The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Dual Deep Neural Networks Cross-Modal Hashing

Zhen-Duo Chen, Wan-Jin Yu, Chuan-Xiang Li, Liqiang Nie, Xin-Shun Xu*
School of Computer Science and Technology, Shandong University
School of Software, Shandong University
{chenzd.sdu, rcwanjinyu, chuanxiang.lee, nieliqiang}@gmail.com, xuxinshun@sdu.edu.cn

Abstract

Recently, deep hashing methods have attracted much attention in multimedia retrieval task. Some of them can even perform cross-modal retrieval. However, almost all existing deep cross-modal hashing methods are pairwise optimizing methods, which means that they become time-consuming if they are extended to large scale datasets. In this paper, we propose a novel tri-stage deep cross-modal hashing method – Dual Deep Neural Networks Cross-Modal Hashing, i.e., DDCMH, which employs two deep networks to generate hash codes for different modalities. Specifically, in Stage 1, it leverages a single-modal hashing method to generate the initial binary codes of textual modality of training samples; in Stage 2, these binary codes are treated as supervised information to train an image network, which maps visual modality to a binary representation; in Stage 3, the visual modality codes are reconstructed according to a reconstruction procedure, and used as supervised information to train a text network, which generates the binary codes for textual modality. By doing this, DDCMH can make full use of inter-modal information to obtain high quality binary codes, and avoid the problem of pairwise optimization by optimizing different modalities independently. The proposed method can be treated as a framework which can extend any single-modal hashing method to perform cross-modal search task. DDCMH is tested on several benchmark datasets. The results demonstrate that it outperforms both deep and shallow state-of-the-art hashing methods.

Introduction

With the explosive growth of various kinds of data, hashing is becoming more and more popular in approximate nearest neighbor (ANN) search due to its fast query speed and low storage cost (Wang et al. 2017b; Tang et al. 2015; Yang et al. 2014; Liu et al. 2016b; Wang et al. 2015; 2017c). Most of the pioneer efforts are specifically proposed for single-modal data; yet, in many scenarios, data exhibit multiple modalities such as text, acoustic and image. In a sense, cross-modal search deserves our attention. For instance, given textual content, we search the images from database. As compared to the single-modal hashing methods, efforts on the multi-modal ones are relatively sparse.

*Corresponding author.

Recently, deep neural networks have been widely used in various fields, e.g., computer vision and data mining (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016). Thereinto, deep Convolutional Neural Networks (CNNs) have demonstrated its theoretical and practical success in feature learning. Inspired by this, some deep hashing methods have been proposed for cross-modal retrieval.

Generally, these deep cross-modal hashing methods can obtain better results than the shallow methods in many real-world datasets. However, there are still some problems that need to be further considered. For example, nearly all the deep cross-modal hashing methods are pairwise optimized, which means that these methods become inefficient if all supervised information is used for training. This makes them hardly be extended to large-scale datasets. Thus, in real applications, they usually sample only a small subset with m points for training where m is typically less than ten thousands even the given training set is very large. Therefore, such methods cannot make full use of all available data samples. In addition, many single-modal hashing methods can obtain good results on single-modal data. Apparently, the binary codes generated by a single-modal hashing method contain useful information of intra-modality. If a model could make use of these binary codes and avoid the pairwise optimization problem, it is expected to obtain better results and become efficient.

Motivated by this, in this paper, we propose a novel tri-stage deep cross-modal hashing method – Dual Deep Neural Networks Cross-Modal Hashing (DDCMH) for cross-modal retrieval, which employs two deep neural networks to generate hash codes for different modalities. Distinct from the existing deep hashing methods, in DDCMH, the two deep networks are optimized independently; therefore, it can avoid the pairwise optimization problem. Without loss of generality, suppose that each sample has two modalities, i.e., text and image. In Stage 1, DDCMH leverages a single-modal hashing method to generate the binary codes of the textual modality of the training set; as to Stage 2, these binary codes are treated as supervised information to train an image network, which maps the image modality to a binary representation; when it comes to Stage 3, these image modality codes are reconstructed according to a reconstruction procedure, and be used as supervised information to train a text network, which generates better hash codes for textual modal-

ity than the single-modal hashing method used in Stage 1. The framework is illustrated in Figure 1.

To summarize, the main contributions of DDCMH include:

- It combines two deep neural networks for cross-modal hashing, avoids pairwise optimization scheme, and gets $O(n)$ complexity by optimizing different modalities independently. In addition, a code reconstruction procedure in this framework is designed for obtaining high-quality hash codes
- It can leverage any single-modal hashing method, which means that the proposed method can be treated as a framework to extend any single-modal hashing methods to cross-modal retrieval task.
- It obtains better results than several state-of-the-art shallow and deep hashing methods. Especially, it obtains much better results than other methods on Image-to-Text retrieval.

The remaining of this paper is organized as follows. First, the related work is discussed. Then it introduces the details of the framework, followed by the experimental results and corresponding analysis. Finally, we conclude the paper.

Related Work

In recent years, cross-modal hashing has been a popular research topic in machine learning and multimedia retrieval.

