as that in their papers.

Implementation Details. We implemented Bi_NCMH with the open source deep learning software library Tensorflow, and adopted the stochastic gradient descent (SGD) as the optimizer [1]. Besides, we initialized deep networks, i.e., CNN-F and VGG19, with parameters pre-trained on the ImageNet, while other parameters are initialized randomly. To determine hyper-parameters, i.e., α, β and γ, we first performed the grid search in a coarse level within a wide range using an adaptive step size. Once we obtained the approximate scope of each parameter, we then performed the fine tuning within a narrow range using a small step size. In addition, we empirically set the batch-size to 128 and the maximum number of iterations as 500 to ensure the convergence.

5.3. Model Comparison

To justify the effectiveness of our proposed model, we first compared it with baseline methods by setting four different lengths of hash codes (i.e., 16, 32, 64, and 128 bits) and two types of image features. Table 1 and 2 display the performance comparison w.r.t. MAP on two datasets. By jointly analyzing them, we have the following observations. (1) Compared with shallow learning methods, deep learning ones generally achieve better performance. Because they integrate the feature learning and hash function learning into an end-to-end framework, therefore making the learnt features optimally match with hash codes. (2) Among deep learning approaches, our proposed model Bi_NCMH shows consistent improvements over DCMH and SSAH. This is because we built a multi-level semantic matrix via bi-direction relation reasoning, rather than utilizing the binary similarity matrix as supervision information to train the network. Moreover, we adopted the normalization operation to bridge the semantic gap and eliminate the negative effect caused by modality distribution difference. (3) Although the gallery set of the NUS-WIDE dataset is relatively larger than that of the MIRFLICKR-25K dataset, the performance improvement of our proposed model Bi_NCMH is more stable on it. This demonstrates that our method is more suitable for the large-scale cross-modal retrieval.

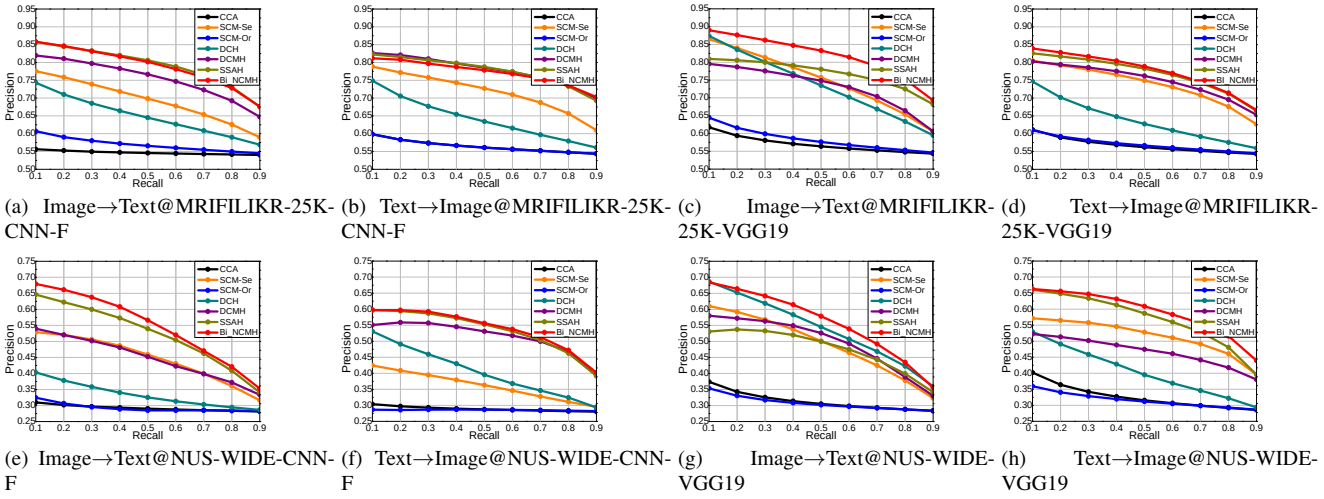
In addition, we further evaluated our method on two datasets using the P-R curve with 64 bits hash code, where CNN-F and VGG19 networks are both utilized. Specifi-

Table 1. The MAP performance comparison between our proposed model and the state-of-the-art baselines on two datasets. The CNN-F features are utilized for shallow learning models, and the best results are highlighted in bold.

