LEARNING A CROSS-MODAL HASHING NETWORK FOR MULTIMEDIA SEARCH

Venice Erin Liong[1,3], *Jiwen Lu*[2], *and Yap-Peng Tan*[3]

[1]Interdisciplinary Graduate School, Nanyang Technological University, Singapore
[2]Department of Automation, Tsinghua University, Beijing, China
[3]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
veniceer001@e.ntu.com.sg; lujiwen@tsinghua.edu.cn; eyptan@ntu.edu.sg;

ABSTRACT

In this paper, we propose a cross-modal hashing network (CMHN) method to learn compact binary codes for cross-modality multimedia search. Unlike most existing cross-modal hashing methods which learn a single pair of projections to map each example into a binary vector, we design a deep neural network to learn multiple pairs of hierarchical non-linear transformations, under which the nonlinear characteristics of samples can be well exploited and the modality gap is well reduced. Our model is trained under an iterative optimization procedure which learns a (1) unified binary code discretely and discriminatively through a classification-based hinge-loss criterion, and (2) cross-modal hashing network, one deep network for each modality, through minimizing the quantization loss between real-valued neural code and binary code, and maximizing the variance of the learned neural codes. Experimental results on two benchmark datasets show the efficacy of the proposed approach.

Index Terms— hashing, cross-modal retrieval, binary code learning

1. INTRODUCTION

Learning-based hashing is an effective technique for large-scale multimedia search and a variety of hashing algorithms have been proposed in recent years [1, 2, 3, 4]. The basic idea of learning-based hashing approaches for large-scale search is to automatically learn a set of hash functions from the training set to map each visual example into a compact binary feature vector so that conceptually similar samples are mapped into similar binary codes. Most existing hashing methods are developed for single-modal retrieval, but in many real-applications, it is easy to access multi-modal data for multimedia retrieval. For example, text data can be used to retrieve an image and vice versa. Hence, it is desirable to develop effective cross-modal fast similarity search methods for efficient multimedia retrieval. In recent years, several cross-modal hashing methods have been proposed in the literature, and most studies are in shallow form in which it only performs a single-layer of linear or nonlinear transformation. These

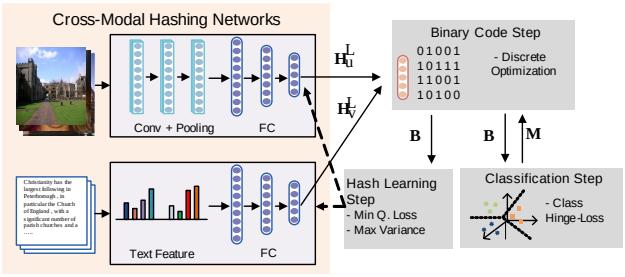


Fig. 1. The basic idea of our proposed approach for multimedia retrieval. We learn a couple of hashing networks and a joint binary code matrix using an alternative optimization procedure to learn the class projection matrix, $\mathbf{M}$, unified binary code, $\mathbf{B}$ and network parameters.

can be classified into two types: *unsupervised* and *supervised*. Unsupervised methods [5, 6, 7, 8, 9] utilize co-occurence information such that the only the image-text pairs which occurred in the same article are known to be of similar semantic. Supervised methods [10, 11, 12, 13] utilize semantic labels to enhance the correlation of cross-modal data. There are few works which have performed deep learning for cross-modal hashing [14, 15, 16, 17] in the form of siamese networks and autoencoders using various metric-based loss functions such as triplet-ranking or cosine-loss. In this paper, we propose a new cross-modal deep hashing method called cross-modal hashing network (CMHN) for cross-modality retrieval. Unlike existing shallow cross-modal hashing methods which learn a single pair of linear or nonlinear projections to map each example into a binary vector, we employ an end-to-end hashing network to learn multiple pairs of hierarchical nonlinear transformations, under which the nonlinear relationship of samples can be well exploited and that the binarized neural codes having same semantic would be similar as possible, and neural codes having different semantic would be dissimilar as possible. Our model is trained under two main steps: First, we perform binary code inference to learn a shared binary code for each cross-modal training pair such that we

obtain a common hamming space for the two modalities and the modality gap can be implicitly reduced.. We perform this in a *discrete* and *discriminative* manner to avoid approximate optimization loss caused by relaxing the binary constraint and strengthen the semantic correlation between modalities using a classification-based hinge loss criterion, respectively. Second, we model the hashing networks such that we minimize the loss between real-valued neural code and binary code, to make the codes as similar as possible, and maximize the variance between neural codes to ensure independent codes. We perform this two steps in a joint and iterative optimization procedure. Figure 1 shows the pipeline of our proposed approach. Experimental results on two benchmark datasets show the efficacy of the proposed approach.

