Deep Binary Reconstruction for Cross-Modal Hashing

Di Hu, Feiping Nie, *Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*—To satisfy the huge storage space and organization capacity requirements in addressing big multimodal data, hashing techniques have been widely employed to learn binary representations in cross-modal retrieval tasks. However, optimizing the hashing objective under the necessary binary constraint is truly a difficult problem. A common strategy is to relax the constraint and perform individual binarizations over the learned real-valued representations. In this paper, in contrast to conventional two-stage methods, we propose to directly learn the binary codes, where the model can be easily optimized by a standard gradient descent optimizer. However, before that, we present a theoretical guarantee of the effectiveness of the multimodal network in preserving the inter- and intra-modal consistencies. Based on this guarantee, a novel multimodal *deep binary reconstruction* model is proposed, which can be trained to simultaneously model the correlation across modalities and learn the binary hashing codes. To generate binary codes and to avoid the tiny gradient problem, a novel activation function first scales the input activations to suitable scopes and, then, feeds them to the tanh function to build the hashing layer. Such a composite function is named *adaptive tanh*. Both linear and nonlinear scaling methods are proposed and shown to generate efficient codes after training the network. Extensive ablation studies and comparison experiments are conducted for the image2text and text2image retrieval tasks; the method is found to outperform several state-of-the-art deep-learning methods with respect to different evaluation metrics.**

***Index Terms*—Cross-modal hashing, binary reconstruction.**

I. Introduction

Each modality can provide a unique way to describe the external environment while efficiently conveying perceptual information such as colorful images in vision, beautiful melody in sound, and elaborate descriptions in text. However, when these modalities focus on the same content, the different generated unimodal information will share the same high-level semantic or content [1]. For example, the speech signal and lip movements are coupled when people talk (vision + sound), and text description is correlated with image content (vision+text). These shared semantics present opportunities for correlating different modalities, which can then be used in multimodal learning. The classical multimodal task of audiovisual speech recognition [2], image-text retrieval [1], and affective analysis [3] utilize the complementary merits of multimodal data, where different modalities can make up for the limitations of the other modalities and jointly provide more valuable information than can the single modality. More works about such combination can also be found in the multimedia community [4]–[6].

In addition to the above tasks, there remains another direction that employs shared semantics as the link between different modalities, thereby providing possibilities for retrieving related items given the query of another modality. A typical scenario is to retrieve relevant images by a text query or vice versa [7], [8]. Recently, new types of cross-modal tasks, such as image2song [9], which attempts to retrieve semantically relevant songs by analyzing the image content, have emerged. Although these diverse cross-modal retrieval tasks provide interesting and effective practical application, they all suffer from the same disadvantages in terms of efficiency. Specifically, the image and text representations generated via deep networks or classical methods are usually high dimensional and real valued, thereby requiring huge storage space and computational resources. Such inefficient retrieval strongly hampers practical usage. Hence, a type of effective and efficient cross-modal retrieval method must be developed.

Fortunately, hashing is an efficient technique for addressing big data, especially in retrieval tasks. More precisely, hashing techniques aim at projecting the original high-dimensional, real-value data into short, binary codes while preserving the nearest neighborhood structure of the data points [10]. After performing the hashing projection, the obtained binary codes can vastly reduce the storage space requirements. Based on logical operations, we can perform the retrieval in milliseconds or less. Most hashing methods focus on unimodal retrieval tasks and relevant applications such as person image retrieval [11], [12], classification [13], and video retrieval [14], [15]. When there is more than one modality, the additional inter-correlation should be considered in the hashing projection. Hence, cross-modal hashing attempts to simultaneously preserve the intra- and inter-modal consistency.

Cross-modal hashing was only recently proposed, but it has become a hot field of study in terms of not only the learning

Manuscript received January 13, 2018; revised April 23, 2018 and July 1, 2018; accepted August 1, 2018. Date of publication August 27, 2018; date of current version March 22, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1107400 and in part by the National Natural Science Foundation of China under Grants 61761130079, 61772427, and 61751202. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xavier Giró-i-Nieto. *(Corresponding author: Di Hu.)*

D. Hu and F. Nie are with the School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: hdui831@mail.nwpu.edu.cn; feipingnie@gmail.com).

X. Li is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China. He was with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@nwpu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2866771

methods but also the relevant applications [16], [17]. In contrast to unimodal hashing, it usually requires different modalities to be represented in the same subspace; then, nearest neighbor searching is performed. More specifically, most cross-modal hashing techniques employ a two-stage framework. Most hashing techniques first generate the real-valued codes in a shared semantic space as classical cross-modal retrieval methods, and then, they binarize the real-valued codes of the different modalities [18]–[20]. However, such methods are usually based on the classical shallow model, where linear projection is a common selection technique for semantic space learning. Hence, the nonlinear correlation across modalities cannot be effectively learned. Fortunately, recent deep networks show a promising ability for nonlinear modeling, which then would contribute to effective high-level representation [21]. Hence, some works choose to learn the common semantic space via a shared layer across the multi-layer nonlinear projection of different modalities [22].

However, two open questions remain after applying deep networks to the cross-modal hashing problem. First, to the best of our knowledge, previous multimodal networks [1], [23] only empirically validate their ability in preserving the intra- and inter-modal consistency. Wang *et al.* [24] consider that such multimodal networks can only preserve the intra-modal correlation. This is unreliable for learning efficient codes with such empirical verification and discrepant consideration, and theoretical analysis is in need. Second, previous networks do not directly generate the binary hashing codes; they are a simple combination of conventional cross-modal networks and binarization. Although such a relaxation strategy simplifies the difficult discrete optimization problem when faced with the binary constraint, the rough binarization in the second stage could destroy the learned semantic space and result in a sub-optimal solution.

In this paper, we propose to strongly exploit the ability of multimodal networks in preserving the nearest neighborhood structure across modalities.[1] To this end, we need to solve the two aforementioned problems, i.e., utilizing a **reliable** multimodal network to **directly** encode the hashing codes of different modalities instead of the separated binarization procedure. Therefore, we make the following contributions:

- We theoretically analyze the *Multimodal Restricted Boltzmann Machine* (MRBM) model under the *Maximum Likelihood Learning* (MLL) objective, which is the **core** unit of existing multimodal networks. We prove that such unit has a promising ability for simultaneously preserving the intra- and inter-consistency across modalities.
- To integrate the classical two-stage method into one unbroken hashing function, we propose a scalable tanh activation framework, which is named *Adaptive Tanh* (ATanh). In contrast to the original tanh function, ATanh is controlled by a learnable parameter, which can to some extent avoid the tiny gradient problem when faced with large inputs and project the inputs into the discrete domain of $\{-1, 1\}$ after training. In practice, we propose two implementations of this learnable function: a linear and nonlinear interpretation. Both interpretations provide possibilities for directly generating binary codes, and the constituted hashing layer can be jointly trained with the whole network.
- Based on the proposed hashing activation function, we utilize the strong nonlinear modeling abilities of deep networks and propose to directly learn the binary hashing codes via a multimodal deep reconstruction network, called *Deep Binary Reconstruction* (DBRC). Within the DBRC, the original multimodal data are reconstructed based on the projected hamming semantic space in an unsupervised manner. The constituted hashing layer makes it possible to simultaneously learn the hashing codes and optimize the deep networks via back-propagation, which can learn more efficient binary codes than can two-stage methods.
- Extensive comparison experiments and ablation studies are conducted on three benchmark datasets. DBRC with the ATanh function achieves better codes compared to various cross-modal hashing methods on different metrics, especially the deep model methods.

We first briefly survey cross-modal hashing methods in Section II. Then, a theoretical analysis of the MRBM with the MLL objective is provided in Section III. Section IV introduces the proposed ATanh hashing activation function and two extensions, linear and nonlinear. Then, we propose the DBRC cross-modal hashing framework in Section V. Experiments are conducted for evaluation in Section VI. Section VII concludes this paper.

II. Related Work

Cross-modal hashing is similar to the classical retrieval task except for the binary constraint on the final representation. Because the discrete constraint makes the models difficult to optimize, most unsupervised hashing methods choose to directly binarize the real-valued representation learned by classical retrieval methods. More specifically, such frameworks involve two steps: First, they project different modalities into a shared low-dimensional (usually the code length) space. Then, the binarization operation (usually thresholding) is performed over the projected real-valued vector to obtain binary codes. According to the projection method, these methods can be categorized into two groups: classical linear modeling and nonlinear modeling based on deep networks [22].

