2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)

Semantic-Guided Hashing for Cross-Modal Retrieval

Zhikui Chen[1,2], Jianing Du[1], Fangming Zhong[1], Shi Chen[3]
[1]School of Software Technology, Dalian University of Technology, Dalian, China
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, China
[3]Department of Computer Science, Emporia State University, Emporia, KS, USA
zkchen@dlut.edu.cn; {zijikanwa,fmzhong}@mail.dlut.edu.cn; schen9@g.emporia.edu

***Abstract*—In the Big Data era, information retrieval across heterogeneous data or multimodal data is a very significant issue. Cross-modal hashing has recently attracted increasing attention for multimodal retrieval with benefits of fast retrieval efficiency and low storage cost. Many supervised cross-modal hashing approaches have been explored to achieve better performance according to label information. However, most of these existing methods take the form of 0/1 binary labels or pairwise relationships as supervised information, resulting in the neglect of valuable semantic correction among different classes. To address this problem, we propose a novel two-step supervised cross-modal hashing approach, termed Semantic-Guided Hashing (SeGH), to obtain the discriminative binary codes. Particularly, in Step 1, our method takes the encoder-decoder paradigm based on label semantics obtained by the word vector of class names to learn the discriminative projection from original feature space to common semantic space. In Step 2, semantic representations of different modalities in the common space are projected into a Hamming space while preserving intra-modality and inter-modality similarity. Extensive experiments compared against several state-of-the-art baselines on two datasets highlight the superiority of the proposed SeGH for cross-modal retrieval, and also demonstrate its effectiveness for zero-shot cross-modal retrieval.**

Keywords-cross-modal hashing; label semantics; zero-shot hashing; discriminative binary codes

I. INTRODUCTION

The information retrieval is very important in the era of big data, especially for the retrieval across heterogeneous or multimodal data such as the multimedia data like images and texts. Data from different modalities usually have semantic relationships, arousing the growing demand for supporting cross-modal retrieval that obtains relevant results of one modality using another modality. Recently, many pioneer efforts for cross-modal retrieval [1-4] have been proposed to explore the semantic correlation between heterogeneous data, and they have achieved remarkable retrieval performance. However, these methods will suffer from high computational complexity when the scale or dimension of data increases.

Motivated by fast retrieval speed and low storage cost of hashing technique, cross-modal hashing has received considerable attention for effectively solving the above problem, which encodes high-dimensional data into compact binary codes, and computes similarity with fast bit-wise XOR operation. Most existing cross-modal hashing methods mainly project data from different modalities into a common semantic space, then generate corresponding hash codes, which can be roughly categorized into two branches, i.e., unsupervised and supervised approaches.

Unsupervised cross-modal hashing methods usually learn hashing function only from original data to preserve the intrinsic structure of data. Collective Matrix Factorization Hashing (CMFH) [5] is the first work to learn hashing function via Matrix Factorization technology, and generates the unified hash codes for different modalities of the same instance. Latent Semantic Sparse Hashing (LSSH) [6] utilizes matrix factorization and sparse coding to extract latent semantic features respectively, then maps them into unified hash codes in a joint abstract space. Although these approaches can extract the relationship between different modalities, the learned hash codes are not discriminative sufficiently in an unsupervised manner.

The supervised ones provide label information of heterogeneous data to boost the retrieval power. Along this line, Supervised Matrix Factorization Hashing (SMFH) [7] extends CMFH by leveraging both label consistency and local information, and yields superior performance. Intra- and Inter-modality Similarity Preserving Hashing (IISPH) [8] maintains the intra- and inter-modality similarity under the low-dimensional Hamming space, and integrates similarity formulations into hashing function learning. Benefiting from the available label, the results of these supervised approaches are promising than unsupervised ones. However, most of the existing supervised methods primarily concentrate on capturing the semantic information from original features to latent common semantic space, while the supervision information is used in the form of 0/1 binary labels (such as one-hot vector) or pairwise relationships, which makes each class independent to others. More importantly, the valuable semantic correlation among labels is completely ignored.

In this work, we address the above problems with the proposed two-step hashing method for cross-modal retrieval, termed Semantic-Guided Hashing (SeGH). Inspired by the excellent ability of word embedding that captures the semantic relationships between categories, we first construct the class-level semantic space by leveraging the word vector obtained by category name, which acts as a guide to learn the common latent semantic space. Moreover, unlike the

978-1-7281-0059-3/19/$31.00 ©2019 IEEE
DOI 10.1109/BigDataService.2019.00031

conventional methods which learn the unidirectional projection from original feature space to semantic embedded space, our proposed method takes an encoder-decoder paradigm with class-level semantic space as the middle layer to learn the projection with retaining all the information in original feature. Specifically, the encoder projects the original feature of different modalities into class-level semantic space, while the decoder aims to reconstruct the original feature precisely.

On the one hand, the common latent semantic space learned subsequently by such a model not only captures the semantic correlations among different categories, but also preserves the original feature, which further enhances the discrimination capacity of the to-be-learned hash codes. On the other hand, our method can also be generalized to solve the retrieval problem of categories that are never seen during training stage (unseen domain), which breaks through the limitations of close-set retrieval in traditional cross-modal retrieval methods. This is because the class-level semantic space builds the connection between seen and unseen classes, which makes the available knowledge be transferred from seen classes to unseen classes. Furthermore, the feature reconstruction demand of this model is generalizable in both seen and unseen domains, and unseen classes can also be projected into the class-level semantic space without domain shift [9]. We will demonstrate this expansion capability in the experiments. Besides, intra-modality and inter-modality similarity preservation is additionally taken into account to improve performance.

The major contributions of the proposed SeGH can be summarized as follows.