Most of the cross-modal hashing methods can be categorized into unsupervised and supervised ones. The former one characterizes and models the intra- and inter-modal relatedness of the given data without supervised information. Typical examples are Inter-Media Hashing (IMH) (Song et al. 2013), Linear Cross-Modal Hashing (LCMH) (Zhu et al. 2013), Latent Semantic Sparse Hashing (LSSH) (Zhou, Ding, and Guo 2014), and Collective Matrix Factorization Hashing (CMFH) (Ding, Guo, and Zhou 2014). By comparison, supervised cross-modal hashing methods take advantage of supervised information, e.g., semantic labels, to obtain better performance. Representative methods are Cross View Hashing (CVH) (Kumar and Udupa 2011), Semantic Correlation Maximization (SCM) (Zhang and Li 2014), Quantized Correlation Hashing (QCH) (Wu et al. 2015), Semi-Relaxation Supervised Hashing(SRSH) (Zhang et al. 2017), Supervised Robust Discrete Multimodal Hashing(SRDMH) (Yan et al. 2016), dictionary learning cross-modal hashing (DLCMH) (Xu 2016) and Semantics Preserving Hashing (SePH) (Lin et al. 2015).

More recently, some deep hashing methods have been proposed. For example, CNNH (Xia et al. 2014) decomposes the hash learning process into a stage of learning approximate hash codes, followed by a deep-network-based stage of simultaneously fine-tuning the image features and hash functions. Lately, in order to achieve simultaneous feature learning and hash-code learning, some end-to-end deep hashing methods (Zhang et al. 2015; Li, Wang, and Kang 2016; Liu et al. 2016a; Erin Liong et al. 2015; Zhao et al. 2015; Wang et al. 2017a) have been proposed, which combine the deep CNN and some specially designed objective function. Apart from that, deep feature based cross-modal hashing methods were proposed recently, including Deep Cross-Modal Hashing (DCMH) (Jiang and Li 2016), Pairwise Relation Guided Deep Hashing (PRDH) (Yang et al. 2017) and Deep Visual-Semantic Hashing (DVSH) (Cao et al. 2016). DCMH integrates both feature learning and hash-code learning into the same deep learning framework with one deep neural network for all modalities. PRDH further exploits different pairwise constraints to enforce the hash codes from not only intra-modality but also inter-modality. DVSH uses CNN and Long Short Term Memory (LSTM) to separately learn hash codes for each modality, but it is constrained to sentences or other sequential texts.

Proposed Method

In this section, we first give the problem formulation; then, show the details of our proposed method including the framework, optimization scheme and its extensions to out-of-sample data and more modalities.

Problem Formulation

Without loss of generality, we assume that there are two modalities for each sample, i.e., text and image[1]. Suppose we have n instances in training set $\mathcal{X} = \{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^{D_x}$ denotes the feature vectors of image modality, which can be the hand-crafted features or the raw pixels of the i-th image; $\mathcal{Y} = \{y_i\}_{i=1}^n$ denotes the feature vectors of text modality, where $y_i \in \mathbb{R}^{D_y}$ is the text tag information of the i-th image. In addition, we define $L = \{l_i\}_{i=1}^n$ where $l_i \in \mathbb{R}^{D_l}$ to denote the semantic labels of images, which can be further used to decide whether two samples are similar. The goal of our method is to learn two modality-specific hash functions to map the image features x and text features y to compact k-bit hash codes. More specifically, we target at learning two functions, e.g., $f_x(\cdot)$ and $f_y(\cdot)$, that satisfy: $b_x = f_x(x), b_y = f_y(y)$, where $b_x, b_y \in \{1, 0\}^k$, such that b_x and b_y preserve the similarity in both intra-modality and inter-modality. In addition, in this paper, $sign(\cdot)$ is an element-wise sign function defined as follows:

$$sign(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0. \end{cases} \tag{1}$$

Dual Deep Neural Networks Cross-Modal Hashing

As illustrated in Figure 1, there are three stages in DDCMH. In Stage 1, DDCMH trains a single-modal hashing method on the textual modality of the samples in training set, and obtains their corresponding binary codes. Thereafter, in Stage 2, these binary codes are treated as supervised information to train an image deep network, which maps the visual modality to a binary representation. When it comes to Stage 3, these visual modality codes are first reconstructed according to a well designed reconstruction procedure; then, these reconstructed binary codes are used as supervised information

[1]We would like to mention that our model can be extended to handle data samples with more than two modalities.

