Neurocomputing (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Diversity Regularized Latent Semantic Match for Hashing

Yong Chen[a], Hui Zhang[a], Yongxin Tong[a], Ming Lu[b,*]

[a] School of Computer Science and Engineering, Beihang University, Beijing 100191, PR China
[b] IT FLEX, Intel China Research Center, Beijing 100080, PR China

ARTICLE INFO

Communicated by Rongrong Ji

Keywords:
Diversity regularizations
Soft orthogonality
Learning to match
Multimodal retrieval
Representation learning
Approximate nearest neighbors

ABSTRACT

Hashing based approximate nearest neighbors (ANN) search has drawn considerable attraction owing to its low-memory storage and hardware-level logical computing which is doomed to be greatly applicable to quantities of large-scale and practical scenarios, such as information retrieval, computer vision and natural language processing. However, most existing hashing methods concentrate either on images only or on pairwise image-texts (labels, short documents) and rarely utilize more common sentences. In this paper, we propose D iversity R egularized L atent S emantic M atch for H ashing (DRLSMH), a new multimodal hashing method that projects images and sentences into a shared latent semantic space with label-supervised semantic constraints to proceed on multimodal retrieval. Notably, soft orthogonality is induced as a novel regularizer to preserve diverse hashing functions for compact and accurate representations; what's more, this kind of regularization also benefits the derivations of closed-form solutions with some proper relaxations under iterative optimization framework. Extensive experiments on two public datasets demonstrate the advantages of our method over some state-of-the-art baselines under cross-modal retrieval both on image-query-image, image-query-text and text-query-image tasks.

1. Introduction

Nearest neighbors (NN) search has acted as a fundamental role in lots of important applications, such as machine learning, computer vision, natural language processing and so forth for decades [1–3]; however, recently ever-changing Internet technologies have already pushed forward the big data era to come: high-dimensional, massive, and heterogeneous data throw a huge challenge on NN. Even for the simplest linearly scanning, it would be impractical and unrealistic for real scenarios now. Hashing, as a new approximate nearest neighbors method, embedding data into the binary hamming space which is capable of preserving similarities between objects makes the memory and computing both extremely effective [4,5]; even ordinary PCs can handle large amounts of data.

Hashing methods can be divided into different classifications according to different views. For example, it can be roughly divided into data-independent methods and data-dependent methods by using the data or not, where LSH [6], KLSH [7] and other LSH-like methods [8] are data-independent and ITQ [9], SpH [10], SSH [11] and MLH [12] are data-dependent ones. From another perspective of using supervised information or not, there could be three kinds: unsupervised [6,13], supervised [5,12] and semi-supervised [11,14,15] methods. Here, we would like to divide the hashing methods into traditional image retrieval methods and current multi-view cross-modal retrieval methods.

Methods mentioned above all belong to the former kind. And as regard to the latter one, there are many new methods emerging in the recent years. Inter-media hashing (IMH) [16] introduces inter-media consistency and intra-media consistency to discovery a common hamming space, and uses regularized linear model to learn view specific hash functions. However, IMH needs to construct the similarity matrix for all the data points, which will impede the effectiveness for large-scale datasets. Latent semantic sparse hashing (LSSH) [17] utilizes the sparse coding to capture the salient structures of images and matrix factorizations to learn the latent concepts from text to perform cross-modal similarity search. However, this kind of learning paradigm, especially the sparsity, makes the training stage consume too much time. Collective matrix factorization hashing (CMFH) [18] learns unified hash codes by collective matrix factorization with latent semantic match model from different modes of one instance, while it's too strict to constraint different modalities to identical hash codes. Semantic topic multimodal hashing (STMH) [19] models text as multiple semantic topics and image as latent semantic structures and then learns the relationship of text and image into their latent semantic spaces. Though STMH has obtained superior performances to some state-of-the-art baselines, we find the extension of out-of-sample need to be simplified.

Although there are many multimodal hashing methods and they all have achieved promising performance in multimodal applications [16–

* Corresponding author.
E-mail addresses: hzhang@buaa.edu.cn (H. Zhang), ming-lu@outlook.com (M. Lu).

http://dx.doi.org/10.1016/j.neucom.2016.11.057
Received 9 April 2016; Received in revised form 5 September 2016; Accepted 30 November 2016
0925-2312/

19], there still needs to be more explorations on models (linear/nonlinear, matrix factorization/probabilistic graphical modes, deep neural network or not), algorithms (convex/nonconvex, distributed parallel gradient-based algorithms) and theories (robustness, sparsity, diversity or low rank), or even for some new formalizations of multi-modal data. In this work, we make full use of the self-characterized image-sentences pairwise data, and map them diversely into a shared latent semantic space via match learning with label-supervised semantic regularizations which is able to preserve similarities between images and sentences, and then put forward a novel method Diversity Regularized Latent Semantic Match for Hashing (DRLSMH). The core contributions of our work can be listed as below:

- We incorporate linear projection instead of direct matrix factorization with learning to match framework, which would definitely lead to two advantages: on one hand, it makes the model look simple (more like convex), and more importantly it would greatly benefit the hashing for out-of-samples just through basic matrix-vector multiplications; on the other hand, this kind of formalizations help to the later closed-form solutions.
- Soft orthogonality is introduced as a novel regularizer for diverse hashing functions, which will provide compact and accurate representations with small fixed number of hash bits. Moreover, closed-form solutions can be easily derived with some relaxations on the regularizations under the iterative framework.
- To the best of our knowledge, this is the pioneer exploration to perform learning to hash for cross-modal retrieval tasks on such kind of datasets: pair-wise image-sentences corpus. Extensive experiments on two public datasets highlight the superiority over some of the state-of-the-art methods for image-query-image, image-query-text and text-query-image missions.

The remainder of this paper is organized as follows. In Section 2, we introduce related work about diversity regularizations, learning to match and deep learning for representations. In Section 3, we define our problem and give necessary notations. In Section 4, we propose our method DRLSMH and present an approximate learning process for match learning and then derive the optimization algorithms. We conduct experiments on three kinds of tasks to evaluate the proposed models in Section 5, and finally draw conclusions in Section 6.

2. Related work

2.1. Diversity regularizations

Very recently, it's quite interesting that there seems a more and more growing attention on the diversity regularizations explored in various aspects of data mining and machine learning, such as ensemble methods, self-paced learning, metric learning, multi-view clustering and so on, without any prior consolations. And lots of superior performances are mined out with the utilizations of diversity constraints in different formalizations. For example, [20] proposed the diversity regularized machine to construct an ensemble of diverse SVMs which lead to an effective reduction on its hypothesis space complexity and better generation ability verified both in theoretical analysis and experiments; [21] threw focus on the preferences both easy and diverse samples into a general non-convex regularizer which would greatly contribute to the self-pace learning; [22] discussed about the tasks of keeping a small number of latent factors meanwhile making them as effective as a large set of factors for the sake of computational efficiency and put forward an diversity constraints with the mean and variance of latent factors, and then learned compact and effective distance metrics for retrieval, clustering and classifications; last but not the least, [23] utilized the Hilbert Schmidt Independence Criterion as a diversity term to explore the complementarity of multi-view representations that could explicitly enforce the learned subspace to be novel with each other for better clustering.

Definitely, diversity is an intuitive and effective idea to be taken advantage of for its compact and effective information presentations in large scale data. However, it's still an open research problem both in wide varieties of tasks, formalizations, algorithms and its theoretical analysis. Here, soft orthogonal constraints are induced on projection matrices as a novel diversity regularizer to obtain diverse hash functions (another perspective different from the former works) for compact representations of both image and text data in our paper. Soft orthogonality not only can achieve comparable effects with a small number of hash functions as that of large sets of hash functions, but also can be made use of for relaxations to derive closed-form solutions which are all of great benefits to multi-modal retrieval.

