Semantics-Reconstructing Hashing for Cross-Modal Retrieval

Peng-Fei Zhang[1], Zi Huang[1(✉)], and Zheng Zhang[2]

[1] School of Information Technology and Electrical Engineering, University of Queensland, Saint Lucia, QLD 4072, Australia
mima.zpf@gmail.com, huang@itee.uq.edu.au
[2] Bio-Computing Research Center, Harbin Institute of Technology, Shenzhen 518055, China
darrenzz219@gmail.com

Abstract. Retrieval on Cross-modal data has attracted extensive attention as it enables fast searching across various data sources, such as texts, images and videos. As one of the typical techniques for cross-model searching, hashing methods project features with high dimension into short-length hash codes, thus effectively improving storage and retrieval efficiency. Recently, many efforts have been made to widely study supervised methods with promising performance. However, there still remain some problems. Conventionally, hash codes and projection functions are learnt by preserving the pairwise similarities between data items, which neglects the discriminative property of class associated with each data item. Most of the existing methods that utilise class labels also undertake the binary codes learning under a classification frame. The relations between binary codes and labels have not been well considered. To tackle these problems, we propose a shallow supervised hash learning method – Semantics-reconstructing Cross-modal Hashing (SCH), which reconstructs semantic representation and learns the hash codes for the entire dataset jointly. For the semantic reconstruction, the learned semantic representation is projected back into label space, extracting more semantic information. By leveraging reconstructed semantic representations, the hash codes are learnt by considering the underlying correlations between labels, hash codes and original features, resulting in a further performance improvement. Moreover, SCH learns the hash codes and functions without relaxing the binary constraints simultaneously, therefore, it further reduces the quantization errors. In addition, the linear computational complexity of its training makes it practicable to big data. Extensive experiments show that the proposed SCH can perform better than the state-of-the-art baselines.

Keywords: Cross-modal hashing · Supervised learning · Discrete optimization

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12085, pp. 315–327, 2020.
https://doi.org/10.1007/978-3-030-47436-2_24

1 Introduction

Recently, the tremendous growing of multimedia data has greatly increased the demand of effective and efficient store and retrieval techniques. Therefore, many hashing-based methods have appealed much attention, mapping instances into binary codes with the short bit-length in a Hamming space and performing the search with the bit-wise XOR operation [1,5,6,10]. Thus, the search becomes much efficient and the storage can be dramatically reduced [4,8,15]. Most pioneer hashing methods are exploited to deal with unimodal searching tasks. However, in real world, multimedia data more often comes with multi-modalities, e.g., a piece of article on many websites often contains some textual contents and a few pictures to attract readers. In many scenarios, people need to retrieve data in different modalities, e.g., searching target images with a certain sentence, or vice versa [16]. Therefore, cross-modal hashing recently has seen a tremendous surge in interest within multimedia community, and many unsupervised and supervised methods have been explored to deal with corresponding tasks. Specifically, without semantic supervised information, unsupervised methods exploit the similarity relationship between original features as the guidance of the binary codes and functions learning. By contrary, supervised ones are able to explore the associated semantic information, e.g., labels/tags, thus performing better than unsupervised ones.

However, there still remain several problem needed to be addressed in existing supervised cross-modal hashing methods. First, some conventional methods learn hash codes and projection functions by preserving the pairwise similarities between data items, neglecting the discriminative property of class associated with each data item and encountering the computationally prohibitive limitation to handle large-scale datasets. Secondly, most of methods that undertake the binary codes learning under a classification frame have not well exploited the relations between the hash codes and the labels. And thirdly, some methods directly discard the discrete constraints during the optimization procedure, which inevitably leads to the large errors of quantization.

To deal with these, in our work, we propose a novel supervised hashing method, namely Semantics-reconstructing Cross-modal Hashing (SCH). It leverages a semantic representation of labels by reconstruction to learn binary codes, In light of this, the sufficient and discriminative semantics are preserved. Moreover, our SCH can effectively obtain the unified binary codes and learn the modality-specific hash functions for the whole dataset simultaneously, such that, the quantization errors can be significantly reduced. In addition, the resulting discrete optimization problem is tackled in a linear computational complexity, such that our hash learning method can be effectively applied to deal with searching tasks for big data. Extensive experiments conducted on three benchmark datasets, i.e., Wiki, MIRFlickr-25K, and NUS-WIDE, demonstrate that SCH obtains promising results and outperforms state-of-the-art cross-modal hashing baselines. To summarize, the main contributions of our work are listed as follows:

– We propose a scalable supervised hashing algorithm, which simultaneously learns the hash codes and functions in one-step learning framework.

– An efficient semantics reconstructing strategy is proposed to preserve supervised semantic information as much as possible, as the result, the performance would be improved.
– An efficient learning scheme is designed to cope with the discrete optimization problem in SCH. The linear time complexity of training making it scalable to large-scale data set.
– Extensive experiments conducted on three widely used datasets demonstrate the superiority of our SCH.

2 Related Work

To better introduce our work, we give a brief overview of some representative hash methods for cross-modal searching which can be coarsely categorised into unsupervised and supervised learning methods.