A. Classical Linear Modeling

Cross-modal hashing can be viewed as a special case of unimodal hashing; hence, some off-the-shelf unimodal methods can be extended to the cross-modal scenario. To the best of our knowledge, the first proposed method, *Cross-view Hashing* (CVH) [26], extends unimodal spectral hashing (SH) [27] by appending the Hamming distance across modalities to the original spectral objective. Hence, the intra- and inter-correlation

[1] The preliminary version [25] has been accepted by ACM Multimedia 2017. In addition, the journal version proposes a unified adaptive activation framework, which extends it to linear and nonlinear conditions. Additional ablation analysis and experiments are also provided.

are both contained in the same Hamming term within this framework. In addition, *Canonical Correlation Analysis* (CCA) [28] performs as a special case of CVH, where the linear projection with maximized correlation between modalities can be learned. Further, there exist two new directions with CVH. On the one hand, in contrast to CVH, *Inter-Media Hashing* (IMH) [29] maintains the original SH objective and strengthens the modal consistency by distinctly learning a tag-semantic subspace, therein achieving impressive improvements over CVH. On the other hand, the orthogonal bases of CCA make it difficult to encode the proximity of samples. Hence, *Predictable Dual-view Hashing* (PDH) [30] proposes to ignore the orthogonal bases and learn more efficient hashing codes in a self-taught manner.

Apart from the above SH-related hashing methods, there is another category based on the matrix operation. In contrast to the subspace projection of CCA, this type of method learns the latent shared concept via matrix factorization [18], [31]. *Collective Matrix Factorization Hashing* (CMFH) [31] decomposes each modality into a modality-specific projection matrix and a latent semantic matrix while linearly reconstructing the original modality from the semantic matrix. Xu *et al.* [32] propose to learn a modality-specific linear projection and take the projected codes as the representative features for discriminative learning with labels. *Latent Semantic Sparse Hashing* (LSSH) [18] employs the factorization over the text modality but utilizes sparse coding to capture the salient structures of images. Then, a shared matrix is learned to maximally correlate the captured semantic information of these two modalities. The hashing codes for both CMFH and LSSH are obtained by applying the sign function over the shared semantic matrix. Differently, *Cross-modal Collaborative Quantization* (CMCQ) [33] proposes to align the quantized representations rather than sharing the same codes across modalities. $\mathrm{SM}^2\mathrm{H}$ [20] interprets the latent concept as the specific dictionary, where different modalities are correlated via the dictionary coefficients. To improve the retrieval quality further, Ding *et al.* [34] propose to employ the rank-order information when projecting the modalities in unified binary codes. However, all the introduced related work above actually employs linear projections for learning the shared semantics across modalities. Such shallow model limits their ability in modeling complex nonlinear correlations, but it is exactly what deep multimodal networks focus on.

B. Deep Nonlinear Modeling

1) Real Valued Representation: Deep networks have demonstrated their strong nonlinear modeling ability [24], [35], [36]. Such merits make them effective at learning sufficient high-level semantics from raw modality data, which can be applied to various applications of different modalities, e.g., object detection [37], speech recognition [38], and text translation [39]. As a result, these unimodal networks provide possibilities for learning reliable correlations across modalities. To the best of our knowledge, the Multimodal Deep Autoencoder (MDAE) [23] is the first method employing deep networks in multimodal learning. This method utilizes two separate branches of stacked *Restricted Boltzmann Machines* (RBMs) to model audio and image. Then, a shared layer is performed over the top layers of the branches. Based on the above structure, MDAE attempts to learn effective joint representations across modalities by minimizing the reconstruction error, therein achieving noticeable performance improvements in the task of *Audiovisual Speech Recognition* (AVSR). Inspired by MDAE, Srivastava *et al.* [1] propose to extend the *Deep Boltzmann Machine* (DBM) [40] into the multimodal scenario by employing the shared layer structure, which is named *Multimodal DBM* (MDBM). In addition, it is the first multimodal network proposed for retrieval tasks. Recently, some works have proposed to strengthen the shared semantic learning within this framework, possibly further improving the retrieval performance. Sohn *et al.* [41] aim to reduce the variant information across modalities, while Hu *et al.* [42] propose semantic similarity learning, which attempts to enhance the similarity between the semantics of different modalities. However, in the above methods, the shared representations are all high dimensional and real valued; hence, they require huge computational resources for comparison during retrieval.

2) Binary Valued Representation: Recently, some works have proposed to apply the hashing technique to multimodal networks and transformed the shared representation into short and binary codes. One intuitive method is to directly binarize the shared representation after training the network, which is actually the principal strategy of existing cross-modal hashing techniques. Wang *et al.* [24] employ the MDAE network for cross-modal hashing, where they impose an orthogonal regularizer on the weights of MDAE to make the joint representation more efficient. Differently, Feng *et al.* [43] and Wang *et al.* [44] propose to employ stacked autoencoders for retrieval, therein minimizing the distance between the high-level features of different modalities to preserve the inter-modal semantic and reconstructing each modality to maintain the intra-modal consistency. In addition, hashing codes are generated by applying an indicator function over the joint representation.

Although these methods utilize the advantages of deep networks for nonlinear modeling, they fail to consider the distinct binary constraints of hashing methods when fine tuning the networks. The simple combination of a multimodal network and binarization can destroy the original joint representation and make the codes inefficient. Although Courbariaux *et al.* [45] aim to make the weights and activations binary, their model still suffers from a difficult optimization process. In contrast to the aforementioned methods, our model can directly generate the hashing codes after training the networks; it is also easy to optimize.

III. Multimodal Maximum Likelihood Learning

To simultaneously preserve the inter- and intra-modal consistency, almost all multimodal deep networks [1], [23], [41] utilize the *Multimodal Restricted Boltzmann Machine* (MRBM) unit. MRBM is a special variant of RBM, which attempts to maximize the joint likelihood of different modalities. Although the effectiveness of MRBM in preserving the correlations has been verified in previous experiments [1], [2], [42], no theoretical analysis has been provided. Here, we will focus on the latter

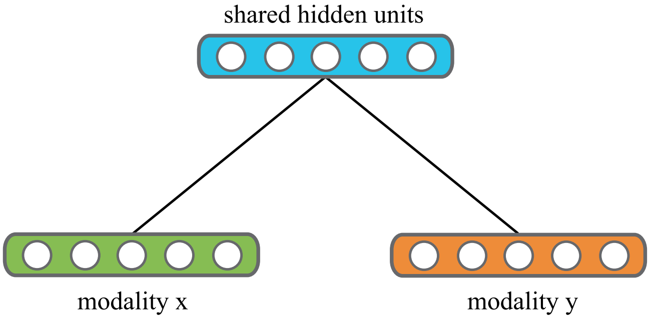


Fig. 1. An illustration of MRBM model.

and theoretically validate the ability of MRBM for cross-modal learning.

As with the energy-based network of RBM, MRBM is an undirected graphical model that consists of different modality layers and one shared hidden layer, as shown in Fig. 1. Based on the shared hidden units, it attempts to maximize the joint distribution over different modalities [1]. Concretely, let x, y, and h denote two modality layers and the hidden layer; then, the joint distribution over the layers can be written

$$P\left(\mathrm{x},\mathrm{y},\mathrm{h}\right) = \frac{1}{Z}\exp\left(-E\left(\mathrm{x},\mathrm{y},\mathrm{h}\right)\right), \tag{1}$$

where Z is the partition function and E is an energy term [1] given by

$$\begin{aligned} E\left(\mathrm{x},\mathrm{y},\mathrm{h}\right) &= -\mathrm{x}^{\mathrm{T}}\mathrm{W}^{\mathrm{x}}\mathrm{h} - \mathrm{y}^{\mathrm{T}}\mathrm{W}^{\mathrm{y}}\mathrm{h} \\ &\quad - \mathrm{x}^{\mathrm{T}}\mathrm{b}^{\mathrm{x}} - \mathrm{y}^{\mathrm{T}}\mathrm{b}^{\mathrm{y}} - \mathrm{h}^{\mathrm{T}}\mathrm{b}^{\mathrm{h}}, \end{aligned} \tag{2}$$

where W^{x} is a matrix of pairwise weights between elements of x and h and similar for W^{y}. b^{x}, b^{y}, and b^{h} are bias vectors for x, y, and h, respectively. Then, the hidden units h can be marginalized out from the distribution to obtain the joint likelihood $P(\mathrm{x},\mathrm{y})$ [1]:

$$P(\mathrm{x},\mathrm{y}) = \sum_{\mathrm{h}}\exp\left(-\mathrm{E}(\mathrm{x},\mathrm{y},\mathrm{h})\right)/\mathrm{Z}. \tag{3}$$

To maximize the joint distribution $P(\mathrm{x},\mathrm{y})$, one common strategy is to use *Contrastive Divergence* (CD) [46] or *Persistent CD* (PCD) [47] to approximate the gradient for optimizing the MRBM and then fine tune it with *Stochastic Gradient Descent* (SGD) under supervision. This is the typical *Maximum Likelihood Learning* (MLL) for MRBM. However, when applying the MLL objective over the current model, what does the model actually model? Can the MLL objective model the cross-modal correlation?

