Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Pairwise Relationship Guided Deep Hashing for Cross-Modal Retrieval

Erkun Yang,[1] **Cheng Deng,**[1] **Wei Liu,**[2] **Xianglong Liu,**[3] **Dacheng Tao,**[4] **Xinbo Gao**[1]
[1] School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2] Tencent AI Lab, Shenzhen, China
[3] Beihang University, Beijing 100191, China
[4] Centre for Artificial Intelligence, University of Technology Sydney, NSW 2007, Australia
ekyang@stu.xidian.edu.cn, {chdeng, xbgao}@mail.xidian.edu.cn, wliu@ee.columbia.edu,
xlliu@nlsde.buaa.edu.cn, dacheng.tao@uts.edu.au

Abstract

With benefits of low storage cost and fast query speed, cross-modal hashing has received considerable attention recently. However, almost all existing methods on cross-modal hashing cannot obtain powerful hash codes due to directly utilizing hand-crafted features or ignoring heterogeneous correlations across different modalities, which will greatly degrade the retrieval performance. In this paper, we propose a novel deep cross-modal hashing method to generate compact hash codes through an end-to-end deep learning architecture, which can effectively capture the intrinsic relationships between various modalities. Our architecture integrates different types of pairwise constraints to encourage the similarities of the hash codes from an intra-modal view and an inter-modal view, respectively. Moreover, additional decorrelation constraints are introduced to this architecture, thus enhancing the discriminative ability of each hash bit. Extensive experiments show that our proposed method yields state-of-the-art results on two cross-modal retrieval datasets.

Introduction

With the fast development of information retrieval techniques and the popularity of social media in the past decades, there exists a tremendous amount of multimodal data being generated on the Internet everyday, such as texts, images, and videos. To take advantage of such massive yet heterogeneous data, a great deal of effort has been invested in approximate nearest neighbor (ANN) search across different modalities (Liu, He, and Lang 2013; Liu et al. 2015; Deng et al. 2013; 2016). Since data from different modalities may have strong semantic correlations, it is essential to support cross-modal retrieval (Song et al. 2013; Zhang and Li 2014; Lin et al. 2015; Yang et al. 2012; 2009) that returns relevant results of one modality when querying another modality, e.g., retrieving images with textual queries. Considering massive volumes and high dimensions of multimodal data, traditional methods designed for single-modal data are not suitable for the cross-modality retrieval scenario. To address this issue, cross-modal retrieval methods relying on hashing techniques have recently attracted much attention in the ANN research community, which compress high-dimensional data instances into compact binary codes with similar binary codes produced for similar data samples. However, due to the heterogeneity across different modalities and the semantic gap between low-level features and high-level semantics, developing effective and efficient cross-modal hashing methods remains a challenge problem.

To date, most of existing Cross-Modal Hashing (CMH) methods focus on embedding instances from different modalities into a unified Hamming code space to conduct search (Ding, Guo, and Zhou 2014; Zhang, Wang, and Si 2011; Kumar and Udupa 2011; Wang et al. 2015; Irie, Arai, and Taniguchi 2015; Liu et al. 2014). Specifically, these methods use shallow architectures to project high-dimensional features into a low-dimensional space, and then generate compact hash codes. A common problem of these CMH methods using shallow architectures is their incapability of capturing heterogeneous correlations effectively to bridge different modalities, so these methods cannot achieve satisfactory search quality. Several recent deep models for multimodal embedding (Frome et al. 2013; Kiros, Salakhutdinov, and Zemel 2014; Long et al. 2015; Karpathy and Fei-Fei 2015; Donahue et al. 2015; Gao et al. 2015; Andreas et al. 2015) have proven that deep learning can discover heterogeneous correlations across different modalities more effectively than shallow learning methods. As a representative work, Deep Cross-Modal Hashing (DCMH) (Jiang and Li 2016) extends traditional deep models for cross-modal retrieval, but it can only capture intra-modal information and ignores inter-modal correlations, which makes the retrieved results suboptimal. Deep Visual-Semantic Hashing (DVSH) (Cao et al. 2016) utilizes Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) to separately learn unified binary codes for each modality. However, the textual modality in DVSH is constrained to sentences or other sequence texts, which is too strict in practice and greatly limits its applications.

In this paper, we propose a novel Pairwise Relationship Guided Deep Hashing (PRDH) method for cross-modal retrieval, which adopts deep CNN models to simultaneously learn feature representations and hash codes for each modality seamlessly in an end-to-end architecture. In this deep architecture, we integrate two types of pairwise la-

bel constraints to guide the hash code learning for intra-modality and inter-modality, respectively. Thus, the learned hash codes can better reflect the intrinsic cross-modal correlations. Moreover, we integrate decorrelation constraints into the unified deep architecture to improve the discriminative ability of each hash bit.