2. CROSS-MODAL HASHING NETWORK

Let $\mathbf{X}_u = [\mathbf{x}_{u1}, \mathbf{x}_{u2}, \cdots, \mathbf{x}_{uN}] \in \mathbb{R}^{d_u \times N}$ and $\mathbf{X}_v = [\mathbf{x}_{v1}, \mathbf{x}_{v2}, \cdots, \mathbf{x}_{vN}] \in \mathbb{R}^{d_v \times N}$ be the training sets from different modalities, where u and v represent two different modalities, N is the number of training samples in each modality, and $\mathbb{R}^{d_u}$ and $\mathbb{R}^{d_v}$ are the feature dimension for each sample in modality u and v, respectively. Cross-modal hashing aims to seek a couple of hash functions to transform each sample in the modality into a compact binary feature vector with K bits:

$$f_u : \mathbb{R}^{d_u} \rightarrow \{-1, 1\}^K, \quad f_v : \mathbb{R}^{d_v} \rightarrow \{-1, 1\}^K \tag{1}$$

Motivated by the advances of deep networks [18, 19, 20], we propose a cross-modal hashing network, to learn multiple pairs of hierarchical non-linear transformations, under which the nonlinear relationship of samples and the relationship of samples from different modalities can be well exploited. Given a pair of training samples $\mathbf{x}_{un}$ and $\mathbf{x}_{vn}$ which are two sampled from different modalities with the same semantic concept, we pass each of them into the corresponding hashing network. The feature at the top layer are the representative real-valued neural codes, $\mathbf{h}^L_{un}$ and $\mathbf{h}^L_{vn}$, which can be embedded into the Hamming space naively using the sign$(\cdot)$ function. We binarize the outputs of the top layers of the networks as follows:

$$\mathbf{b}_{un} = \text{sgn}(\mathbf{h}^L_{un}), \mathbf{b}_{vn} = \text{sgn}(\mathbf{h}^L_{vn}) \tag{2}$$

But because we want to reduce the modality gap between the binary codes obtained from each modality as much as possible, it is desired to have a single binary code, $\mathbf{b}_n = \mathbf{b}_{un} = \mathbf{b}_{vn}$, to represent the cross-modal pairs during training. By doing so, the neural codes of different modalities but of similar semantics is encoded to a single binary code, implicitly reducing the modality gap. Hence, our CMHN model aim to learn a unified binary code matrix $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots \mathbf{b}_N] \in \{-1, 1\}^{N \times K}$, and the parameters of the cross-modal hashing networks $\theta_u \triangleq \{\mathbf{W}^l_u, \mathbf{c}^l_u\}^L_{l=1}$ and $\theta_u \triangleq \{\mathbf{W}^l_v, \mathbf{c}^l_v\}^L_{l=1}$ simultaneously. Where $\mathbf{W}^l$ and $\mathbf{c}^l$is the projection matrix and bias vector for the l-th layer, and L is the top-most layer in the network. Unlike other binary code learning techniques which relaxes the binary constraints for an easy single optimization, we learn the binary codes and hash networks alternatively in a discrete manner so that we preserve the binary constraints avoiding the approximation loss caused by relaxation. Also, to make our binary codes discriminative such that samples that are semantically relevant (irrelevant) have similar (dissimilar) binary codes as much as possible, we exploit the label information to enforce maximizing the semantics correlation using a classification-based hinge-loss criterion. We now present the joint formulation and how to perform optimization in the proceeding subsections.

Learning Unified Binary Codes. To learn discriminative unified binary codes that are semantically correlated, we follow the assumption that these codes should be able to perform well on a multi-classification problem. Hence, we create a hinge-loss function such that we minimize the loss between label information and a unified binary feature. Given label data $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots \mathbf{y}_N] \in \{1, 0\}^{N \times C}$ where C is the number of class labels and $\mathbf{y}_{n,j} = 1$ if the n-th sample belongs to class j and 0 otherwise, we learn a multi-class projection matrix defined as $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \cdots \mathbf{m}_C] \in \mathbb{R}^{K \times C}$ such that the hinge loss is minimized. This is possible with the slack variable, $\xi_n \geq 1$. We learn K independent class matrices which solve the two-class hinge-loss problem as follows:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{M}} J_1 \quad &= \quad \|\mathbf{M}\|^2_F + \sum_n^N \xi_n \\ &\forall n, j \;\; \mathbf{y}_{n,j}(\mathbf{m}^\top_j \mathbf{b}_n) \geq 1 - \xi_n \end{aligned} \tag{3}$$