To answer the aforementioned questions, we attempt to decouple the original objective. Let $P_\theta(\mathrm{x},\mathrm{y})$ denote the MRBM joint distribution parameterized by $\theta = \{\mathrm{W}^*, \mathrm{b}^*\}$, and let $P_D(\mathrm{x},\mathrm{y})$ denote the data-generating distribution. The MLL objective of MRBM can be re-written as follows:

$$\begin{aligned} MLL &= E_{P_D(\mathrm{x},\mathrm{y})}\left[\log P_\theta\left(\mathrm{x},\mathrm{y}\right)\right] && (4) \\ &= E_{P_D(\mathrm{x})}\left[E_{P_D(\mathrm{y}|\mathrm{x})}\log P_\theta\left(\mathrm{x}\right)\right] && (5) \\ &\quad + E_{P_D(\mathrm{x})}\left[E_{P_D(\mathrm{y}|\mathrm{x})}\log P_\theta\left(\mathrm{y}|\mathrm{x}\right)\right] && (6) \\ &= -E_{P_D(\mathrm{x})}\left[E_{P_D(\mathrm{y}|\mathrm{x})}\log\frac{P_D\left(\mathrm{x}\right)}{P_\theta\left(\mathrm{x}\right)}\right] && (7) \\ &\quad - E_{P_D(\mathrm{x})}\left[E_{P_D(\mathrm{y}|\mathrm{x})}\log\frac{P_D\left(\mathrm{y}|\mathrm{x}\right)}{P_\theta\left(\mathrm{y}|\mathrm{x}\right)}\right] + C && (8) \\ &= -E_{P_D(\mathrm{x})}\left[E_{P_D(\mathrm{y}|\mathrm{x})}\log\frac{P_D\left(\mathrm{y}|\mathrm{x}\right)}{P_\theta\left(\mathrm{y}|\mathrm{x}\right)}\right] && (9) \\ &\quad - E_{P_D(\mathrm{x})}\left[\log\frac{P_D\left(\mathrm{x}\right)}{P_\theta\left(\mathrm{x}\right)}\right] + C && (10) \\ &= -\underbrace{E_{P_D(\mathrm{x})}\left[KL(P_D\left(\mathrm{y}|\mathrm{x}\right)||P_\theta\left(\mathrm{y}|\mathrm{x}\right))\right]}_{\text{cross modalities}} && (11) \\ &\quad - \underbrace{KL(P_D\left(\mathrm{x}\right)||P_\theta\left(\mathrm{x}\right))}_{\text{single modality}} + C && (12) \end{aligned}$$

where the constant $C = E_{P_D(\mathrm{x})}\left[E_{P_D(\mathrm{y}|\mathrm{x})}\log P_D\left(\mathrm{x}\right)\right] + E_{P_D(\mathrm{x})}\left[E_{P_D(\mathrm{y}|\mathrm{x})}\log P_D\left(\mathrm{y}|\mathrm{x}\right)\right]$ is independent of θ. Note that the above formula is decoupled w.r.t. modality x, which can also be re-written w.r.t. y. According to the above equations, we can find that the MLL objective can be decoupled into two terms: one related to the *Kullback-Leibler* (KL) divergence between the distributions of modality x and the other being the conditional probability of the cross-modalities under the expectation of P_D. Hence, maximizing the joint distribution $P_\theta(\mathrm{x},\mathrm{y})$ under the expectation of $P_D\left(\mathrm{x},\mathrm{y}\right)$ is equal to simultaneously modeling the unimodal and cross-modal data distribution. In other words, when minimizing the two terms of the Kullback-Leibler divergence, MRBM actually learns to preserve both the intra- and inter-modal consistency. Hence, both the experimental results and theoretical analysis verify that the multimodal networks based on MRBM have the ability to satisfy the objectives of cross-modal hashing. Note that the analysis w.r.t. the multimodal MLL objective does not solely focus on the MRBM network. Specifically if the properties of the networks meet the requirements of joint likelihood modeling and conditional independence structure, the analysis is also suitable for them.

Moreover, the MLL analysis also provides insights in designing the algorithm and refining the model. For example, to enhance the shared information across modalities, we can pay greater attention to the KL term of the conditional probability by regularizing the similarity across modalities [42] or reduce the unimodal importance by ignoring the KL term of the marginal distribution when applying the MLL objective [41].

IV. Binary Activation Function

Although the theoretical analysis confirms the multimodal learning capacity of MRBM, the joint representation is real valued instead of binary valued. Hence, this multimodal unit is not completely suitable for cross-modal hashing. In this section, we

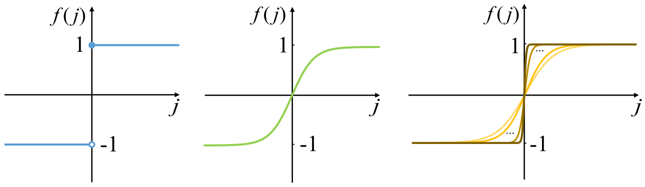


Fig. 2. The comparison among sign, tanh, and ATanh functions. ATanh is composed of a scaling function and the standard tanh function, which can scale the inputs into different scopes so that the final tanh takes various patterns.

introduce a hashing layer upon the joint representation, which can be trained with the whole network and directly generate binary codes. Such merits come from a novel binary activation framework, and two practical implementations within the framework are introduced in the following sections.

A. *Adaptive Tanh*

A hashing technique requires the codes to be binary, i.e., $\{-1, 1\}$. However, multimodal networks under binary constraints are difficult to optimize. Hence, most previous work [24], [44] chose to relax the constraints and separately applied the sign function over the joint representation after training the network. For example, the activation value of the tanh function is binarized via the mask of the sign function.[2] However, numerous activations actually fall into the interval of $(-1, 1)$, and the resulting quantization error could destroy the learned multimodal structure [25]. On the other hand, if the activation is sufficiently large, there will be a very small gradient passed to the lower layers, which makes it difficult to optimize the network [48]. These two problems make it difficult to obtain efficient hashing codes.

Fig. 2 shows a comparison of the sign function and tanh function. Intuitively, a difference only exists in the transition interval between -1 and 1. One technique is to scale up the inputs; then, the tanh function can approach the sign function. Similarly, the inputs can also be scaled down to reduce the influence of a small gradient. Hence, we can first employ a scaling function to transform the inputs to an appropriate scale and then take it as the new inputs to tanh, as in the example in Fig. 2. Formally, a composite activation function is proposed for generating the hashing codes, which is written as follows:

$$f\,(\mathrm{s}) = \tanh(g(\mathrm{s};\alpha)), \tag{13}$$

where s is the activations of the previous layer and $g(\mathrm{s};\alpha)$ is the scaling function parameterized by α. The proposed composite function is named *Adaptive Tanh* (ATanh), which can be trained with the whole network.

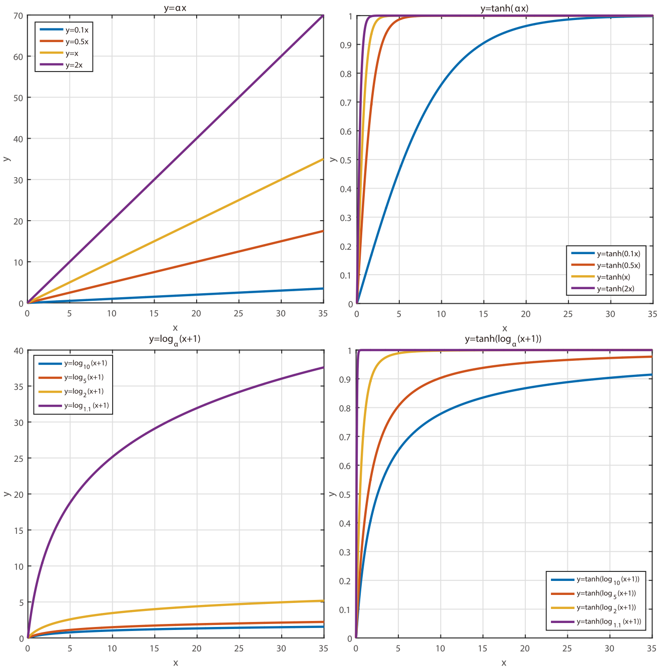


Fig. 3. An illustration of the linear and nonlinear composite function of tanh (only the first quadrant is shown). The top-left figure is an illustration of the linear scaling function, and the top-right figure is the corresponding modified tanh function. When the scaling parameter α decreases, $\tanh(\alpha\mathrm{s})$ becomes closer to the sign function, as is the case for the nonlinear ones in the bottom two figures.