- We propose a Semantic-Guided Hashing (SeGH) for cross-modal retrieval, which builds a class-level semantic space according to the semantic representations of class names generated by GloVe model. In such space, the semantic correlations among different classes are captured.
- A model of encoder-decoder paradigm based on class label semantics is developed to learn the projection from original feature space to common latent space, such that all the information contained in the original feature will be preserved to the projection. It does not only enhance the discriminability of the subsequently learned hash codes, but also can be extended to unseen domain.
- The proposed SeGH method has been extensively evaluated on two benchmark datasets and the results show that it achieves superior performance in traditional cross-modal retrieval. In addition, the extended experiments also demonstrate the effectiveness of our method for cross-modal retrieval in the unseen domain, i.e., zero-shot cross-modal retrieval.

The remainder of this paper is organized as follows. Section II briefly introduces some related works on cross-modal hashing and zero-shot hashing. In section III, we elaborate the details of the proposed SeGH, followed by the experimental results and extensive evaluations on two datasets in Section IV. Finally, the conclusion of this work is given in Section V.

II. Related Work

In this section, we will give a brief introduction of the existing work related to our method mainly including cross-modal hashing and zero-shot hashing.

A. Cross-Modal Hashing

Recently, cross-modal hashing has attracted considerable attention and various research works based on it have been proposed. In terms of the utilization of label information, cross-modal hashing approaches can be grouped into two categories: unsupervised and supervised ones.

For unsupervised cross-modal hashing methods, they usually preserve the intrinsic structure by utilizing the co-occurrence information of training data. Song et al. [10] proposed inter-media hashing (IMH) that learns the common Hamming space by maintaining intra-media and inter-media consistency. In addition, Collective Matrix Factorization Hashing (CMFH) [5] and Latent Semantic Sparse Hashing (LSSH) [6] have also been developed to generate the identical hash codes for different modalities of one instance. In particular, CMFH employs collective matrix factorization to project heterogeneous data into a common semantic space and then generates unified binary codes, while LSSH obtains the unified hash codes by mapping two isomorphic semantic feature spaces of extracted image and text feature respectively into a joint abstraction space. Although without available supervised information, these methods can also achieve the impressive performance for cross-modal retrieval.

The supervised ones learn hashing functions by leveraging the class label information, which can further boost the retrieval performance. Zhang et al. [11] proposed Semantic Correlation Maximization (SCM) for large-scale training data, which utilizes the semantic labels to maximize the semantic correlations. Based on CMFH approach, Supervised Matrix Factorization Hashing (SMFH) [7] takes both the label and local structure information into consideration, while Intra- and Inter-modality Similarity Preserving Hashing (IISPH) [8] considers both the intra-modality and inter-modality preservation under the low-dimensional Hamming space. It is worth noting that these methods generate the binary codes by discarding the discrete constraints, which results in large quantization errors and less effective hash codes. To overcome this issue, Discrete Cross-modal Hashing (DCH) [12] and Cross-Modal Discrete Hashing (CMDH) [13] propose the discrete optimization framework to learn the optimized binary codes in a bit-wise manner.

However, the supervision information of these methods is limited to the form of 0/1 labels or pairwise relationships, which neglects the semantic correlation among labels. The proposed SeGH tackles the above problem with class semantics as a guideline, and narrows the semantic gap between independent labels.

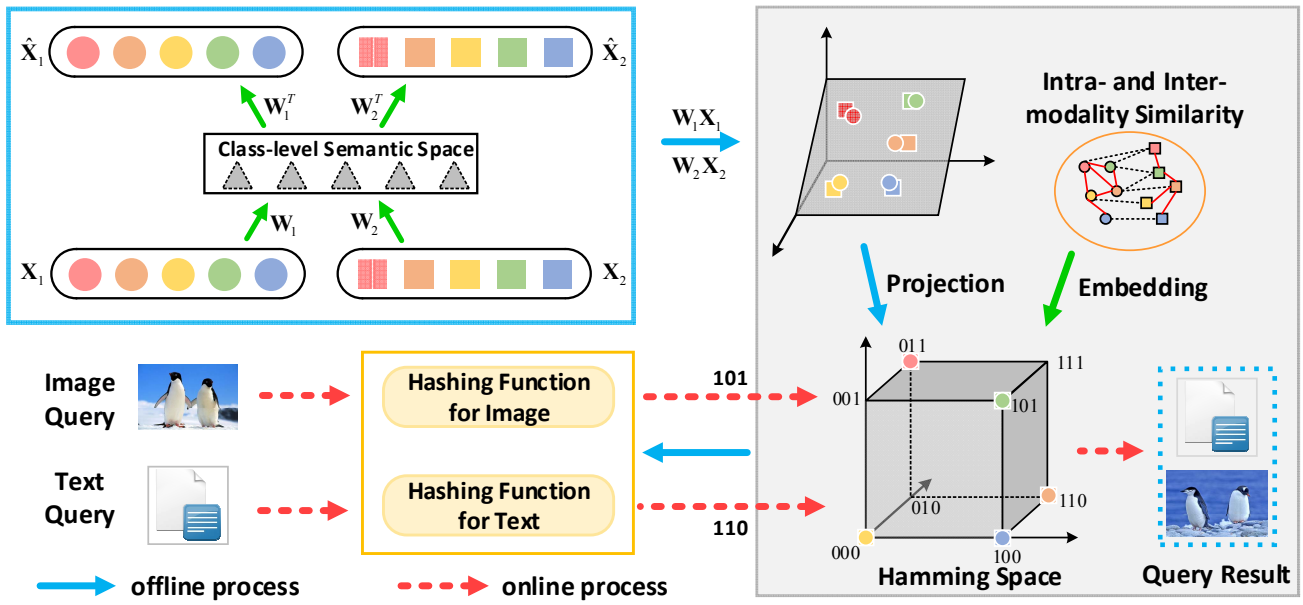


Fig. 1 The overall architecture of the proposed SeGH.

B. *Zero-Shot Hashing*

Zero-shot hashing incorporates the merits of both hashing-based retrieval and zero-shot learning, which can generate the hash codes for newly-emerging categories (unseen classes) by exploiting limited training categories (seen classes).