The main contributions of this work can be summarized as follows:

- The proposed method works under an end-to-end hashing mechanism, which essentially integrates feature learning and hash code learning into a unified deep learning architecture.
- The proposed method exploits different pairwise constraints to enforce the hash codes from intra-modality and inter-modality, which can effectively discover the heterogeneous correlations across different modalities and greatly preserve the semantic similarities among them.
- The experimental results on real datasets highlight the advantages of our method and demonstrate that the proposed PRDH method outperforms several state-of-the-art approaches.

Related Work

In recent years, various hashing methods have been proposed for CMH, which can be roughly classified into unsupervised methods (Ding, Guo, and Zhou 2014; Song et al. 2013; Zhou, Ding, and Guo 2014) and supervised methods (Bronstein et al. 2010; Kumar and Udupa 2011). For a comprehensive survey, we refer the readers to (Wang et al. 2016).

Unsupervised hashing methods generally learn projection functions from original feature space to Hamming space. Inter-media hashing (IMH) (Song et al. 2013) is reported to explore intra-view and inter-view consistency, and the hash functions are learned with the help of a linear regression model. Furthermore, collective matrix factorization hashing (CMFH) (Ding, Guo, and Zhou 2014) employs collective matrix factorization to learn two view-specific hash functions and then projects multi-source data into unified hash codes. Besides, latent semantic sparse hashing (LSSH) (Zhou, Ding, and Guo 2014) uses sparse coding to learn the latent semantic representation for each modality, and then embeds these learned features into a joint space to obtain the unified hash codes.

Supervised hashing methods can explore the semantic information to enhance the data correlation from different modalities and reduce the semantic gap. Hence supervised methods usually achieve superior performance compared with the unsupervised ones. CMSSH (Bronstein et al. 2010) adopts Adaboost scheme, and optimizes each bit by minimizing a weighted distance between semantic similarity and the dot product of learned hash codes iteratively. CVH (Kumar and Udupa 2011) intends to minimize a similarity-weighted cumulative Hamming distance between pairwise data to learn a low-dimensional linear embedding function.

Most of the previous hashing methods based on shallow architectures cannot describe the complicated nonlinear correlations among different modalities. Latest deep models for multimodal embedding show that deep architectures can better capture the heterogeneous correlations for image captioning and cross-modal reasoning. Inspirited by this idea, we develop a hybrid deep architecture integrating composite pairwise label constraints and a decorrelation constraint to discover the semantic correlations and enhance the discriminative ability of each hash bit.

Notations and Problem Definition

Notations

In this paper, calligraphic uppercase letters, such as $\mathcal{X}$, are used to denote sets; bold face uppercase letters, such as $\mathbf{A}$, represents matrices; bold face lowercase letters, such as $\mathbf{b}$, are vectors. Moreover, X_{ij} denotes the (i,j)th element of $\mathbf{X}$, and the ith row of $\mathbf{X}$ is defined as $\mathbf{X}_{*i}$. We use $\mathbf{1}$ to denote a vector with all elements being 1. $tr(\cdot)$ and $\|\cdot\|_F$ denote the trace and the Frobenius norm of a matrix, respectively.

Problem Definition

Assuming that $\mathcal{O}= \{o_i\}_{i=1}^N$ is the training set containing N instances. $\mathbf{X}$ and $\mathbf{Y}$ correspond to two modalities, such as image and text. $\mathbf{S}_{N\times N}$ is a similarity matrix of training data. $S_{ij} = 1$ if o_i and o_j are similar, and $S_{ij} = 0$ otherwise.

When given the training data and its similarity matrix $\mathbf{S}$, the proposed method learns two modality-specific hash functions, i.e., $h^x(\cdot)$ for image and $h^y(\cdot)$ for text. The learned hash functions can be utilized to generate c-bit hash codes for query and database instances in both modalities.

Pairwise Relationship Guided Deep Cross-Modal Hashing

The deep architecture of the proposed PRDH model is illustrated as Figure 1.

Deep Architecture

We apply two deep neural networks to extract features for two modalities, respectively.

For the image modality, we apply the VGG-F (Chatfield et al. 2014) network due to its excellent performance on object classification. The original VGG-F model consists of five convolutional layers $(conv1 - conv5)$ and three fully-connected layers $(fc6 - fc8)$. We replace the $fc8$ layer with a new fch hash layer with c hidden nodes, which embeds the learned deep features into a low-dimensional Hamming space. For the textual modality, we adopt the multilayer perceptrons (MLP) to comprise three fully connected layers. Similar to the image modality, we also replace the last layer with a new fch hash layer with c hidden nodes. One carefully designed objective function based on the pairwise label constraints is used to combine the hash code learning procedure across different modalities.