2.2. Learning to match

Relevance has always been considerably important in search and will always be, and match is a key factor for similarity, especially in the contemporary heterogeneous, multi-view, associated big data era. Learning to match (match learning) [24–27] is a sharp sword in such scenarios including question answering, recommender systems, machine translation, cross-language information retrieval, online advertising, image annotation, drug design and couple pairing. In recent research, [28] leveraged both clicks and content to learn to match heterogeneous objects via shared latent structures for web search. Likewise, image annotations [29], recommendation systems [30], and Cross-modal Search [17] all mapped different modals or views (i.e. keywords v.s. images, users v.s. products, images v.s. texts etc.) into a shared latent high-level semantic space with low dimensions and bridged them each other for better and effective relevance.

However, in this paper, the datasets explored are formed with images and sentences pair-wisely; therefore we can naturally connect them into a common latent semantic space from two distinct image and sentence spaces with the assumptions that they both describe the same object/thing with just different languages.

2.3. Deep learning for representations

Deep learning (deep machine learning, or deep neural network learning, or hierarchical learning, or sometimes DL) is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers with complex structures, or otherwise composed of multiple non-linear transformations [31–33]. A large amount of exploration and research on AutoEncoder, CNN, LSTM, and other types of DNNs have brought unprecedented changes in fields such as image understanding and recognition, speech recognition, and distributed representations and language processing in recent few years since 2006, when Hinton and Salakhutdinov gave a second birth to the traditional neural network [34]. In this paper, we would like to focus on two well-used DL tools VGG-16 [35] and Sentence2vec [36,37] for image and text representations respectively, which will be prepared for the next match learning parts illustrated in the middle of Fig. 1.

3. Problem statement

Suppose that $O = \{o_s\}s = 1^N$ is a set of multimodal instances, which consists of an image and its corresponding texts (sentences), i.e. $o_s = (D_s^i, D_s^t)$, where $D_s^i \in R^{M_1}$ is an M_1-dimensional image descriptor extracted from VGG-16[1] and PCA, and $D_s^t \in R^{M_2}$ is an M_2-dimensional text feature obtained from Sentence2vec[2] (usually $M_1 \neq M_2$). Given the bits length K, the purpose of DRLSMH is to learn an integrated binary

[1] http://www.robots.ox.ac.uk/vgg/research/very_deep/

[2] https://github.com/klb3713/sentence2vec

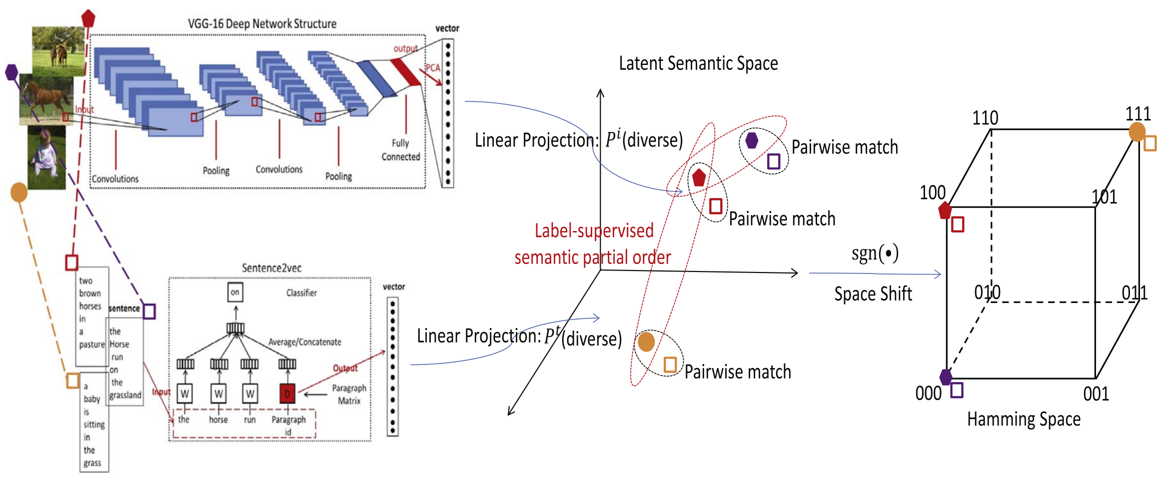

Fig. 1. A brief framework of DRLSMH illustrated with toy examples. (1)Left: Using deep network tools VGG-16 and Sentence2vec to vectorize images and its corresponding descriptive sentences. (2)Middle: Linearly project (diverse projections) images and sentences into a common latent semantic space, which can integrate pairwise match learning and label-supervised semantic partial order regularizations (notably, labels can be extracted from sentences to supervise image representations in latent semantic space). (3)Right: Space shift from latent semantic space to hamming space with elementwise operator $\mathrm{sgn}(\cdot)$, which is beneficial for tasks on image-query-image, image-query-text and text-query-image retrieval.

code $h_s \in [0, 1]^K$ for o_s, $s = 1, 2, \ldots, N$, such that h_s and h_t preserve the semantic similarity between o_s and o_t with high probabilities. More specifically, if o_s and o_t are two objects with similar semantic, h_s and h_t should have a small Hamming distance, and vice versa.

As illustrated in Fig. 1, there are three stages from original objects (iamges, sentences) to the final bits coded in our proposed method. Firstly, images can be represented as a 4096-dimensional vector with trained VGG-16 model on ImageNet[3] and then further reduced to a 128-dimensional vector by PCA for saving computing and storage resources. Meanwhile, sentences would be embedded into a 100-D vector through trained Sentence2vec on millions of MSCOCO[4] sentences. Secondly, images and texts are mapped into a shared latent semantic space with match learning by linear diverse projections P^i and P^t respectively and semantic regularizations. Finally, elementwise operator $\mathrm{sgn}(\cdot)$ implements the space shift from latent semantic space to the hamming space for binary codes. Table 1 summarizes the necessary notations and explanations.

4. Diversity regularized latent semantic match for hashing

This section details the proposed DRLSMH model for cross-modal similarity search. Without loss of generality, we restrict the discussion to bimodal instances consisting of images and texts (sentences) because they are the most common and important scene in real world.

4.1. Diversity regularized match learning

Learning to match is a very useful framework in modeling multi-modal datasets with the assumptions that different views of data can be bridged each other for the same/similar semantic relatedness. Take image-sentences corpus for example, image and its corresponding sentences, two distinct modals, would both talk about the same or similar topics or other things and it's naturally to map them to a common latent semantic space for connection which is the soul of match learning. Therefore, in regard to the pairwise image-sentences datasets, we design a three-stage hashing model whose framework is shown in Fig. 1. Represented images and sentences are transformed into V^i and V^t respectively in a shared latent semantic space through linear projections P^i and P^t correspondingly. Since they have tight relatedness, we can formalize it based on the learning to match framework as follows:

Table 1
Math Notations.

Notations	Explanations
N	Number of multimodal instances in the dataset $O = \{o_s\}_{s=1}^N$, S is the index from 1 to N.
M_1	Dimensions of image descriptor extracted from VGG−16 and PCA.
M_2	Dimensions of text feature obtained from Sentence2vec.
K	Number of code bits or hashing functions.
$D^i \in R^{M_1 \times N}$	Image representation of the dataset and each column $D_s^i \in R^{M_1}$ represents an image.
$D^t \in R^{M_2 \times N}$	Text representation of the dataset and each column $D_s^t \in R^{M_2}$ represents a sentence.
$P^i \in R^{K \times M_1}$	K linear projections from image space to the common latent semantic space and each row represents one projection.
$P^t \in R^{K \times M_2}$	K linear projections from text space to the common latent semantic space and each row represents one projection.
$V^i \in R^{K \times N}$	The corresponding representations of images in the shared latent semantic space.
$V^t \in R^{K \times N}$	The corresponding representations of texts in the shared latent semantic space.
$\mathrm{sgn}(\cdot)$	Elementwise operator: if $x > 0$, then $\mathrm{sgn}(x) = 1$; else $\mathrm{sgn}(x) = 0$.
$\alpha, \beta, \gamma, \varepsilon$	Hyperparameters of DRLSMH as balance factors between loss, match and semantic similarity.

$$\min\ \alpha \underbrace{\|P^i D^i - V^i\|_F^2}_{\#1} + \beta \underbrace{\|P^t D^t - V^t\|_F^2}_{\#2} + \gamma \underbrace{\|V^i - V^t\|_F^2}_{\#3}, \tag{1}$$

where part #1 and #2 are designed to learn semantic bases (P^i and P^t) and representations (V^i and V^t) simultaneously both for images and sentences through linear projections instead of matrix factorizations. Part #3 characterizes the semantic similarities between the image and sentences which is the key to bridge heterogeneous data. α, β and γ are regulators to balance intra-semantics and inter-match respectively. Notably, P^i and P^t can also be viewed as hashing functions and each row represents one.