Without supervised information like tags available, unsupervised hasing methods learn hash codes for the original samples. One typical method is IMH [14], which learn to find a common Hamming space so that they can consistently connect and represent different types of media data. To avoid time-consuming graph construction for large-scale datasets, in LCMH [21], authors proposed to find a small number of cluster centers to represent the original data points for hash codes and functions learning. Besides, CMFH [2] generates hash codes unified for media data from heterogeneous data sources by collective matrix factorization strategy, which can enable cross-modal retrieval and improve searching performance.

In contrast, supervised ones are able to explore the associated semantic information, e.g., labels/tags, to obtain the hash codes or the hash functions. For instance, in order to learn each bit of the binary codes well, in CRH [19], authors design a learning algorithm called boosted co-regularization and also defines the modality-specific large-margin with labels to further improve performance. SePH [9] learns a probability distribution for original data points, and then approximates it with the binary codes. The final hash codes can be obtained by minimizing the KL-divergence on probability distribution and binary codes. DCH [17] propose a novel algorithm to directly learn the hash projection functions specific for each modality and the discriminative hash codes without discarding the discrete binary constraints. SDMCH [12] combines the nonlinear manifold learning with hashing learning, and constructs the correlation across data of multiple modalities to improve the performance.

3 Semantics-Reconstructing Hashing

3.1 Notations

For simplicity, we suppose each instance contains two modalities. However, it can be easily extended to deal with the conditions of more modalities, as shown later in this paper. The training dataset is $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{n}$, where $\mathbf{x}_i^{(1)} \in \mathbb{R}^{d_1}$ and

$\mathbf{x}_i^{(2)} \in \mathbb{R}^{d_2}$ denote the d_1-dimension image feature vector and the d_2-dimension text feature vector of the i-th instance, respectively. Their matrix representations are $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively. $\mathbf{Y} = \{0,1\}^{n\times l}$ is the ground-truth label matrix where $\mathbf{Y}_{ij} = 1$ indicates the i-th sample is in class j and 0 otherwise. Given the training data, the purpose of our method is to learn the unified hash codes $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^n$ for different modalities, where $\mathbf{b}_i = \{0,1\}^k$, k is the bit length.

3.2 Semantics Reconstructing

For purpose of making use of the full label information and make the optimization problem easy to be solved, we first introduce an semantic representation $\mathbf{F}$ which can be learned under a classification framework and the semantic labels are set as the guidance. In light of this, we define the problem as follows:

$$\min_{\mathbf{F},\mathbf{U}} \|\mathbf{Y} - \mathbf{F}\mathbf{U}\|_F^2, \quad s.t. \ \ \mathbf{F} \in \mathbb{R}^{n\times k}, \tag{1}$$

where $\mathbf{U}$ is a projection matrix.

To further reduce the errors, we assume the learned semantic representation can be reconstructed from the label matrix $\mathbf{Y}$. Then, the problem is reformulated as follows:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{F}} \alpha \|\mathbf{Y} - \mathbf{F}\mathbf{U}\|_F^2 + \beta \|\mathbf{F} - \mathbf{Y}\mathbf{V}\|_F^2, \quad s.t. \ \ \mathbf{F} \in \mathbb{R}^{n\times k}, \tag{2}$$

where $\mathbf{U}$ and $\mathbf{V}$ represent the projection matrices, $\alpha > 0$ and $\beta > 0$ are balance parameters. In light of this, we can reconstruct the semantic representation $\mathbf{F}$ from labels so as to adequately extract discriminative semantic information from the labels.

Thereafter, we suppose the hash codes can be learned from the semantic representation $\mathbf{F}$ with a rotation matrix. For this purpose, we define the following optimization problem:

$$\min_{\mathbf{B},\mathbf{R}} \|\mathbf{B} - \mathbf{F}\mathbf{R}\|_F^2, \quad s.t. \ \ \mathbf{F} \in \mathbb{R}^{n\times k}, \mathbf{B} \in \{-1,1\}^{n\times k}, \mathbf{R}\mathbf{R}^\mathsf{T} = \mathbf{I}. \tag{3}$$

It is worth noting that Eq. (2) and Eq. (3) can be merged into one equation if we replace the semantic representation $\mathbf{F}$ with the hash code matrix $\mathbf{B}$, which is also able to directly learn the hash codes. However, we have to encounter some problems. First, the optimization problem becomes troublesome to deal with. Although some strategies like discrete cyclic coordinate descent (DCC) in the work SDH [13] have been use to solve similar discrete optimization iteratively, such bit-wise optimization is time-consuming. Secondly, it is not robust to noise when directly using the hash codes for the projection matrix learning which maps the samples from the original feature space into the hash space.

3.3 Hash Functions Learning

To gain efficient binary projection functions for multi-modal data, we need to consider how to preserve the similarity relationships across various modalities.