Learning Deep Cross-modal networks To learn the parameters θ_u and θ_v for our cross-modal network, we employ the following formulation:

$$\begin{aligned} \min_{\theta_u,\theta_v} J_2 \quad &= \quad \left(\|\mathbf{B} - \mathbf{H}^L_u\|^2_F + \|\mathbf{B} - \mathbf{H}^L_v\|^2_F\right) \\ &\quad -\alpha \left(\text{tr}(\mathbf{H}^L_u \mathbf{H}^{L\top}_u) + \text{tr}(\mathbf{H}^L_v \mathbf{H}^{L\top}_v)\right) \end{aligned} \tag{4}$$

$\mathbf{H}^L_u$ and $\mathbf{H}^L_v$ are the neural codes obtained from $\mathbf{X}_u$ and $\mathbf{X}_v$ given parameter θ_u and θ_v. As can be seen, we learn the deep hashing networks from a single unified binary code where we minimize the quantization loss between neural codes and binary code. By doing so, the energy of the samples can be well preserved in the learned hashing networks [21]. In addition, we add a regularizer term parameterized by α which maximizes the variances of the zero-centered neural code from different modalities. By doing so, the binary bits are as independent as possible for each modality [1, 22].

Iterative Optimization. Combining the two objectives previously mentioned, we are able to formulate the following optimization objective to learn compact binary codes $\mathbf{B}$, the classification projection matrix $\mathbf{M}$, and the parameters for the cross-modal hashing networks θ_u and θ_v:

$$\min_{\mathbf{B},\mathbf{M},\theta_u,\theta_v} J \quad = \quad J_1 + \lambda_1 J_2 \tag{5}$$

where λ_1 is a constant parameter to balance the effects of the two terms in the objective function.

The optimization problem in Eq.(5) is non-convex due to the binary constraints, which makes it difficult to solve. However, it can be addressed using an iterative approach where we keep other variables fixed and solve one alternatively and iteratively. We learn the binary code, multi-class projection matrix and hash functions simultaneously as follows:

Classification Step: Fixing $\mathbf{B}$ and parameters θ, we are left with a support vector machine (SVM) formulation which can solved through a standard solver[1] to learn the classification matrix $\mathbf{M}$ in Eq.(3).

Binary Code Step: Fixing $\mathbf{M}$ and parameters θ, we can simplify Eq.(5) as follows:

$$\begin{aligned} \min_{\mathbf{b}_i} J(\mathbf{b}_n) &= -\sum_{j=1}^{C} \mathbf{m}_{c_n,j}^{\top} \mathbf{b}_n \\ &+ \lambda_1(\|\mathbf{b}_n - \mathbf{h}_{un}^L\|_F^2 + \|\mathbf{b} - \mathbf{h}_{vn}^L\|_F^2) \\ \text{subject to} \quad & \mathbf{b}_i \in \{-1, 1\}^{1\times K} \end{aligned} \tag{6}$$

Because the function $\mathbf{B}$ from Eq.(2) is non-differentiable due to the employed sgn function, we learn $\mathbf{B}$ through a discrete optimization technique.Eq.(6) becomes a binary quadratic problem which can be solved through a linear gradient technique similar to [23] as follows:

$$\mathbf{b}_n = \text{sgn}(\mathbf{y}_n \mathbf{M}^{\top} + \lambda_1(\mathbf{h}_{un}^L + \mathbf{h}_{vn}^L)) \tag{7}$$

Hash Function Step: To solve the network parameters we fix $\mathbf{B}$ and $\mathbf{M}$ from Eq.(5) and obtain the resulting formulation:

$$\begin{aligned} \min_{\theta} J(\theta) &= \lambda_1(\|\mathbf{B} - \mathbf{H}_u^L\|_F^2 + \|\mathbf{B} - \mathbf{H}_v^L\|_F^2) \\ &- \lambda_1 \alpha \left(\text{tr}(\mathbf{H}_u^L \mathbf{H}_u^{L\top}) + \text{tr}(\mathbf{H}_v^L \mathbf{H}_v^{L\top})\right) \end{aligned} \tag{8}$$

We employ the batch-wise gradient descent method to learn parameters θ of the networks of modalities u and v. We get the gradient of J in Eq.(5) with respect to the neural code. For each layer of the network, the gradients can easily be computed through the chain rule during backpropagation. The parameters of the networks are updated using these gradients based on a given learning rate, momentum and weight decay. For new instances or query data, we simply use the learned hashing networks to obtain the neural codes and finally binarize them using the sign$(\cdot)$ function as shown in Eq.(2).