Most of existing zero-shot hashing methods mainly utilize the word embedding of concepts or human-defined attributes to recognize the unseen classes. For instance, one of the well-known pioneer works was proposed by Yang et al. [14], named Zero-Shot Hashing (ZSH), which transfers the supervised knowledge of seen classes to unseen classes by semantic embedding representation obtained via word2vec model. Inspired by CNN-based hashing, a multi-task framework termed Discrete Similarity Transfer Network (SitNet) [15] was proposed to simultaneously consider the semantic embedding loss and regularized center loss. Instead of leveraging the word vectors, Attribute Hashing (AH) [16] exploits the semantically-rich attribute as the semantic representation, and captures the correlations among features, hash codes, semantic labels. Nevertheless, these zero-shot hashing methods are designed based on single-modality data, thus it is difficult to extend to the cross-modal domain.

To the best of our knowledge, up to now, only one work has been developed to tackle the cross-modal zero-shot hashing retrieval, named Attribute-Guided Network (AgNet) [17], which utilizes attributes to guide the hashing function learning. However, AgNet only solves the tasks of image-based image retrieval (IBIR) and text-based image retrieval (TBIR). The image-based text retrieval (IBTR) task has not been resolved. Therefore, strictly speaking, it does not belong to the zero-shot cross-modal hashing approach, while our SeGH can be extended to zero-shot cross-modal hashing, which can perform both TBIR and IBTR tasks.

III. SEMANTIC-GUIDED HASHING

The details of our proposed semantic-guided hashing are elaborated in this section. For simplicity, we describe the SeGH method with bimodal data consisting of image and text, which can be readily extended to multi-modal scenario.

A. *Problem Definition*

Let $\mathbf{X}_1 \in \mathbb{R}^{d_1 \times n}$ and $\mathbf{X}_2 \in \mathbb{R}^{d_2 \times n}$ be the training data of two modalities which represent the same object. d_1, d_2 are the dimensions of image and text feature, respectively, and n is the number of objects. In addition, we define $\mathbf{Y} \in \{0,1\}^{c \times n}$ as the binary label matrix, where c is the number of categories. Given the length of hash code k, the purpose of our SeGH is to generate the unified binary codes for different modalities of the same objects, and to learn two hashing functions, i.e., $h_1(\mathbf{x}_1): \mathbb{R}^{d_1} \mapsto \{-1,1\}^k$ for image and $h_2(\mathbf{x}_2): \mathbb{R}^{d_2} \mapsto \{-1,1\}^k$ for text.

B. *Overall Framework*

The overall framework of the proposed SeGH approach is illustrated in Fig. 1, including offline and online process. Specifically, offline process aims at hash codes generation and hashing functions learning for out-of-sample data, which consists of two steps, namely discriminative semantic-guided projection learning and hash codes learning. In the first step, according to the off-the-shelf GloVe model, the class-level semantic space is firstly built according to the word vectors of category names. Then, the discriminative projection is learned based on encoder-decoder paradigm guided by class label semantics. In the second step, original data are firstly projected into the common latent semantic space by

leveraging the projections learned in the previous step. Then, the binary codes of heterogeneous data are generated in the Hamming space, while preserving the intra-modality and inter-modality similarity. For online process that performs cross-modal retrieval, one image or text query can be encoded into binary codes by hashing function with the threshold strategy. The relevant results are returned by calculating the Hamming distance between hash codes of query and database.

C. Discriminative Semantic-Guided Projection Learning

Most cross-modal hashing methods generally take advantage of 0/1 form of binary labels or pairwise relationships as supervision information, which implicitly neglects the semantic correlation among different classes. Inspired by the superior capability of word embedding, we consider embedding each category into a 300-dimensional word vector which is extracted with the off-the-shelf GloVe model [22]. Then, the class-level semantic space is constructed by the word vector obtained above of category names. Instead of the independent label matrix described above, the label matrix is represented by the class semantic matrix $\mathbf{S} \in \mathbb{R}^{300\times n}$ in the following part.

Next, the model of encoder-decoder paradigm is developed based on class semantics. Specifically, the matrices $\mathbf{W}_1 \in \mathbb{R}^{300\times d_1}$ and $\mathbf{W}_2 \in \mathbb{R}^{300\times d_2}$ are obtained from feature space $\mathbf{X}_1$ and $\mathbf{X}_2$ to class-level semantic space, respectively. Meanwhile, the semantic space is mapped back to the original space with two projection matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$. Based on the work of [18], we tie the weights to simplify the model, i.e., $\mathbf{W}_1^* = \mathbf{W}_1^T$ and $\mathbf{W}_2^* = \mathbf{W}_2^T$. Given the class semantic matrix $\mathbf{S}$ under the class-level semantic space, this model can be achieved as follows.

$$\begin{aligned} &\min_{\mathbf{W}_1,\,\mathbf{W}_2} \| \mathbf{X}_1 - \mathbf{W}_1^T\mathbf{W}_1\mathbf{X}_1 \|_F^2 + \| \mathbf{X}_2 - \mathbf{W}_2^T\mathbf{W}_2\mathbf{X}_2 \|_F^2 \\ &s.t.\ \ \mathbf{W}_1\mathbf{X}_1 = \mathbf{S},\ \mathbf{W}_2\mathbf{X}_2 = \mathbf{S} \end{aligned} \tag{1}$$

Considering that it is difficult to solve the hard constraints $\mathbf{W}_1\mathbf{X}_1 = \mathbf{S}$ and $\mathbf{W}_2\mathbf{X}_2 = \mathbf{S}$, Equation. (1) is rewritten by relaxing the constraint as stated below.

$$\begin{aligned} J_1 = \min_{\mathbf{W}_1,\,\mathbf{W}_2} &\| \mathbf{X}_1 - \mathbf{W}_1^T\mathbf{S} \|_F^2 + \alpha_1 \| \mathbf{W}_1\mathbf{X}_1 - \mathbf{S} \|_F^2 \\ &+ \| \mathbf{X}_2 - \mathbf{W}_2^T\mathbf{S} \|_F^2 + \alpha_2 \| \mathbf{W}_2\mathbf{X}_2 - \mathbf{S} \|_F^2 \end{aligned} \tag{2}$$