To address this, we project data from different feature spaces into a common subspace and define the objective function as follows:

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{W}_t} \sum_{t=1}^{2} \lambda_t \left\| \mathbf{F} - f_t(\mathbf{X}^{(t)}) \right\|_F^2 + \sum_{t=1}^{2} \gamma \left\| \mathbf{W}_t \right\|_F^2, \\ &s.t.\ \ \mathbf{F} \in \mathbb{R}^{n \times k}, f_t(\mathbf{X}^{(t)}) = \phi(\mathbf{X}^{(t)})\mathbf{W}_t, \sum_{t=1}^{2} \lambda_t = 1, \end{aligned} \tag{4}$$

where $\mathbf{F}$ is the semantic representation, matrix $\mathbf{X}^{(t)}$ represent the features of the t-th modality, and $\lambda_t > 0$ and $\gamma > 0$ are balance parameters. $f_t(\mathbf{X}^{(t)}) = \phi(\mathbf{X}^{(t)})\mathbf{W}_t$ is the mapping function, $\mathbf{W}_t$ indicates the projecting matrix for the t-th modality, and $\phi(\mathbf{X})^{(t)}$ is a nonlinear embedding of $\mathbf{X}^{(t)}$, In our work, we choose the RBF kernel, In particular, $\phi(x) = [exp(\frac{-\|x-\hat{x}_1\|_2^2}{2\sigma^2}), ..., exp(\frac{-\|x-\hat{x}_c\|_2^2}{2\sigma^2})]$, where $\{\hat{x}_j\}_{j=1}^{c}$ are c anchor samples randomly selected from the training instances $\{x_i\}_{i=1}^{n}$ and σ is the kernel number.

3.4 Final Objective Function

Integrating the above Eq. (2), (3) and (4) together, we obtain the final objective function:

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{U},\mathbf{V},\mathbf{W}_t,\mathbf{R}} \alpha \left\| \mathbf{Y} - \mathbf{F}\mathbf{U} \right\|_F^2 + \beta \left\| \mathbf{F} - \mathbf{Y}\mathbf{V} \right\|_F^2 + \mu \left\| \mathbf{B} - \mathbf{F}\mathbf{R} \right\|_F^2 \\ &\quad + \sum_{t=1}^{2} \lambda_t \left\| \mathbf{F} - f_t(\mathbf{X}^{(t)}) \right\|_F^2 + \rho \ell(\mathbf{U}, \mathbf{V}, \sum_{t=1}^{2} \mathbf{W}_t), \\ &s.t.\ \ \mathbf{F} \in \mathbb{R}^{n \times k}, \mathbf{B} \in \{-1, 1\}^{n \times k}, \sum_{t=1}^{2} \lambda_t = 1, f_t(\mathbf{X}^{(t)}) = \phi(\mathbf{X}^{(t)})\mathbf{W}_t, \mathbf{R}\mathbf{R}^{\mathsf{T}} = I, \end{aligned} \tag{5}$$

where $\alpha > 0$, $\beta > 0$, $\rho > 0$ and $\mu > 0$ are balance parameters. By reconstructing the semantic representation from labels, the first two terms can make the semantic representation contain the substantial semantic information of labels. By building the projection from semantic representation to the hash codes with the third term, we can directly obtain the hash codes without relaxation so that the quantization errors may be reduced. The fourth one is utilized to generate the modality-specific hash functions; more specifically, it maps the samples from multiple data sources into a common space, and preserves the similarity between them. The last is a regularizer which is defined as follows:

$$\ell(\mathbf{U}, \mathbf{V}, \sum_{t=1}^{2} \mathbf{W}_t) = \left\| \mathbf{U} \right\|_F^2 + \left\| \mathbf{V} \right\|_F^2 + \sum_{t=1}^{2} \left\| \mathbf{W}_t \right\|_F^2. \tag{6}$$

3.5 Optimization Algorithm

We design an iterative scheme to solve the discrete optimization problem of Eq. (5), which is composed of six steps as shown below.

Step 1: Updating F with other variables fixed.

After fixing other variables, we rewrite Eq. (5) as the following one,

$$\begin{aligned}\min_{\mathbf{F}} \alpha \left\|\mathbf{Y}-\mathbf{F}\mathbf{U}\right\|_F^2 + \beta \left\|\mathbf{F}-\mathbf{Y}\mathbf{V}\right\|_F^2 + \mu \left\|\mathbf{B}-\mathbf{F}\mathbf{R}\right\|_F^2 \\ + \sum_{t=1}^{2} \lambda_t \left\|\mathbf{F}-\phi(\mathbf{X}^{(t)})\mathbf{W}_t\right\|_F^2, \quad s.t. \ \ \mathbf{F}\in\mathbb{R}^{n\times k}.\end{aligned} \tag{7}$$

To solve it, we further simplify Eq. (7) as follows by expanding each item and then removing irrelevant items:

$$\begin{aligned}\min_{\mathbf{F}} -2Tr(\mathbf{F}(\alpha\mathbf{U}\mathbf{Y}^\mathsf{T}+\mu\mathbf{R}\mathbf{B}^\mathsf{T})) - 2Tr(\mathbf{F}^\mathsf{T}(\beta\mathbf{Y}\mathbf{V}+\sum_{t=1}^{2}\lambda_t\phi(\mathbf{X}^{(t)})\mathbf{W}_t)) \\ + \alpha\left\|\mathbf{F}\mathbf{U}\right\|_F^2 + (\beta+1)\left\|\mathbf{F}\right\|_F^2 + \mu\left\|\mathbf{F}\mathbf{R}\right\|_F^2, \quad s.t. \ \ \mathbf{F}\in\mathbb{R}^{n\times k}.\end{aligned} \tag{8}$$

By setting the derivation of Eq. (8) w.r.t. $\mathbf{F}$ equal to zero, we can get the solution:

vspace*vspace*-2mm

$$\mathbf{F} = (\alpha\mathbf{Y}\mathbf{U}^\mathsf{T}+\beta\mathbf{Y}\mathbf{V}+\mu\mathbf{B}\mathbf{R}^\mathsf{T}+\sum_{t=1}^{2}\lambda_t\phi(\mathbf{X}^{(t)})\mathbf{W}_t)(\alpha\mathbf{U}\mathbf{U}^\mathsf{T}+\mu\mathbf{R}\mathbf{R}^\mathsf{T}+(\beta+1)\mathbf{I})^{-1}. \tag{9}$$

Step 2: Updating U with other variables fixed.