Furthermore, different formalizations of diversity have been widely considered and explored in various fields (information retrieval [22], ensemble methods [20], self-paced learning [21], clustering [23] etc.) of machine learning and obtained promising performance in many applications. In this paper, we also put forward a novel diversity regularizer named soft orthogonality on projections for diverse hashing functions to preserve compact and accurate binary codes. Specifics are listed as below:

$$\| P^i (P^i)^T - I_K \| \le T_1, \tag{2}$$

$$\| P^t (P^t)^T - I_K \| \le T_2, \tag{3}$$

[3] http://www.image-net.org/
[4] http://mscoco.org/home/

where T_1 and T_2 are two predefined thresholds to control the diversity of hash functions; and the smaller, the more diverse. I_K denotes a $K \times K$ unit diagonal matrix.

Therefore, we can summarize this subsection as the following optimization problems:

$$\min \alpha \parallel P^i D^i - V^i \parallel_F^2 + \beta \parallel P^t D^t - V^t \parallel_F^2 + \gamma \parallel V^i - V^t \parallel_F^2 \quad s.\ t. \begin{cases} \parallel P^i (P^i)^T - I_K \parallel \le T_1 \\ \parallel P^t (P^t)^T - I_K \parallel \le T_2 \end{cases}. \tag{4}$$

4.2. Semantic similarity preserving

Now that we have managed to find a proper way to bridge different types of media data, i.e. exploring the inter-media consistency with match learning. Many previous state-of-the-art hashing methods [38,16,19] have shown that compact binary codes should make the similar data points closer than that of dissimilar pairs within a short hamming distances in a single data type. So inspired by these works, we also intend to seek for the intra-semantic similarity especially for images i.e. V^i should be regularized with some kind of semantic supervisions.

Take image and its descriptive sentences for analysis, if two images are of similar semantics, there would probably be more shared words in their corresponding texts. Accordingly, we can design the similarities for each pair by the intersection and union operations of their text's word sets. More specifically, a set of words would be obtained for each image through natural language processing tools (NLTK,[5] ANSJ[6]) such as word-participle, part-of-speech analysis and word stemming. Set $S_m^i = \{word_{m1}, word_{m2}, \ldots, word_{mp}\}$ and $S_n^i = \{word_{n1}, word_{n2}, \ldots, word_{nq}\}$ for the word-set of the m-th and n-th image, then the similarity W_{mn} between them is formalized as follows:

$$W_{mn} = \frac{Card\,(S_m^i \cap S_n^i)}{Card\,(S_m^i \cup S_n^i)}, \tag{5}$$

where $Card\,(\cdot)$ denotes the number of the given set within the braces.

With respect to the m-th and n-th image, more shared words means bigger similarities W_{mn} which indicates smaller distances in latent semantic space, i.e. ${V_m}^i$ and ${V_n}^i$ should be semantic relatedness accordingly. For each pair images in this dataset, we can construct an $N \times N$ similarity matrix W. To preserve the semantic similarities among images, an optimization problem can be drawn as below:

$$\min \sum_{m=1}^{N} \sum_{n=1}^{N} W_{mn} \parallel V_m^i - V_n^i \parallel_2^2. \tag{6}$$

By introducing a diagonal $N \times N$ matrix G, whose entries are given by $G_{ss} = \sum_{j=1}^{N} W_{sj}$ and others zero, Eq. (6) can be rewritten as:

$$\min tr\,\{V^i (G - W)(V^i)^T\} = tr\,\{V^i L (V^i)^T\}, \tag{7}$$

where L is the graph Laplacian defined on the image data, and $tr\,(\cdot)$ is the trace function. By minimizing this term, the similarity between different images can be preserved in the learned codes.

4.3. Overall objective function

The overall objective function, combining the diversity regularized match learning in Eq. (4) and the semantic similarity preserving in Eq. (7), is written as below.

$$\min_{P^i, P^t, V^i, V^t} \mathrm{L_f}(P^i, P^t, V^i, V^t) s.\ t. \begin{cases} \parallel P^i (P^i)^T - I_K \parallel \le T_1 \\ \parallel P^t (P^t)^T - I_K \parallel \le T_2 \end{cases}, \tag{8}$$

where

$$\mathrm{L}_f = \alpha \parallel P^i D^i - V^i \parallel_F^2 + \beta \parallel P^t D^t - V^t \parallel_F^2 + \gamma \parallel V^i - V^t \parallel_F^2 + \varepsilon tr\,\{V^i L (V^i)^T\} \tag{9}$$

and ε is the hyper parameter to regulate the importance of semantic similarities for images.

One more point, I think, should be added is that binary codes can be easily computed by the elementwise operator $\mathrm{sgn}(\cdot)$ implementing space shift from latent semantic space to hamming space after solving the optimization problems (8).

4.4. Optimization algorithm

The optimization problem (8) is non-convex with respect to four matrices P^i, P^t, V^i, V^t. However, it can be convex with respect to any one of the four matrices while fixing the other ones. Following the practice in Sparse Coding [39], we optimize the objective function in (8) by alternately minimizing it with respect to P^i, P^t, V^i, V^t. This procedure is elaborated in the following parts.

Algorithm 1. Diversity Regularized Latent Semantic Match for Hashing (DRLSMH).

Input: images $D^i \in R^{M_1 \times N}$, texts $D^t \in R^{M_2 \times N}$, binary codes number K, parameters $\alpha, \beta, \gamma, \varepsilon$
Output: Hash codes H, and matrices P^i, P^t, V^i, V^t
random initialize $P^i(0)$, $P^t(0)$, $V^i(0)$ and $V^t(0)$;
for s=1:S **do**
  $P^i(s) \leftarrow UpdateP^i(P^i(s-1), P^t(s-1), V^i(s-1), V^t(s-1))$;
  $P^t(s) \leftarrow UpdateP^t(P^i(s), P^t(s-1), V^i(s-1), V^t(s-1))$;
  $V^i(s) \leftarrow UpdateV^i(P^i(s), P^t(s), V^i(s-1), V^t(s-1))$;
  $V^t(s) \leftarrow UpdateV^t(P^i(s), P^t(s), V^i(s), V^t(s-1))$;
  if *convergence is satisfied* **then**
    //here convergence can be defined as the limited //absolute error:
    $//max(max(abs(P^i(s) - P^i(s-1)))) < eps$, //(e.g. eps=1e-4).
    //Same with other matrices;
    break;
  end
end
$H = (H^i, H^t) = (\mathrm{sgn}(V^i), \mathrm{sgn}(V^t))$;
return H and P^i, P^t, V^i, V^t;

(1) **Update of Matrix P^i**: Holding matrix $P^i(s-1)$, $P^t(s-1)$, $V^i(s-1)$, $V^t(s-1)$ fixed, the update of $P^i(s)$ amounts to solving the following optimization problem:

[5] http://www.nltk.org/
[6] http://www.nlpcn.org/

$$P^i(s) = \underset{P^i}{\text{argmin}}: L_f = \alpha \parallel P^iD^i - V^i(s-1)\|_F^2 \ s.\ t. \quad \parallel P^i(P^i)^T - I_K \|_F^2 \leq T_1. \tag{10}$$

