SPECIAL SECTION ON BIOLOGICALLY INSPIRED IMAGE PROCESSING CHALLENGES AND FUTURE DIRECTIONS
IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 21, 2019, accepted February 7, 2019, date of publication February 14, 2019, date of current version March 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899536

Deep Multi-Level Semantic Hashing for Cross-Modal Retrieval

ZHENYAN JI[1], WEINA YAO[1], WEI WEI[2], (Senior Member, IEEE), HOUBING SONG[3], (Senior Member, IEEE) AND HUAIYU PI[1]

[1]School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China
[2]School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710054, China
[3]Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA

Corresponding author: Zhenyan Ji (zhyji@bjtu.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB0203900, and in part by the Key Research and Development Program of Shaanxi Province under Grant 2018ZDXM-GY-036.

ABSTRACT With the rapid growth of multimodal data, the cross-modal search has widely attracted research interests. Due to its efficiency on storage and computing, hashing-based methods are broadly used for large scale cross-modal retrieval. Most existing hashing methods are designed based on binary supervision, which transforms complex relationships of multi-label data into simple similar or dissimilar. However, few methods have explored the rich semantic information implicit in multi-label data to improve the accuracy of searching results. In this paper, the multi-level semantic supervision generating approach is proposed by exploring the label relevance. And a deep hashing framework is designed for multi-label image-text cross retrieval tasks. It can simultaneously capture the binary similarity and the complex multi-level semantic structure of data in different forms. Moreover, the effects of three different convolutional neural networks, CNN-F, VGG-16, and ResNet-50, on the retrieval results are compared. The experimental results on an open source cross-modal dataset show that our approach outperforms several state-of-the-art hashing methods, and the retrieval result on the CNN-F network is better than VGG-16 and ResNet-50.

INDEX TERMS Cross-modal retrieval, deep learning, hashing method, multi-label learning.

I. INTRODUCTION

With the development of mobile Internet, multimodal data such as texts, images and videos are rapidly increasing, which results in the fast growth of cross-modal retrieval. Cross-modal retrieval, also called cross-media retrieval, models the relationship among different modalities. It aims at computing cross-modal similarities and retrieving relevant instances of different modal types [1], for example, searching images by text queries, searching videos by image queries, etc. Unlike traditional multimodal retrieval [2], the key to cross-modal search lies in that mapping the data of different modalities to a public feature space. Traditional cross-modal retrieval usually uses handcrafted features that rely on domain knowledge, and ''semantic gap'' is still a thorny problem in this field. In recent years, a great number of studies have been made in the representation of multimedia information and many breakthroughs have been made on applying deep learning to cross-modal retrieval [3]–[6]. Among them, the combination of deep learning and hashing methods used in large-scale similarity search is currently a popular research direction.

The hashing method was first applied to the Approximate Nearest Neighbor(ANN) problem to speed up hierarchical tree-decomposition based methods [7]. The basic idea is to hash points in the database so that close points are more possibly collided than those far apart. The cross-modal hashing method maps the original high-dimensional data of different modalities into a set of binary codes in a unified form, while keeping the similarity of the data in the original space. The compact binary codes are not only efficient for large-scale data storage, but also saves a lot of computing resources. Existing cross-modal hashing methods can be divided into 2 categories, shallow hashing and deep hashing. Representative works of shallow hashing are done by Ding *et al.* [8], Zhang and Li [9], Lin *et al.* [10], Wang *et al.* [11], etc. A common point of these methods is that they all use shallow architectures for multimodal embedding. That is not efficient in capturing features of data in different types. Thus shallow hashing cannot achieve satisfactory retrieval results.

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

To solve this problem, several deep learning based approaches are proposed to depict the hidden complicated correlation of different data modalities, which are called deep hashing. Representative works are done by Jiang and Li [12], Cao *et al.* [13], Shen *et al.* [14], Yang *et al.* [15], Zhang *et al.* [16], etc. The powerful feature learning and representing ability of deep learning significantly improves search quality. Moreover, the structure of neural networks enables seamlessly integrating feature learning and hash code learning parts into a unified end-to-end framework.

However, most of these deep hashing methods are designed based on binary supervision. They can only deal with issues related to single label data, and the constraint information is simple similar or dissimilar relationships. For the data with multiple labels, the complex multi-level semantic structure has not been well explored yet. For example, two data points with more common labels should be regarded as an instance with higher similarity than those with less common labels. To solve this problem, here we propose a deep multi-level semantic hashing(DMSH) method for learning hash functions that preserve multi-level semantic similarity between multi-label data. The main contributions are outlined as follows:

- A semantic similarity matrix based on label co-occurrence is proposed to preserve the rich semantic information embedded in multi-label data.
- We proposed a deep hashing framework that not only take binary supervision but also multi-level supervision into account to guide the hashing code learning.
- Experimental results show that our method can not only learn the compact hash code of both image and text modalities but also preserves more semantic information than binary supervision based methods.