With other variables fixed, Eq. (5) is reformulated as follows:

$$\min_{\mathbf{U}} \alpha\left\|\mathbf{Y}-\mathbf{F}\mathbf{U}\right\|_F^2 + \rho\left\|\mathbf{U}\right\|_F^2. \tag{10}$$

After expanding each item and then removing irrelevant items, we further simplify Eq. (10) to the following one:

$$\min_{\mathbf{U}} \alpha(-2Tr(\mathbf{F}\mathbf{U}\mathbf{Y}^\mathsf{T}) + \left\|\mathbf{F}\mathbf{U}\right\|_F^2) + \rho\left\|\mathbf{U}\right\|_F^2. \tag{11}$$

By setting the derivation of Eq. (11) w.r.t. $\mathbf{U}$ equal to zero, we can obtain the following solution:

$$\mathbf{U} = (\mathbf{F}^\mathsf{T}\mathbf{F}+\frac{\rho}{\alpha}\mathbf{I})^{-1}\mathbf{F}^\mathsf{T}\mathbf{Y}. \tag{12}$$

Step 3: Updating V with other variables fixed.

Similarly, with other variables fixed, Eq. (5) becomes:

$$\min_{\mathbf{V}} \beta\left\|\mathbf{F}-\mathbf{Y}\mathbf{V}\right\|_F^2 + \rho\left\|\mathbf{V}\right\|_F^2. \tag{13}$$

Removing irrelevant items, we can rewrite Eq. (13) as follows:

$$\min_{\mathbf{V}} \beta(-2Tr(\mathbf{F}^{\mathsf{T}}\mathbf{YV}) + \|\mathbf{YV}\|_F^2) + \rho\|\mathbf{V}\|_F^2 . \tag{14}$$

Setting the derivation of Eq. (14) w.r.t. $\mathbf{V}$ equal to zero, we can get:

$$\mathbf{V} = (\mathbf{Y}^{\mathsf{T}}\mathbf{Y} + \frac{\rho}{\beta}\mathbf{I})^{-1}\mathbf{Y}^{\mathsf{T}}\mathbf{F}. \tag{15}$$

Step 4: Updating $\mathbf{W}_t$ with other variables fixed. By fixing other variables, the objective function can be simplified as follows:

$$\min_{\mathbf{W}^{(t)}} \sum_{t=1}^{2} \lambda_t \left\|\mathbf{F} - \phi(\mathbf{X}^{(t)})\mathbf{W}_t\right\|_F^2 + \sum_{t=1}^{2} \gamma \parallel \mathbf{W}_t \parallel_F^2 . \tag{16}$$

We first simplify Eq. (16) as follows:

$$\min_{\mathbf{W}^{(t)}} \sum_{t=1}^{2} \lambda_t(-2Tr(\mathbf{W}_t\mathbf{F}^{\mathsf{T}}\phi(\mathbf{X}^{(t)})) + \left\|\phi(\mathbf{X}^{(t)})\mathbf{W}_t\right\|_F^2) + \sum_{t=1}^{2} \gamma \parallel \mathbf{W}_t \parallel_F^2 . \tag{17}$$

By setting the derivation of Eq. (17) w.r.t. $\mathbf{W}_t$ equal to zero, we can obtain:

$$\mathbf{W}_t = (\phi(\mathbf{X}^{(t)})^{\mathsf{T}}\phi(\mathbf{X}^{(t)}) + \frac{\lambda_t}{\gamma}\mathbf{I})^{-1}\phi(\mathbf{X}^{(t)})^{\mathsf{T}}\mathbf{F}. \tag{18}$$

Step 5: Updating R with other variables fixed.

Fixing other variables are fixed, we rewrite Eq. (5) as follows:

$$\min_{\mathbf{R}} \mu\|\mathbf{B} - \mathbf{FR}\|_F^2, \quad s.t. \ \ \mathbf{RR}^{\mathsf{T}} = I. \tag{19}$$

Inspired by the work [3], we first compute the singular-value decomposition (SVD) of the $k \times k$ matrix $\mathbf{B}^{\mathsf{T}}\mathbf{F} = \mathbf{S}\,\Omega\,\mathbf{P}^{\mathsf{T}}$ and then we can obtain the solution of Eq. (19), i.e.,

$$\mathbf{R} = \mathbf{PS}^{\mathsf{T}}. \tag{20}$$

Step 6: Updating B by fixing other variables.