Hash Code Learning

For efficient nearest neighbor search, assuming that two instances o_i and o_j are semantically similar, their corresponding hash codes should also be similar in Hamming space, and vice versa. To better preserve the semantic similarities

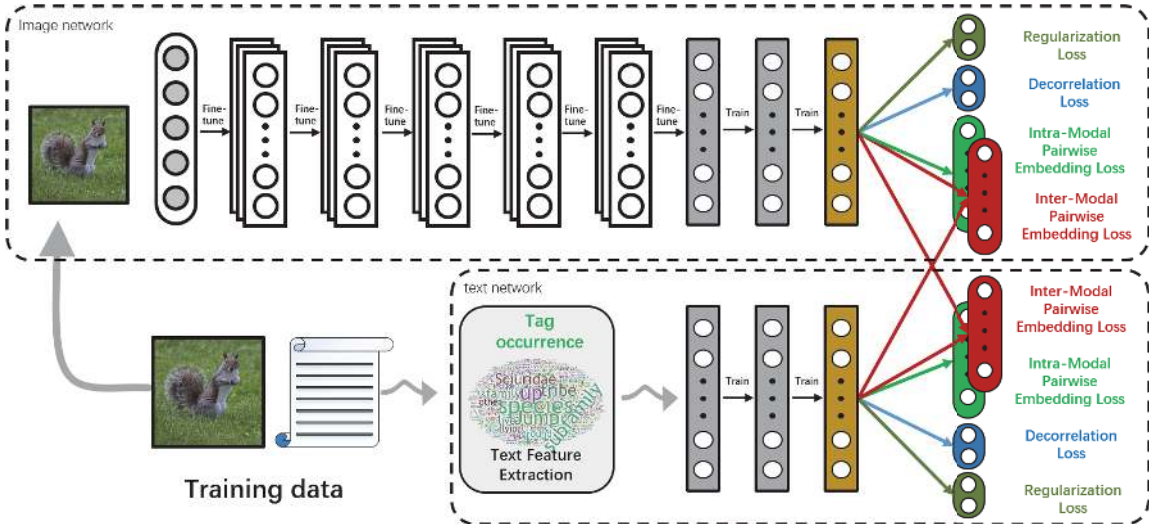

Figure 1: Framework of the proposed PRDH.

of training instances, our objective function comprises four parts: (1) inter-modal pairwise embedding loss; (2) intra-modal pairwise embedding loss; (3) decorrelation loss; and (4) regularization loss.

Pairwise embedding loss enhances the correlation between a pair of semantic similar instances and reduces the similarity between semantic dissimilar instances. Concretely, we use the negative log likelihood of similarity between pairwise points to measure such relationship.

Since our goal is to conduct efficient cross-modal retrieval, it is intuitive to add inter-modal pairwise embedding loss. Based on hash codes of image modality output from the image neural network $\mathbf{U}^x = \{U_i^x\}_{i=1}^N$, hash codes of textual modality output from the textual neural network $\mathbf{U}^y = \{U_i^y\}_{i=1}^N$, and the pairwise labels $\mathbf{S} = \{S_{ij}\}$, the likelihood function is defined as:

$$p(S_{ij}|\mathbf{U}_{*i}^x, \mathbf{U}_{*j}^y) = \begin{cases} \sigma(\Omega_{ij}^{xy}) & S_{ij} = 1 \\ 1 - \sigma(\Omega_{ij}^{xy}) & S_{ij} = 0, \end{cases} \tag{1}$$

where $\Omega_{ij}^{xy} = \frac{1}{2}\mathbf{U}_{*i}^{x\top}\mathbf{U}_{*j}^y$, and $\sigma(\Omega_{ij}^{xy}) = \frac{1}{1+e^{-\Omega_{ij}^{xy}}}$. $\mathbf{U}_{*i}^x = f^x(x_i, \theta_x)$, $\mathbf{U}_{*j}^y = f^y(y_j, \theta_y)$. Hence, the inter-modal pairwise embedding loss is formulated as:

$$\begin{aligned} \mathcal{J}_1 &= -\log p(\mathbf{S}|\mathbf{U}^{xy}) = -\sum_{S_{ij}\in\mathbf{S}} \log p(S_{ij}|\mathbf{U}^{xy}) \\ &= -\sum_{S_{ij}\in\mathbf{S}} \left(S_{ij}\Omega_{ij}^{xy} - \log(1 + e^{\Omega_{ij}^{xy}})\right). \end{aligned} \tag{2}$$

It is easy to find that optimizing the above loss will reduce the Hamming distance between two similar instances, and enlarge the Hamming distance between two dissimilar instances. Therefore, we can preserve the semantic similarities of instances from different modalities.