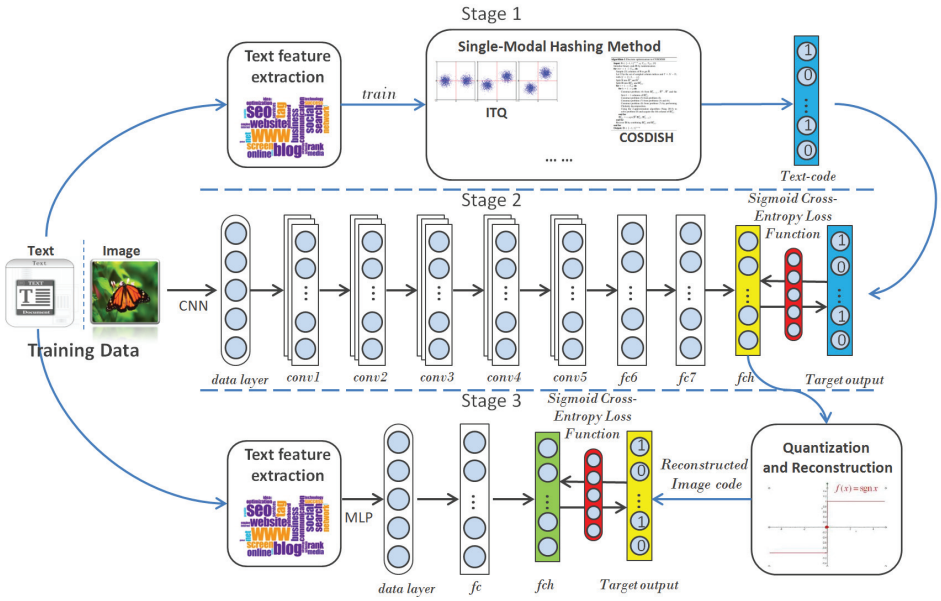

Figure 1: Schematic illustration of our proposed DDCMH.

to train a text deep network which can generate the final binary codes for textual modality. The details of all stages are described in the following paragraphs.

What should be emphasized is that there are substantial divergences between single-modal and cross-modal hashing methods because cross-modal methods need to consider inter-modality similarity. Therefore, it is hard for a single-modal hashing method to work on multi-modal data. However, DDCMH can extend any single-modal hashing methods to work on multi-modal data. From this point of view, DDCMH can be treated as a framework to extend single-modal hashing methods to perform cross-modal retrieval.

Stage 1: Initial Hash Codes Generation for Textual Modality

As mentioned previously, in this stage, DDCMH first tries to generate the initial binary codes of the textual modality. Note that, in some bi-stage methods, e.g., CNNH and SePH, they all first use approximate hash codes generation methods to obtain approximate binary codes, followed by further learning hash functions based on these approximate binary codes. Thus, to obtain the initial binary codes of textual modality, we can also adopt these approximate hash code generation methods. However, considering that the initial binary codes could affect the quality of final binary codes, we choose a different strategy here.

Recently, some excellent single-modal hashing methods have been proposed, e.g., ITQ (Gong et al. 2011), COSDISH (Kang, Li, and Zhou 2016), and SDH (Shen et al. 2015), which have achieved state-of-the-art performance. These methods can generate high-quality and efficient hash codes. Thus, we can choose one of such existing single-modal hashing methods, train it with the text modality to obtain the initial binary codes for text modality. In this paper, we choose COSDISH as the single-modal hash method in Stage 1. However, other single-modal hashing methods can also be used here. Due to space limitation, we do not discuss much about COSDISH. The detail about it can be found in the original paper. After training, we obtain a trained single-modal hashing model and the textual modality hash codes, e.g., $B^{y1} = \{b_i^{y1}\}_{i=1}^n$.

Stage 2: Image Network Training

After obtaining the textual modality hash codes B^{y1} in Stage 1, we subsequently utilize it as the supervised information and the original images to train a deep neural network. It is essentially similar to a multi-label classification task, with the acquired textual modality hash codes as labels to map the visual modality $\mathcal{X}$ of data to the image hash codes $B^x = \{b_i^x\}_{i=1}^n$.

Considering the impressive image feature learning power of CNNs, here we adopt a CNN model as the image network. Specifically, we modify AlexNet (Krizhevsky, Sutskever, and Hinton 2012), a famous and widely used CNN model. The original AlexNet model consists of five convolutional layers (*conv1-conv5*) and three fully-connected layers (*fc6-fc8*), and is pre-trained on the ImageNet-1000 dataset (Deng et al. 2009). To obtain the image hash codes, we replace the *fc8* with a new hashing layer *fch* with k nodes, each of which corresponds to one bit in the target hash code. With the *fch* layer, the *fc7* layer representation is transformed to a k-dimensional representation. The deep architecture of the image network is demonstrated in Figure 1.

More specifically, let $z_i^x = h_x(x_i; \theta_x)$ be the output of the image network, where x_i is the input image and θ_x is the parameter of the image network. Since our goal is to train an image network according to the text codes B^{y1}, we define the following likelihood function:

$$p(b_{ij}^{y1}|z_{ij}^x) = \begin{cases} \sigma(z_{ij}^x) & b_{ij}^{y1} = 1 \\ 1 - \sigma(z_{ij}^x) & b_{ij}^{y1} = 0, \end{cases} \tag{2}$$

where b_{ij}^{y1} is the hash code corresponding to the j-th bit of the i-th sample in B^{y1}, z_{ij}^{x} is the output of the j-th node in *fch* layer of the i-th sample, and $\sigma(\cdot)$ is a sigmoid function, i.e.,

$$\sigma(z_{ij}^{x}) = \frac{1}{1+e^{-z_{ij}^{x}}} . \tag{3}$$

Then, the loss function of the image network can be formulated as:

$$\begin{aligned} L_x &= -\frac{1}{nk}\log p(B^{y1}|Z^x) \\ &= -\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}\log p(b_{ij}^{y1}|z_{ij}^{x}) \\ &= -\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}[b_{ij}^{y1}\log p_{ij}^{x} + (1-b_{ij}^{y1})\log(1-p_{ij}^{x})], \end{aligned} \tag{4}$$

where n is the number of training samples, k is the number of bits in each hash code, and $p_{ij}^{x} = \sigma(z_{ij}^{x})$.