The optimization problem (10) can be transformed into the equivalent one as follows:

$$P^i(s) = \underset{P^i}{\text{argmin}}: L_f = \|P^iD^i - V^i(s-1)\|_F^2 + \eta_0\|P^i(P^i)^T - I_K\|_F^2, \tag{11}$$

where η_0 denotes a predefined hyper-parameter to control the diversity of hashing functions; and the bigger, the more diverse. In order to get a closed-form solution (usually closed-form solutions would greatly contribute to the reduced computation instead of the time-consuming iterative updates), we further make some relaxations on $\boldsymbol{P^i}$, and Eq. (11) is approximately converted to solve the following optimization problem:

$$P^i(s) = \underset{P^i}{\text{argmin}}: L_f = \|P^iD^i - V^i(s-1)\|_F^2 + \eta_0\|P^i(P^i(s-1))^T - I_K\|_F^2. \tag{12}$$

Let $\frac{\partial L_f}{\partial P^i} = 0$, then we obtain:

$$P^i(s) = [V^i(s-1)(D^i)^T + \eta_0 P^i(s-1)] \times [D^i(D^i)^T + \eta_0[P^i(s-1)]^T[P^i(s-1)]]^{-1}. \tag{13}$$

(2) **Update of Matrix $\boldsymbol{P^t}$**: It is easy to find the symmetry between $\boldsymbol{P^i}$ and $\boldsymbol{P^t}$, and the processing is also the same as that of $\boldsymbol{P^i}$; therefore, we can directly write the final solutions:

$$P^t(s) = [V^t(s-1)(D^t)^T + \eta_1 P^t(s-1)] \times [D^t(D^t)^T + \eta_1[P^t(s-1)]^T[P^t(s-1)]]^{-1}, \tag{14}$$

where η_1 represents a predefined hyper-parameter to control the diversity of hashing functions; and the bigger, the more diverse. Generally, η_0 and η_1 can be set to the same value for simplifications.

(3) **Update of Matrix $\boldsymbol{V^i}$**: Holding matrix $P^i(s)$, $P^t(s)$, $V^i(s-1)$, $V^t(s-1)$ fixed, the Update of matrix $V^i(s)$ is equivalent to solving the following optimization problems:

$$V^i(s) = \underset{V^i}{\text{argmin}}: L_f = \alpha\|P^i(s)D^i - V^i\|_F^2 + \gamma\|V^i - V^t(s-1)\|_F^2 + \varepsilon tr(V^iL(V^i)^T). \tag{15}$$

Let $\frac{\partial L_f}{\partial V^i} = 0$, then we achieve:

$$V^i(s) = [\alpha P^i(s)D^i + \gamma V^t(s-1)][(\alpha+\gamma)I_N + \varepsilon L]^{-1}. \tag{16}$$

(4) **Update of Matrix $\boldsymbol{V^t}$**: Similar as update of matrix $\boldsymbol{V^i}$, we can get:

$$V^t(s) = \frac{1}{\beta+\gamma}[\beta P^t(s)D^t + \gamma V^i(s)]. \tag{17}$$

Based on the above derivations and analysis, the algorithm is summarized in Algorithm 1.

4.5. Out-of-sample extension

In practice, the components of a new query can be quite diverse, now we discuss it in the following three situations.

(1) **Image Only:** Let $d^i \in R^{M_1}$ be the image-query feature, then its hash code $h^i \in R^K$ can be easily obtained by

$$h^i = \text{sgn}(P^id^i). \tag{18}$$

(2) **Text Only:** Similarly, let $d^t \in R^{M_2}$ be the text-query feature, then the corresponding hash code $h^t \in R^K$ can be easily obtained by

$$h^t = \text{sgn}(P^td^t). \tag{19}$$

(3) **Both Image and Text:** We can use the same way to get hash codes described in Image only or Text only.

Here a clear advantage is exposed by the above subsection: DRLSMH is capable of dealing with large scale and online out-of-samples for its simplest matrix-vector hashing operations which could be easily distributed and paralleled efficiently. As regard to the training stage, we find it's much faster than LSSH [17] and comparable with CMFH [18], STMH [19], which is probably beneficial from the designed closed-form solutions, in our experiments.

5. Experiments

To evaluate the effectiveness of our proposed DRLSMH, pioneer experiments on two public multi-modal corpora UIUC and Flickr8k consisting of images and sentences are elaborately conducted on three retrieval missions: image-query-image, image-query-text and text-query-image over some state-of-the-art hashing methods.

5.1. Experimental setup

We use the UIUC [40] and Flickr8K [41] datasets in our experiments. Each image in these datasets is annotated with 5 sentences using Amazons Mechanical Turk.

5.1.1. Datasets

Flickr8k can be downloaded from here.[7] It provides JSON files for the dataset, the source code for extracting VGG-16 features [35] for Flickr8K. Therefore, each image is represented by a 4096-dimensional CNN feature vector and then further reduced to a 128-dimensional vector via PCA for saving computing and memory resources. Meanwhile, we utilize Paragraph Vector model (Sentence2vec) [37] to obtain the sentence representation for each image description. For each image, we use the average value of its corresponding 5 sentence vectors obtained by Sentence2vec as the final image description. Here, the default parameter setting for sentence2vec is used and thus each image description is represented by a 100-dimensional vector. Note that in order to obtain better sentence representation, the corpus of MSCOCO [42] from both training and validation sentence data has been utilized as training set for Sentence2vec. For Flickr8K, we use all 6000 pairwise image-sentences for training, 1000 for validation, and the rest 1000 for testing. Finally, a returned point is considered to be a true neighbor if they share at lease one common label in their corresponding descriptive sentences.

UIUC is a small dataset that randomly sampled from PASCAL VOC 2008 training and validation data with 20 object categories. And there are 50 image-sentences for each category. In our experiment, we randomly select 40 image-sentences from each category for training, 5 for validations and the remaining for testing. The feature extraction of image/sentences follows the same setting as Flickr8K does.

5.1.2. Baseline methods

According to different retrieval tasks, baselines can be divided into two categories: traditional image-query-image assignment and current cross-modal search. Therefore, IMH [16], LSSH [17], CMFH [18], and STMH [19] are selected as comparisons for image-query-text and text-query-image missions, while apart from these four state-of-the-art hashing methods, more ones such as PCAH, PCA-RR and ITQ [9], CBE-opt [43], LSH [6], SH [44], SKLSH [8], DSH [45], SpH [10], SELVE [4], BRE [46] are prepared for the traditional image-query-image search. For all the compared methods, the codes are kindly provided by the authors and the model parameters are tuned and utilized as suggested in their papers. When comparing with the baselines, we set the parameters which yield the best MAP on validation sets for our method on UIUC ($\alpha = \beta = 1$, $\gamma = 50$, $\varepsilon = 5e-2$, and $\eta_0 = \eta_1 = 0.1$) and Flickr8k ($\alpha = \beta = 1$, $\gamma = 20$,

[7] http://cs.stanford.edu/people/karpathy/deepimagesent/

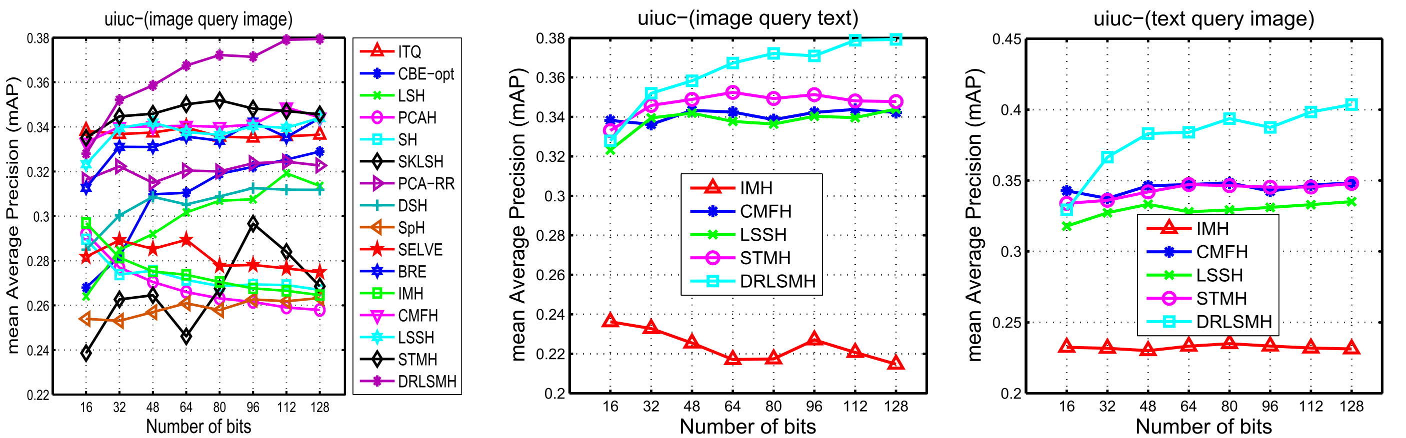

Fig. 2. mAP curves on UIUC for retrieval tasks varying code length.