In CMH, good hash codes from different modalities should preserve semantic similarity efficiently. Moreover, they should also have good discriminative abilities in their own modality intrinsically to preserve semantic information. On the other hand, effective hash codes in each modality are beneficial to improve the performance of cross-modal retrieval. So, it is necessary to add the intra-modal pairwise embedding loss for image modality and textual modality, respectively.

For image modality, according to the output of the image neural network $\mathbf{U}^x = \{U_i^x\}_{i=1}^N$ and the pairwise labels $\mathbf{S} = \{S_{ij}\}$, the pairwise embedding loss is formulated as follows:

$$\begin{aligned} \mathcal{J}_2 &= -\log p(\mathbf{S}|\mathbf{U}^x) = -\sum_{S_{ij}\in\mathbf{S}} \log p(S_{ij}|\mathbf{U}^x) \\ &= -\sum_{S_{ij}\in\mathbf{S}} \left(S_{ij}\Omega_{ij}^x - \log(1 + e^{\Omega_{ij}^x})\right), \end{aligned} \tag{3}$$

where $\Omega_{ij}^x = \frac{1}{2}\mathbf{U}_{*i}^{x\top}\mathbf{U}_{*j}^x$.

Analogously, pairwise embedding loss for textual modality is formulated as follows:

$$\mathcal{J}_3 = -\sum_{S_{ij}\in\mathbf{S}} \left(S_{ij}\Omega_{ij}^y - \log(1 + e^{\Omega_{ij}^y})\right), \tag{4}$$

where $\Omega_{ij}^y = \frac{1}{2}\mathbf{U}_{*i}^{y\top}\mathbf{U}_{*j}^y$.

Note that if some different hash bits have high correlation, for example, if $\mathbf{U}_{*i}^x$ and $\mathbf{U}_{*j}^x$ vary together for all instances, there will be redundant information between these two hash bits (Cogswell et al. 2015). To maximize information provided by each bit, we add decorrelation constraints for both modalities to reduce the correlations between different bits:

$$\begin{aligned} \mathcal{J}_4 = &\frac{1}{2}(\| \mathbf{C}^x \|_F^2 - \| diag(\mathbf{C}^x) \|_F^2) \\ &+ \frac{1}{2}(\| \mathbf{C}^y \|_F^2 - \| diag(\mathbf{C}^y) \|_F^2), \end{aligned} \tag{5}$$

where $\mathbf{C}^x = \frac{1}{T}\sum_{n=1}^{T}(U_{in}^x - \mu_i)(U_{jn}^x - \mu_j)$ is the covariance matrix of hash bit i and hash bit j over the batch from image

modality, $i, j \in \{1, 2...c\}$, $\mu_i = \frac{1}{T}\sum_{n=1}^{T} U_{in}^x$ is the instance mean of feature i over the batch, and T is the batch size. $\mathbf{C}^y$ is the covariance matrix for textual modality, which is similar to the matrix for image modality. To enable back propagation in the networks, we relax the elements of $\mathbf{U}^x$ and $\mathbf{U}^y$ to be continuous values for both modalities.

To understand the effect of decorrelation constraints further, we consider the gradient of this part w.r.t. a particular hash bit a for a particular instance m in image modality:

$$\begin{aligned}\frac{\partial \mathcal{J}_4}{\partial U_{am}^x} =& \frac{1}{T}\sum_{j\neq a}[\frac{1}{T}\sum_{n=1}^{T}(U_{an}^x - \mu_a)(U_{jn}^x - \mu_j)] \\ & \cdot (U_{jm}^x - \mu_j).\end{aligned} \tag{6}$$

When denoting the rightmost term in Eq. (6) by $I^x(j,m) = (U_{jm}^x - \mu_j)$, and noticing that the term on the left in the gradient expression is simply the covariance between hash bit a and hash bit j, the gradient can be rewritten as:

$$\frac{\partial \mathcal{J}_4}{\partial U_{am}^x} = \frac{1}{N}\sum_{j\neq a} C_{aj}^x \cdot I^x(j,m). \tag{7}$$

Intuitively, we can consider $I^x(j,m)$ as a weight for the element of covariance matrix. A large $I^x(j,m)$ means jth hash bit is important for mth instance. If j is correlated with another hash bit a, the gradient w.r.t. hash bit a will be large. Optimizing this part by Stochastic Gradient Descent (SGD), the activation of hash bit a will be suppressed. Thus, we can learn more representative hash codes.