3. EXPERIMENTS

Datasets: We employed two cross-modal datasets in our experiments: IAPRTC12 and NUS-WIDE. The IAPR TC-12 dataset[2] contains 19627 images with corresponding sentence descriptions. These image-sentence pairs present various semantics such as landscape, action and people categories. Similar to [16], we use the top 22 frequent labels from the 275 concepts obtained generated from the segmentation task. For the text features, we pre-process the sentence data removing the stop words and extract a bag-of-words (BoW) representation with a dimension of 500. We randomly select 100 pairs per class as the query set and the remaining data as the gallery set. The NUS-Wide dataset[3] contains 269648 images which were annotated by 81 concept tags. Following the same settings in previous works [24, 25], we selected the 10 most frequent concepts and constructed a subset which contains 186577 images-tag pairs. In our experiments, each text is represented by a 1000-dimensional feature vector which is computed by the bag-of-words model. We randomly selected 99% samples to form the database and the rest as query samples.

Table 1. mAP performance of different cross-modal hashing methods on the IAPRTC12 dataset.

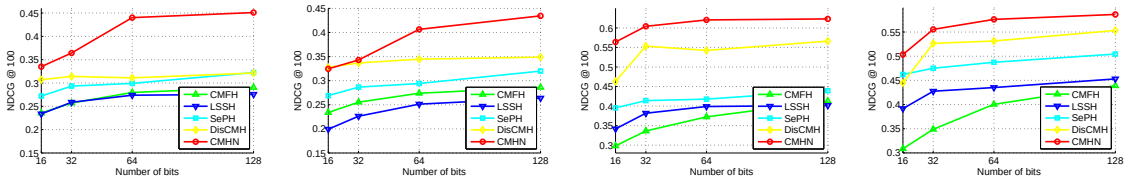
	Method	16 bits	32 bits	64 bits	128 bits
$I \to T$	CMFH [7]	0.5601	0.5829	0.6079	0.6179
	LSSH [6]	0.5440	0.5769	0.5964	0.5985
	SePH - *km* [10]	0.6177	0.6447	0.6500	0.6781
	DisCMH [12]	0.6174	0.6596	0.6503	0.6594
	DNH-C [26]	0.5250	0.5592	0.5902	0.6339
	DVSH [16]	0.5696	0.6321	0.6964	0.7236
	CMHN	**0.6483**	**0.7274**	**0.7974**	**0.8251**
	CMHN (o)	0.5768	0.7062	0.7780	0.8060
$T \to I$	CMFH [7]	0.5592	0.5834	0.6084	0.6187
	LSSH [6]	0.4868	0.5264	0.5547	0.5724
	SePH - *km* [10]	0.6105	0.6340	0.6404	0.6730
	DisCMH [12]	0.6532	0.6910	0.6921	0.6949
	DNH-C [26]	0.4692	0.4838	0.4905	0.5053
	DVSH [16]	0.6037	0.6395	0.6806	0.6751
	CMHN	**0.6687**	**0.6925**	**0.7535**	**0.7925**
	CMHN (o)	0.6716	0.6615	0.6677	0.6490

Implementation Details: Our deep architecture was implemented under the MatConvNet [20] framework. For the *image hashing network*, our deep model used the pre-trained VGG-net [27] as our initial convolution and pooling layers up to FC7, and stack the number of new FC layers with dimensions of $[4096 \to 500 \to 200 \to K]$ for all datasets. We perform end-to-end learning by having the learning rate at the new fully connected layers to be 0.01. To avoid overfitting and ruining the representative abstract features already learned during the pre-training, we reduce the learning rate of the remaining convolution and FC layers to be 0.0001. For the *text hashing network*, we designed fully-connected networks and use the pre-processed text features, given by each experiment, as input. We set the FC layers as $[1386 \to 500 \to K]$, and $[1000 \to 500 \to K]$, for the IAPRTC12, and NUS-WIDE dataset, respectively. For both image and text network, we used the ReLU activation [28] as the nonlinear activation function for the new fully connected layers except for the last

[1] we use LibSVM: http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[2] http://imageclef/photodata, http://imageclef/SIAPRdata.