However, it remains difficult to guarantee that the activations of ATanh fall into the binary codes after training the deep networks. We hope that the inputs are shrunk at the beginning; then, they are adaptively amplified when training the network. Finally, ATanh can approach the sign function and generate the binary hashing codes. Hence, a regularization term is applied to the composite function:

$$f\,(\mathrm{s}) = \tanh(g(\mathrm{s};\alpha)) + \zeta(\alpha), \tag{14}$$

where $\zeta(\alpha)$ is the regularization term providing a convenient way to control the magnitude of α, which prompts the function to be closer to the sign function. In practice, the function $g(\mathrm{s};\alpha)$ can be implemented with respect to the scaling method. In the following sections, we introduce two different methods that can be effectively employed for scaling.

B. *Linearized ATanh*

The simplest and most direct scaling method is linear scaling. According to the zoom coefficient, the transition area of tanh around zero can be stretched or shrunk. Hence, the scaling function is written as $g(\mathrm{s}) = \alpha\mathrm{s}$; then, Eq. (13) becomes

$$f(\mathrm{s}) = \tanh(\alpha\mathrm{s}). \tag{15}$$

The standard tanh function can be viewed as a special case of Eq. (15) when $\alpha = 1$. In Fig. 3, the top two figures show different scaled tanh functions. It is easy to find that the linearized tanh with larger α is closer to the y-axis. Hence, it is necessary to enlarge the value of α when training the network. Then, Eq. (14) becomes

$$f(\mathrm{s}) = \tanh(\alpha\mathrm{s}) + \lambda\left\|\alpha^{-1}\right\|_2^2, \tag{16}$$

where λ is a regularization constant. By minimizing the regularization term of $\left\|\alpha^{-1}\right\|_2^2$, α can be gradually increased so that the

[2] When the activation function is a sigmoid, there is an offset of -0.5 for the activation values.

final activation of $f(\mathrm{s}) = \tanh(\alpha \mathrm{s})$ approaches the sign function and has the ability to generate binary codes. In addition, the linearized ATanh is an element-wise function; thus, different α can be learned for different bits. In other words, 32 functions can be simultaneously learned for 32 $bits$ hashing codes, which makes the codes more adaptable compared with the constant sign function.

The linearized ATanh can be jointly trained with other layers via back-propagation. Hence, the partial derivative with respect to α_i can be simply derived by the chain rule:

$$\frac{\partial \varepsilon}{\partial \alpha_i} = \frac{\partial \varepsilon}{\partial f(\mathrm{s}_i)} \frac{\partial f(\mathrm{s}_i)}{\partial \alpha_i}, \tag{17}$$

where ε denotes the objective function (e.g., reconstruction error). Eq. (17) consists of two gradient terms: the gradient to the current hashing layer and the derivative of the ATanh function. The derivative rule for the composite function is employed for the second term:

$$\frac{\partial f(\mathrm{s}_i)}{\partial \alpha_i} = \left(1 - \tanh^2(\alpha_i \mathrm{s}_i)\right) \mathrm{s}_i - 2\lambda \alpha_i^{-3}. \tag{18}$$

At this point, we can train the hashing layer with the whole network by employing a standard SGD optimizer. After training the multimodal network, the binary hashing codes can be directly generated in the hashing layer.

C. Non-Linearized ATanh

Although the linearized ATanh can scale the inputs, it still suffers from the problem of tiny gradients. This is because the non-binary area (i.e., the transition area) of the tanh function mainly lies in the interval $(-3, 3)$ on the x-axis. Regardless of the size of the scaling parameter α, the non-binary area can only be linearly extended to a limited size, and the external area remains unchanged. Hence, can we employ a nonlinear projection to solve this problem?

The nonlinear scaling function should project all the original inputs into the non-binary area of the tanh function; then, almost all activations can utilize the gradient passed from the higher layers. Moreover, the resulting composite function should also have the ability to generate binary codes. Hence, we propose to employ the logarithmic function as the scaling function, where the scaling parameter α is applied as the base:

$$g(\mathrm{s}) = \log_\alpha(\mathrm{s} + 1), \quad \alpha > 1. \tag{19}$$

Here, we restrict α to be greater than 1. As shown in Fig. 3, the bottom two figures show different tanh functions scaled by different logarithms. When α increases, the inputs far away from the zero point can still use the gradients compared with the linear scaling method. Similarly, the composite ATanh function should also approach the sign function after training. In contrast to the linear ATanh function, the nonlinear ATanh function requires that α gradually decrease. Hence, we append the l_2 norm to the parameter α; then, Eq. (14) becomes

$$f(\mathrm{s}) = \tanh(\log_\alpha(\mathrm{s}{+}1)) + \lambda \left\|\alpha\right\|_2^2. \tag{20}$$

Compared with the linear case, the activations in Eq. (20) have an offset of 1 due to the property of the logarithmic function.

The optimization of Eq. (20) shares the same procedure as the linear case. The gradient to the current hashing layer is first derived based on the chain rule, and then, the derivative w.r.t. α_i in Eq. (20) is derived as follows:

$$\frac{\partial f(\mathrm{s}_i)}{\partial \alpha_i} = -\left[1 - \tanh^2(g(\mathrm{s}_i))\right] \frac{\ln(\mathrm{s}_i + 1)}{\alpha_i \ln^2 \alpha_i} + 2\lambda \alpha_i. \tag{21}$$

Note that the constraint of $\alpha > 1$ is implemented via $\log_{\alpha+1}(\mathrm{s} + 1)$ in practice. To this end, we can optimize the hashing layer based on Eq. (21) within the multimodal networks.

V. Deep Binary Reconstruction Network

The performance of the MRBM multimodal unit in preserving the intra- and inter-modal consistency has been theoretically guaranteed, and it has become possible to directly generate binary codes after training the networks. Hence, to integrate the above two contributions for cross-modal hashing, we propose a novel multimodal *Deep Binary Reconstruction* (DBRC) network that can directly project original real-valued, high-dimensional multimodal data into binary, short hashing codes, as shown in Fig. 4. Specifically, to capture the unimodal manifold, multiple layers of nonlinear projections are first employed for modeling each modality. This modality-specific network can encode the unimodal data into low-dimensional representations [21], [49]. Then, the joint representation across the unimodal representations is learned via the MRBM model (i.e., Fig. 1). To binarize the real-valued representation of MRBM, one hashing layer is appended above the joint representation, where the element-wise ATanh function is performed over the activations of the shared layer. Based on the shared hashing layer across modalities, we attempt to reconstruct the original modalities. Hence, the overall objective of DBRC becomes

$$L = \left\|\tilde{\mathrm{x}} - \mathrm{x}\right\|_2^2 + \left\|\tilde{\mathrm{y}} - \mathrm{y}\right\|_2^2, \tag{22}$$

where $\tilde{\mathrm{x}}$ and $\tilde{\mathrm{y}}$ denote the reconstructed modality x and y, respectively. Because the ATanh function is derivable, the whole network can be trained via back-propagation, and we directly generate the embedded binary codes.[3]

Cross-modal hashing attempts to generate binary codes for each modality and then perform the retrieval by comparing the codes of different modalities. Since all the modalities are available in the training phase, we can directly utilize the property of DBRC to generate identical hashing codes. However, there is only one modality available in the testing scenario and thus, the complete structure of DBRC is not suitable. Inspired by Ngiam *et al.* [23] and Hu *et al.* [42], it is possible to reconstruct both modalities based on only one modality input, which is named the unimodal reconstruction structure. Because this structure retains the MRBM model and applies the multimodal learning in the reconstruction part, is still effective in learning the joint representation [23], [42]. Concretely, to generate the hashing codes for a specific modality, we first set the other

[3] Although the proposed ATanh function is very close to the sign function after training, very few activations fall into the interval $(-1, 1)$. Hence, we simply perform binarization over these activations.