We also add a regularization term to reduce the quantization loss and keep the learned hash codes balanced:

$$\begin{aligned}R =& \parallel \mathbf{B} - \mathbf{U}^x \parallel_F^2 + \parallel \mathbf{B} - \mathbf{U}^y \parallel_F^2 \\ & + \parallel \mathbf{U}^x \cdot \mathbf{1} \parallel_F^2 + \parallel \mathbf{U}^y \cdot \mathbf{1} \parallel_F^2.\end{aligned} \tag{8}$$

where $\mathbf{B}$ is the unified hash codes from both modalities.

The overall objective function, combining the pairwise embedding loss in each modalities $\mathcal{J}_1$, $\mathcal{J}_2$ and across different modalities $\mathcal{J}_3$, the decorrelation loss $\mathcal{J}_4$, and the regularization term R together, is written as below:

$$\begin{aligned}&\mathcal{J} = (\mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3) + \lambda\mathcal{J}_4 + \gamma R \\ &s.t.\ B \in \{-1, +1\}^{c\times N},\end{aligned} \tag{9}$$

where λ and γ are tradeoff parameters to control the weight of each part.

Optimization Algorithm

The first seven layers of the CNN module for image modality are fined-tuned from the VGG-F model, the new fch layer and the multilayer perceptrons from textual modality are jointly trained with mini-batch SGD method.

The optimization problem in Eq. (9) can be solved by using an alternating learning strategy. One parameter is optimized with others fixed each time. The model is updated by the following steps iteratively until convergency or the preset maximum number of iterations is reached. We summarize the whole alternating learning procedure in Algorithm 1.

1. Fix θ_x and θ_y, optimize B.

Since θ_x and θ_y are fixed, the objective function can be reformulated as follows:

$$\begin{aligned}&\max_{\mathbf{B}} tr(\mathbf{B}^\top(\gamma(\mathbf{U}^x + \mathbf{U}^y))) = tr(\mathbf{B}^\top\mathbf{V}) = \sum\nolimits_{ij} B_{ij}V_{ij} \\ &s.t.\ \mathbf{B} \in \{-1, +1\}^{c\times N},\end{aligned} \tag{10}$$

where $\mathbf{V} = \gamma(\mathbf{U}^x + \mathbf{U}^y)$. We can derive that the optimized B_{ij} should have the same sign as V_{ij}:

$$\mathbf{B} = sign(\mathbf{V}) = sign(\gamma(\mathbf{U}^x + \mathbf{U}^y)). \tag{11}$$

2. Fix θ_y and $\mathbf{B}$, optimize θ_x.

For each instance output U_i^x, we can derive the gradient of the loss w.r.t. the CNN output as:

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial \mathbf{U}_{*i}^x} =& \frac{1}{2}\sum_{j=1}^{N}\left(\sigma(\Omega_{ij}^{xy})\mathbf{U}_{*i}^y - S_{ij}\mathbf{U}_{*i}^y\right) \\ & + \sum_{j=1}^{N}\left(\sigma(\Omega_{ij}^{x})\mathbf{U}_{*i}^x - S_{ij}\mathbf{U}_{*i}^x\right) \\ & + \lambda\frac{1}{N}\sum_{j\neq a}\mathbf{C}_{*j}^x I^x(j,i) \\ & + 2\gamma(\mathbf{U}_{*i}^x - \mathbf{U}_{*i}^y + \mathbf{F}\cdot\mathbf{1}).\end{aligned} \tag{12}$$

3. Fix θ_x and $\mathbf{B}$, optimize θ_y.

We also use SGD to optimize the neural network parameter θ_y of the textual modality. For each sampled point U_i^y, we can derive the gradient of the loss w.r.t. the CNN output as:

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial \mathbf{U}_{*i}^y} =& \frac{1}{2}\sum_{j=1}^{N}\left(\sigma(\Omega_{ij}^{xy})\mathbf{U}_{*i}^x - S_{ij}\mathbf{U}_{*i}^x\right) \\ & + \sum_{j=1}^{N}\left(\sigma(\Omega_{ij}^{y})\mathbf{U}_{*i}^y - S_{ij}\mathbf{U}_{*i}^y\right) \\ & + \lambda\frac{1}{N}\sum_{j\neq a}\mathbf{C}_{*j}^y I^y(j,i) \\ & + 2\gamma(\mathbf{U}_{*i}^x - \mathbf{U}_{*i}^y + \mathbf{F}\cdot\mathbf{1}).\end{aligned} \tag{13}$$

It should be noticed that the generating manner of pairwise instances in our proposed model is different from the traditional methods. Concretely, assuming that the batch size is T and the dataset size is N, traditional methods use instances from the same batch in one iteration and can only generate up to $\frac{T(T-1)}{2}$ pairwise data points, and the information from the rest of the training dataset will be neglected. To take advantage of the whole training dataset, we store the result of the training dataset in a matrix, and combine instances from the mini-batch and instances from the training dataset to form pairwise data points, by which, up to $(TN - \frac{T(T+1)}{2})$ pairs can be generated. Since $N \gg T$, more pairwise instances will be generated in one iteration with the same batch size. Hence the optimization will be more effective and robust to noise and outliers. The used matrix will be updated in every iteration.