Fixing other variables, we simplify Eq. (5) as follows:

$$\min_{\mathbf{B}} \mu\|\mathbf{B} - \mathbf{FR}\|_F^2, \quad s.t. \ \ \mathbf{B} \in \{-1, 1\}^{n\times k}. \tag{21}$$

Then, we reformulate Eq. (21) as:

$$\begin{aligned} &\min_{\mathbf{B}} \sum_{i=1}^{2} \mu Tr((\mathbf{B} - \mathbf{FR})^{\mathsf{T}}(\mathbf{B} - \mathbf{FR})), \\ &= \|\mathbf{B}\|_F^2 - \mu(2Tr(\mathbf{B}^{\mathsf{T}}\mathbf{FR}) - \|\mathbf{FR}\|_F^2), \quad s.t. \ \ \mathbf{B} \in \{-1, 1\}^{n\times k}, \end{aligned} \tag{22}$$

Algorithm 1. Semantics-reconstructing Hashing.

Input: Training data matrices $\mathbf{X}^{(t)}$, $t = 1, 2$; label matrix $\mathbf{Y}$; parameters α, β, ρ, γ and μ; bit length of hash code k.
Output: Hash codes $\mathbf{B}$; semantic representation $\mathbf{F}$; mapping matrix $\mathbf{W}_t$, $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{R}$.
Procedure:
1. Randomly initialize $\mathbf{F}, \mathbf{U}, \mathbf{V}, \mathbf{R}, \mathbf{W}_t$ and $\mathbf{B}$;
Reapt:
2. Fix $\mathbf{B}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{R}$ and $\mathbf{W}_t$, update $\mathbf{F}$ using Eqn. (9);
3. Fix $\mathbf{B}$, $\mathbf{F}$, $\mathbf{V}$, $\mathbf{R}$ and $\mathbf{W}_t$, update $\mathbf{U}$ using Eqn. (12);
4. Fix $\mathbf{B}$, $\mathbf{F}$, $\mathbf{U}$, $\mathbf{R}$ and $\mathbf{W}_t$, update $\mathbf{V}$ using Eqn. (15);
5. Fix $\mathbf{B}$, $\mathbf{F}$, $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}_t$, update $\mathbf{R}$ using Eqn. (20);
6. Fix $\mathbf{B}$, $\mathbf{F}$, $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{R}$, update $\mathbf{W}_t$ using Eqn. (18);
7. Fix $\mathbf{F}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{R}$ and $\mathbf{W}_t$, update $\mathbf{B}$ using Eqn. (24);
until convergency.
Return: $\mathbf{B}$, $\mathbf{F}$, $\mathbf{U}$, $\mathbf{V}$,$\mathbf{R}$ and $\mathbf{W}_t$;

where $Tr(\cdot)$ is the trace norm. Apparently, $\|\mathbf{B}\|_F^2$ and $\|\mathbf{FR}\|_F^2$ are constants. Therefore, Eq. (22) is equivalent to the following problem:

$$\min_{\mathbf{B}} -Tr(\mathbf{B}^\mathsf{T}(\mu\mathbf{FR})), \quad s.t. \quad \mathbf{B} \in \{-1, 1\}^{n\times k}. \tag{23}$$

The solution to Eq. (23) is :

$$\mathbf{B} = sgn(\mu\mathbf{FR}). \tag{24}$$

The learning algorithm iteratively optimizes each variable until it converges or meets the maximum iteration number. We summarize the overall learning scheme in Algorithm 1.

3.6 Extension

For ease of representation, we restrain the discussion of SCH to bimodal case. Importantly, it can be conveniently extended to multi-modal data, as shown below.

$$\begin{aligned}
&\min_{\mathbf{F},\mathbf{U},\mathbf{V},\mathbf{W}_t,\mathbf{R}} \alpha\left\|\mathbf{Y}-\mathbf{FU}\right\|_F^2+\beta\left\|\mathbf{F}-\mathbf{YV}\right\|_F^2+\mu\left\|\mathbf{B}-\mathbf{FR}\right\|_F^2\\
&\quad+\sum_{t=1}^{m}\lambda_t\left\|\mathbf{F}-f_t(\mathbf{X}^{(t)})\right\|_F^2+\rho L(\mathbf{U},\mathbf{V},\sum_{t=1}^{M}\mathbf{W}_t),\\
&s.t.\ \ \mathbf{F}\in\mathbb{R}^{n\times k},\mathbf{B}\in\{-1,1\}^{n\times k},\sum_{t=1}^{m}\lambda_t=1,f_t(\mathbf{X}^{(t)})=\phi(\mathbf{X}^{(t)})\mathbf{W}_t,\mathbf{RR}^\mathsf{T}=I,
\end{aligned} \tag{25}$$

where $M \geq 2$ denotes the number of modalities. We can see the extension to more modalities is simple and easy, and it can also be solved by adapting the Algorithm 1.