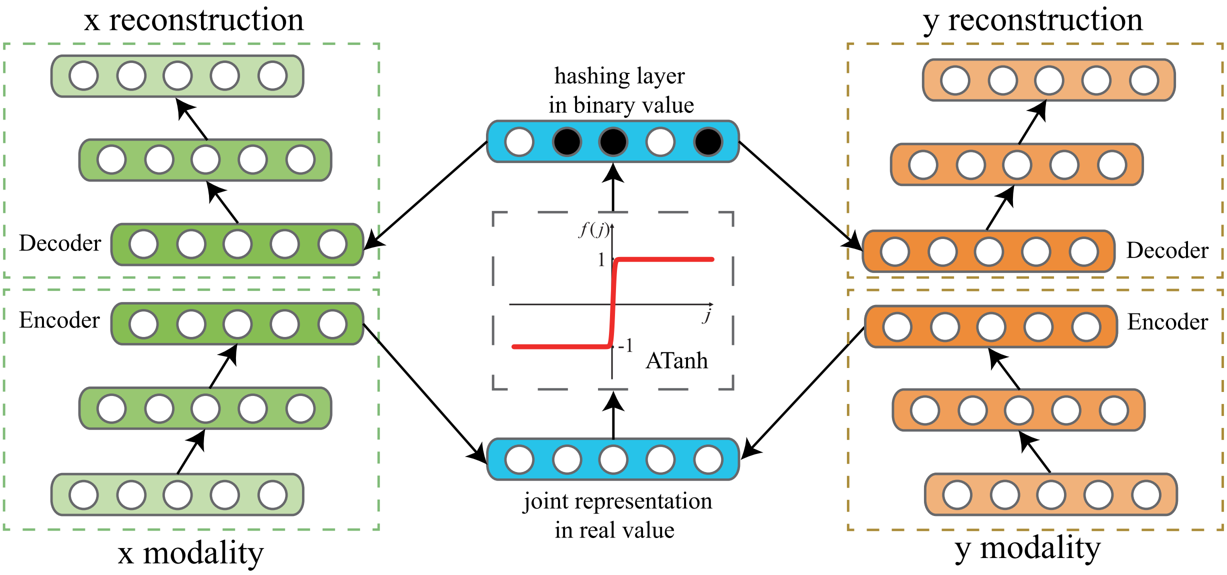


Fig. 4. The proposed deep binary reconstruction network. The joint representation across the two modality-specific networks is adaptively binarized into the hashing layer by performing the ATanh activation. The whole network is trained to minimize the reconstruction error based on the shared binary representation.

modality to zero; then, we combine these two modalities (actually, only one modality) as the inputs to the DBRC network. Eq. (22) is finally applied as the objective to reconstruct both original modalities. In the testing phase, the hashing codes for query items can be directly generated by feeding them into the above network and extracting the outputs of the intermediate hashing layer. In practice, we find that the pre-trained model with both available modalities in the training phase provides better initialization for such reconstruction model, as it captures the manifold structure of the absent modality. Hence, we choose to initialize the unimodal reconstruction model from the complete DBRC model and then fine tune it with the unimodal input data.

Concretely, there are two pathways within the proposed model: one for the image and one for the text. Each pathway consists of an encoder and a decoder, where the encoder is a 3-layer network (n-128-512, n is the dimension of a raw feature) and the decoder takes on a 512-128-n settings. These two pathways are connected by the shared layers, whose unit number is set to the code length. ReLU is selected as the activation function of the whole network except for the hashing layer, where the proposed ATanh function is chosen for learning binary codes. The hyper-parameter of λ is set to 0.001 for all the datasets. Since both the linear and nonlinear version of ATanh are derivable, RMSprop is adopted as the optimizer, which adaptively rescales the step size for the update according to the gradient history, where the parameters are set as the learning rate $l = 0.001$ and the weight decay $\rho = 0.9$.

VI. Experiments

In this section, we focus on the image2text (I2T) and text2image (T2I) cross-modal hashing task. Two sets of experiments are performed for evaluating the proposed DBRC: one is an ablation study focusing on the comparison among activation functions, and the other is a study with other hashing methods. Different code lengths are also considered.

A. Dataset

Three benchmark datasets, Wiki[4] [50], FLICKR-25K[5] [51], and NUS-WIDE[6] [52], are chosen for the evaluation.

Wiki is an image-text dataset collected from Wikipedia's "featured article". There are 2,866 pairs in the dataset. For each pair, the image modality is represented as 128-dimensional SIFT descriptor histograms, and text is expressed as 10-dimensional semantic vectors via latent Dirichlet allocation model. These pairs are annotated with one of 10 topic labels. In this paper, we choose 25% of the dataset as the query set and the remainder as the retrieval set.

FLICKR-25K is an image collection from Flickr, where 25,000 images are associated with multiple textual tags (text). The average number of tags for each image is approximately 5.1 [1]. In addition, these image-tag pairs are annotated by 24 provided labels. Following the setting in [53], we select the textual tags that appear more than 20 times and retain the valid pairs. The left images are represented with a 150-dimensional edge histogram, and the texts are expressed as a 500-dimensional tagging vector. Here, we take 5% of the dataset as the query set and the remainder as the training set.

NUS-WIDE dataset consists of 269,648 multi-label images. Each image is also associated with multiple tags (6 on average). The image-tag pairs are annotated with 81 concept labels. Among these concepts, the 10 most common concepts are considered in our experiments. The images are represented as 500-dimensional bag-of-words based on the SIFT descriptor.

[4] http://www.svcl.ucsd.edu/projects/crossmodal/
[5] http://press.liacs.nl/mirflickr/
[6] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

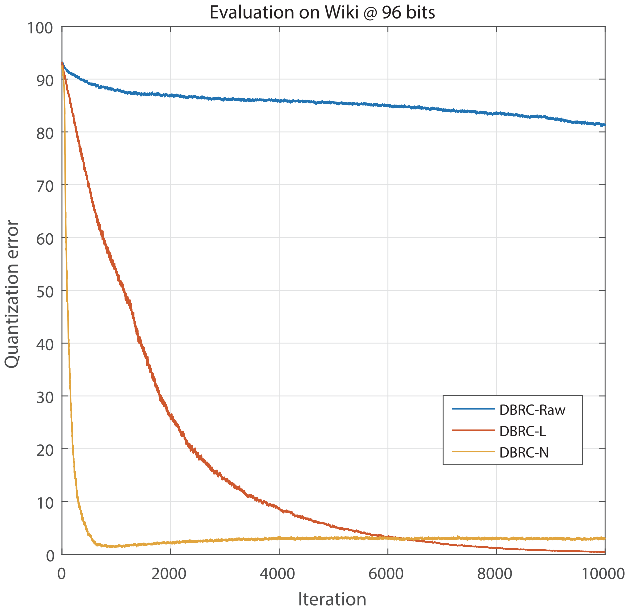


Fig. 5. The comparison results in reducing the quantization error of binarization, where the quantization error is calculated by binarizing the activations in the hashing layer after each iteration. Obviously, the proposed DBRC-L and DBRC-N achieve faster convergence than the DBRC of the raw tanh function.

The textual tags are expressed with a 1000-dimensional tag occurrence vector. A total of 4000 image-tag pairs are uniformly sampled as the query set, and the remaining pairs serve as the training set.

B. Metric

Two evaluation metrics are adopted, e.g., Hamming ranking and hash lookup. Specifically, Hamming ranking focuses on the quality of all the retrieved items, while hash lookup simply addresses the top retrieved items. Hence, different evaluation metrics are considered, where *Mean Average Precision* (MAP) is computed based on the Hamming distance to a query for the Hamming ranking, and the retrieved items within a Hamming ball of radius 2 to the query are evaluated with the $\{Precision, F-measure\}$ metrics for the hash lookup. In addition, the ground truth of relevant items to a query is defined as whether they have at least one common label.

C. Ablation Study

1) Convergence: The proposed ATanh activation function attempts to directly generate the binary codes by progressively approaching the sign function when training the network. Hence, it is necessary to analyze the approaching speed of such functions. To quantify the convergence, the quantification error is calculated after each epoch, which is obtained by applying the sign function over the activations of the current hashing layer. In addition, the proposed DBRC with linear ATanh (DBRC-L) and nonlinear ATanh (DBRC-N) is analyzed, and the original tanh function (DBRC-Raw) is included.

Fig. 5 shows a comparison of the convergence performances among the DBRCs with different hashing activation functions, which is applied to the Wiki dataset with 96-bit hashing codes. In Fig. 5, the nonlinear ATanh enjoys the fastest convergence, the linear function is the second fastest, and the tanh function is the slowest. The superior performance of DBRC-N comes from the efficient regularization term in Eq. (20). Although DBRC-L also applies a penalty term to α, its magnitude is substantially smaller than those in DBRC-N. This is because the $2\lambda\alpha_i$ term in Eq. (21) contributes larger gradients than does $2\lambda\alpha_i^{-3}$ in Eq. (18) in regulating α. Moreover, the other elements in both equations also provide similar contributions. Concretely, these elements in Eq. (21) and Eq. (18) lie in the opposite directions in regularizing the parameter α, and the absolute value of the first term in Eq. (21) is substantially smaller than the corresponding one in Eq. (18) in the early stage of the optimization; hence, DBRC-N achieves faster convergence than DBRC-L. Although the naive tanh function also reduces the quantization error, it is much slower than the other functions and has a serious impact on generating the final binary codes. In addition, we find that DBRC-L ultimately achieves a lower quantization error than DBRC-N. Such phenomenon has two causes. On the one hand, the nonlinear scaling function of the logarithm first compresses the inputs to a lower scope, which allows them to best utilize the non-binary area of the tanh function. The linear ATanh function fails to apply such projection; hence, its activations mostly fall into binary values and take lower quantization errors compared with the nonlinear function, as shown in Fig. 3. On the other hand, along with the network optimization, the scaling parameter α of DBRC-N decreases. Hence, the regularization term $2\lambda\alpha_i$ contributes a smaller gradient in regularizing α, which results in the stranded quantization error. The following experiments show that such cases degrade the final performance but still outperform other methods.