Algorithm 1: The learning algorithm for PRDH

Input: Image set $\mathbf{X}$, text set $\mathbf{Y}$, and pairwise similarity matrix $\mathbf{S}$, parameters λ, μ and γ, bit length c.

Output: Hash codes $\mathbf{B}$, parameters θ_x and θ_y of the deep neural networks for both modalities.

Procedure:

Initialize network parameters θ_X and θ_Y, mini-batch size $N_x = N_y = 128$, and iteration number $t_x = \frac{N}{N_x}$, $t_y = \frac{N}{N_y}$.

repeat

  Update $\mathbf{B}$ with to Eq. (11);

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

    Randomly sample N_x instances from $\mathbf{X}$;

    Calculate the outputs and update the matrix $\mathbf{U^x}$ by $\mathbf{U^x_{*i}} = f(x_i; \theta_x)$;

    Back propagate the neural network according to Eq. (12) and update θ_x.

  end

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

    Randomly sample N_y instances from $\mathbf{Y}$;

    Calculate the outputs and update the matrix $\mathbf{U^y}$ by $\mathbf{U^y_{*i}} = f(y_i; \theta_y)$;

    Back propagate the neural network according to Eq. (13) and update θ_y.

  end

until *a fixed number of iterations*;

Out-of-Sample Extension

For a new point that is not in the training set, its hash code can be generated as long as one of its modalities is observed. In particular, given one instance p when only the image modality is available, we treat it as the input of the image network, and forward propagate the network to generate hash codes as follows:

$$b_p^x = sign(h^x(x_p; \theta_x)).$$

Similarly, if only the textual modality features is observed, the hash code can also be generated as follows:

$$b_p^y = sign(h^y(y_p; \theta_y)),$$

where $sign(\cdot)$ is an element-wise sign function.

Experiments

To evaluate the performance of the proposed PRDH method, extensive experiments are implemented on two popular datasets. In the following part, we first introduce these two datasets used in our experiments, and then carefully discuss the parameter setting. After that, we compare the experimental results of our method with several state-of-the-art approaches.

Datasets

MIRFlickr (Huiskes and Lew 2008): It originally consists of 25,000 instances, each with an image and its associated tags. Every instance belongs to at least one of the 24 provided labels. In our experiment, those textual tags appearing less than 20 times are removed, and we only keep the instances with textual tags and labels. Subsequently, we get 20015 instances. For each instance, the textual modality is represented as a 1386-dimensional bag-of-words vector. For traditional methods based on shallow architectures, a 512-dimensional SIFT feature vector is used as its image modality representation. For DCMH and the proposed deep hash-

Table 1: Comparison with baselines on MIRFlickr in terms of MAP. The best accuracy is shown in boldface.

Task	method	Code Length		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CCA	0.5634	0.5630	0.5626
	CMFH	0.5804	0.5790	0.5797
	SCM	0.6153	0.6279	0.6288
	LSSH	0.5784	0.5804	0.5797
	STMH	0.5876	0.5951	0.5942
	CVH	0.6067	0.6177	0.6157
	SePH	0.6441	0.6492	0.6508
	DCMH	0.7056	0.7035	0.7140
	PRDH	**0.7126**	**0.7128**	**0.7201**
Text Query v.s. Image Database	CCA	0.5639	0.5631	0.5627
	CMFH	0.5782	0.5778	0.5779
	SCM	0.6102	0.6184	0.6192
	LSSH	0.5898	0.5927	0.5932
	STMH	0.5763	0.5877	0.5826
	CVH	0.6026	0.6041	0.6017
	SePH	0.6455	0.6474	0.6506
	DCMH	0.7311	0.7487	0.7499
	PRDH	**0.7467**	**0.7540**	**0.7505**

Table 2: Comparison with baselines on NUS-WIDE in terms of MAP. The best accuracy is shown in boldface.