II. RELATED WORK

Of late, there has been increasing interest in deep learning based cross-modal hashing(CMH). It develops rapidly based on single-modal deep hashing methods. For image-text cross retrieval, many deep hashing image retrieval achievements are adopted to improve the cross retrieval results. Xia *et al.* [17] trained hash codes of images from the supervised information based on a two-stage method. A similarity matrix is decomposed to generate the initial codes, then the image representation and hash functions are learned by CNN based on the approximate hash codes. Zhu *et al.* [18] proposed a supervised deep hashing network(DHN) to simultaneously handle the image feature learning and hash code learning step to avoid quantization errors. It uses 3 fully connected layers to generate hash codes and combines a pairwise cross-entropy loss and a quantization loss to ensure the high quality of learned codes. Lai *et al.* [19] designed a novel deep neural network with divide-and-encode module to map images to binary codes. It uses a triplet ranking loss to capture the semantic similarity in hamming space. Zhao *et al.* [20] proposed a deep semantic ranking based hashing(DSRH) framework which uses the list-wise ranking supervision to preserve multi-level semantic similarity between multi-label images. In addition, there are many other studies focusing on network structure design, quantization methods and optimization strategies of deep hashing single-modal retrieval algorithms [13], [21]–[23].

For the cross-modal tasks, known techniques in this field can be divided into supervised methods and unsupervised methods. Supervised deep hashing methods take the semantic information of class labels into consideration to explore the cross-modal correlation. Deep cross modal hashing (DCMH) [12] launched a pioneering end-to-end framework which can perform simultaneous feature learning and hash code learning. It uses two deep neural networks for feature extraction of each modality. The learning process is guided by the constraint that semantically similar points are closer than those dissimilar in hamming space. However, it only measures the inter-modal similarity and ignores the intra-modal correlation, which reduces the retrieval accuracy. As an improvement, Pairwise Relationship Guided Deep Hashing(PRDH) [15] utilized both inter-modal and intra-modal similarities to effectively discover the heterogeneous correlations across different modalities. These two methods perform well on datasets with discrete tags as the text modality. The searching accuracy declined when it comes to continuous sentences. Deep Visual-Semantic Hashing(DVSH) [13] and Textual-Visual Deep Binaries (TVDB) [14] are developed to solve image-sentence cross retrieval problems, which capture the spatial dependency of images and temporal dynamics of text sentences for feature learning and cross-modal embedding. More remarkable research on supervised deep cross-modal hashing are discussed in [24]–[31].

Unsupervised methods deal with the case that all training data is unlabeled. A representative study on unsupervised deep hashing is proposed by Zhang *et al.* [16]. Inspired by recent progress of generative adversarial network(GAN), they develop a unsupervised generative adversarial cross-modal hashing(UGACH) framework to capture the cross-modal correlation in an unsupervised fashion. It makes full use of GAN's ability for unsupervised representation learning to exploit the underlying manifold structure of cross-modal data. Zhang *et al.* [32] employed attention-aware mechanism to the adversarial hashing network to enhance the measurement of content similarities by selectively focusing on informative parts of multi-modal data. It keeps exploring the solution of discovering content similarities between different data modalities. Li *et al.* [33] proposed a SSAH model to first introduce adversarial learning to cross-modal retrieval. Two different adversarial networks are used to learn high dimensional features and hash code of different modals. Zhang *et al.* [34] addressed the problem of (1) rely on large labeled data and (2) ignore information in rich unlabeled data. They proposed to use semi-supervised hashing approach by generative adversarial network. Different from the methods above, we aim to make full use of labeled data and better supervise the proposed network.Except for deep

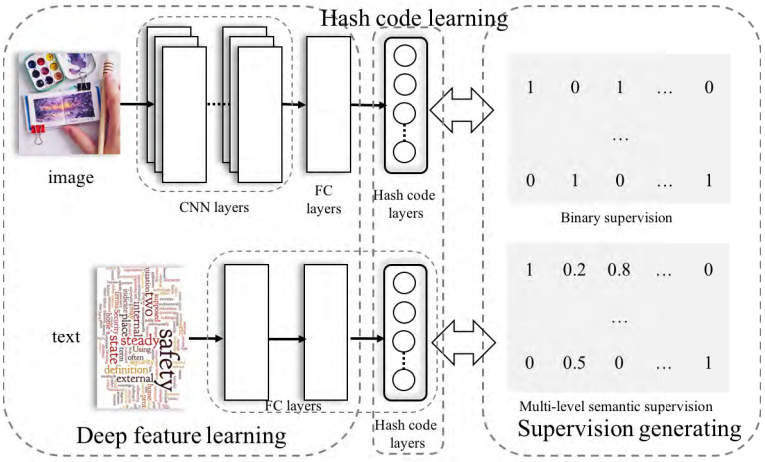

FIGURE 1. Framework of DMSH.

convolutional neural networks, other deep structures like deep Boltzmann machines and auto-encoders are introduced to unsupervised deep cross-modal hashing frameworks to improve the search results [35]–[38].