D. Hash Codes Learning

As we can see, the projection matrices $\mathbf{W}_1$ for image and $\mathbf{W}_2$ for text from feature space to common latent semantic space are obtained by solving the problem in (2). Therefore, the common latent semantic representations can be learned based on $\mathbf{W}_1$ and $\mathbf{W}_2$, which are then projected into a Hamming space to generate the hash codes by the projection matrix $\mathbf{P} \in \mathbb{R}^{k\times 300}$. Consequently, the objective function of hash codes learning can be stated as follows.

$$\min_{\mathbf{P},\mathbf{H}} \beta_1 \| \mathbf{P}\mathbf{W}_1\mathbf{X}_1 - \mathbf{H} \|_F^2 + \beta_2 \| \mathbf{P}\mathbf{W}_2\mathbf{X}_2 - \mathbf{H} \|_F^2 + \lambda R(\mathbf{P}) \tag{3}$$

To avoid overfitting, a regularization term $R(\cdot)$ is introduced. Finally, the hash codes can be generated by the sign function, i.e., $\mathbf{B} = sign(\mathbf{H})$.

E. Intra- and Inter-modality Similarity Embedding

In order to learn more fine-grained and discriminative unified hash codes, both intra-modality and inter-modality similarities are embedded into the learning procedure of hash codes and hashing functions.

Firstly, we consider the intra-modality similarity preservation for each modality. Specifically, two nearest neighbor affinity matrices $\mathbf{A}^{(m)}(m = 1, 2)$ are constructed to explore the local geometric structure for different modalities, defined as follows.

$$A_{ij}^{(m)} = \begin{cases} 1, \text{ if } \mathbf{x}_i^{(m)} \in N_k(\mathbf{x}_j^{(m)}) \text{ or } \mathbf{x}_j^{(m)} \in N_k(\mathbf{x}_i^{(m)}) \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

where $N_k(\cdot)$ is defined as the set of k nearest neighbors.

Moreover, the label information is incorporated to maintain the similarity between different modalities. Hence, the similarity matrix $\mathbf{A}^{\text{inter}}$ of heterogeneous data $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_j^{(2)}$ can be defined as:

$$A_{ij}^{\text{inter}} = \begin{cases} 1, \text{ if } \mathbf{x}_i^{(1)} \textit{ and } \mathbf{x}_j^{(2)} \textit{ belong to the same class} \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

Combining the above similarity matrices involving intra-modality and inter-modality, we formulate the similarity preservation as the following form:

$$\begin{aligned} J_{se}(\mathbf{H}) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \| \mathbf{h}_i - \mathbf{h}_j \|^2 (A_{ij}^{(1)} + A_{ij}^{(2)} + A_{ij}^{\text{inter}}) \\ &= tr(\mathbf{H}\mathbf{D}\mathbf{H}^T) - tr(\mathbf{H}\mathbf{A}^{total}\mathbf{H}^T) = tr(\mathbf{H}\mathbf{L}\mathbf{H}^T) \end{aligned} \tag{6}$$

where $tr(\cdot)$ indicates the trace of the matrix. $\mathbf{A}^{total} = \mathbf{A}^{(1)} + \mathbf{A}^{(2)} + \mathbf{A}^{\text{inter}}$ and $\mathbf{D}$ is a diagonal matrix, which can be computed as $D_{ii} = \sum_i A_{ij}^{total}$.

F. Overall Objective Function and Optimization

As we mentioned in the previous section, our SeGH method is composed of two steps. The objective function of the first step that learns the discriminative semantic-guided projection is represented as J_1, as shown in (1). To optimize J_1, we can take the derivative of J_1 with respect to $\mathbf{W}_1$ and $\mathbf{W}_2$ to zero. Then we obtain:

$$\begin{aligned} \mathbf{S}\mathbf{S}^T\mathbf{W}_1 + \mathbf{W}_1\alpha_1\mathbf{X}_1\mathbf{X}_1^T - \mathbf{S}\mathbf{X}_1^T - \alpha_1\mathbf{S}\mathbf{X}_1^T = 0 \\ \mathbf{S}\mathbf{S}^T\mathbf{W}_2 + \mathbf{W}_2\alpha_2\mathbf{X}_2\mathbf{X}_2^T - \mathbf{S}\mathbf{X}_2^T - \alpha_2\mathbf{S}\mathbf{X}_2^T = 0 \end{aligned} \tag{7}$$

which is the well-known Sylvester equation [19] with the following formulation of $\mathbf{AW}+\mathbf{WC}+\mathbf{D}=0$ and it can be solved using the *lyap* function in MATLAB.

In the second step, the objective function combining hash codes learning in (3) and similarity embedding in (6) is formulated as follows:

$$J_2 = \min_{\mathbf{P},\mathbf{H}} \beta_1 \| \mathbf{W}_1\mathbf{X}_1 - \mathbf{H} \|_F^2 + \beta_2 \| \mathbf{W}_2\mathbf{X}_2 - \mathbf{H} \|_F^2 + \gamma tr(\mathbf{HLH}^T) + \lambda \mathrm{R}(\mathbf{P}) \tag{8}$$

where $\beta_1, \beta_2, \gamma, \lambda$ are balance parameters. However, it is hard to directly resolve due to two unknown variables. Here, an iterative method with the following steps is adopted to optimize this formulation.