Given the above loss function, we learn the parameter θ_x of the network using the Back-Propagation (BP) algorithm with stochastic gradient descent (SGD). Specifically, we first take the derivative of the loss function:

$$\begin{aligned} \frac{\partial L_x}{\partial z_{ij}^{x}} &= \frac{\partial L_x}{\partial p_{ij}^{x}}\frac{\partial p_{ij}^{x}}{\partial z_{ij}^{x}} \\ &= -\frac{1}{nk}(b_{ij}^{y1}\frac{1}{p_{ij}^{x}} - \frac{1-b_{ij}^{y1}}{1-p_{ij}^{x}})(p_{ij}^{x}(1-p_{ij}^{x})) \\ &= -\frac{1}{nk}(p_{ij}^{x} - b_{ij}^{y1}). \end{aligned} \tag{5}$$

Thereafter, we can obtain $\partial L_x/\partial\theta_x$ with $\partial L_x/\partial z_{ij}^{x}$ using the chain rule, based on which the BP can be used to update the parameter θ_x. After training, we obtain a deep CNN model for image modality and the corresponding image hash codes $B^x = \{b_i^x\}_{i=1}^n$ of the training set, where $b_i^x = sign(h_x(x_i;\theta_x))$.

Stage 3: Code Reconstruction and Text Network Training

Thus far we have already trained two models that can generate binary codes for textual and visual modalities, i.e., the single-modal hashing in Stage 1 and the image deep network in State 2. We can use them to generate the binary codes of multi-modal samples. However, the procedure of the text hash code generation in Stage 1 is independent of the visual modality; consequently, it may result in a suboptimal result if we directly use it to generate the hash code of a query data point and perform the retrieval task among the image hash codes. To solve this problem, we further incorporate Stage 3 into the proposed framework.

In Stage 3, we design a multi-layer perceptron (MLP) model[2] with three fully-connected layers for textual modality, i.e., the text network. The architecture of the text network is also illustrated in Figure 1. Its architecture is similar to the last three layers of the image network, except for the first layer being replaced with an input layer, and the text-modal feature vectors are used as the input.

[2]Other deep models can also be used here.

Analogous to Stage 2, we can train the text network with the binary codes of images generated in Stage 2. However, this could lead to suboptimal because the procedure of generating image hash codes B^x in Stage 2 is essentially a multi-label classification task without explicit similarity preserving. Thus, the distribution of image hash codes B^x in the Hamming space is not as good as the text hash codes B^{y1}, which are generated by a well-designed hashing method and the quality of them can be guaranteed. Consequently, if we directly utilize these suboptimal codes to train the text network, the result could be worse. In addition, our method is not a pairwise one; we thus cannot adopt pairwise optimization to catch the similarity between data points like other deep cross-modal hashing methods do.

To consider the above problems, we further propose a code reconstruction method to optimize the image hash codes B^x according to semantic labels, which is demonstrated as follows:

$$b_i^{xr} = \frac{1}{\|l_i L^\top\|_1} l_i L^\top B^x, \tag{6}$$

where l_i is the semantic label vector of the i-th image, B^x is the hash codes of image modality generated in stage 2 and $\|\cdot\|_1$ is the 1-norm. It is obvious that, for the reconstructed hash codes $B^{xr} = \{b_i^{xr}\}_{i=1}^n$, the data points with the same semantic labels will have the same code because they are constructed with the same code from B^x based on their label information. This reconstruction method can optimize the distribution of image hash codes in the Hamming space. In the experiment section, we demonstrate that it can effectively improve the performance of Text-to-Image retrieval task. In addition, the time complexity of code reconstruction procedure is also $O(n)$, which guarantees its efficiency.

Once given the reconstructed codes B^{xr}, Stage 3 works similarly as Stage 2: Utilizing the B^{xr} as supervised information, and the textual feature vector as input to train the text network, which maps the textual modality Y of data to hash codes after training, i.e., $B^y = \{b_i^y\}_{i=1}^n$. Specifically, let $z_i^y = h_y(y_i;\theta_y)$ be the output of the text network, where y_i is the input text and the θ_y is the parameter vector of the text network, the loss function of the text network can be formulated as:

$$\begin{aligned} L_y &= -\frac{1}{nk} logp(B^{xr}|Z^y) \\ &= -\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} logp(b_{ij}^{xr}|z_{ij}^{y}) \\ &= -\frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}[b_{ij}^{xr} logp_{ij}^{y} + (1-b_{ij}^{xr})log(1-p_{ij}^{y})]. \end{aligned} \tag{7}$$

Similarly, the partial derivative of L_y with respect to z_{ij}^{y} is:

$$\frac{\partial L_y}{\partial z_{ij}^{y}} = \frac{\partial L_y}{\partial p_{ij}^{y}}\frac{\partial p_{ij}^{y}}{\partial z_{ij}^{y}} = -\frac{1}{nk}(p_{ij}^{y} - b_{ij}^{xr}). \tag{8}$$

Then, $\partial L_y/\partial \theta_y$ can be obtained with $\partial L_y/\partial z_{ij}^y$ by using the chain rule, based on which the BP can be used to update the parameter θ_y.