III. DEEP MULTI-LEVEL SEMANTIC HASHING

In this paper, we only use images and texts as the multi-modal instances to discuss our method. The framework of proposed DMSH are shown in Figure 1 The model mainly consists of three modules: deep feature learning module, supervision generating module and hash code learning module. The whole architecture of DMSH and the details about multi-level similarity constraint are illustrated in this section.

We use some notations shown in Table 1 for concise expression.

TABLE 1. Notations.

n	Number of data points in training set
c	Hash code length
E	A vector with all elements being 1
$G = \{g_i\}_{i=1}^{n}$	Image set
$X = \{x_i\}_{i=1}^{n}$	Text set
$T = \{t_i\}_{i=1}^{n}$	Label set
$D = \{(g_i, x_i)\}_{i=1}^{n}$	Data points in the training set
$C^{(g)}, C^{(x)}$	Hash code for image and text
$S^{(b)} = \left\{S_{ij}^{(b)}\right\}$	Binary similarity matrix
$S^{(l)} = \left\{S_{ij}^{(l)}\right\}$	Multi-level similarity matrix

A. DEEP FEATURE LEARNING

DMSH separately uses two deep neural networks to extract image and text features. For image modality, we adopt a refined CNN-F model. CNN-F [39] was inspired by Alexnet [40]. It consists of 5 convolutional layers and 3 fully connected layers. The adapted CNN-F changes the number of neurons in the last fc layer from 1000 to the hash code length. Details of the deep architecture are shown in table 2. We use bag-of-words(BOW) to represent text modality and adopt 3 fully connected layers for text feature learning. The length of the first layer is the same as the length of BOW vectors, the middle layer consists of 4096 nodes and the length of last layer is the same as the code length. The two modules are correlated at the output end by a carefully designed multi-level semantic objective function.

TABLE 2. Architecture of image feature learning module.

Layer name	Architecture	Output size
conv1	$11 \times 11 \times 64$, st.4, pad0, LRN, max pooling	27×27
conv2	$5 \times 5 \times 256$, st.1, pad2, LRN, max pooling	13×13
conv3	$3 \times 3 \times 256$, st. 1, pad1	13×13
conv4	$3 \times 3 \times 256$, st. 1, pad1	13×13
conv5	$3 \times 3 \times 256$, st. 1, pad1, 2 max pooling	6×6
fc6	4096, Dropout	1×1
fc7	4096, Dropout	1×1
fc8	c, linear	1×1

B. SUPERVISION GENERATING

The supervision generating module consists of binary supervision generating and multi-level semantic supervision generating. They both produce a cross-modal similarity matrix . For example, if image i is similar with text j and otherwise. The similarity between different modality is measured by class labels. That is, image i and text j are similar if they share the same class label; otherwise they are dissimilar. This method works well with single-label data. However, it ignores the rich semantic information for multi-label data. Therefore, we adopt a semantic similarity matrix calculation method based on label co-occurrence to obtain the multi-level semantic similarity matrix. Here we use the notation $S^{(b)}$

and $S^{(l)}$ to represent binary supervision and multi-level semantic supervision respectively. The following describes the generation methods.

The binary cross-modal similarity is defined as:

$$S_{ij}^{(b)} = \begin{cases} 1, & \text{if point } i \text{ and } j \text{ share at least 1 label} \\ 0, & \text{else} \end{cases} \tag{1}$$

For two class labels t_i, t_j, we define the label similarity as:

$$s(t_i, t_j) = e^{-d(t_i,t_j)} \tag{2}$$

where denotes the semantic distance of two labels and are calculated as below.

$$d(t_i, t_j) = \frac{max(logN_{t_i}, logN_{t_j}) - logN_{t_i,t_j}}{logN_c - min(logN_{t_i}, logN_{t_j})} \tag{3}$$

N_{t_i}, N_{t_j} represents the number of t_i, t_j in the training set, N_{t_i,t_j} indicates the times t_i and t_j co-occurred, N_c is the number of all labels in the training set.