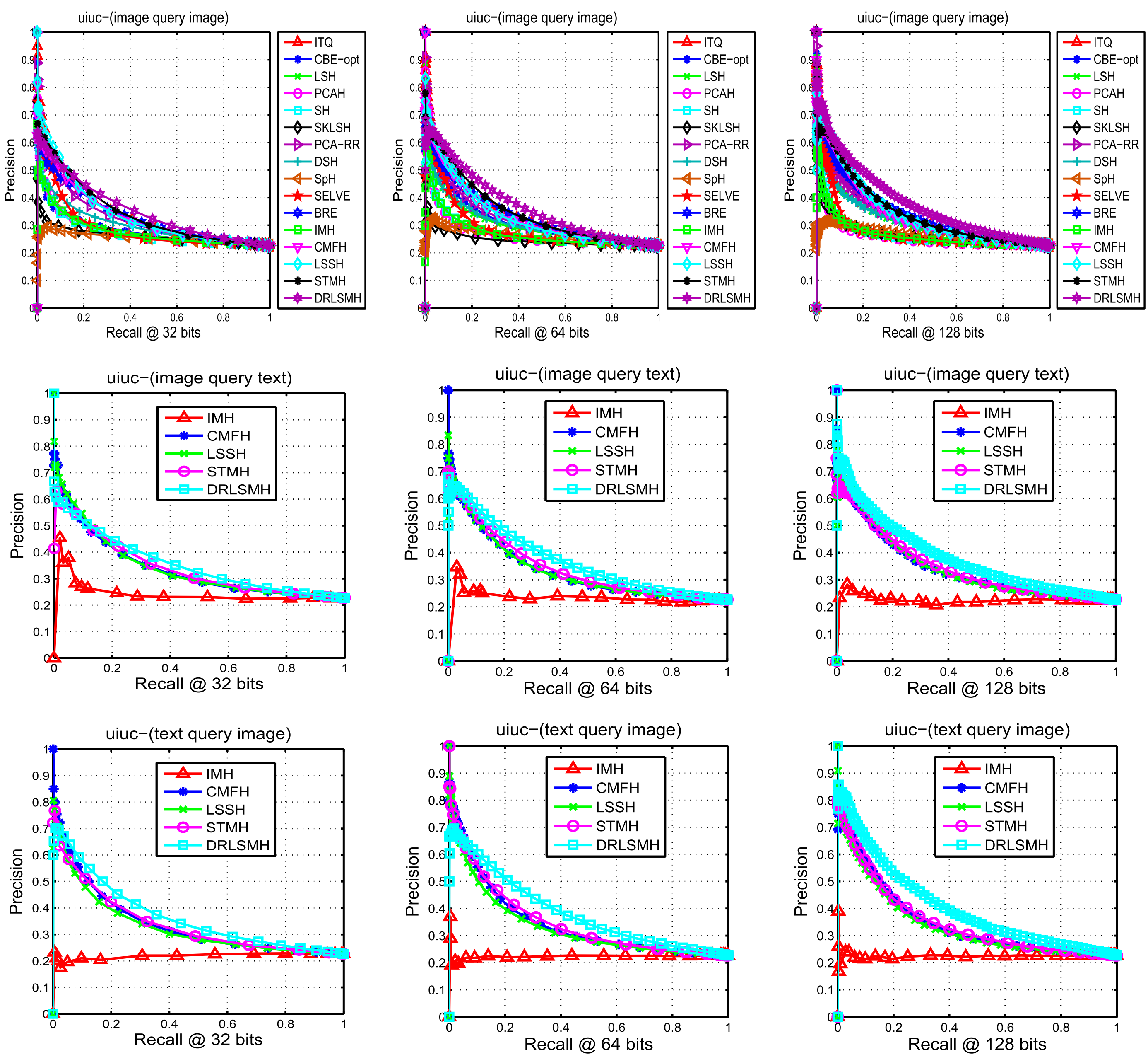

Fig. 3. PR curves on UIUC for retrieval tasks varying code length.

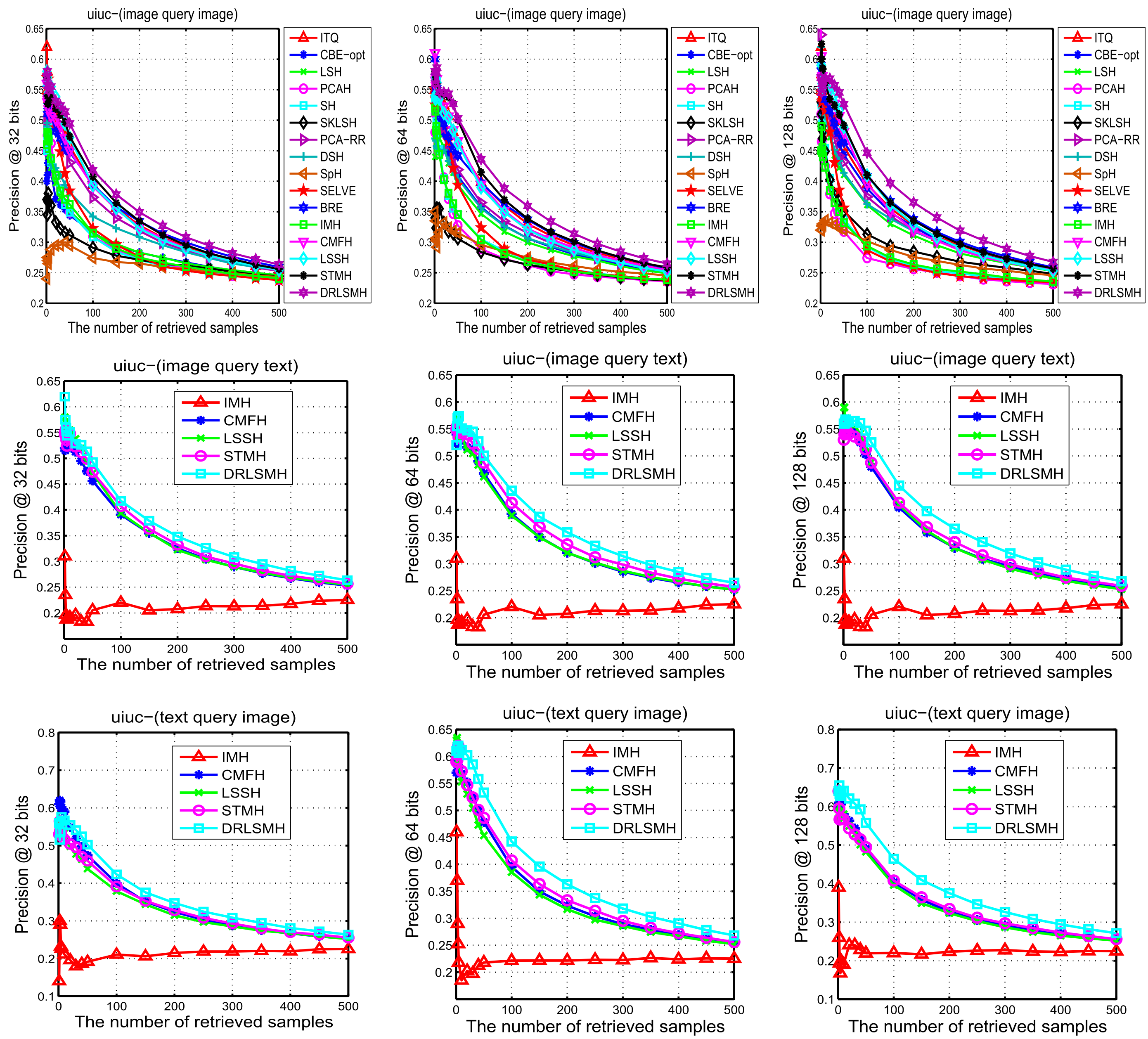

Fig. 4. Precision@topN curves on UIUC for retrieval tasks varying code length.