Out-of-Sample Extension

For a new instance which is not in the training set, we can easily generate its hash code as long as we can get one of its modalities. For example, given a query data point with visual modality x_q, we directly use it as the input of the image network, then forward propagate the network to generate its hash code as follows:

$$b_q^x = sign(h_x(x_q; \theta_x)). \tag{9}$$

Similarly, we can also use the text network to generate the hash code of a data point with only textual modality y_q:

$$b_q^y = sign(h_y(y_q; \theta_y)). \tag{10}$$

Extension to More Modalities

We have demonstrated how the proposed framework works on data with two modalities. We have to point it out that it is very easy to extend our model to data with more than two modalities. For example, we can append additional networks for more modalities in the framework, and train them on the corresponding modalities of training data with the binary codes generated by the preceding network as supervised information. After that, we can use them to obtain the hash codes of the corresponding modalities.

Experiments

To justify our proposed method, we carried out extensive experiments on two public benchmark datasets, i.e., MIRFlickr-25K (Huiskes and Lew 2008) and NUS-WIDE (Chua et al. 2009), and compared DDCMH with several state-of-the-art hashing methods for cross-modal search task including deep and shallow hashing methods.

Datasets

MIRFlickr-25K: It originally consists of 25,000 image samples collected from the Flickr website. Each image is annotated by one or more labels selected from 24 labels and some textual tags. In our experiment, we only retained the instances which have at least 20 tags. We then got 20,015 instances. The textual modality of each instance is represented as a 1,386-dimensional bag-of-words vector. For traditional methods based on shallow architectures, each image is represented by a 150-dimensional SIFT feature vector. For deep hashing methods, we directly utilized raw pixels as the image modality input. On this dataset, we sampled 2,000 instances as the test set (query), and the remaining as the retrieval set (database). In addition, as mentioned previously, existing deep hashing methods are based on pairwise optimization, which cannot efficiently work on large-scale datasets. For fairness, for all methods, we randomly sample 5,000 instances from the retrieval set as the training set.

NUS-WIDE: It contains 260,648 images from a public web image dataset. There are 81 ground truth concepts manually annotated for search evaluation. In our experiment, we selected 186,577 image-text pairs that belong to some of the 10 most frequent concepts. The text modality of each instance is represented as a 1,000-dimensional bag-of-words vector. For traditional methods based on shallow architectures, each image is represented by a 500-dimensional bag-of-words vector. For deep hashing methods, we directly used raw pixels as the image modality inputs. On this dataset, we chose 1% of the dataset as the test set (query) and the rest as the retrieval set (database). In addition, we also randomly sampled 5,000 instances from the retrieval set to construct the training set.

Baselines

We compared our proposed method with eight state-of-the-art cross-modal hashing methods, namely, CVH (Kumar and Udupa 2011), IMH (Song et al. 2013), SCM$_{seq}$ (Zhang and Li 2014), LSSH (Zhou, Ding, and Guo 2014), CMFH (Ding, Guo, and Zhou 2014), SePH$_{km}$ (Lin et al. 2015), DCMH (Jiang and Li 2016), and PRDH (Yang et al. 2017). DCMH and PRDH are deep hashing methods; others are shallow hashing methods. Generally, they can be further divided into two categories: Specifically, IMH, LSSH and CMFH as unsupervised methods; CVH, SCM$_{seq}$, SePH$_{km}$, DCMH and PRDH as supervised ones.

All parameters are set to those suggested in original papers or selected by a validation process. In addition, as mentioned previously, DDCMH adopts COSDISH as the single-modal hashing method in Stage 1 to generate the initial hash codes of textual modality. The image and text are considered to be similar if they share at least one common label. Otherwise, they are considered to be dissimilar.

Results and Discussions

The performance of all methods are measured by Mean Average Precision (MAP), which is a widely used metric for evaluating the accuracy of hashing. We also plotted the precision-recall curves of some cases.

Experimental results on MIRFlickr-25K are summarized in Table 1. From this table, we can observe that:

- All deep hashing methods, i.e., DCMH, PRDH and DDCMH, generally obtain better performance than those of the shallow hashing methods.
- DDCMH on all cases outperforms all the baselines no matter the deep or shallow ones.
- DDCMH obtains much better results on Image-to-Text cases than other methods.

To gain deep insights into DDCMH and all baselines, we plotted the precision-recall curves of the cases with 32 and 64 bits in Figure 2. From Figure 2, we can also observe similar results to those in Table 1.