2) Activation Comparison: To validate the effectiveness of the proposed ATanh function in learning the binary codes, different variants of ATanh are compared. Among these variants, DBRC-C [54] utilizes a fixed sequence of α for training the network step by step, where the network trained with the previous α provides better initialization for the training with the next α. The final α is the largest (linear) or smallest (nonlinear), which transforms the standard tanh function into an approximated sign function. In addition, DBRC-P is the same as DBRC but does not include the regularization term, i.e., Eq. (13). Hence, DBRC and DBRC-P contain learnable hashing activation functions, while DBRC-C does not. Both the linear and nonlinear versions of these variants are considered to provide a thorough comparison. The experiments are conducted on the Wiki and FLICKR-25K dataset, and the Hamming ranking is adopted as the evaluation metric. Different code lengths are also considered.

Table I shows the comparison results among DBRC, DBRC-C, and DBRC-P, where 'L' and 'N' denote the linear and nonlinear implementation, respectively. There are three points on which we should focus. First, DBRC and DBRC-P usually outperform DBRC-C in the two tasks. This performance indicates that the learnable activation function contributes more efficient codes for the reconstruction model. This is because ATanh can adaptively learn element-wise binarization functions according to the low-dimensional shared subspace across modalities.

TABLE I
THE ABLATION COMPARISON AMONG DIFFERENT VARIANTS OF ATANH, WHICH ARE EXPLOITED IN THE DBRC NETWORK. ALL THE EXPERIMENTS ARE CONDUCTED ON THE WIKI AND FLICKR-25K DATASETS WITH VARYING CODE LENGTHS, WHERE THE HAMMING RANKING PERFORMANCE (IN MAP) IS EMPLOYED FOR THE EVALUATION

Task	Dataset	Wiki					FLICKR-25K				
	Code Length	8 $bits$	16 $bits$	48 $bits$	64 $bits$	96 $bits$	8 $bits$	16 $bits$	48 $bits$	64 $bits$	96 $bits$
I2T	DBRC-C-L	0.2219	0.2199	0.2234	0.2379	0.2483	0.5804	0.5816	0.5877	0.5868	0.5878
	DBRC-P-L	0.2308	0.2500	0.2616	0.2565	0.2632	0.5803	0.5855	0.5880	0.5894	0.5880
	DBRC-L	0.2327	**0.2534**	**0.2674**	**0.2686**	**0.2736**	**0.5866**	0.5873	**0.5914**	0.5902	0.5923
	DBRC-C-N	0.2134	0.2150	0.2146	0.2301	0.2201	0.5793	0.5871	0.5893	0.5857	0.5806
	DBRC-P-N	0.2227	0.2287	0.2345	0.2394	0.2615	0.5845	0.5858	0.5896	0.5916	0.5931
	DBRC-N	**0.2346**	0.2338	0.2576	0.2502	0.2653	0.5863	**0.5922**	0.5905	**0.5954**	**0.5954**
T2I	DBRC-C-L	0.4342	0.5165	0.5419	0.5351	0.5424	0.5841	0.5873	0.5808	0.5943	0.5901
	DBRC-P-L	0.4650	0.5382	0.5508	0.5336	0.5469	0.5785	0.5826	0.5859	0.5883	0.5867
	DBRC-L	**0.4868**	**0.5439**	**0.5538**	**0.5476**	**0.5520**	**0.5941**	0.5883	**0.5941**	**0.5962**	0.5951
	DBRC-C-N	0.4532	0.4941	0.5333	0.5401	0.5368	0.5875	0.5894	0.5871	0.5915	0.5860
	DBRC-P-N	0.4539	0.4839	0.5238	0.5308	0.5317	0.5803	0.5924	0.5873	0.5927	0.5931
	DBRC-N	0.4734	0.5070	0.5361	0.5376	0.5437	0.5935	**0.5938**	0.5901	0.5938	**0.5964**

TABLE II
THE COMPARISON RESULTS OF DIFFERENT CROSS-MODAL METHODS WITH RESPECT TO HAMMING RANKING (MAP). THREE BENCHMARK DATASETS ARE EMPLOYED FOR THE EVALUATION

Task	Dataset	Wiki				FLICKR-25K				NUS-WIDE			
	Code Length	16 $bits$	32 $bits$	64 $bits$	128 $bits$	16 $bits$	32 $bits$	64 $bits$	128 $bits$	16 $bits$	32 $bits$	64 $bits$	128 $bits$
I2T	IMH [28]	0.1593	0.1477	0.1420	0.1291	0.5621	0.5643	0.5649	0.5642	**0.4187**	0.3975	0.3778	0.3668
	CVH [25]	0.1993	0.1889	0.1803	0.1782	0.5815	0.5756	0.5710	0.5677	0.3888	0.3744	0.3621	0.3537
	CMFH [18]	0.2126	0.2208	0.2322	0.2337	0.5721	0.5740	0.5739	0.5736	0.3443	0.3438	0.3454	0.3461
	LSSH [17]	0.2122	0.2260	0.2155	0.2297	0.5779	0.5795	0.5848	0.5878	0.3891	0.3910	0.3977	0.3949
	Corr-Full-AE [41]	0.1802	0.1937	0.1911	0.2014	0.5557	0.5551	0.5583	0.5553	0.3468	0.3468	0.3470	0.3410
	DMHOR [23]	0.1919	0.1841	0.1847	0.1877	0.5848	0.5810	0.5842	0.5851	0.3657	0.3620	0.3678	0.3590
	DBRC-L	**0.2534**	**0.2648**	**0.2686**	**0.2878**	0.5873	0.5898	**0.5902**	0.5907	0.3939	**0.4087**	**0.4166**	**0.4165**
	DBRC-N	0.2338	0.2518	0.2502	0.2570	**0.5922**	**0.5922**	0.5854	**0.5910**	0.3927	0.4038	0.4097	0.4023
T2I	IMH [28]	0.1417	0.1297	0.1243	0.1105	0.5624	0.5643	0.5651	0.5648	0.4053	0.3892	0.3758	0.3627
	CVH [25]	0.1652	0.1582	0.1512	0.1469	0.5817	0.5761	0.5715	0.5681	0.3822	0.3697	0.3592	0.3519
	CMFH [18]	0.4830	0.5147	0.5338	0.5370	0.5673	0.5693	0.5681	0.5682	0.3506	0.3509	0.3524	0.3547
	LSSH [17]	0.4992	0.5245	0.5326	0.5395	0.5874	0.5926	0.5957	0.5964	0.4115	0.4162	0.4229	0.4198
	Corr-Full-AE [41]	0.1410	0.1262	0.1366	0.1483	0.5576	0.5545	0.5576	0.5567	0.3385	0.3438	0.3390	0.3382
	DMHOR [23]	0.4272	0.4874	0.4916	0.4818	0.5664	0.5622	0.5540	0.5653	0.3724	0.3613	0.3498	0.3401
	DBRC-L	**0.5439**	**0.5377**	**0.5476**	**0.5488**	0.5883	**0.5963**	**0.5962**	**0.5975**	0.4249	**0.4294**	**0.4381**	**0.4427**
	DBRC-N	0.5070	0.5264	0.5376	0.5406	**0.5938**	0.5952	0.5938	0.5903	**0.4251**	0.4205	0.4277	0.4359

Second, DBRC performs better than DBRC-P for both the linear and nonlinear version. This is because the regularization term is not used within DBRC-P, which results in the large quantization error, as the results show in Fig. 5. In addition, direct binarization over the representation can destroy the learned joint distribution. Hence, DBRC can learn better hashing codes. Third, the linear function tends to generate better codes than the nonlinear function under all variants of DBRC. Although the latter variant achieves faster convergence while reducing the quantization error, the converged error still decreases the final performance.

D. Comparison Experiments

In this section, the proposed DBRC is compared with several unsupervised cross-modal hashing methods; an analysis of the sensitivity of the hyper-parameter λ is also provided.