As for out-of-sample extension, the hash codes can be easily generated for new samples with the learned parameters. For example, given a query instance $\mathbf{x}_i^{(o)} \in \mathbb{R}^d$, we can get its binary representation by:

$$b_i^{(o)} = sgn(\phi(\mathbf{x}_i^{(o)})\mathbf{W}_t\mathbf{R}). \tag{26}$$

3.7 Complexity Analysis

In this section, we give the detailed analysis of the computational cost of the training of SCH. Specifically, the time complexity of Step 1, 2 and 3 in Algorithm 1, is $O(nk^2 + nkl + lk^2 + k^3 + k^2)$, $O(nk^2 + nkl + k^3 + k^2)$ and $O(nl^2+nkl+l^3+k^2)$, respectively. Similarly, it is $O(nc^2, nck+c^3+c^2)$, $O(nk^2+k^3)$ and $O(nk^2, nk)$ for Step 4, 5 and 6, respectively. Therefore, the overall training cost of the proposed SCH is $O(n(k^2 + k + kl + l^2 + c^2 + ck)$. c indicate the number of anchors; k denotes the bit length of binary codes and l represents number of classes. Usually, they are much smaller than n for a large-scale dataset. In addition, SCH is able to converge within several iterations as shown in the experiments section. Therefore, the overall training cost is $O(n)$, scalable for large-scale datasets.

4 Experiments

4.1 Datasets

Wiki: It consists of 2,866 training pairs of image and text, each pair belongs to at least one of 10 semantic classes. 2173 pairs separated from the dataset for training and the remaining 693 pairs for testing. In addition, the visual modality and the textual one of each instance is represented by a 128-dimension bag-of-visual SIFT feature vector and a 10-dimension topic vector, respectively.

MIRFlickr-25K: The data set contain 25,000 images with corresponding textual tags which are collected from Flickr. There are 24 unique labels totally. They use 150-dimension edge histogram to represent each image and its textual content is represented as a 500-dimension feature vector derived from PCA on its binary tagging vector w.r.t the remaining textual tags.

NUS-WIDE: There are totally 269,648 images associated with textual tags in the dataset. There are 81 ground-truth labels to annotate data pairs. In our experiments, we choose top 10 most commonly used categories and the associated 186,577 images as the dataset for train and test. We annotate each image-text with at least 1 of 10 concepts, and represent each image and text by a 500-dimension bag-of-visual SIFT and a 1,000-dimension vector, respectively.

Considering the computational efficiency, we randomly select 5,000 samples from the original MIRFlickr-25K and 10,000 samples from NUS-WIDE dataset for training, while for testing, 1% samples of the each dataset are selected as the testing samples.

Table 1. The MAP results of all methods on three datasets. The best results are shown in boldface.

Task	Method	Wiki				MIRFlickr-25K				NUS-WIDE			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
Image-to-Text	IMH	0.1644	0.1684	0.1736	0.1744	0.5649	0.5685	0.5691	0.5698	0.3553	0.3539	0.3670	0.3583
	SCM-seq	0.2577	0.2785	0.2157	0.2935	0.6512	0.6617	0.6688	0.6718	0.5129	0.5200	0.5263	0.5287
	LSSH	0.1958	0.2108	0.2061	0.2063	0.5582	0.5644	0.5699	0.5693	0.3399	0.3594	0.3640	0.3772
	CMFH	0.1203	0.1222	0.1252	0.1232	0.5708	0.5703	0.5712	0.5713	0.3549	0.3540	0.3547	0.3544
	CCQ	0.2048	0.2118	0.2127	0.2130	0.5680	0.5681	0.5681	0.5679	0.3421	0.3421	0.3431	0.3429
	SePH-km	0.2796	0.2820	0.3076	0.3137	0.6843	0.6873	0.6882	0.6874	0.5369	0.5440	0.5449	0.5510
	DCH	0.3349	0.3620	0.3762	0.3799	0.6849	0.6976	0.6937	0.7121	0.5970	0.5826	0.5909	0.6100
	SDMCH	0.3183	0.3402	0.3621	0.3669	0.6530	0.6476	0.7249	0.7053	0.5193	0.6138	0.6246	0.6084
	SCH	**0.3387**	**0.3860**	**0.3844**	**0.3893**	**0.7014**	**0.7175**	**0.7255**	**0.7282**	**0.6092**	**0.6286**	**0.6385**	**0.6408**
Text-to-Image	IMH	0.1362	0.1395	0.1436	0.1398	0.5635	0.5675	0.5671	0.5684	0.3553	0.3539	0.3670	0.3583
	SCM-seq	0.3690	0.4064	0.4301	0.4316	0.6524	0.6670	0.6766	0.6807	0.4979	0.5079	0.5183	0.5218
	LSSH	0.4286	0.4654	0.4901	0.5029	0.4286	0.4654	0.4901	0.5029	0.3466	0.3541	0.3725	0.3772
	CMFH	0.1280	0.1309	0.1351	0.1331	0.5732	0.5732	0.5738	0.5742	0.3580	0.3565	0.3574	0.3573
	CCQ	0.2731	0.2859	0.2869	0.2863	0.5746	0.5753	0.5755	0.5755	0.3633	0.3651	0.3657	0.3658
	SePH-km	0.6379	0.6451	0.6662	0.6706	0.7389	0.7457	0.7476	0.7497	0.6203	0.6358	0.6405	0.6391
	DCH	0.6624	0.7040	0.7241	0.7203	0.7513	0.7664	0.7716	0.7967	0.7041	0.6995	0.7085	0.7355
	SDMCH	0.7085	0.7272	0.7513	0.7533	0.7154	0.6818	0.7920	0.7843	0.6199	0.7364	0.7454	0.7339
	SCH	**0.7267**	**0.7570**	**0.7606**	**0.7614**	**0.7723**	**0.7851**	**0.8028**	**0.8158**	**0.7390**	**0.7605**	**0.7694**	**0.7739**