Jointly analyzing Table 1 and Figure 2, we can observe that the results of DDCMH are a little different from those of other compared methods. For example, for other compared methods, the results on Text-to-Image are better than their corresponding results on Image-to-Text; however, DDCMH obtains better results on Image-to-Text than those on Text-to-Image. We think the main reason is that DDCMH adopts

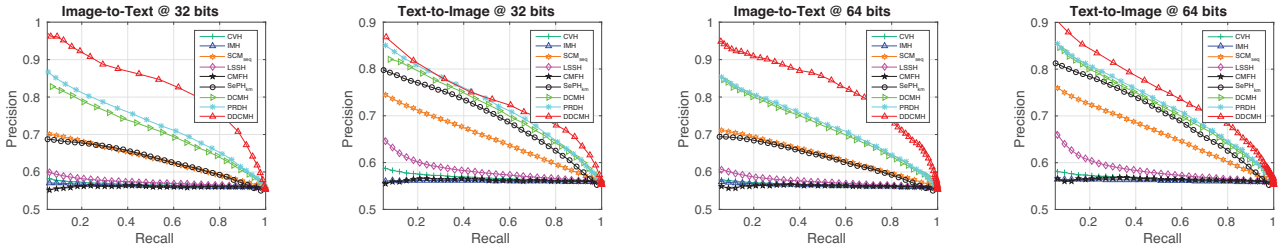

Figure 2: Precision-Recall curves on MIRFlickr-25K.

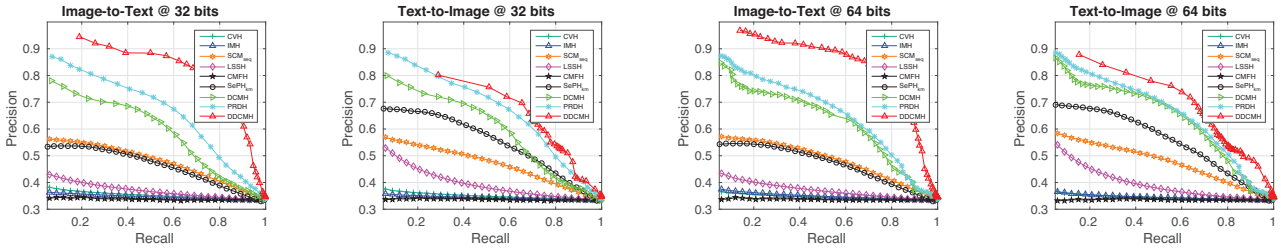

Figure 3: Precision-Recall curves on NUS-WIDE.

Table 1: MAP comparison on MIRFlickr-25K. The best results are in boldface.

Method	$I \rightarrow T$			$T \rightarrow I$		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
CVH	0.5741	0.5706	0.5682	0.5735	0.5705	0.5686
IMH	0.5655	0.5640	0.5658	0.5630	0.5622	0.5642
SCM_{seq}	0.6405	0.6576	0.6613	0.6335	0.6527	0.6594
LSSH	0.5826	0.5824	0.5795	0.5867	0.5917	0.5941
CMFH	0.5660	0.5673	0.5666	0.5659	0.5682	0.5676
SePH_{km}	0.6844	0.6872	0.6883	0.7333	0.7402	0.7440
DCMH	0.7056	0.7035	0.7140	0.7311	0.7487	0.7499
PRDH	0.7126	0.7128	0.7201	0.7467	0.7540	0.7505
DDCMH	**0.8208**	**0.8434**	**0.8551**	**0.7731**	**0.7766**	**0.7905**

Table 2: Perfromance (MAP) comparison on NUS-WIDE. The best results are in boldface.

Method	$I \rightarrow T$			$T \rightarrow I$		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
CVH	0.3662	0.3572	0.3522	0.3604	0.3535	0.3497
IMH	0.3524	0.3515	0.3547	0.3495	0.3498	0.3523
SCM_{seq}	0.5455	0.5576	0.5583	0.4864	0.5006	0.5080
LSSH	0.3899	0.3926	0.3917	0.4157	0.4206	0.4186
CMFH	0.3382	0.3371	0.3381	0.3382	0.3370	0.3392
SePH_{km}	0.5350	0.5409	0.5487	0.6177	0.6334	0.6418
DCMH	0.6141	0.6167	0.6427	0.6591	0.6487	0.6847
PRDH	0.6348	0.6529	0.6506	0.6808	0.6961	0.6943
DDCMH	**0.8023**	**0.8271**	**0.8316**	**0.6867**	**0.7012**	**0.7412**

COSDISH to generate initial binary codes for text modality. We further analyzed this in the following subsection.

The results on NUS-WIDE are displayed in Table 2 and Figure 3. Specifically, Table 2 lists the MAP values of all methods on different cases; Figure 3 illustrates the precision-recall curves of all methods on cases with 32 and 64 bits. From Table 2 and Figure 3, we have the following observations:

- Deep methods also perform better on this dataset.
- DDCMH outperforms all baselines on all cases.
- DDCMH obtains much better results on all cases of Image-to-Text retrieval than other methods.
- DDCMH also obtains better results on Image-to-Text retrieval than those on Text-to-Image retrieval, which is similar to that on MIRFlickr-25K.

To sum up, from the results on MIRFlickr-25K and NUS-WIDE, we can find that DDCMH can obtain better results than all baselines, which confirms its effectiveness.