From the above definition, we can know that $s(t_i, t_j) \in [0, 1]$. It shows that the similarity of two labels is larger if they are shared with more data points. Based on the label similarity, the multi-level semantic similarity of two data points D_m, D_n is defined as:

$$S^{(l)}(m, n) = \frac{\sum_i^{|t_m|} \sum_j^{|t_n|} s(t_m(i), t_n(j))}{|t_m| \times |t_n|} \tag{4}$$

where t_m, t_n denote the label sets of data points D_m, D_n, and $|t_m|$, $|t_n|$ are the number of labels in t_m, t_n respectively. By the definition, we know that D_m and D_n are more similar if their label sets t_m, t_n have more relevant labels, and $S^{(l)}(m, n) \in [0, 1]$. If D_m and D_n share the same label set, $S^{(l)}(m, n)$ reaches the maximum value 1. If labels in t_m are all irrelevant to labels in t_n, $S^{(l)}(m, n)$ takes the minimum value 0. Therefore, the multi-label based semantic similarity matrix $S^{(l)}$ can be used as the hash learning supervision information. Compared with the binary $S^{(b)}$, $S^{(l)}$ expands the cross-modal similarity from discrete $\{0, 1\}$ to continuous $[0, 1]$ intervals which preserves more abundant semantic information.

C. HASH CODE LEARNING

We use ${F^{(g)}}_{*i} = f^{(g)}(g_i; \varphi_g)$ to represent the output of the CNN model, which is the image feature vector of a data point D_i; ${F^{(x)}}_{*j} = f^{(x)}(x_j; \varphi_x)$ denotes the output of the deep neural network for text modality, which corresponds to the text feature of data point D_j. φ_g, φ_x denote the parameters of the CNN for image and the multi-layer perceptron network for text, respectively.

To preserve the binary cross-modal similarity in $S_{(b)}$, we use the sigmoid cross entropy loss:

$$\min_{\varphi_g,\varphi_x} \mathscr{L}_c = -\sum_{i,j=1}^{n} S_{ij}^{(b)} log(\sigma(\Phi_{ij})) + (1 - S_{ij}^{(b)}) log(1 - \sigma(\Phi_{ij})) \tag{5}$$

where $\Phi_{ij} = \frac{1}{2} {F^{(g)}}_{*i}^T {F^{(x)}}_{*j}$, $\sigma(\Phi_{ij}) = \frac{1}{1+e^{-\Phi_{ij}}}$. To ensure stability and avoid overflow, the implementation uses this equivalent formulation:

$$\min_{\varphi_g,\varphi_x} \mathscr{L}_c = \sum_{i,j=1}^{n} max(\Phi_{ij}, 0) - S_{ij}^{(b)} \Phi_{ij} + log(1 + e^{-\Phi_{ij}}) \tag{6}$$

Based on this, we introduce a multi-level semantic loss $\mathscr{L}_m$ to make our model preserve the rich semantic information embedded in $S^{(l)}$. It is also derived from sigmoid cross entropy loss, here we directly give its formulation:

$$\min_{\varphi_g,\varphi_x} \mathscr{L}_m = \sum_{i,j=1}^{n} max(\Phi_{ij}, 0) - S_{ij}^{(l)} \Phi_{ij} + log(1 + e^{-\Phi_{ij}}) \tag{7}$$

therefore, the objective function of DMSH is defined as follows:

$$\begin{aligned} \min_{\varphi_g,\varphi_x,C^{(g)},C^{(x)}} & \mathscr{L} = \mathscr{L}_c + \mu \mathscr{L}_m + \rho(\left\| C^{(g)} - F^{(g)} \right\|_F^2 \\ & + \left\| C^{(x)} - F^{(x)} \right\|_F^2) + \tau(\left\| F^{(g)} E \right\|_F^2 + \left\| F^{(x)} E \right\|_F^2) \\ s.t.\ & C^{(g)}, C^{(x)} \in \{-1, +1\}^{c \times n} \\ & F^{(g)} = f(G; \varphi_g) \\ & F^{(x)} = f(X; \varphi_x) \end{aligned} \tag{8}$$

where $C^{(g)} = sign(F^{(g)})$, $C^{(x)} = sign(F^{(x)})$, and $sign(\cdot)$ is a sign function defined as:

$$sign(x) = \begin{cases} 1, & x \geqslant 0 \\ -1, & x < 0 \end{cases} \tag{9}$$

$\|\cdot\|_F^2$ denotes the Frobenius norm.

$F^{(g)}$ and $F^{(x)}$ are the learned features of image and text modality, which keep the semantic information in $S^{(b)}$ and $S^{(l)}$. The third regularization term is adopted to ensure that the learning hash code $C^{(g)}$, $C^{(x)}$ preserve the information contained in $F^{(g)}$ and $F^{(x)}$. Thus the semantic similarity can be transferred to the learned hash codes. The forth term $\tau(\left\| F^{(g)} E \right\|_F^2 + \left\| F^{(x)} E \right\|_F^2)$ is used to balance the number of "+1" and "−1" for each bit of the hash code on all training set so that the information provided by each bit reaches the maximum.