[3] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm



(a) IAPRTC12 ($I \rightarrow T$) (b) IAPRTC12 ($T \rightarrow I$) (c) NUSWIDE ($I \rightarrow T$) (d) NUSWIDE ($T \rightarrow I$)

Fig. 2. NDCG performance of different cross-modal hashing methods for the IAPRTC12 and NUSWIDE database.

Table 2. mAP performance of different cross-modal hashing methods on the NUS-WIDE dataset.

	Method	16 bits	32 bits	64 bits	128 bits
$I \rightarrow T$	CMFH [7]	0.4772	0.5301	0.5763	0.6258
	LSSH [6]	0.5547	0.5734	0.5980	0.5968
	SePH - *km* [10]	0.6177	0.6447	0.6500	0.6781
	DisCMH [12]	0.6826	0.7583	0.7752	0.7605
	CAH [15]	0.4920	0.5084	0.5407	0.5628
	DCMH [17]	0.6249	0.6355	0.6720	-
	CMHN	**0.7893**	**0.8170**	**0.8236**	**0.8289**
	CMHN (o)	0.6558	0.7480	0.7818	0.7614
$T \rightarrow I$	CMFH [7]	0.4965	0.5432	0.5995	0.6405
	LSSH [6]	0.5857	0.6242	0.6293	0.6464
	SePH - *km* [10]	0.6604	0.6766	0.7043	0.7024
	DisCMH [12]	0.6519	0.7378	0.7535	0.7511
	CAH [15]	0.5019	0.5135	0.5451	0.5800
	DCMH [17]	0.6791	0.6829	0.6906	-
	CMHN	**0.6829**	**0.7469**	**0.7651**	**0.7772**
	CMHN (o)	0.6643	0.6950	0.7170	0.7062

layer. We use the hyperbolic tangent (tanh) function for the top layer of the network because it is able to squeeze the representation to a {-1,1} range which ensures that the quantization loss can be reduced as much as possible. The parameters in the new fully connected layers are initialized using the Xavier initialization [29]. The momentum, and weight decay were set to 0.9, and 0.0001, respectively. The parameters λ_1 and α were set to 0.2 and 0.001, respectively, which were obtained by cross-validation using $K = 16$ bits.

Comparisons with State-of-the-arts: We compared our CMHN with the different state-of-the-art cross-modal hashing methods which can be grouped to unsupervised (LSSH, CMFH) and supervised (SePH, DisCMH).[4] In addition, we also compare out method with current deep cross-modal hashing methods[5]. For the shallow methods, we make use of CNN features extracted at the FC7 layer for the images from the pre-trained model initially used by our CMHN method. We use the whole gallery as training data to learn the hashing functions. For each dataset, we performed two cross-modal retrieval tasks: image-to-text retrieval and text-to-image retrieval. Tables 2 show the mAP performance by Hamming Ranking. It can be observed that our method provided the best performance compared to the shallow cross-modal hashing methods. The SePH performed nonlinear transformations but was done explicitly through kernels which cannot really maximize the information from raw data. The DisCMH method gave competitive results with our CMHN at lower bits, but did not consistently improve as the bit size increased. Because only the shallow methods perform unified code learning, we also test our method in an out-of-sample extension, named CMDH(o), in which we use the learned hashing networks to obtain the binary codes, and not the learned unified binary code, as gallery. It can still be seen, that our deep model gave best and competitive results compared to the deep models. This is because the CAH method still used handcrafted image features as input while our method performed a complete network learning from raw images. Second, the DCMH method performed end-to-end learning but exploited the label information directly to the neural code output of the hash networks, and not the binary code which may have lead to some approximation loss. Finally, DVSH and DNH-C both performed end-to-end supervised metric-based network training in the form of cosine hinge loss and triplet ranking loss, respectively, which may not fully obtain discriminative binary codes compared to our classification-based hinge loss learning. Figures 2 show the NDCG [30] performance, a ranking metric, for both IAPRTC12 and NUS-WIDE experiment. Unlike other methods that gave the same weight if samples have least one similar label between them during training, it can be seen that our method shows the best results by a large margin which shows that our method addressed the ranking problem well by exploiting the label information fully.

[4]Authors provided their code except for DisCMH in which we implemented ourselves.

[5]Results are obtained from the respective author's papers. We used the same experimental setup as mentioned in their papers.