Other Analysis

Necessity of Stage 3 and Code Reconstruction: In DDCMH, Stage 3 is added into the framework to learn the hash codes of textual modality. A code reconstruction procedure is used in Stage 3. In order to demonstrate that both Stage 3 and the code reconstruction procedure are necessary for DDCMH, we design two variants of the proposed approach, i.e., DDCMH-1 and DDCMH-2. The first one is a bi-stage method without Stage 3, which directly uses the single-modal hashing method in Stage 1, rather than the text net-

Table 3: Performance (MAP) comparison of the variants of DDCMH on the Text-to-Image retrieval of two datasets. The best results are in boldface.

Method	MIRFlickr-25K			NUS-WIDE		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
DDCMH-1	0.7227	0.7367	0.7336	0.6839	0.6979	0.7008
DDCMH-2	0.7159	0.7322	0.7295	0.6723	0.6785	0.6849
DDCMH	**0.7731**	**0.7766**	**0.7905**	**0.6867**	**0.7012**	**0.7412**

work, to generate the binary codes for the query data points. DDCMH-2 has all the three stages; however, the code reconstruction procedure is removed in Stage 3, which means that the output of the image network is directly used as supervised information to train the text network in Stage 3 after quantization.

We conducted experiments on both MIRFlickr-25k and NUS-WIDE. The MAP results of both variants and DDCMH are shown in Table 3. Note that these variants mainly affect the results of Text-to-Image retrieval, Table 3 only shows the results of Text-to-Image retrieval. From Table 3, we have the following observations:

- DDCMH outperforms DDCMH-1 on all cases of both datasets, which confirms that Stage 3 is necessary for DDCMH.
- DDCMH outperforms DDCMH-2 by 3.1% to 5.4 % on all cases of both datasets, which confirms that DDCMH also benefits from the code reconstruction procedure.
- The performance of DDCMH-2 is a little worse than that of DDCMH-1. This further demonstrates the necessity and effectiveness of code reconstruction procedure to optimize the codes generated by the image network.

Impact of Single-Modal Hashing: As shown in Figure 1, i.e., the architecture of DDCMH, we leverage a single-modal hashing method in Stage 1 to generate initial binary codes of textual modality. Intuitively, the performance of the single-modal hashing method could affect the performance of DDCMH, i.e., the better performance of the single-modal hashing method is, the better performance DDCMH achieves. To confirm this, we further design a new variant of DDCMH, i.e., DDCMH-ITQ, which adopts ITQ as the single-modal hashing method in Stage 1, instead of COSDISH used in previous experiments. ITQ is a well-known single-modal hashing method, but it is an unsupervised method and its performance is generally worse than that of COSDISH. We carried out experiments on MIRFlick-25K. The MAP results are shown in Table 4, in which DDCMH-ITQ means ITQ is used in Stage 1 and COSDISH is used in DDCMH-COSDISH. Note that the last three columns are the results of ITQ and COSDISH, which means that the MAP values are computed based on the hash codes generated by ITQ and COSDISH on Text-to-Text retrieval. From this table, we have the following observations:

- COSDISH outperforms ITQ on all cases of the Text-to-Text retrieval task.
- DDCMH-COSDISH outperforms DDCMH-ITQ in most cases except the case of 64 bits of Text-to-Image retrieval. This justifies that DDCMH can obtain better results by using a single hashing model with better performance.
- DDCMH-ITQ obtains better results on Text-to-Image retrieval than those on Image-to-Text retrieval. This is different from that of DDCMH-COSDISH. We thus can conclude that DDCMH obtains better results on Image-to-Text retrieval than those on Text-to-Image retrieval observed in Table 1 and 2 is because that COSDISH generates different initial binary codes.

Conclusion and Future Work

In this paper, we present a novel deep cross-modal hashing methods for cross-modal retrieval task, i.e., Dual Deep Neural Networks Cross-Modal Hashing, DDCMH, which consists of three stages. By optimizing different modalities independently, DDCMH avoids pairwise optimization problem. Moreover, DDCMH can be treated as a framework to extend any single-modal hashing methods to cross-modal retrieval task. Extensive experiments are conducted on two benchmark datasets. The results demonstrate that the proposed method outperforms several state-of-the-art baselines for cross-modal retrieval task.

In this work, DDCMH first generates initial binary codes of the textual modality so that we can make full use of the powerful representation of CNN in Stage 2. In our future work, we plan to explore whether it can also work well if it first generates the binary codes of visual modality in Stage 1. Besides that, we only train the image network and the text network once separately because of the limitation of time. Since the experiments demonstrated that it helps to freeze one model and use it as a supervisor while the other trains and vice versa, we may try to continue this process over and over until the performance saturates. We will further explore the balance between the time consumed and performance gained by continuing the process in the future work.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (61173068, 61573212), Program for New Century Excellent Talents in University of the Ministry of Education, Key Research and Development Program of Shandong Province (2016GGX101044).

References

Cao, Y.; Long, M.; Wang, J.; Yang, Q.; and Yu, P. S. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *KDD*, 1445–1454.

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 *CIVR*, 48C–56.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Table 4: Performance comparison of DDCMH with different single-modal hashing methods, i.e., ITQ and COSDISH.