1) Compared Methods: The compared methods involve the linear and nonlinear models, where the linear models are IMH [29], CVH [26], CMFH (UCMFH) [19], and LSSH [18], and the nonlinear deep models are Corr-Full-AE [43] and Deep Multimodal Hashing with Orthogonal Regularization (DMHOR) [24]. Among these methods, we utilize the source codes of IMH, CVH, CMFH, and LSSH provided by the authors for comparison, while the remaining two deep models are carefully implemented by the authors. Corr-Full-AE is a full version of Corr-AE [43], but the concrete network architecture and optimization method are not provided in the original paper. For fairness, we employ the same unimodal network for comparison. Different strategies are employed for the optimization, and the best strategy is selected. Since Corr-Full-AE suffers from issues of falling into local optima and tends to

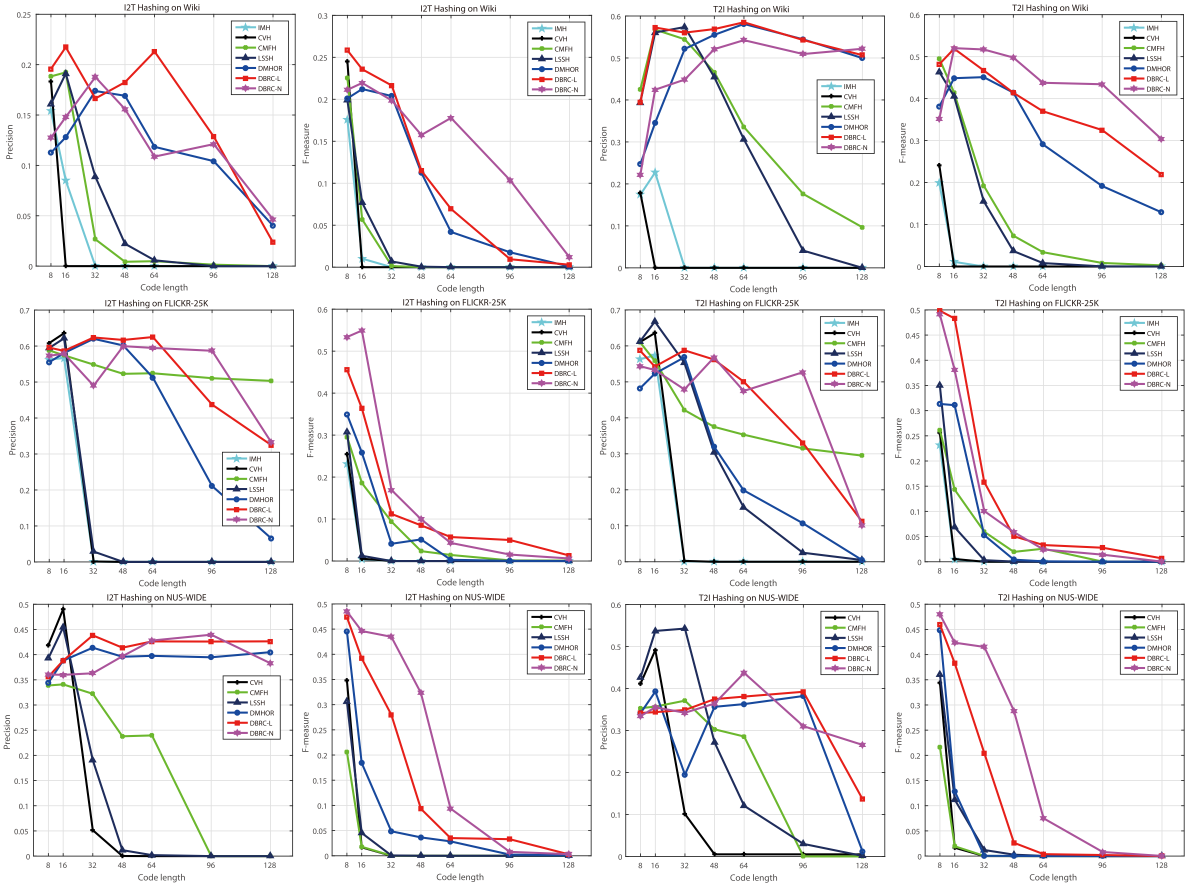


Fig. 6. The hash lookup performance (in terms of Precision and F-measure) of different cross-modal hashing methods on the Wiki, FLICKR-25K, and NUS-WIDE datasets. Two cross-modal retrieval tasks, I2T and T2I, are applied, and different code lengths are considered. Clearly, our proposed DBRC-L and DBRC-N almost achieve the top two performances in all comparisons.

learn similar codes for all the data, only the Hamming ranking is adopted for evaluation. For DMHOR, the network and hyper-parameter settings are well determined and utilized in the experiments.

2) Experimental Results: In the experiments, we focus on the image2text and text2image retrieval task. Table II shows the comparison performance in terms of the Hamming ranking. Four points should be indicated according to the MAP results. First, the spectral-based models (IMH and CVH) are comparable to the linear matrix-factorization models (CMFH and LSSH) on the FLICKR-25K and NUS-WIDE datasets, but they perform much worse on the Wiki dataset. This situation may come from the difficulty in capturing the intrinsic manifold structure of the small size of the Wikipedia data. By contrast, our proposed DBRC can effectively model and encode the data into efficient codes. Second, although deep models have a stronger ability for modeling complex images and text data, Corr-Full-AE and DMHOR do not outperform the shallow models in many cases. The primary reason for this is that the shared layers within these deep models are not fully activated, which results in an inevitable quantization error. On the other hand, they also suffer from unbalanced activations [21], which means that the activation probabilities do not follow a symmetric distribution. Hence, the directly applied binarization over these activations results in inefficient codes [55]. Third, our proposed DBRC-L and DBRC-N perform the best among these methods except in a few case. This is because DBRC integrates the conventional two-stage strategy into one step for generating the binary codes; this unified framework can seek out a better solution than the individual strategies. Fourth, we can find that DBRC-L achieves a better MAP score than DBRC-N in most instances. As discussed in the ablation study, DBRC-N achieves efficient convergence in generating binary codes but fails to reduce the quantization error to absolute zero, as shown in Fig. 5. Hence, the additional binarization has to be considered for the activations, although DBRC-N still shows a noticeable superiority over other models.

In addition to the Hamming ranking evaluation, the hash lookup metric is also considered. Fig. 6 shows the comparison results in terms of Precision and F-measure.[7] Although the two proposed models remain top performers among all the methods, some other results still should be noted and analyzed. First, in contrast to the cases of the Hamming ranking, the nonlinear methods outperform the linear methods, especially in terms

[7] Due to the limited memory of our desktop personal computer, the reported MAP score of IMH in [22] is adopted, but the hash lookup results are not supported in [22]; hence, IMH is not considered in the hash lookup evaluation on the NUS-WIDE dataset.

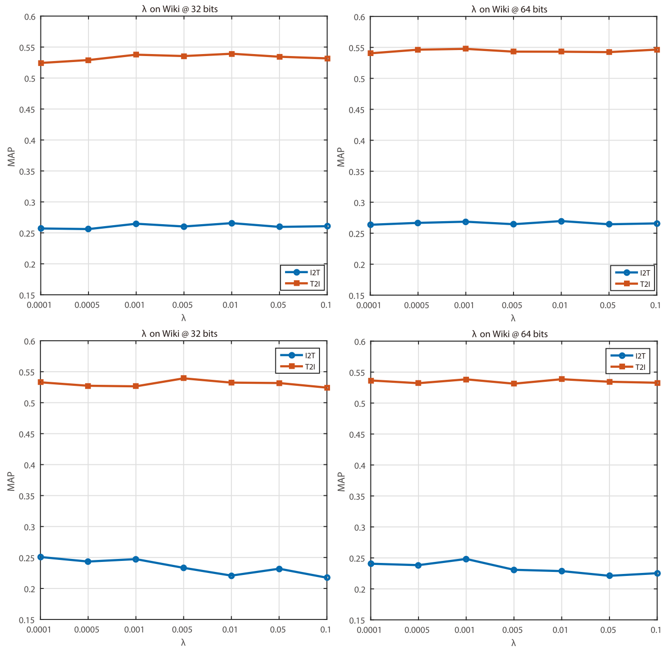


Fig. 7. The parameter sensitivity analysis of λ in the ATanh activation function. The top two figures are the results of DBRC-L, and the bottom two figures are the results of DBRC-N.

of F-measure. The distinction comes from the different evaluation metrics. Specifically, the Hamming ranking addresses the whole retrieved sequence, while the hash lookup only focuses on the top retrieved items within a specific Hamming ball. Hence, the nonlinear models are skilled at encoding the most discriminative items but tend to mix the remaining items with the items from other modalities, which is further influenced by the separated binarization. Second, it is obvious that both the Precision and F-measure measures are highly influenced by the code length. The performance of most cross-modal hashing methods decreases sharply with the growing code length. This is because the increasing code length results in a sparser Hamming space when the samples are fixed; it then becomes difficult to capture the intrinsic structure of the intra- and inter-modalities. However, the deep models can reduce the substantial negative impact to some extent, especially the proposed models. For example, the Precision metric of DBRC-L and DBRC-N remains stable on the three datasets for the two retrieval tasks, which indicates that our models can effectively model multiple modalities, even when faced with sparse encoding spaces. Third, DBRC-L tends to perform better than DBRC-N in terms of Precision but worse in terms of F-measure. This situation indicates that DBRC-L can retrieve more exact items (with high precision), while DBRC-N can obtain more proper items (with high recall).