Step 1: Fix $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{H}$, let $\frac{\partial J_2}{\partial \mathbf{P}} = 0$, we have:

$$\mathbf{P} = (2\beta_1\mathbf{H}\mathbf{X}_1^T\mathbf{W}_1^T + 2\beta_2\mathbf{H}\mathbf{X}_2^T\mathbf{W}_2^T)(\beta_1\mathbf{W}_1\mathbf{X}_1\mathbf{X}_1^T\mathbf{W}_1^T + \beta_2\mathbf{W}_2\mathbf{X}_2\mathbf{X}_2^T\mathbf{W}_2^T + 2\lambda\mathbf{I})^{-1} \tag{9}$$

Step 2: Fix $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{P}$, let $\frac{\partial J_2}{\partial \mathbf{H}} = 0$, then we can obtain:

$$\mathbf{H} = (2\beta_1\mathbf{P}\mathbf{W}_1\mathbf{X}_1 + 2\beta_2\mathbf{P}\mathbf{W}_2\mathbf{X}_2)[2(\beta_1+\beta_2)\mathbf{I} + \gamma(\mathbf{L}^T + \mathbf{L})]^{-1} \tag{10}$$

The whole procedure of the proposed SeGH is summarized in Algorithm 1. Once the projection matrices $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{P}$ are obtained, the hashing function for different modalities can be easily generated according to the following equation, i.e., $h_1(\mathbf{x}_1) = sign(\mathbf{P}\mathbf{W}_1\mathbf{x}_1)$ for image and $h_2(\mathbf{x}_2) = sign(\mathbf{P}\mathbf{W}_2\mathbf{x}_2)$ for text.

Moreover, the objective function for our extension version with unseen domain as mentioned in Section I is same to all the formulations above. The only difference from traditional cross-modal hashing is that the training sets comprise of instances of seen classes, while the query sets are only from unseen classes.

G. Complexity Analysis

The computational complexity of our SeGH is discussed as follows. In the training phase, the time consuming mainly includes semantic-guided projection learning, graph Laplacian matrix construction and hash codes learning. Typically, two projection $\mathbf{W}_1$ and $\mathbf{W}_2$ can be computed at a cost of $O(v^3)$, where $v = 300$. The complexity for computing graph Laplacian matrix $\mathbf{L}$ is $O(dn^2)$, where $d = \max\{d_1, d_2\}$. For hash code learning, it involves the alternating updates of $\mathbf{P}$ and $\mathbf{H}$. Solving (9) for $\mathbf{P}$ and (10) for $\mathbf{H}$ requires the time complexity of $O(t(d(k+v)n + kvd + v^2d + v^3 + kv^2))$ and $O(t(n^3 + kn^2 + kdn + kvd))$ respectively, where t is the number of iterations. Because k, d and $v \ll n$, the overall time cost is approximately $O(dn^2 + t(n^3 + d(k+v)n))$. In the testing phase, the time complexity of each query is constant with $O(kd)$.

Algorithm 1 Semantic-Guided Hashing

Input: Feature matrices $\mathbf{X}_1$ for image and $\mathbf{X}_2$ for text, class semantic matrix $\mathbf{S}$, binary label $\mathbf{Y}$, hash code length k, and parameters $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma, \lambda$.

Output: Binary hash codes $\mathbf{B} \in \{-1,1\}^{k \times n}$, projection matrices $\mathbf{W}_1, \mathbf{W}_2, \mathbf{P}$.

1: Randomly initialize $\mathbf{W}_1, \mathbf{W}_2$, respectively.
2: Compute $\mathbf{W}_1$ and $\mathbf{W}_2$ by solving (7).
3: Randomly initialize $\mathbf{P}, \mathbf{H}$, and construct the graph Laplacian matrix $\mathbf{L}$ by (4) and (5).
4: **repeat**
5: Fix $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{H}$, update $\mathbf{P}$ by (9).
6: Fix $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{P}$, update $\mathbf{H}$ by (10).
7: **until** convergence
8: $\mathbf{B} = sign(\mathbf{H})$

Compared to the time complexity of training most baseline methods such as CMFH, SMFH and IISPH, which are at least $O(n^2)$, our method is relatively higher due to the inverse calculation for graph Laplacian matrix. The other baseline approaches such as SCM_Orth and SCM_Seq mainly focus on the complexity problem and get lower time cost, but their retrieval performance may be less optimistic on some datasets.

Although our method has a relatively higher time complexity, considering its great advantages in overall performance, we can learn that SeGH can be competitive with the baselines.

IV. Experiment

In this section, we carry out the extensive experiments on two public benchmark datasets, and validate the performance of our method in comparison with several state-of-the-art cross-modal hashing approaches. In addition, the extended experiments are also conducted to demonstrate the applicability and effectiveness of our method on zero-shot cross-modal retrieval tasks.

A. Datasets

LabelMe [20] consists of 2686 images with annotated several tags, which can be grouped into 8 outdoor scene categories. Each image is represented by 512-dimensional GIST feature, and each text is depicted with 366-demensional index vectors of tags. In our experiments, we randomly select 2014 samples as the training set, the rest as testing set. Moreover, all categories are randomly split into 5 seen and 3 unseen categories for each round in the extended experiment.

TABLE I. The mAP results on LabelMe and Pascal VOC 2007 datasets. The best result is shown in boldface.

Task	Method	LabelMe					Pascal VOC 2007				
		8 bits	*16 bits*	*32 bits*	*64 bits*	*128 bits*	*8 bits*	*16 bits*	*32 bits*	*64 bits*	*128 bits*
Img to Txt	SCM_Orth	0.1502	0.1494	0.1456	0.1471	0.1486	0.2031	0.1565	0.1362	0.1287	0.1232
	SCM_Seq	0.1956	0.2554	0.3253	0.2451	0.3388	0.1956	0.2554	0.3253	0.2451	0.3388
	CMFH	0.3991	0.3908	0.3827	0.3669	0.3880	0.1677	0.1861	0.1853	0.1777	0.1728
	LSSH	0.5816	0.5882	0.6360	0.6580	0.6522	0.2293	0.2532	0.2705	0.2762	0.2834
	IISPH	0.3961	0.3811	0.3739	0.3666	0.3720	0.1649	0.1889	0.1855	0.1768	0.1714
	SMFH	0.4710	0.5903	0.6238	0.6386	0.6898	0.1865	0.2233	0.2324	0.2390	0.2345
	SeGH	**0.6154**	**0.7201**	**0.7775**	**0.8116**	**0.8263**	**0.2752**	**0.3301**	**0.3510**	**0.3727**	**0.4042**
Txt to Img	SCM_Orth	0.1333	0.1333	0.1323	0.1313	0.1288	0.2387	0.1982	0.1484	0.1197	0.1006
	SCM_Seq	0.2037	0.2989	0.4108	0.2652	0.4531	0.2037	0.2989	0.4108	0.2652	0.4531
	CMFH	0.5348	0.4973	0.4771	0.4572	0.4773	0.3648	0.4838	0.5412	0.5173	0.4960
	LSSH	0.5771	0.5994	0.6238	0.6522	0.6496	0.4423	0.5396	0.6059	0.6198	0.6311
	IISPH	0.5282	0.4937	0.4750	0.4573	0.4578	0.3512	0.4880	0.5458	0.5192	0.4939
	SMFH	0.6719	0.8124	0.8121	0.8010	0.8088	0.3729	0.5837	0.6556	0.6518	0.6140
	SeGH	**0.8315**	**0.8952**	**0.9194**	**0.9260**	**0.9336**	**0.7019**	**0.8289**	**0.8608**	**0.8831**	**0.9011**