Method	$I \rightarrow T$			$T \rightarrow I$			Method	$T \rightarrow T$		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits		16 bits	32 bits	64 bits
DDCMH-ITQ	0.6302	0.6339	0.6357	0.6822	0.7334	**0.8047**	ITQ	0.6087	0.6128	0.6052
DDCMH-COSDISH	**0.8208**	**0.8434**	**0.8551**	**0.7731**	**0.7766**	0.7905	COSDISH	**0.7766**	**0.7892**	**0.8041**

Ding, G.; Guo, Y.; and Zhou, J. 2014. Collective matrix factorization hashing for multimodal data. In *CVPR*, 2075–2082.

Erin Liong, V.; Lu, J.; Wang, G.; Moulin, P.; and Zhou, J. 2015. Deep hashing for compact binary codes learning. In *CVPR*, 2475–2483.

Gong, Y.; Lazebnik, S.; Gordo, A.; and Perronnin, F. 2011. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. In *CVPR*, 817–824.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huiskes, M. J., and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *MM*, 39–43.

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

Kang, W.-C.; Li, W.-J.; and Zhou, Z.-H. 2016. Column sampling based discrete supervised hashing. In *AAAI*, 1230–1236.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Kumar, S., and Udupa, R. 2011. Learning hash functions for cross-view similarity search. In *IJCAI*, 1360–1365.

Li, W.; Wang, S.; and Kang, W. 2016. Feature learning based deep supervised hashing with pairwise labels. In *IJCAI*, 1711C–1717.

Lin, Z.; Ding, G.; Hu, M.; and Wang, J. 2015. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, 3864–3872.

Liu, H.; Wang, R.; Shan, S.; and Chen, X. 2016a. Deep supervised hashing for fast image retrieval. In *CVPR*, 2064–2072.

Liu, X.; Huang, L.; Deng, C.; Lang, B.; and Tao, D. 2016b. Query-adaptive hash code ranking for large-scale multi-view visual search. *IEEE Transactions on Image Processing* 25:4514–4524.

Shen, F.; Shen, C.; Liu, W.; and Shen, H. T. 2015. Supervised discrete hashing. In *CVPR*, 37–45.

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 *SIGMOD*, 785–796.

Tang, J.; Li, Z.; Wang, M.; and Zhao, R. 2015. Neighborhood discriminant hashing for large-scale image retrieval. *IEEE Transactions on Image Processing* 24(9):2827–2840.

Wang, M.; Li, W.; Liu, D.; Ni, B.; Shen, J.; and Yan, S. 2015. Facilitating image search with a scalable and compact semantic mapping. *IEEE transactions on cybernetics* 45(8):1561–1574.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017a. Adversarial cross-modal retrieval. In *MM*, 154–162.

Wang, J.; Zhang, T.; Song, J.; Sebe, N.; and Shen, H. T. 2017b. A survey on learning to hash. *IEEE transactions on pattern analysis and machine intelligence*.

Wang, M.; Fu, W.; Hao, S.; Liu, H.; and Wu, X. 2017c. Learning on big graph: Label inference and regularization with anchor hierarchy. *IEEE Transactions on Knowledge and Data Engineering* 29(5):1101–1114.

Wu, B.; Yang, Q.; Zheng, W.-S.; Wang, Y.; and Wang, J. 2015. Quantized correlation hashing for fast cross-modal search. In *IJCAI*, 3946–3952.

Xia, R.; Pan, Y.; Lai, H.; Liu, C.; and Yan, S. 2014. Supervised hashing for image retrieval via image representation learning. In *AAAI*, 2156C–2162.

Xu, X.-S. 2016. Dictionary learning based hashing for cross-modal retrieval. In *MM*, 177–181.

Yan, T.-K.; Xu, X.-S.; Guo, S.; Huang, Z.; and Wang, X. 2016. Supervised robust discrete multimodal hashing for cross-media retrieval. In *CIKM*, 1271–1280.

Yang, Y.; Zha, Z.-J.; Gao, Y.; Zhu, X.; and Chua, T.-S. 2014. Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Transactions on Multimedia* 16:1677–1689.

Yang, E.; Deng, C.; Liu, W.; Liu, X.; Tao, D.; and Gao, X. 2017. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, 1618–1625.

Zhang, D., and Li, W. 2014. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, 2177–2183.

Zhang, R.; Lin, L.; Zhang, R.; Zuo, W.; and Zhang, L. 2015. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Transactions on Image Processing* 24(12):4766–4779.

Zhang, P.-F.; Li, C.-X.; Liu, M.-Y.; Nie, L.; and Xu, X.-S. 2017. Semi-relaxation supervised hashing for cross-modal retrieval. In *MM*, 1762–1770.

Zhao, F.; Huang, Y.; Wang, L.; and Tan, T. 2015. Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR*, 1556–1564.

Zhou, J.; Ding, G.; and Guo, Y. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *SIGIR*, 415–424.

Zhu, X.; Huang, Z.; Shen, H. T.; and Zhao, X. 2013. Linear cross-modal hashing for efficient multimedia search. In *MM*, 143C–152.