Experimental results in [12] show that better performance can be achieved when another constraint $C^{(g)} = C^{(x)} = C$ is added to the loss function in the training process. We adopt this achievement to our model and the refined objective function is as below:

$$\begin{aligned} \min_{\varphi_g,\varphi_x,C} & \mathscr{L} = \mathscr{L}_c + \mu \mathscr{L}_m + \rho(\left\| C - F^{(g)} \right\|_F^2 + \left\| C - F^{(x)} \right\|_F^2) \\ & + \tau(\left\| F^{(g)} E \right\|_F^2 + \left\| F^{(x)} E \right\|_F^2) \\ s.t.\ & C \in \{-1, +1\}^{c \times n} \\ & F^{(g)} = f(G; \varphi_g) \\ & F^{(x)} = f(X; \varphi_x) \end{aligned} \tag{10}$$

Since φ_g, φ_x, C are the 3 parameters to be optimized, we use an alternating learning strategy that fixing two parameters and updating the left one at a time. The training procedure is shown in Algorithm 1 [12].

Algorithm 1 Hash Code Learning Algorithm

Input: Image dataset G, text dataset X, code length c, parameters μ, ρ, τ;

Output: Hash codes C, parameters of DNNs for two modalities φ_g, φ_x;

Initialize φ_g, φ_x, mini-batch size $b_g = b_x = 128$, iteration number $n_g = n/b_g$, $n_x = n/b_x$. Compute $S^{(b)}$ and $S^{(l)}$ according to Eq.(1)-(4);
repeat
  for epoch=1, 2, $\cdots$, n_g **do**
    Randomly select b_g images from G to construct a mini-batch;
    For each sampled point g_i in mini-batch, calculate ${F^{(g)}}_{*i}$ by forward propagation;
    Calculate derivative $\frac{\partial \mathscr{L}}{\partial {F^{(g)}}_{*i}}$ according to Eq.(10);
    Update φ_g with back-propagation;
  end for
  for epoch=1, 2, $\cdots$, n_x **do**
    Randomly select b_x texts from X to construct a mini-batch;
    For each sampled point x_j in mini-batch, calculate ${F^{(x)}}_{*j}$ by forward propagation;
    Calculate derivative $\frac{\partial \mathscr{L}}{\partial {F^{(x)}}_{*j}}$ according to Eq.(10);
    Update φ_x with back-propagation;
  end for
  Optimize C;
until A fixed number of iterations.

IV. EXPERIMENTS

To evaluate the proposed DMSH, we compare the retrieval performance with several state-of-the-art algorithms on a benchmark cross-modal dataset MIRFlickr-25K. We use the open source TensorFlow [41] framework and the experiments are deployed on a server with 64G RAM and NVIDIA K80 GPU. The effects of 3 different convolutional neural networks, CNN-F, VGG-16 [42] and ResNet-50 [43], on the retrieval results are also compared with the same configuration.

A. DATASET

The MIRFlickr-25K [44] open evaluation project consists of 25000 instances and 24 class labels downloaded from the social photography site Flickr. Each data point consists of an image and corresponding textual tags. We select the points with textual tags no less than 20, and get 20015 instances for our experiments. The text is coded into a 1386-demensional bag-of-words vector. For deep learning based methods, the raw pixels are directly used as the input of image modality. While a 512-dimensional scale-invariant feature transform(SIFT) feature vector provided by the dataset is adopted for hand-crafted feature based methods.

B. BASELINES

To evaluate the effectiveness, DMSH is compared with several state-of-the-art cross-modal hashing methods:

CCA [45] is a multivariate statistical analysis method. It uses the correlation between two basis vectors to reflect the overall correlation between two sets of variables. The basic principle of CCA is to find subspaces that maximize the correlation of a set of related heterogeneous data. CCA has been widely used as a benchmark method for cross-modal retrieval. It reflects the linear correlation between two groups of heterogeneity variables.

CMFH [8] learns hash codes by collective matrix factorization with latent model from multimodal information sources. It directly learns unified code from multi-modal data other than traditional combining or concatenating codes learned from different views.

STMH [11] explores semantic topics of text and image by clustering and matrix factorization, respectively. Then the relations of the two modalities in a common subspace is learned for cross-modal hashing embedding.

SCM [9] proposes a supervised multimodal hashing method with high scalability. It integrates semantic labels into the hashing learning procedure by maximizing the semantic correlations.

SePH [10] transforms cross-modal distances in both semantic space and hamming space into two probability distribution, and learns the hashing codes by minimizing the Kullback-Leibler divergence between them. The semantic distance between cross-modal instances are measured by the labels of training data, which ensures the learnt hash codes preserving the semantic affinities.