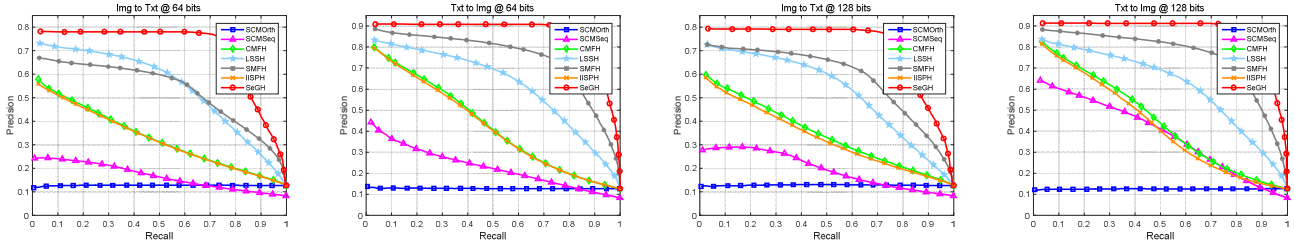


Fig. 2. Comparison of Precision-recall curves with hash codes @ 64 and 128 bits on both tasks of LabelMe dataset.

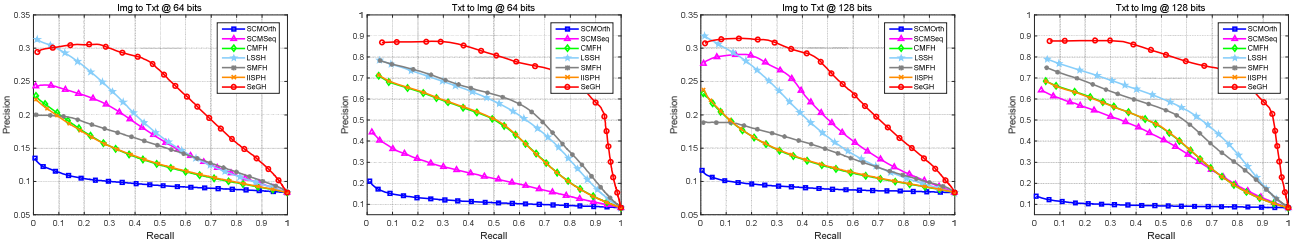


Fig. 3. Comparison of Precision-recall curves with hash codes @ 64 and 128 bits on both tasks of Pascal VOC 2007 dataset.

Pascal VOC 2007 [21] was downloaded from Flickr with 5011 training and 4952 testing image-text pairs. The images with multi-labels are discarded, resulting in 2808 images for training and 2841 for testing. Each image-text pair is labeled with one of 20 semantic categories. For each instance, the image is detailed with 512-dimensional GIST feature and the text is represented by 399-dimensional word frequency vector. In the extended experiment, 4 categories are randomly selected as unseen categories and the others as seen ones for each round.

B. Protocols and Baseline Methods

The performance of cross-modal hashing methods is measured on two different retrieval tasks, namely 'Txt to Img' and 'Img to Txt'. In both tasks, we adopt two types of evaluation metrics, i.e., mean average precision (mAP) and precision-recall curves. The mAP is the mean of the average precision (AP) for all the query samples, and AP is computed as $\mathrm{AP}=\frac{1}{T}\sum_{r=1}^{R}P(r)\delta(r)$, where T denotes the number of relevant instances in top R retrieved results and $P(r)$ indicates the precision of top r retrieved instances. $\delta(r)=1$ means the r-th result is related to the query, and 0 otherwise. Furthermore, the precision-recall curves reflect the variation of precision with respect to different recall, which are widely adopted to evaluate the performance of retrieval tasks.

Our method is compared against six state-of-the-art cross-modal hashing approaches which are CMFH [5], LSSH [6], SCM_Orth [11], SCM_Seq [11], IISPH [8], SMFH [7], respectively. They can be classified into two group: CMFH and LSSH are unsupervised methods, and the remaining are supervised ones. For the extended experiment

in unseen domain, two zero-shot hashing methods including ZSH [14] and AH [16] are added to comprehensively evaluate the retrieval performance. The parameter settings of these methods in our experiment are consistent with those in their original papers.

C. Experiment Results

1) Results on LabelMe

Table I reports the mAP values of SeGH and six baseline approaches on LabelMe dataset with 8, 16, 32, 64, 128 bits. From Table I, we easily observe that SeGH achieves the best mAP scores on both retrieval tasks, which demonstrates the effectiveness and superiority of our method. Note that most of the methods have higher mAP score of Txt-to-Img task than that of Img-to-Txt task. This may mean that it is more difficult to capture the latent semantic information in images than texts.

The precision-recall curves on LabelMe dataset with 64 bits and 128 bits are plotted in Fig. 2. Similar to the results of mAP, it can be observed that SeGH significantly outperforms all the baseline methods on different tasks. Additionally, we also find that our SeGH performs better with longer hash codes, this is because the fact that more discriminative information can be encoded into binary codes as the code length increases.