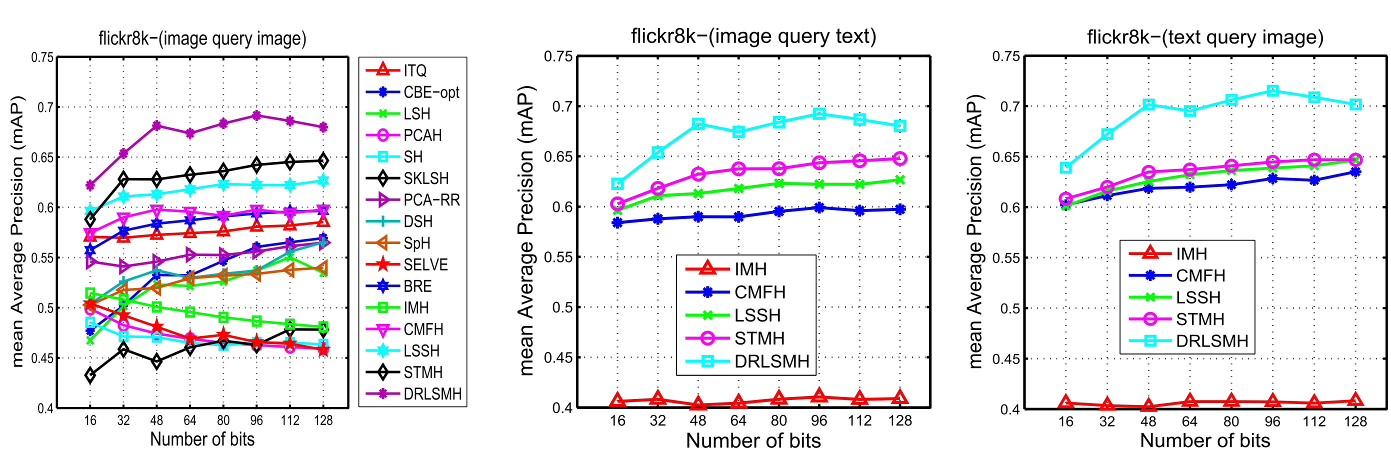

Fig. 5. mAP curves on Flickr8k for retrieval tasks varying code length.

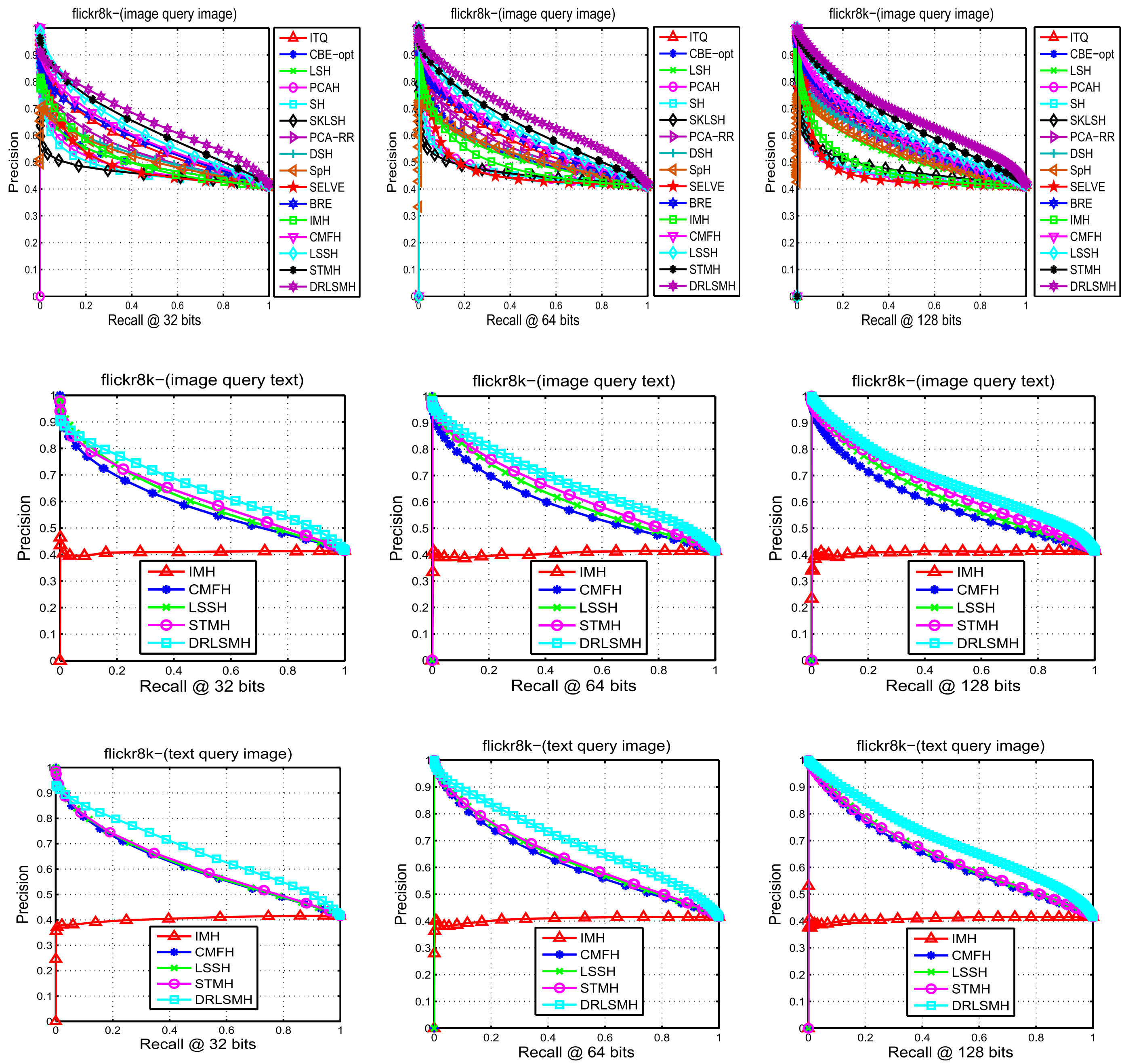

Fig. 6. PR curves on Flickr8k for retrieval tasks varying code length.

$\varepsilon = 5e - 3$, and $\eta_0 = \eta_1 = 0.8$).

5.1.3. Evaluation metrics

We adopt the mean of average precision (mAP), precision-recall, and top-N precision as the evaluation metrics for similarity search effectiveness in our experiments. More details can be referred to [17].

5.2. Results and discussions

5.2.1. Results on UIUC

The mAP Curves for DRLSMH against corresponding baseline hashing methods for different kinds of retrieval tasks varying code length are reported in Fig. 2. The precision-recall and topN precisions curves are plotted in Figs. 3 and 4 respectively. We can observe that DRLSMH outperforms all baseline methods on both image-query-image, image-query-text and text-query-image tasks varying code length overall (except the comparable performance at mAP @16bits with other methods), which on the whole verifies the effectiveness of our proposed hashing method.

More specifically, cross-modal hashing methods (DRLSMH, STMH, LSSH, CMFH, IMH) are mostly above the traditional image-query-image hashing models (such as BRE, SELVE, CBE, ITQ etc.) from different metric curves; this is mainly because more contextual texts other than traditional labels or none are utilized to supervise better semantic image codes. When considering cross-modal retrieval, our proposed method also performs better than the selected four state-of-the-art hashing models by the three metrics; and even compared with the best baseline (STMH), DRLSMH owns an averaged increase of 5.0%, 4.6%, and 10.9% for the image-query-image, image-query-text and text-query-image tasks respectively with the mAP. The main reason would probably come to the diversity regularizations and full utilization of sentence semantics.