Method	MIRFLICKR-25K								NUS-WIDE							
	Image→Text				Text→Image				Image→Text				Text→Image			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CCA	0.553	0.545	0.548	0.547	0.554	0.583	0.549	0.548	0.306	0.299	0.294	0.290	0.301	0.295	0.290	0.287
SCM-Or	0.594	0.580	0.572	0.560	0.605	0.590	0.567	0.555	0.330	0.311	0.300	0.289	0.313	0.298	0.286	0.281
SCM-Se	0.686	0.691	0.691	0.694	0.698	0.727	0.713	0.716	0.428	0.434	0.442	0.449	0.362	0.364	0.362	0.363
DCH	0.638	0.642	0.662	0.669	0.636	0.643	0.659	0.638	0.331	0.330	0.339	0.347	0.397	0.399	0.419	0.424
DCMH	0.730	0.741	0.748	0.726	0.759	0.767	0.775	0.749	0.426	0.413	0.440	0.446	0.477	0.491	0.498	0.524
SSAH	0.767	0.775	0.782	0.772	**0.767**	0.774	0.753	0.739	0.486	0.501	0.512	0.529	0.506	0.520	0.525	0.531
Bi_NCMH	**0.770**	**0.781**	**0.796**	**0.780**	0.760	**0.776**	**0.780**	**0.781**	**0.511**	**0.528**	**0.540**	**0.557**	**0.526**	**0.542**	**0.545**	**0.546**

Table 2. The MAP performance comparison between our proposed model and the state-of-the-art baselines on two datasets. The VGG19 features are utilized for shallow learning models, and the best results are highlighted in bold.

Method	MIRFLICKR-25K								NUS-WIDE							
	Image→Text				Text→Image				Image→Text				Text→Image			
	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
CCA	0.603	0.586	0.574	0.565	0.599	0.583	0.571	0.562	0.351	0.336	0.319	0.308	0.377	0.354	0.332	0.316
SCM-Or	0.621	0.602	0.587	0.569	0.617	0.590	0.573	0.560	0.352	0.329	0.312	0.301	0.380	0.343	0.318	0.304
SCM-Se	0.728	0.741	0.746	0.750	0.710	0.727	0.733	0.737	0.457	0.482	0.486	0.494	0.485	0.510	0.507	0.516
DCH	0.725	0.719	0.756	0.749	0.637	0.632	0.654	0.652	0.496	0.515	0.515	0.574	0.386	0.397	0.412	0.428
DCMH	0.690	0.697	0.742	0.735	0.722	0.722	0.744	0.745	0.453	0.465	0.488	0.501	0.442	0.473	0.464	0.486
SSAH	0.790	0.799	0.754	0.748	**0.760**	**0.771**	0.759	0.750	0.503	0.501	0.468	0.378	0.543	0.560	0.560	0.522
Bi_NCMH	**0.790**	**0.800**	**0.808**	**0.881**	0.750	0.760	**0.765**	**0.771**	**0.515**	**0.526**	**0.555**	**0.562**	**0.550**	**0.573**	**0.574**	**0.583**



(a) Image→Text@MRIFILIKR-25K-CNN-F (b) Text→Image@MRIFILIKR-25K-CNN-F (c) Image→Text@MRIFILIKR-25K-VGG19 (d) Text→Image@MRIFILIKR-25K-VGG19

(e) Image→Text@NUS-WIDE-CNN-F (f) Text→Image@NUS-WIDE-CNN-F (g) Image→Text@NUS-WIDE-VGG19 (h) Text→Image@NUS-WIDE-VGG19

Figure 3. The P-R curves of different methods on two datasets, where CNN-F and VGG19 networks are utilized and the hash code length is 64 bits.

cally, we calculated the precision of returned retrieval results given different recall rate, ranging from 0.1 to 0.9 with a step size of 0.1. From Fig. 3, we can easily find that the performance are consistent with those in Table 1 and 2. Our proposed model Bi_NCMH generally surpasses all baselines and always obtains the highest precision for the specific recall rate, which justifies the validity of our proposed model from another perspective.