2) Results on Pascal VOC 2007

The mAP values of SeGH and all the baselines on Pascal VOC 2007 dataset are presented in Table I. It can be seen from the table, the mAP results of the proposed SeGH are superior to other baseline methods on both tasks. In particular, compared with the second best method, SeGH gains a significant increment of 20.5% to 27% for Txt-to-Img task. Furthermore, with the increasing of hash code length, some methods such as SCM_Orth, CMFH and IISPH decrease to some extent, while our method achieves the continuous performance improvement.

Fig. 3 shows the precision-recall curves on Pascal VOC 2007 under the setting of 64 and 128 bits. It is clearly to find that our SeGH achieves superior performance against most of the baselines apart from LSSH, which is consistent with the above results on LabelMe dataset. It is worth noting that the unsupervised method LSSH is almost comparable to or even outperforms all the supervised approaches on Pascal VOC 2007, while the proposed SeGH still gains the best results in Txt-to-Img task. However, LSSH has an advantage over SeGH at the beginning stage of Img-to-Txt task. We conjecture that a large number of label information to constrain the binary codes may be too strict for the supervised approaches on Pascal VOC 2007 dataset.

Considering the advantages of our approach in all experiments, it can be concluded that the proposed method has the promising ability to deal with cross-modal retrieval tasks, and can be competitive with several state-of-the-art approaches.

D. Extended Experiments for Unseen Domain

To validate the effectiveness of our method for zero-shot cross-modal retrieval, the extended experiments are carried out in this section. The mAP score is adopted to evaluate the retrieval performance in unseen domain. Since unseen classes are randomly sampled for each time, the average results over 20 times are taken as final results to avoid unstable results in all methods. The mAP values of SeGH and all comparative methods on LabelMe and Pascal VOC 2007 datasets are shown in Fig. 4. It should be noticed that SeGH achieves the highest mAP scores on both two datasets with all code length cases consistently, while other approaches show relatively poor performance because they cannot capture the common characteristics of seen and unseen classes. Moreover, similar to phenomenon of previous experiments for traditional cross-modal retrieval, the mAP values of our SeGH gradually increase as the code length varies from 8 to 128 bit. Besides, an interesting observation is that single-modal zero-shot hashing methods such as AH and ZSH drastically outperform supervised cross-modal hashing methods such as SCM_Orth and SCM_Seq on both tasks of LabelMe dataset. This confirms that traditional close-set retrieval approaches may suffer from serious performance degradation in the scenarios dealing with unseen classes, which also indicates that the proposed SeGH has the ability to apply to zero-shot problem.

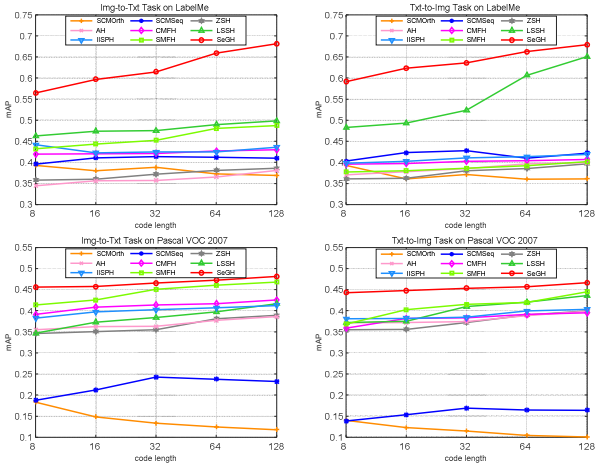


Fig. 4. The mAP values of different methods for zero-shot cross-modal hashing retrieval on LabelMe and Pascal VOC 2007 datasets with both tasks.

E. Parameter Sensitivity Analysis

From the overall objective function mentioned above, the proposed SeGH has six parameters, namely $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma, \lambda$, respectively. In this section, we only analyze the effect of different parameter settings for both tasks on Pascal VOC 2007 dataset because LabelMe dataset has the similar results for different parameters. In particular, the length of hash codes is fixed as 64, and the experiments on one parameter are performed by keeping the value of other parameters unchanged. Fig. 5 shows the mAP results of SeGH under different setting of six parameters on both tasks of Pascal VOC 2007 dataset. It can be observed that the proposed SeGH is insensitive to all the parameters and also validates that SeGH can achieve outstanding results over a wide range of parameter values. To compare with other baselines, we

empirically set α_1 and α_2 to 10^3, β_1 and β_2 both to 10^{-3}, γ and λ both to 10^{-2} in all the previous experiments.

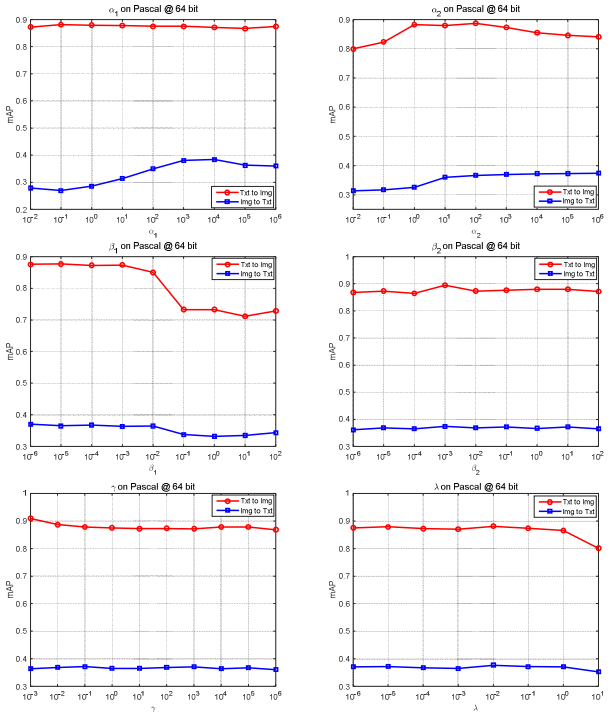
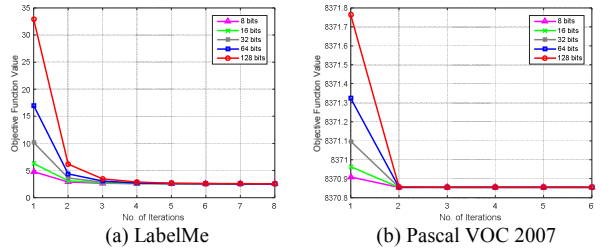


Fig. 5. Parameter sensitivity analysis on Pascal VOC 2007 dataset.