3) Parameter Sensitivity: There is one key hyper-parameter within the ATanh activation function: λ. To evaluate its influence on the hashing learning, we conduct a parameter analysis on this parameter. Fig. 7 shows the Hamming ranking performance of the two proposed models on the Wiki dataset with 32- and 64-bit codes. Since λ controls the importance of the regularization term, it has a significant influence on reducing the quantization error. The influence of λ is mainly reflected in the convergence speed but is limited on the differences of the quantization error. Concretely, when λ increases, DBRC will achieve faster convergence, and the units in the hashing layer will be quickly transformed into binary values. In these situations, if the shared subspace across modalities has not been effectively learned, the premature binarization with large λ could result in inefficient hashing codes. Fortunately, the employed end-to-end DBRC model still possesses a strong ability for modeling complex data distributions under such situations. Hence, we can find that DBRC-N shows a tolerable decline with increasing λ, as shown in Fig. 7. By contrast, DBRC-L remains stable to some extent. This is because the scaling parameter α takes the negative power in the regularization term of Eq. (17), which makes the network insensitive to variations in λ. In this paper, we choose $\lambda = 0.001$ for both DBRC-L and DBRC-N in all cases, thereby providing a proper balance between convergence speed and model performance.

VII. Conclusion

In this paper, we focus on the difficult optimization problem of cross-modal hashing objectives with binary constraints. To learn efficient binary codes, the complex correlation across modalities should first be effectively captured; hence, deep networks with strong nonlinear modeling abilities are considered. A theoretical guarantee is also provided to validate its capacity in preserving the inter- and intra-modal consistencies. Then, a deep binary reconstruction network is proposed to directly learn the binary codes. The superiority comes from a novel shared hashing layer across unimodal networks, which consists of a composite activation function that adaptively scales the input activations into the nonlinear scope of the tanh function. In addition, linear and nonlinear scaling extensions, which can be jointly trained with the whole network, are provided in the paper. Extensive experiments are conducted on three benchmark datasets, the results of which showing that the proposed models are quite capable of directly producing more efficient binary codes.

References

[1] N. Srivastava and R. R. Salakhutdinov, “Multimodal learning with deep Boltzmann machines,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[2] D. Hu, X. Li, and X. Lu, “Temporal multimodal learning in audiovisual speech recognition,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3574–3582.

[3] L. Pang, S. Zhu, and C.-W. Ngo, “Deep multimodal learning for affective analysis and retrieval,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[4] L. Baraldi, C. Grana, and R. Cucchiara, “Recognizing and presenting the storytelling video structure with deep multimodal networks,” *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 955–968, May 2017.

[5] Y. Huang, W. Wang, and L. Wang, “Unconstrained multimodal multi-label learning,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1923–1935, Nov. 2015.

[6] Z. Zhao *et al.*, “Social-aware movie recommendation via multimodal network learning,” *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 430–440, Feb. 2018.

[7] J. Dong, X. Li, and D. Xu, “Cross-media similarity evaluation for web image retrieval in the wild,” 2017, Preprints, arXiv:1709.01305.

[8] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, Jan. 2018.

[9] X. Li, D. Hu, and X. Lu, "Image2song: Song retrieval via bridging image content and lyric words," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5649–5656.

[10] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.

[11] X. Lu, Y. Chen, and X. Li, "Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 106–120, Jan. 2018.

[12] X. Lu, X. Zheng, and X. Li, "Latent semantic minimal hashing for image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 355–368, Jan. 2017.

[13] Y. Gong, S. Kumar, H. A. Rowley, and S. Lazebnik, "Learning binary codes for high-dimensional data using bilinear projections," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 484–491.

[14] J. Song, L. Gao, L. Liu, X. Zhu, and N. Sebe, "Quantization-based hashing: A general framework for scalable image and video retrieval," *Pattern Recognit.*, vol. 75, pp. 175–187, 2018.

[15] J. Song *et al.*, "Self-supervised video hashing with hierarchical binary auto-encoder," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3210–3221, Jul. 2018.

[16] Y. Hao *et al.*, "Stochastic multiview hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 1–14, Jan. 2017.

[17] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep video hashing," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1209–1219, Jun. 2017.

[18] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 415–424.

[19] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Nov. 2016.

[20] F. Wu *et al.*, "Sparse multi-modal hashing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.

[21] R. Salakhutdinov and G. Hinton, "Semantic hashing," *Int. J. Approx. Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[22] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," 2016, Preprints, arXiv:1607.06215.

[23] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[24] D. Wang, P. Cui, M. Ou, and W. Zhu, "Learning compact hash codes for multimodal representations using orthogonal deep structure," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1404–1416, Sep. 2015.

[25] X. Li, D. Hu, and F. Nie, "Deep binary reconstruction for cross-modal hashing," in *Proc. 25th ACM Multimedia Conf.*, 2017, pp. 1398–1406.

[26] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, vol. 2, 2011, pp. 1360–1365.

[27] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inform. Process. Syst.*, 2009, pp. 1753–1760.

[28] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[29] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 785–796.

[30] M. Rastegari, J. Choi, S. Fakhraei, H. Daumé III, and L. S. Davis, "Predictable dual-view hashing," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28, 2013, pp. 1328–1336.

[31] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 2075–2082.

[32] X. Xu, F. Shen, Y. Yang, and H. T. Shen, "Discriminant cross-modal hashing," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 305–308.

[33] T. Zhang and J. Wang, "Collaborative quantization for cross-modal similarity search," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2036–2045.

[34] K. Ding, B. Fan, C. Huo, S. Xiang, and C. Pan, "Cross-modal hashing via rank-order preserving," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 571–585, Mar. 2017.

[35] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[36] Q. Dong, M. Shu, H. Cui, H. Xu, and Z. Hu, "Learning stratified 3D reconstruction," *Sci. China Inf. Sci.*, vol. 61, no. 2, 2018, Art. no. 023101.

[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[38] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 6645–6649.

[39] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[40] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proc. 12th Int. Conf. Artif. Intell. Statist.*, 2009, pp. 448–455.

[41] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2141–2149.

[42] D. Hu, X. Lu, and X. Li, "Multimodal learning via exploring deep semantic similarity," in *Proc. ACM Multimedia Conf.*, 2016, pp. 342–346.

[43] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[44] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," in *Proc. VLDB Endowment*, 2014, vol. 7, no. 8, pp. 649–660.

[45] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or –1," 2016, Preprints, arXiv:1602.02830.

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

[47] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1064–1071.

[48] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[49] T.-T. Do, A.-Z. Doan, and N.-M. Cheung, "Discrete hashing with deep neural network," 2015, Preprints, arXiv:1508.07148.

[50] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[51] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.

[52] T.-S. Chua *et al.*, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, Art. no. 48.

[53] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3864–3872.

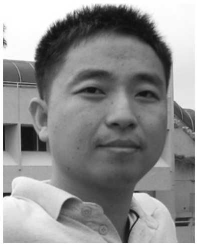
[54] Z. Cao, M. Long, J. Wang, and P. S. Yu, "HashNet: Deep learning to hash by continuation," 2017, Preprints, arXiv:1702.00758.

[55] X. Li, D. Hu, and F. Nie, "Large graph hashing with spectral rotation," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2203–2209.

Di Hu is currently working toward the Ph.D. degree at the Northwestern Polytechnical University, Xi'an, P. R. China, under the supervision of Feiping Nie and Xuelong Li.

He has authored or co-authored several papers in top conferences, such as CVPR, ICCV, AAAI, and ACM MM. His research interests include multimodal machine learning and relevant applications, such as audiovisual understanding, cross-modal retrieval, etc.

Mr. Hu was a PC member for AAAI, CVPR, and ACCV.



Feiping Nie (M'17) received the Ph.D. degree in computer science from Tsinghua University, Beijing, P. R. China, in 2009.

He is currently a Full Professor with the Northwestern Polytechnical University, Xi'an, P. R. China. He has authored or co-authored more than 100 papers in the following journals and conferences: TPAMI, IJCV, TIP, TNNLS, TKDE, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, and ACM MM. His papers have been cited more than 10 000 times and the H-index is 53. His research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Dr. Nie is an Associate Editor and a PC member for several prestigious journals and conferences in the related fields.

Xuelong Li (M'02–SM'07–F'12) is a Full Professor with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. He was with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an.