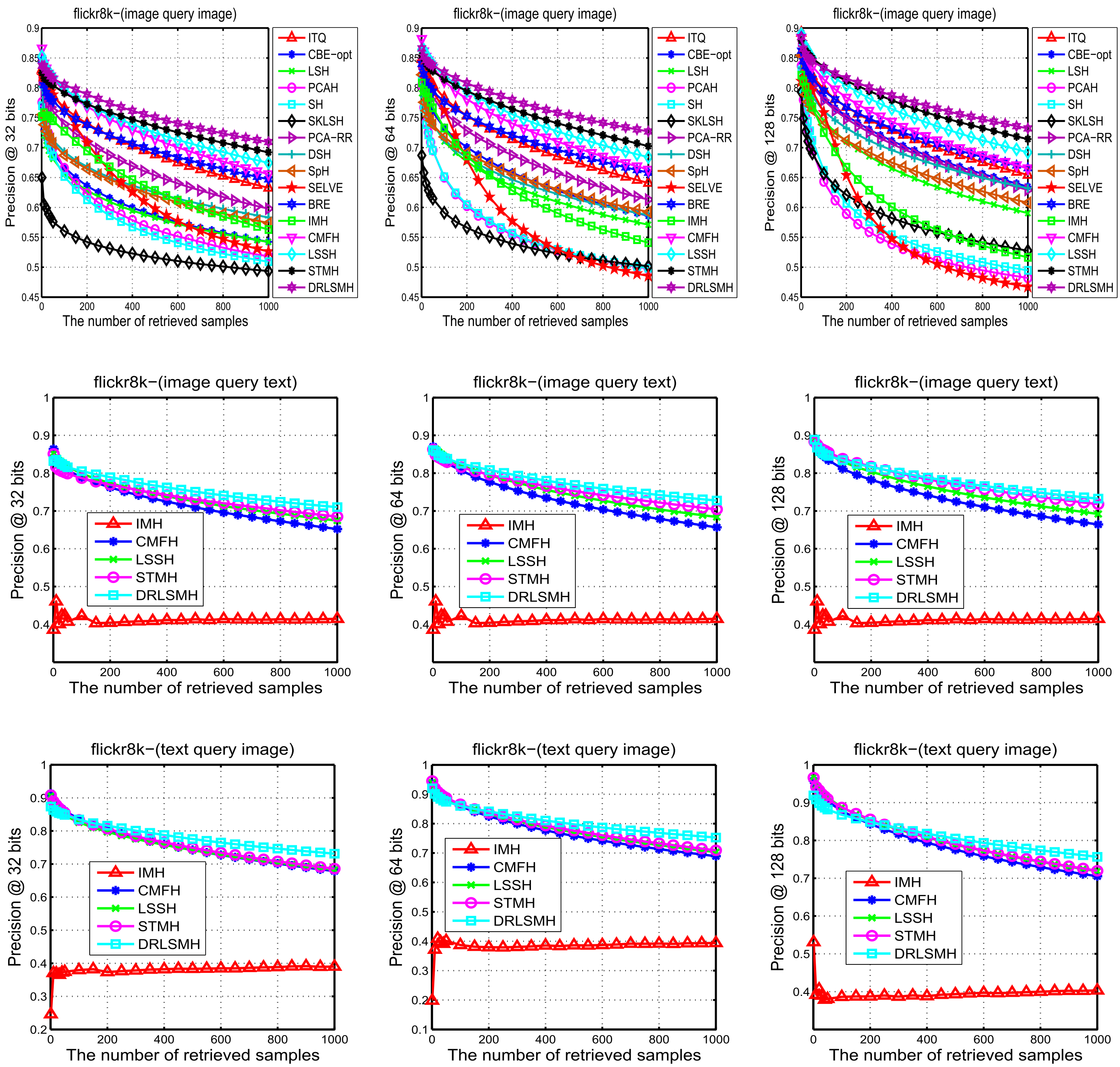

Fig. 7. Precision@topN curves on Flickr8k for retrieval tasks varying code length.

Furthermore, DRLSMH also shows better performances with longer code on different retrieval missions from displayed figures. This is reasonable because longer hash codes can encode more information and thus can improve the mAP, PR and Precision@topN performance. In addition, I would like to emphasize the superior topN (top-400/500 for example) precision over other state-of-the-art methods with consistent phenomenon (also the longer, the better) in these three search missions which indicates a great promising application in real-world.

5.2.2. Results on Flickr8k

Similar as the former subsection, we further make a thorough inquiry about the DRLSMH's effectiveness on more scaled corpus Flickr8k with mAP, precision-recall and topN precision retrieval metrics. The results for these three evaluation metrics are displayed in Figs. 5, 6 and 7 respectively on both traditional image retrieval and current cross-modal search tasks varying code length. As shown in figures, we can clearly figure out that a significant better performance than other 10+ baselines in all these retrieval missions demonstrates the superiority of our proposed model. More specifically, DRLSMH performs an averaged increase of 6.4%, 6.1% and 9.0% for image-query-image, image-query-text and text-query-image tasks with mAP metric respectively even over the best preformed baseline STMH method. Furthermore, better results with longer code bits as well as consistent performance in three different search tasks echo the corresponding former experiments on UIUC which further verify our model's effectiveness. Moreover, topN (top-400/500) precision figures exhibits a probable promise in multi-tasks and multi-modal information retrieval for DRLSMH in real world for the contemporary.

5.2.3. Summary

From the above two designed experiments, three common points can be easily summarized as below.

- Generally speaking, current cross-modal hashing methods (e.g. CMFH, LSSH, STMH, DRLSMH) most probably have a superior performance against traditional image-query-image hashing models (e.g. BRE, SELVE, CBE-opt, ITQ, LSH etc.) because of the utilization of more semantic contextual information.
- Our proposed method, DRLSMH, shows the best performance on the whole over other 10+ baselines including traditional and current hashing methods. This is most probably beneficial from the designed framework (Fig. 1): deep representations (images and sentences), semantic match learning and space shift for hashing.
- In light of the Precision@topN metric, DRLSMH are all above others from all the plotted curves no matter what kind of retrieval tasks and how long the bits code are, especially at the focus of top-400/500, which indicates a promising application for real world search engine.

6. Conclusions

In this paper, we put forward a new hashing method, referred to as Diversity Regularized Latent Semantic Match for Hashing, for cross-modal retrieval between images and texts (sentences). More specifically, we map the feature vectors extracted from deep learning models of images and sentences to a shared latent semantic space with label-supervised graph Laplacians for intra-media consistency in the match learning framework, where soft orthogonality is induced as a novel regularizer on projections for diverse hashing functions to preserve compact and accurate data representations. Then elementwise operator $\mathrm{sgn}(\cdot)$ is utilized to implement space shift from the latent semantic space to the final hamming space. Notably, proper relaxations on diversity regularization greatly contribute to the closed-form solutions for the iterative algorithms which make the training fast and efficient.

Pioneer extensive experiments on two public multi-modal corpora consisting of images and sentences show the superior performance against several state-of-the-art cross-view hashing methods both on image-query-image, image-query-text and text-query-image retrieval tasks, especially with longer hash codes.

Acknowledgment

This research was supported by National High Technology Research and Development Program (863 Program of China) under Grant 2014AA021504.