F. *Convergence Study*

Since the iterative procedure is employed to optimize the objectives in Algorithm 1, here we analyze its convergence property on all datasets. Fig. 6 (a) and (b) plot the convergence curves of objective function value with different code lengths on LabelMe and Pascal VOC 2007 datasets, respectively. It can be observed that our SeGH converges quickly on both datasets. Particularly, SeGH can converge within only 6 iterations for LabelMe dataset, and for Pascal dataset, it converges at 3-th iteration in the case of all code lengths, which indicates the high efficiency of the proposed method. Combined with the previous experimental results, it demonstrates that the proposed SeGH can achieve remarkable performance with fast convergence rate.



(a) LabelMe (b) Pascal VOC 2007

Fig. 6. Convergence study on two datasets.

V. Conclusion

In this paper, a novel two-step supervised hashing approach named Semantic-Guided Hashing (SeGH), has been proposed to solve the cross-modal retrieval problem, which aims at gaining the discriminative binary codes guided by label semantics. Specifically, the class-level semantic space is firstly constructed by using the word vector obtained by category name. Based on label semantics under this space, a model of encoder-decoder paradigm is introduced to learn the projection matrix, and further to obtain the common latent space. Finally, the discriminative binary codes can be generated by mapping common representations into Hamming space. Extensive experiments on two public datasets validated that the proposed method can achieve superior performance for cross-modal retrieval, and also demonstrated that SeGH has the ability to deal with unseen domain problem.

Acknowledgment

This work is jointly supported by the National Key Research and Development Program of China under Grant 2018YFC0831305 and the Nature Science Foundation of China under Grant 61672123.

References

[1] J. C. Pereira et al., "On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 3, pp. 521–535, Mar. 2014.

[2] C. Chaudhary, P. Goyal, J. Ruben Antony Moniz, N. Goyal, and Y.-P. P. Chen, "Linguistic Patterns and Cross Modality-based Image Retrieval for Complex Queries," in Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, New York, NY, USA, 2018, pp. 257–265.

[3] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval," in Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, New York, NY, USA, 2018, pp. 19–27.

[4] P. Xu et al., "Cross-modal subspace learning for fine-grained sketch-based image retrieval," Neurocomputing, vol. 278, pp. 75–86, Feb. 2018.

[5] G. Ding, Y. Guo, and J. Zhou, "Collective Matrix Factorization Hashing for Multimodal Data," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2083–2090.

[6] J. Zhou, G. Ding, and Y. Guo, "Latent Semantic Sparse Hashing for Cross-modal Similarity Search," in Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, New York, NY, USA, 2014, pp. 415–424.

[7] J. Tang, K. Wang, and L. Shao, "Supervised Matrix Factorization Hashing for Cross-Modal Retrieval," IEEE Transactions on Image Processing, vol. 25, no. 7, pp. 3157–3166, Jul. 2016.

[8] Z. Chen, F. Zhong, G. Min, Y. Leng, and Y. Ying, "Supervised Intra- and Inter-Modality Similarity Preserving Hashing for Cross-Modal Retrieval," IEEE Access, vol. 6, pp. 27796–27808, 2018.

[9] E. Kodirov, T. Xiang, and S. Gong, "Semantic Autoencoder for Zero-Shot Learning," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4447–4456.

[10] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media Hashing for Large-scale Retrieval from Heterogeneous Data Sources," in Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 2013, pp. 785–796.

[11] D. Zhang and W.-J. Li, “Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization,” in Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[12] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, “Learning Discriminative Binary Codes for Large-scale Cross-modal Retrieval,” IEEE Transactions on Image Processing, vol. 26, no. 5, pp. 2494–2507, May 2017.

[13] V. E. Liong, J. Lu, and Y.-P. Tan, “Cross-Modal Discrete Hashing,” Pattern Recognition, vol. 79, pp. 114–129, Jul. 2018.

[14] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen, “Zero-Shot Hashing via Transferring Supervised Knowledge,” in Proceedings of the 24th ACM International Conference on Multimedia, New York, NY, USA, 2016, pp. 1286–1295.

[15] Y. Gao, Y. Guo, G. Ding, and J. Han, “SitNet: Discrete Similarity Transfer Network for Zero-shot Hashing,” in Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 2017, pp. 1767-1773.

[16] Y. Xu, Y. Yang, F. Shen, X. Xu, Y. Zhou, and H. T. Shen, “Attribute hashing for zero-shot image retrieval,” in 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 133–138.

[17] Z. Ji, Y. Sun, Y. Yu, Y. Pang, and J. Han, “Attribute-Guided Network for Cross-Modal Zero-Shot Hashing,” arXiv:1802.01943 [cs], Feb. 2018.

[18] M. aurelio Ranzato, Y. -la. Boureau, and Y. L. Cun, “Sparse Feature Learning for Deep Belief Networks,” in Advances in Neural Information Processing Systems 20, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1185–1192.

[19] S.-G. Lee and Q.-P. Vu, “Simultaneous solutions of Sylvester equations and idempotent matrices separating the joint spectrum,” Linear Algebra and its Applications, vol. 435, no. 9, pp. 2097–2109, Nov. 2011.

[20] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: A Database and Web-Based Tool for Image Annotation,” Int J Comput Vis, vol. 77, no. 1, pp. 157–173, May 2008.

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge,” Int J Comput Vis, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[22] J. Pennington, R. Socher, C. Manning,“Glove: Global Vectors for Word Representation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar,2014, pp. 1532–1543.