6. Conclusion And Future Work

In this paper, we present a novel deep normalized cross-modal hashing approach with bi-direction relation reasoning. Specifically, we explore the irrelevant information, i.e., the categories that samples do not belong to, and build the multi-level semantic similarity matrix by considering the consistent and inconsistent directions, separately. To bridge the modality gap and eliminate the negative effect caused by modality distribution difference, we devise a normalization operation on hash representations so as to better represent similarity relations among instances. Moreover, we integrate three loss indicators, named inter-modal, intra-modal and regularization loss, to respectively constrain the inter-modal, intra-modal and relaxation similarity. Extensive experiments have been conducted on two datasets and the results demonstrate the effectiveness of the proposed scheme.

In this work, we assume that each label is independent

when modeling the bi-direction relation. However, it is possible that there are semantic associations among semantics of labels, which may influence the estimation of similarity. In the future, we plan to explore such potential correlations among labels, therefore constructing the relation among samples more precisely. Moreover, we will design new objective functions from multiple perspectives to optimize the training process of the model.

Acknowledgements. This research was partially supported by NSF SCH-2123521, NeTS-2109982 and the gift donation from Cisco. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

References

[1] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(3):185–196, 1993. 6

[2] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. Cross-modal hamming hashing. In *Proceedings of the European Conference on Computer Vision*, pages 207–223, 2018. 1

[3] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *Computer Research Repository*, abs/1405.3531:1–11, 2014. 3

[4] Zhen-Duo Chen, Yongxin Wang, Hui-Qiong Li, Xin Luo, Liqiang Nie, and Xin-Shun Xu. A two-step cross-modal hashing by exploiting label correlations and preserving similarity in both steps. In *Proceedings of the ACM International Conference on Multimedia*, pages 1694–1702, 2019. 1

[5] Zhikui Chen, Fangming Zhong, Geyong Min, Yonglin Leng, and Yiming Ying. Supervised intra- and inter-modality similarity preserving hashing for cross-modal retrieval. *IEEE Access*, 6:27796–27808, 2018. 1

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 48–48, 2009. 5

[7] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[9] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2083–2090, 2014. 1, 2

[10] Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Transactions on Image Processing*, 25(11):5427–5440, 2016. 1, 2

[11] Fei Dong, Xiushan Nie, Xingbo Liu, Leilei Geng, and Qian Wang. Cross-modal hashing based on category structure preserving. *Journal of Visual Communication and Image Representation*, 57:28–33, 2018. 1

[12] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the ACM International Conference on Multimedia*, pages 7–16, 2014. 4

[13] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 817–824, 2011. 2, 6

[14] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the ACM SIGMM International Conference on Multimedia Information Retrieval*, pages 39–43, 2008. 5

[15] Go Irie, Hiroyuki Arai, and Yukinobu Taniguchi. Alternating co-quantization for cross-modal hashing. In *Proceedings of the Fast-Gaussian SIFT for Fast and Accurate Feature Extraction IEEE International Conference on Computer Vision*, pages 1886–1894, 2015. 1, 2

[16] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3278, 2017. 1, 2, 3, 4, 6

[17] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4242–4251, 2018. 1, 2, 4, 6

[18] Zijia Lin, Guiguang Ding, Jungong Han, and Jianmin Wang. Cross-view retrieval via probability-based semantics-preserving hashing. *IEEE Transactions on Cybernetics*, 47(12):4342–4355, 2017. 2, 3