4.2 Baselines and Evaluation Metrics

We compared the proposed SCH with the sate-of-the-art shallow baselines, including four supervised methods, i.e., SCM-seq [18], CVH [7], SePH-km [9], DCH [17], SDMCH [12] and four unsupervised methods, i.e., LSSH [20], CCQ [11], IMH [14], and CMFH [2]. The parameters of SCH were selected by a validation procedure, i.e., $\alpha = 4.5$, $\beta = 0.01$, $\mu = 0.5$, $\lambda_1 = 0.3$, $\lambda_2 = 0.7$, $\rho = 0.01$, and $\gamma = 0.01$.

We chose Mean Average Precision (MAP), precision-recall and top-N precision curves as performance metrics to evaluate the proposed SCH and all the compared method.

4.3 Results and Discussions

MAP Results. We reported the MAP results of SCH and all of the compared methods on there datasets with bit length varying from 16 bits to 128 bits in Table 1, including the results of the Image-to-Text and Text-to-Image search tasks. From these results, we have the following observations. Firstly, SCH outperforms all supervised and unsupervised baselines in all cases. In terms of quantitative comparison, our method achieves about 4.6% and 6% overall improvements over DCH and SDMCH which have better performance compared with other baselines, respectively. These well demonstrate the effectiveness of SCH. One of the main reasons for the superiority of our SCH is that it can capture more similarity and discriminative information constructing the semantic representation and embed the information into the binary codes. Another reason is

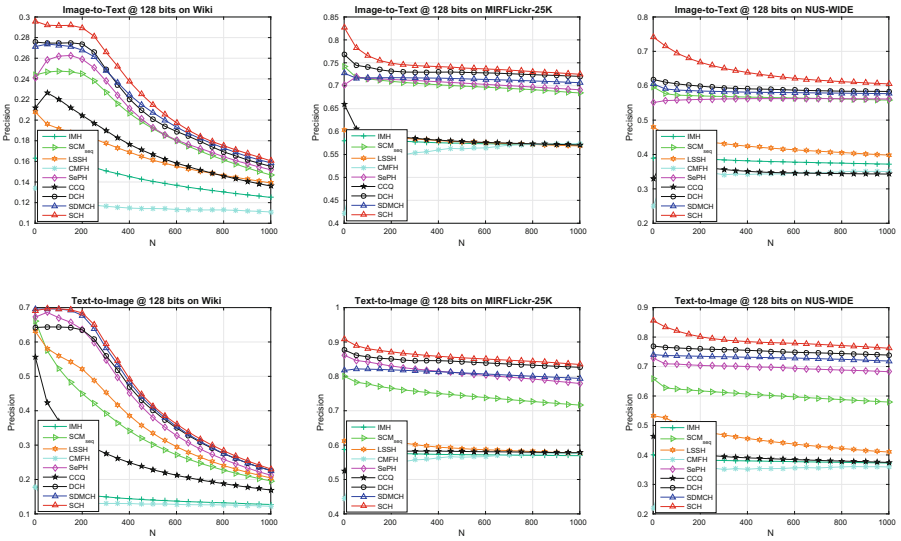

Fig. 1. Top-N precision curves with 128-bit on three datasets.

that it solves the optimization problem discretely and learns the binary codes directly, reducing the quantization errors. Secondly, Generally speaking, with code length increasing, the performance of all methods keeps increasing, which means that utilizing longer hash codes can contain more semantic information. Lastly, Most of the methods have better performance when searching images with the given text query than the other retrieval task. The main reason is that the text features can better describe the content information of an image-text pair than that of the image features.

Top-N Precision and Precision-Recall Curves. The top-N precision and precision-recall curves of the cases with 128 bits are plotted in Fig. 1 and 2. From the figure, we can find that SCH has the best overall performance. In addition, we can also observe that most of the supervised methods outperform the unsupervised ones, reflecting the importance of supervised information in the learning of binary codes. Moreover, From the top-N precision curves, we can see that SCH performs much better than all the compared methods, especially at the early stage. This implies SCH returns more samples close to queries when N is small, which is very important in a retrieval task.

To summarize, from the comparison between our SCH and other methods on Wiki, MIRFlickr-25K and NUS-WIDE, we can have the conclusion that the proposed SCH can work well on these datasets, and outperform other state-of-the-art cross-modal hashing methods.