Task	method	Code Length		
		16 bits	32 bits	64 bits
Image Query v.s. Text Database	CCA	0.3742	0.3667	0.3617
	CMFH	0.3825	0.3858	0.3890
	SCM	0.4904	0.4945	0.4992
	LSSH	0.3900	0.3924	0.3962
	STMH	0.4344	0.4461	0.4534
	CVH	0.3687	0.4182	0.4602
	SePH	0.5314	0.5340	0.5429
	DCMH	0.6141	0.6167	0.6427
	PRDH	**0.6348**	**0.6529**	**0.6506**
Text Query v.s. Image Database	CCA	0.3731	0.3661	0.3613
	CMFH	0.3915	0.3944	0.3990
	SCM	0.4595	0.4650	0.4691
	LSSH	0.4286	0.4248	0.4248
	STMH	0.3845	0.4089	0.4181
	CVH	0.3646	0.4024	0.4339
	SePH	0.5086	0.5055	0.5710
	DCMH	0.6591	0.6487	0.6847
	PRDH	**0.6808**	**0.6961**	**0.6943**

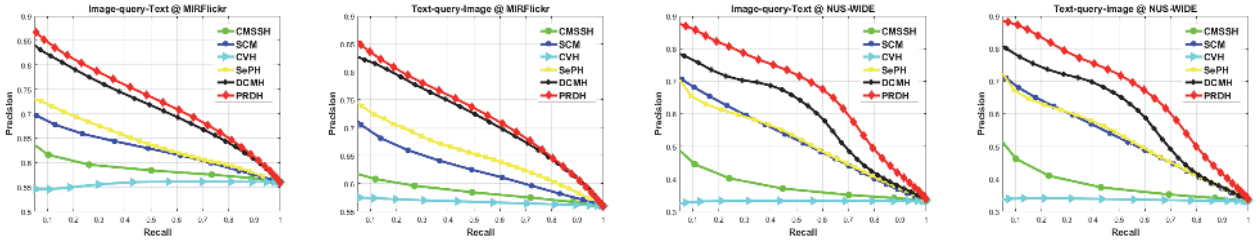

Figure 2: Precision-recall curves (the code length is 32).

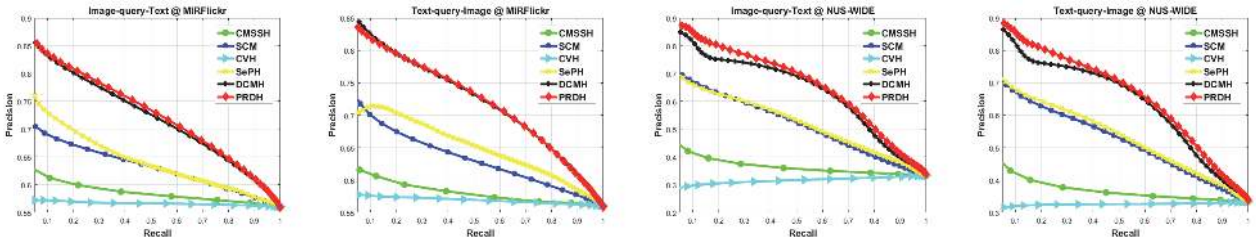

Figure 3: Precision-recall curves (the code length is 64).

ing method, we directly use raw pixels as the image modality inputs. Instances are considered to be similar if they share at least one common label, otherwise, they are dissimilar.

NUS-WIDE (Chua et al. 2009): It is a real-world web image dataset containing 269648 instances, each including an image and its associated textual tags. Instances are annotated with no less than one label from provided 81 concept labels. The most frequent 10 concepts are kept, hence we get 186577 text-image pairs. For each instance, the textual modality is represented as a 1000-dimensional bag-of-words vector, the image modality for the shallow hashing methods is represented as a 500 dimensional bag-of-words vector. For DCMH and the proposed deep hashing methods, we directly use raw pixels as the image modality inputs. Instances are considered to be similar if they share at least one common label, otherwise, they are considered to be dissimilar.

Baselines

Our method is compared with various state-of-the-art cross-modal hashing methods. Specifically CCA (Hotelling 1936), CMFH (Ding, Guo, and Zhou 2014), SCM (Zhang and Li 2014), LSSH (Zhou, Ding, and Guo 2014), STMH (Wang et al. 2015), CVH (Kumar and Udupa 2011), SePH (Lin et al. 2015), and DCMH (Jiang and Li 2016) are adopted as baselines. Source codes of most baselines are kindly provided by the authors, except for DCMH, CMFH and CCA. Since SePH is a kernel-based method, we use RBF kernel and randomly select 500 points as kernel bases following by the authors' suggestions. It also should be noticed that in SePH two strategies are proposed to construct the hash codes for retrieval instances according to whether all the modalities of a query point are observed or not. Since we focus on cross-modal retrieval, only one modality is utilized to construct the hash codes for the retrieval points. Parameters for all baselines are carefully tuned and set according to the original papers.