[19] Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baochang Zhang. Cross-modality binary code learning via fusion similarity hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6345–6353, 2017. 2

[20] Xin Liu, An Li, Ji-Xiang Du, Shu-Juan Peng, and Wentao Fan. Efficient cross-modal retrieval via flexible supervised collective matrix factorization hashing. *Multimedia Tools and Applications*, 77(21):28665–28683, 2018. 1

[21] Xu Lu, Lei Zhu, Zhiyong Cheng, Xuemeng Song, and Huaxiang Zhang. Efficient discrete latent semantic hashing for scalable cross-modal retrieval. *Signal Processing*, 154:217–231, 2019. 1

[22] Lei Ma, Hongliang Li, Fanman Meng, Qingbo Wu, and King Ngi Ngan. Global and local semantics-preserving based deep hashing for cross-modal retrieval. *Neurocomputing*, 312:49–62, 2018. 1

[23] Devraj Mandal, Kunal N. Chaudhury, and Soma Biswas. Generalized semantic preserving hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(1):102–112, 2019. 1

[24] Jonathan Masci, Michael M. Bronstein, Alexander M. Bronstein, and Jürgen Schmidhuber. Multimodal similarity-

preserving hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):824–830, 2014. 1

[25] Sean Moran and Victor Lavrenko. Regularised cross-modal hashing. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 907–910, 2015. 1

[26] Dimitrios Rafailidis and Fabio Crestani. Cluster-based joint matrix factorization hashing for cross-modal retrieval. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 781–784, 2016. 2, 4

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, pages 1–14, 2015. 3

[28] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the International Conference on Management of Data*, pages 785–796, 2013. 1, 2

[29] Changchang Sun, Xuemeng Song, Fuli Feng, Wayne Xin Zhao, Hao Zhang, and Liqiang Nie. Supervised hierarchical cross-modal hashing. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 725–734, 2019. 1, 3, 4

[30] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, and Qing Zhang. LBMCH: learning bridging mapping for cross-modal hashing. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 999–1002, 2015. 1

[31] Lin Wu, Yang Wang, and Ling Shao. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(4):1602–1612, 2019. 1

[32] Yiling Wu, Shuhui Wang, and Qingming Huang. Learning semantic structure-preserved embeddings for cross-modal retrieval. In *Proceedings of the ACM Multimedia Conference on Multimedia Conference*, pages 825–833, 2018. 4

[33] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, Li He, and Jingkuan Song. Cross-modal retrieval with label completion. In *Proceedings of the ACM Conference on Multimedia Conference*, pages 302–306, 2016. 4

[34] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, 26(5):2494–2507, 2017. 3, 6

[35] Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, and Yueting Zhuang. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 395–404, 2014. 1

[36] Dongqing Zhang and Wu-Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2177–2183, 2014. 1, 2, 6

[37] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian. Pl-ranking: A novel ranking method for cross-modal retrieval. In *Proceedings of the ACM Conference on Multimedia Conference*, pages 1355–1364, 2016. 4

[38] Lei Zhang, Yongdong Zhang, Richang Hong, and Qi Tian. Full-space local topology extraction for cross-modal retrieval. *IEEE Transactions on Image Processing*, 24(7):2212–2224, 2015. 1

[39] Pengfei Zhang, Chuan-Xiang Li, Meng-Yuan Liu, Liqiang Nie, and Xin-Shun Xu. Semi-relaxation supervised hashing for cross-modal retrieval. In *Proceedings of the ACM on Multimedia Conference*, pages 1762–1770, 2017. 2

[40] Peichao Zhang, Wei Zhang, Wu-Jun Li, and Minyi Guo. Supervised hashing with latent factor models. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 173–182, 2014. 2

[41] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 1

[42] Jile Zhou, Guiguang Ding, and Yuchen Guo. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 415–424, 2014. 1, 2

[43] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the Fast-Gaussian SIFT for Fast and Accurate Feature Extraction ACM Multimedia Conference*, pages 143–152, 2013. 1