DCMH [12] is a remarkable binary supervision based deep hashing method which details are shown in section 2.

C. IMPLEMENTATION DETAILS

The model is implemented based on the open source TensorFlow framework. For training, we take 10000 instances of MIRFlickr-25K as the training set. For testing, we take 2000 instances as the test queries, and the rest as retrieval set. We set $\mu = 0.5$, $\rho = \tau = 1$ in our experiments. The batch size is fixed to 128, and the algorithm runs 1000 iterations. The CNN-F module is pre-trained on ImageNet [46] and it is fine-tuned during the training of DMSH model.

Mean average precision(mAP) is adopted to directly present the performance of all compared methods. mAP for a set of queries is the mean of the average precision values for each query. It is defined as:

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{11}$$

$$AveP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{number\ of\ relevant\ documents} \tag{12}$$

where Q is the number of queries, $P(k)$ is the precision at cut-off k in the list, $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant one, 0 otherwise. mAP is an indicator that shows the ranking quality of all retrieval results. Table 3 shows the mAP of all baselines and our method. The best accuracy is shown in boldface. From the results we can know that deep learning based methods achieves better performance than all the others using hand-crafted features, while proposed DMSH outperforms binary supervision based DCMH. The code length has significant influence on the final result. More information is preserved as the code length increase. However it will lead to the overfitting problem when the code length goes too long. In our experiments, our model makes the best performance with 64bit code length.

TABLE 3. mAP of all compared methods on MIRFlickr-25K.

Method	Task					
	text2image			image2text		
	c=16	c=32	c=64	c=16	c=32	c=64
CCA	0.563	0.563	0.563	0.562	0.562	0.562
CMFH	0.578	0.579	0.583	0.576	0.58	0.581
STMH	0.614	0.62	0.622	0.624	0.629	0.631
SCM	0.607	0.609	0.611	0.628	0.629	0.631
SePH	0.713	0.726	0.731	0.709	0.711	0.717
DCMH	0.755	0.757	0.77	0.715	0.72	0.73
DMSH	**0.755**	**0.763**	**0.775**	**0.726**	**0.737**	**0.75**

We also use precision-recall(PR) curves to measure the accuracy of results returned under certain hamming radius. Precision is the fraction of the data points retrieved that are relevant to the queries. Recall indicates how many true positive samples are successfully retrieved. Figure 2 shows the curves at 64bit code length. The horizontal coordinate denotes recall while the vertical coordinate represents precision values. The figure on left is the PR curve of searching images by text queries and the right one denotes the curve of retrieving texts by image queries. Results of each method are represented by lines with distinct nodes and colors as the icons show in the figures: red for DMSH, purple for DCMH, green for SePH, yellow for SCM, blue for STMH, sky for CMFH, pink for CCA.

From the curves we can also see that the deep hashing methods obviously outperform shallow ones for both image query text tasks and text query image tasks. For the two compared deep hashing methods, our DMSH with multi-level

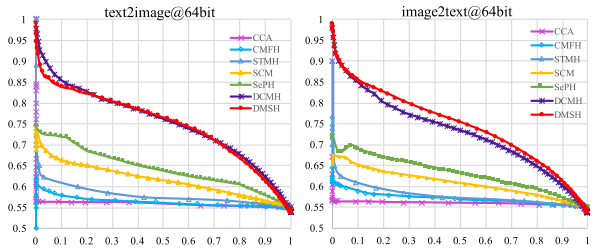

FIGURE 2. Precision-recall curves with code length 64 on MIRFlickr-25K.

semantic information is obviously better than DCMH in image query text database tasks, while in the text query image tasks they both achieve almost the same search performance. We also compare the process time of DCMH and DMSH in generating 64 bit text and image hash codes. The table below shows the average time of generating one 64 bit text code and image code using the two methods. It can be seen that our method is more efficient on both tasks.

TABLE 4. Comparison of process time.

Method	Time/ms	
	text@64bit	image@64bit
DCMH	8.90722	8.82807
DMSH	8.79811	8.76655

In order to analyze the effect of deep networks on the retrieval results, the CNN-F network in the original DMSH model was changed to VGG-16 and ResNet-50 respectively for training. Therefore the three models are denoted as DMSH-C, DMSH-V, and DMSH-R. At the same time, the same experiments are conducted on the deep learning based DCMH to ensure the reliability of the conclusion. The corresponding three models are referred to as DCMH-C, DCMH-V, and DCMH-R.