Settings and Performance Comparisons

For MIRFlickr, we take 2000 instances as the test set, and the rest as the retrieval set. To reduce computational costs, the training set include 5000 instances which are randomly sampled from the retrieval set. For NUS-WIDE, we take 1% of the dataset as the test set and the remaining as the retrieval set. We also randomly sampled 5000 instances from the retrieval set to construct the training set.

We use a validation set to choose the hyperparameter λ and γ. According to the results in the validation set, we set $\lambda = \gamma = 1$ in our experiments. The batch size is fixed to be 128 and the iteration number of the outer-loop in Algorithm 1 is set to be 1000.

We adopt two widely used evaluation measures: Mean Average Precision(MAP) and *precision-recall* curves. Table 1 and Table 2 report the MAP scores of all compared methods. Figure 2 and Figure 3 show the *precision-recall* curves of some representative methods on MIRFlickr and NUS-WIDE with 32 and 64 bits, respectively. It can be observed that our PRDH performs better than all baselines.

Conclusion

In this paper, we presented a novel hashing method, called as pairwise relationship guided deep hashing (PRDH) for large-scale cross-modal similarity search. Compact hash codes of image and text are generated in an end-to-end deep learning architecture. To sufficiently discover the heterogenous relationships and preserve the semantic similarities of the learned hash codes among different modalities, PRDH explores two types of pairwise constraints from intra-modality and inter-modality jointly. In addition, decorrelation constraints are also added in this deep architecture to

enhance the discriminative ability of each hash bit. Experiments on two datasets show that our PRDH model yields state-of-the-art performance in cross-modal retrieval tasks.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61572388, 61402026, and 61432014), Australian Research Council Projects FT-130101457, DP-140102164 and LE140100061, Program for Changjiang Scholars and Innovative Research Team in University (No.IRT13088), and the Fund of State Key Lab of Software Development Environment (SKLSDE-2016ZX-04).

References

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2015. Deep compositional question answering with neural module networks. *arXiv preprint arXiv:1511.02799*.

Bronstein, M. M.; Bronstein, A. M.; Michel, F.; and Paragios, N. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, volume 1, 5.

Cao, Y.; Long, M.; Wang, J.; Yang, Q.; and Yu, P. S. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1445–1454.

Chatfield, K.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. 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*, 48.

Cogswell, M.; Ahmed, F.; Girshick, R.; Zitnick, L.; and Batra, D. 2015. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*.

Deng, C.; Ji, R.; Liu, W.; Tao, D.; and Gao, X. 2013. Visual reranking through weakly supervised multi-graph learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2600–2607.

Deng, C.; Tang, X.; Yan, J.; Liu, W.; and Gao, X. 2016. Discriminative dictionary learning with common label alignment for cross-modal retrieval. *IEEE Transactions on Multimedia* 18(2):208–218.

Ding, G.; Guo, Y.; and Zhou, J. 2014. Collective matrix factorization hashing for multimodal data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2075–2082.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2625–2634.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2121–2129.

Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; and Xu, W. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, 2296–2304.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28:321–377.

Huiskes, M. J., and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39–43.

Irie, G.; Arai, H.; and Taniguchi, Y. 2015. Alternating co-quantization for cross-modal hashing. In *Proceedings of the IEEE International Conference on Computer Vision*, 1886–1894.

Jiang, Q.-Y., and Li, W.-J. 2016. Deep cross-modal hashing. *arXiv preprint arXiv:1602.02255*.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Kumar, S., and Udupa, R. 2011. Learning hash functions for cross-view similarity search. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, 1360.

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

Liu, X.; He, J.; Deng, C.; and Lang, B. 2014. Collaborative hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2139–2146.

Liu, X.; Huang, L.; Deng, C.; Lu, J.; and Lang, B. 2015. Multi-view complementary hash tables for nearest neighbor search. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 1107–1115.

Liu, X.; He, J.; and Lang, B. 2013. Reciprocal hash tables for nearest neighbor search. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA.*, 626–632.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of The 32nd International Conference on Machine Learning*, 97–105.

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

Wang, D.; Gao, X.; Wang, X.; and He, L. 2015. Semantic topic multimodal hashing for cross-media retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3890–3896.

Wang, J.; Liu, W.; Kumar, S.; and Chang, S.-F. 2016. Learning to hash for indexing big datała survey. *Proceedings of the IEEE* 104(1):34–57.

Yang, Y.; Xu, D.; Nie, F.; Luo, J.; and Zhuang, Y. 2009. Ranking with local regression and global alignment for cross media retrieval. In *International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, October*, 175–184.

Yang, Y.; Nie, F.; Xu, D.; and Luo, J. 2012. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4):723–742.

Zhang, D., and Li, W.-J. 2014. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, volume 1, 7.

Zhang, D.; Wang, F.; and Si, L. 2011. Composite hashing with multiple information sources. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 225–234.

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