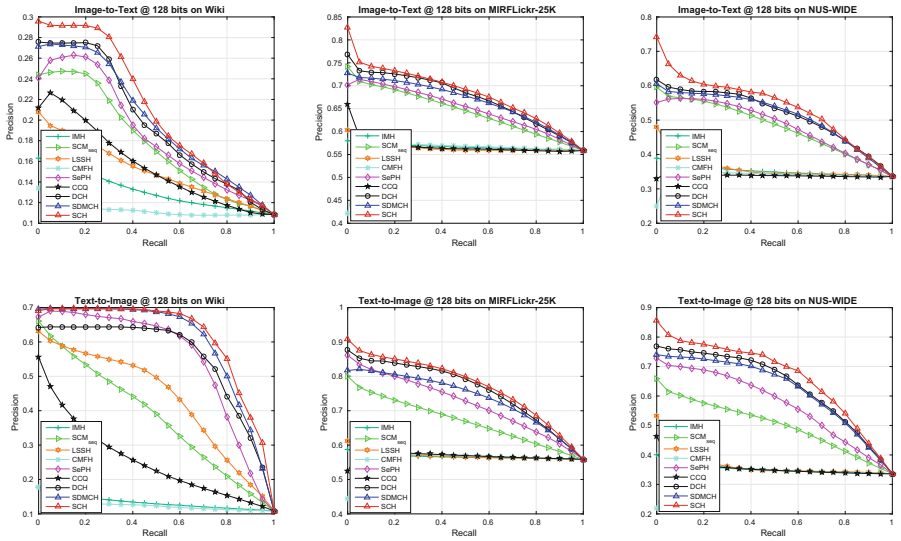

Fig. 2. Precision-recall curves with 128-bit on three datasets.

5 Conclusion and Future Work

In this paper, we propose a scalable supervised hashing method for cross-modal retrieval, i.e., Semantics-reconstructing Hashing for Cross-modal Retrieval. It learns efficient and effective hash codes semantically consistent with semantic information by reconstructing semantic representation with labels. Moreover, with the semantic representation, it constructs the correlations between the original features, the labels and the binary codes for the entire dataset. Furthermore, it simultaneously learns the hash codes and the hash functions without any relaxation, reducing the quantization errors and makes the optimization easy to be solved by an iterative algorithm. Extensive experiments on three widely used datasets demonstrate that SCH outperforms eight state-of-the-art shallow baselines for cross-modal search.

In our work, we concentrate on the design of the loss function and the discrete optimization scheme. And we believe that SCH can be combined with a deep model to generate an end-to-end deep hashing method. We leave this as our future work.

Acknowledgements. This work was partially supported by Australian Research Council Discovery Project (ARC DP190102353), and China Scholarship Council.

References

1. Cao, Y., Long, M., Wang, J., Yang, Q., Yu, P.S.: Deep visual-semantic hashing for cross-modal retrieval. In: SIGKDD, pp. 1445–1454 (2016)

2. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: CVPR, pp. 2075–2082 (2014)
3. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. TPAMI **35**(12), 2916–2929 (2013)
4. Gui, J., Li, P.: R2SDH: robust rotated supervised discrete hashing. In: SIGKDD, pp. 1485–1493 (2018)
5. Huang, Q., Ma, G., Feng, J., Fang, Q., Tung, A.K.: Accurate and fast asymmetric locality-sensitive hashing scheme for maximum inner product search. In: SIGKDD, pp. 1561–1570 (2018)
6. Jiang, Q.Y., Li, W.J.: Scalable graph hashing with feature transformation. In: IJCAI, pp. 2248–2254 (2015)
7. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: IJCAI, pp. 1360–1365 (2011)
8. Lian, D., et al.: High-order proximity preserving information network hashing. In: SIGKDD, pp. 1744–1753 (2018)
9. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-preserving hashing for cross-view retrieval. In: CVPR, pp. 3864–3872 (2015)
10. Liu, X., Deng, C., Lang, B., Liu, W.: Query-adaptive reciprocal hash tables for nearest neighbor search. TIP **25**(2), 907–919 (2016)
11. Long, M., Cao, Y., Wang, J., Yu, P.S.: Composite correlation quantization for efficient multimodal retrieval. In: SIGIR, pp. 579–588 (2016)
12. Luo, X., Yin, X.Y., Nie, L., Song, X., Wang, Y., Xu, X.S.: SDMCH: Supervised discrete manifold-embedded cross-modal hashing. In: IJCAI, pp. 2518–2524 (2018)
13. Shen, F., Shen, C., Liu, W., Tao Shen, H.: Supervised discrete hashing. In: CVPR, pp. 37–45 (2015)
14. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: SIGMOD, pp. 785–796 (2013)
15. Tang, J., Li, Z., Wang, M., Zhao, R.: Neighborhood discriminant hashing for large-scale image retrieval. TIP **24**(9), 2827–2840 (2015)
16. Wang, D., Cui, P., Ou, M., Zhu, W.: Deep multimodal hashing with orthogonal regularization. In: IJCAI, pp. 2291–2297 (2015)
17. Xu, X., Shen, F., Yang, Y., Shen, H.T., Li, X.: Learning discriminative binary codes for large-scale cross-modal retrieval. TIP **26**(5), 2494–2507 (2017)
18. Zhang, D., Li, W.J.: Large-scale supervised multimodal hashing with semantic correlation maximization. In: AAAI, pp. 2177–2183 (2014)
19. Zhen, Y., Yeung, D.Y.: Co-regularized hashing for multimodal data. In: NIPS, pp. 1376–1384 (2012)
20. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hashing for cross-modal similarity search. In: SIGIR, pp. 415–424 (2014)
21. Zhu, X., Huang, Z., Shen, H.T., Zhao, X.: Linear cross-modal hashing for efficient multimedia search. In: MM, pp. 143–152 (2013)