4. CONCLUSION

In this paper, we have proposed a cross-modal hashing network (CMHN) method for cross-modal multimedia search. Our method constructs two deep neural network to learn two sets of hierarchical nonlinear transformations so that compact binary codes are learned. We learn these deep networks through a joint binary code inference and back-propagation optimization which exploits class label information. Experimental results on two multimedia datasets have shown the effectiveness of the proposed approach.

5. REFERENCES

[1] Yunchao Gong and Svetlana Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *CVPR*, 2011, pp. 817–824.

[2] F. Shen, C. Shen, W. Liu, and H. Shen, "Supervised discrete hashing," in *CVPR*, 2015, pp. 37–45.

[3] Wei Liu, Jun Wang, Sanjiv Kumar, and Shih-Fu Chang, "Hashing with graphs," in *ICML*, 2011, pp. 1–8.

[4] Ke Jiang, Qichao Que, and Brian Kulis, "Revisiting kernelized locality-sensitive hashing for improved large-scale image retrieval," in *CVPR*, 2015, pp. 1–10.

[5] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao, "Linear cross-modal hashing for efficient multimedia search," in *ACM MM*, 2013, pp. 143–152.

[6] Jile Zhou, Guiguang Ding, and Yuchen Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *ACM SIGIR*, 2014, pp. 415–424.

[7] Guiguang Ding, Yuchen Guo, and Jile Zhou, "Collective matrix factorization hashing for multimodal data," in *CVPR*, 2014, pp. 2083–2090.

[8] Botong Wu and Yizhou Wang, "Neighborhood-preserving hashing for large-scale cross-modal search," in *ACM MM*, 2016.

[9] Di Wang, Xinbo Gao, Xiumei Wang, and Lihuo He, "Semantic topic multimodal hashing for cross-media retrieval," in *IJCAI*, 2015, pp. 3890–3896.

[10] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang, "Semantics-preserving hashing for cross-view retrieval," in *CVPR*, 2015, pp. 3864–3872.

[11] Dongqing Zhang and Wu-Jun Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *AAAI*, 2014, pp. 2177–2183.

[12] Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen, "Discriminant cross-modal hashing," in *ICMR*, 2016, pp. 305–308.

[13] Botong Wu, Qiang Yang, Wei-Shi Zheng, Yizhou Wang, and Jingdong Wang, "Quantized correlation hashing for fast cross-modal search," in *IJCAI*, 2015, pp. 25–31.

[14] Jonathan Masci, Michael M Bronstein, Alexander M Bronstein, and Jürgen Schmidhuber, "Multimodal similarity-preserving hashing," *TPAMI*, vol. 36, no. 4, pp. 824–830, 2014.

[15] Yue Cao, Mingsheng Long, Jianmin Wang, and Han Zhu, "Correlation autoencoder hashing for supervised cross-modal search," in *ICMR*, 2016, pp. 197–204.

[16] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *KDD*, 2016.

[17] Qing-Yuan Jiang and Wu-Jun Li, "Deep cross-modal hashing," *arXiv preprint arXiv:1602.02255*, 2016.

[18] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[19] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *ICML*, 2009, pp. 609–616.

[20] Andrea Vedaldi and Karel Lenc, "Matconvnet: Convolutional neural networks for matlab," in *ACM MM*, 2015, pp. 689–692.

[21] Yair Weiss, Antonio Torralba, and Rob Fergus, "Spectral hashing," in *NIPS*, 2008, pp. 1753–1760.

[22] Jun Wang, Sanjiv Kumar, and Shih-Fu Chang, "Semi-supervised hashing for large-scale search," *TPAMI*, vol. 34, no. 12, pp. 2393–2406, 2012.

[23] Wei Liu, Cun Mu, Sanjiv Kumar, and Shih-Fu Chang, "Discrete graph hashing," in *NIPS*, 2014, pp. 3419–3427.

[24] Shaishav Kumar and Raghavendra Udupa, "Learning hash functions for cross-view similarity search," in *IJCAI*, 2011, vol. 22, p. 1360.

[25] Mohammad Rastegari, Jonghyun Choi, Shobeir Fakhraei, Daume Hal, and Larry Davis, "Predictable dual-view hashing," in *ICML*, 2013, pp. 1328–1336.

[26] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *CVPR*, 2015, pp. 3270–3278.

[27] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[28] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[29] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *ICAIS*, 2010, pp. 249–256.

[30] Kalervo Järvelin and Jaana Kekäläinen, "Ir evaluation methods for retrieving highly relevant documents," in *ACM SIGIR*, 2000, pp. 41–48.