TABLE 5. Comparison of different networks in terms of mAP.

method	task	
	text2image	image2text
DCMH-C	0.77	0.73
DCMH-V	0.717	0.674
DCMH-R	0.699	0.641
DMSH-C	0.775	0.75
DMSH-V	0.722	0.685
DMSH-R	0.705	0.659

Table 5 shows the mAP values for each model with hash code length of 64 bits. As shown in Table 5, the accuracy of CNN-F, VGG-16, and ResNet-50 networks decreases in turn, and this trend is exhibited in both the baseline DCMH and the proposed DMSH. This indicates that due to insufficient data, the training process has encountered overfitting problem with the complexity of the network increasing. That the loss of the training set is smaller than that of the test set also verifies this conclusion. The CNN-F, which has fewer parameters and lower network complexity, is more suitable for training tasks of the current data volume. The features learned by the network have better generalization capabilities, and thus achieve the best results.

V. CONCLUSION

In this paper, we propose a deep multi-level semantic hashing method for cross-modal retrieval. It solves a common problem of existing supervised deep cross-modal hashing methods that rich semantic information in multi-label data

is not sufficiently used. The multi-level semantic supervision based on label co-occurrence is adopted to ensure that the learned hash codes preserve the accurate semantic similarities, that is, data points with more common class labels are more similar than those with less common labels. The proposed model comprises of deep feature learning, supervision generating, and hash code learning parts. DMSH learns image and text features by two carefully designed deep neural networks respectively. Then compact hash codes and the discriminative features of each modality are learned simultaneously in one framework. This end-to-end structure guarantees the learned features optimum for the specific cross-modal retrieval tasks. Experiments on a benchmark cross-modal dataset MIRFlickr-25K are conducted and results are compared with CCA, CMFH, STMH, SCM, SePH, and DCMH. The influence of deep networks on hashing based cross-modal retrieval methods is also explored by comparing the results of 3 deep neural networks: CNN-F, VGG-16 and ResNet-50. The results show that DMSH outperforms all these state-of-the-art hashing methods, which demonstrates the superiority of incorporating multi-level semantic information in the supervision matrix for hash code learning. And the best results are achieved on CNN-F among the compared 3 networks due to its reasonable complexity.

In this paper, we only focus on the image-text cross retrieval tasks. Further studies can be performed in taking more modalities into consideration to design a unified, flexible and adaptable cross-modal hashing framework.

ACKNOWLEDGMENT

(Zhenyan Ji and Weina Yao contributed equally to this work.)

REFERENCES

[1] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[2] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 153–162.

[3] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[4] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proc. IJCAI*, vol. 367, 2015, pp. 2291–2297.

[5] F. Feng, R. Li, and X. Wang, "Deep correspondence restricted Boltzmann machine for cross-modal retrieval," *Neurocomputing*, vol. 154, pp. 50–60, Apr. 2015.

[6] 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, Jun. 2016.

[7] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," *VLDB*, vol. 99, no. 6, pp. 518–529, 1999.

[8] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2083–2090.

[9] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. AAAI*, vol. 1. 2014, p. 7.

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

[11] D. Wang, X. Gao, X. Wang, and L. He, "Semantic topic multimodal hashing for cross-media retrieval," in *Proc. IJCAI*, 2015, pp. 3890–3896.

[12] Q.-Y. Jiang and W.-J. Li. (2016). "Deep cross-modal hashing." [Online]. Available: https://arxiv.org/abs/1602.02255

[13] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1445–1454.

[14] Y. Shen, L. Liu, L. Shao, and J. Song, "Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4097–4106.

[15] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. AAAI*, 2017, pp. 1618–1625.

[16] J. Zhang, Y. Peng, and M. Yuan. (2017). "Unsupervised generative adversarial cross-modal hashing." [Online]. Available: https://arxiv.org/abs/1712.00358

[17] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing for image retrieval via image representation learning," in *Proc. AAAI*, vol. 1, 2014, p. 2.

[18] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep hashing network for efficient similarity retrieval," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2415–2421.

[19] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3270–3278.

[20] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1556–1564.

[21] Z. Zhang, Q. Zou, Q. Wang, Y. Lin, and Q. Li. (2018). "Instance similarity deep hashing for multi-label image retrieval." [Online]. Available: https://arxiv.org/abs/1803.02987

[22] C. Zhong, Y. Yu, S. Tang, S. Satoh, and K. Xing, "Deep multi-label hashing for large-scale visual search based on semantic graph," in *Proc. Asia–Pacific Web (APWeb) Web-Age Inf. Manage. (WAIM) Joint Conf. Web Big Data*. Springer, 2017, pp. 169–184. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-63579-8_14

[23] T. Li, S. Gao, and Y. Xu, "Deep multi-similarity hashing for multi-label image retrieval," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 2159–2162.

[24] Y. Cao, M. Long, J. Wang, and H. Zhu, "Correlation autoencoder hashing for supervised cross-modal search," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2016, pp. 197–204.

[25] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[26] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, "Effective deep learning-based multi-modal retrieval," *VLDB J.*, vol. 25, no. 1, pp. 79–101, 2016.