References

[1] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval ideas, influences, and trends of the new age, ACM Comput. Surv. 40 (2) (2008) 5.
[2] J. He, S. Kumar, S.-F. Chang, On the difficulty of nearest neighbor search, in: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012.
[3] J. Moraleda, G. Shakhnarovich, T. Darrell, P. Indyk, Nearest-neighbors methods in learning and vision theory and practice, Pattern Anal. Appl. 11 (2) (2008) 221–222.
[4] X. Zhu, L. Zhang, Z. Huang, A sparse embedding and least variance encoding approach to hashing, IEEE Trans. Image Process. 23 (9) (2014) 3737–3750.
[5] H. Liu, R. Ji, Y. Wu, W. Liu, G. Hua, Supervised matrix factorization for cross-modality hashing, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA , 2016, pp. 1767–1773.
[6] M. Datar, N. Immorlica, P. Indyk, V.S. Mirrokni, Locality-sensitive hashing scheme based on p-stable distributions, in: Proceedings of the 20th ACM Symposium on Computational Geometry, Brooklyn, New York, USA, 2004, pp. 253–262.
[7] B. Kulis, K. Grauman, Kernelized locality-sensitive hashing for scalable image search, in: Proceedings of the IEEE12th International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 2130–2137.
[8] M. Raginsky, S. Lazebnik, Locality-sensitive binary codes from shift-invariant kernels, in: Proceedings of the Advances in neural information processing systems, 2009, pp. 1509–1517.
[9] Y. Gong, S. Lazebnik, Iterative quantization: A procrustean approach to learning binary codes, in: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 2011, pp. 817–824.
[10] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, S.-E. Yoon, Spherical hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 2957–2964.
[11] J. Wang, O. Kumar, S.-F. Chang, Semi-supervised hashing for scalable image retrieval, in: Proceedings of the twenty third IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 3424–3431.
[12] M. Norouzi, D.J. Fleet, Minimal loss hashing for compact binary codes, in: Proceedings of the 28th International Conference on Machine Learning, Bellevue, Washington, USA, 2011, pp. 353–360.
[13] J. Wang, X.-S. Xu, S. Guo, L. Cui, X.-L. Wang, Linear unsupervised hashing for ann search in euclidean space, Neurocomputing 171 (2016) 283–292.
[14] D. Wang, X. Gao, X. Wang, Semi-supervised constraints preserving hashing, Neurocomputing 167 (2015) 230–242.
[15] B. Gholami, A. Hajisami, Kernel auto-encoder for semi-supervised hashing, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, 2016, pp. 1–8.
[16] J. Song, Y. Yang, Y. Yang, Z. Huang, H.T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 2013, pp. 785–796.
[17] J. Zhou, G. Ding, Y. Guo, Latent semantic sparse hashing for cross-modal similarity search, in: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, QLD, Australia, 2014, pp. 415–424.
[18] G. Ding, Y. Guo, J. Zhou, Collective matrix factorization hashing for multimodal data, in: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 2083–2090.
[19] D. Wang, X. Gao, X. Wang, L. He, Semantic topic multimodal hashing for cross-media retrieval, in: Twenty-Proceedings of the Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015, pp. 3890–3896.
[20] Y. Yu, Y.-F. Li, Z.-H. Zhou, Diversity regularized machine, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, 2011, pp. 1603–1608.
[21] L. Jiang, D. Meng, S.-I. Yu, Z.-Z. Lan, S. Shan, A. G. Hauptmann, Self-paced learning with diversity, in: Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, 2014, pp. 2078–2086.
[22] P. Xie, Learning compact and effective distance metrics with diversity regularization, in: Proceedings of the Machine Learning and Knowledge Discovery in Databases - European Conference, Porto, Portugal Proceedings, 2015, pp. 610–624.
[23] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 586–594.
[24] H. Li, J. Xu, Semantic Matching in Search, Foundations and Trends in Information Retrieval.
[25] J. Shang, T. Chen, H. Li, Z. Lu, Y. Yu, A parallel and efficient algorithm for learning to match, in: Proceedings of the IEEE International Conference on Data Mining, Shenzhen, China, 2014, pp. 971–976.
[26] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Proceedings of the Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, 2014, pp. 2042–2050.
[27] Z. Lu, H. Li, A deep architecture for matching short texts, in: Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, United States, 2013, pp. 1367–1375.
[28] W. Wu, Z. Lu, H. Li, Learning bilinear model for matching queries and documents, J. Mach. Learn. Res. 14 (1) (2013) 2519–2548.
[29] Y. Feng, M. Lapata, Automatic image annotation using auxiliary text information, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, USA, 2008, pp. 272–280.
[30] S. Purushotham, Y. Liu, Collaborative topic regression with social matrix factorization for recommendation systems, in: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012.
[31] Y. Bengio, Learning deep architectures for ai, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.
[32] L. Deng, D. Yu, Deep learning: Methods and Applications, Foundations and Trends in Signal Processing.
[33] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[34] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[35] Simonyan, Karen, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: arXiv preprint arXiv:1409.1556arXiv:1409.1556, 2014.
[36] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, United States., 2013, pp. 3111–3119.
[37] Q.V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31th International Conference on Machine Learning, Beijing, China, 2014, pp. 1188–1196.
[38] D. Zhang, F. Wang, L. Si, Composite hashing with multiple information sources, in: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, 2011, pp. 225–234.
[39] Z. Lu, X. Gao, L. Wang, J.-R. Wen, S. Huang, Noise-robust semi-supervised learning by large-scale sparse coding, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, Texas, USA, 2015, pp. 2828–2834.
[40] Rashtchian, Young, Hodosh, Hockenmaier, Collecting image annotations using amazon's mechanical turk, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010.
[41] Hodosh, Young, Hockenmaier, Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, Journal of Artificial Intelligence Research.
[42] T.-Y. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision - ECCV 2014 - Proceedings of the 13th European Conference, Zurich, Switzerland, Proceedings, Part V, 2014, pp. 740–755.
[43] F.X. Yu, S. Kumar, Y. Gong, S.-F. Chang, Circulant binary embedding, in: Proceedings of the 31th International Conference on Machine Learning, Beijing, China, 2014, pp. 946–954.
[44] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in: Proceedings of the Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2008, pp.

1753–1760.
[45] Z. Jin, C. Li, Y. Lin, D. Cai, Density sensitive hashing, IEEE Trans. Cybern. 44 (8) (2014) 1362–1371.
[46] B. Kulis, T. Darrell, Learning to hash with binary reconstructive embeddings, in: Proceedings of the Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2009, pp. 1042–1050.

Yong Chen received the B.Sc. degree in Computer Science and Technology from CAFUC (Civil Aviation Flight University of China) in 2011, and the MEng degree in School of Computer Science and Engineering from BUAA (Beihang University) in 2014. He is currently a PhD candidate in the State Key Lab of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, PRChina. His research interests include machine learning, data mining and big data. For more personal research information, you can refer to http://www.scholarmate.com/scmwebsns/pv/alpha.

Zhang Hui Ph.D. Professor, Deputy Director of State Key Software Development Environment of School of Computer Science of Beihang University. He received his MS and Ph.D. degrees in Computer Science from Beihang University in 1994 and 2009 respectively. He had been working in University of Chicago and Argonne National Laboratory, USA from 2007 to 2008 as a guest researcher. He has been teaching Computer Networks for undergraduate students of School of Computer Science for more than 10 years. He is also an Associate Director of National Engineering Technology Center of S & T Resources Sharing Services which organizing and managing the e-Science data of Chinese government.

His main research interests include e-Science, data management and data mining, web information retrieval and cloud computing. He published over fifty research papers in these fields on journals and conferences. Contact: telephone 82338088 (O).

Ming Lu received his M.S degree in the Image Processing Center, School of Astronautics at Beihang University, Beijing 100191, PR China. Before that, he received the B.S degree in Computer Science and Technology from East China University of Science and Technology (ECUST) in 2007, Shanghai, China. His research interests include Computer Vision and Machine Learning.

Yongxin Tong received his Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2014. He is currently an associate professor in the School of Computer Science and Engineering, Beihang University, Beijing, China. Before that, he served as a Research Assistant Professor and a Postdoctoral fellow at HKUST. His research interests include crowdsourcing, uncertain data mining and management and social network analysis. His research interests include crowdsourcing, uncertain data processing and social network analysis. He has published more than 20 papers in highly refereed database and data mining journals and conferences such as SIGMOD, SIGKDD, VLDB, ICDE, TKDE, and TOIS. He was awarded the Microsoft Research Asia Fellowship 2012, and received the Excellent Demonstration Award and the Best Paper Award conferred by the VLDB 2014 and WAIM 2016 conferences, respectively. He has also been a reviewer for leading academic journals, such as TKDE, and has served in the program committees of prestigious international conferences, such as IJCAI 2015. He is a member of IEEE, ACM, and CCF.