[27] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, and Q. Dai, "Supervised hash coding with deep neural network for environment perception of intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 284–295, Jan. 2018.

[28] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, and Q. Dai, "Effective uyghur language text detection in complex background images for traffic prompt identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 220–229, Jan. 2018.

[29] C. Yan *et al.*, "A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 573–576, May 2014.

[30] C. Yan *et al.*, "Efficient parallel framework for HEVC motion estimation on many-core processors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 12, pp. 2077–2089, Dec. 2014.

[31] C. Yan *et al.*, "An effective Uyghur text detector for complex background images," *IEEE Trans. Multimedia*, to be published.

[32] X. Zhang *et al.* (2017). "HashGAN: Attention-aware deep adversarial hashing for cross modal retrieval." [Online]. Available: https://arxiv.org/abs/1711.09347

[33] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4242–4251.

[34] J. Zhang, Y. Peng, and M. Yuan. (2018). "SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network." [Online]. Available: https://arxiv.org/abs/1802.02488

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

[36] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," *Proc. VLDB Endowment*, vol. 7, no. 8, pp. 649–660, 2014.
[37] H. Zhang, Y. Yang, H. Luan, S. Yang, and T.-S. Chua, "Start from scratch: Towards automatically identifying, modeling, and naming visual attributes," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 187–196.
[38] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.
[39] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/1405.3531
[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[41] M. Abadi *et al.* (2016). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems." [Online]. Available: https://arxiv.org/abs/1603.04467
[42] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[44] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr.*, 2008, pp. 39–43.
[45] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.
[46] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

ZHENYAN JI received the Ph.D. degree from the Institute of Software, Chinese Academy of Sciences. She was with the Norwegian University of Science and Technology and Mid Sweden University. She is currently an Associate Professor with Beijing Jiaotong University. Her main research interests include data mining, image registration, and distributed computing. She is a member of the Theoretical Computer Science Technical Committee, CCF.

WEINA YAO received the B.S. degree from the Department of Electronic Engineering, Xidian University, China, in 2014, and the M.S. degree in software engineering from Beijing Jiaotong University, Beijing, China, in 2018. Her research interests include image retrieval and deep learning.

WEI WEI (SM'17) received the M.S. and Ph.D. degrees from Xian Jiaotong University, in 2011 and 2005, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China. He ran many funded research projects as a Principal Investigator and Technical Member. He has published around 100 research papers in international conferences and journals. His research interests include wireless networks, wireless sensor networks application, image processing, mobile computing, distributed computing, and pervasive computing, the Internet of Things, and sensor data clouds. He is an Editorial Board Member of FGCS, AHSWN, IEICE, and KSII. He is a TPC Member of many conferences and regular reviewer of the IEEE TPDS, TIP, TMC, TWC, and many other Elsevier journals.

HOUBING SONG (M'12–SM'14) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2012, and the M.S. degree in civil engineering from The University of Texas, El Paso, TX, USA, in 2006.

In 2007, he was an Engineering Research Associate with the Texas A&M Transportation Institute. He served on the faculty of West Virginia University, from 2012 to 2017. In 2017, he joined the Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA, where he is currently an Assistant Professor and the Director of the Security and Optimization for Networked Globe Laboratory (SONG Lab, www.SONGLab.us). He is an Editor of four books, including *Smart Cities: Foundations, Principles, and Applications*, (Hoboken, NJ: Wiley, 2017), *Security and Privacy in Cyber-Physical Systems: Foundations, Principles, and Applications*, (Chichester, U.K.: Wiley-IEEE Press, 2017), *Cyber-Physical Systems: Foundations, Principles and Applications*, (Boston, MA: Academic Press, 2016), and *Industrial Internet of Things: Cybermanufacturing Systems*, (Cham, Switzerland: Springer, 2016). He is the author of more than 100 articles. His research interests include cyber-physical systems, cybersecurity and privacy, the Internet of Things, edge computing, big data analytics, unmanned aircraft systems, connected vehicle, smart and connected health, and wireless communications and networking.

Dr. Song is a Senior Member of ACM. He was a recipient of the prestigious Air Force Research Laboratory's Information Directorate Visiting Faculty Research Fellowship, in 2018, and the very first recipient of the Golden Bear Scholar Award, the highest campus-wide recognition for research excellence with the West Virginia University Institute of Technology, in 2016. He serves as an Associate Technical Editor for the *IEEE Communications Magazine*.

HUAIYU PI received the B.S. degree from the School of Software Engineering and Communication Engineering, Jiangxi University of Finance and Economy. He is currently pursuing the degree with the School of Software Engineering, Beijing Jiaotong University. His research interests include recommender systems, machine learning, and social networks.