Deep Joint Semantic-Embedding Hashing

Ning Li¹, Chao Li¹, Cheng Deng¹*, Xianglong Liu², Xinbo Gao¹
¹ School of Electronic Engineering, Xidian University, Xi’an 710071, China
² Beihang University, Beijing 100191, China
ningli2017@gmail.com, li_chao@stu.xidian.edu.cn, {chdeng, xbgao}@mail.xidian.edu.cn, xliu@nlsde.buaa.edu.cn

Abstract

Hashing has been widely deployed to large-scale image retrieval due to its low storage cost and fast query speed. Almost all deep hashing methods do not sufficiently discover semantic correlation from label information, which results in the learned hash codes less discriminative. In this paper, we propose a novel Deep Joint Semantic-Embedding Hashing (DSEH) approach that consists of LabNet and ImgNet. Specifically, LabNet is explored to capture abundant semantic correlation between sample pairs and supervise ImgNet from both semantic level and hash codes level, which is conductive to the generated hash codes being more discriminative and similarity-preserving. Extensive experiments on three benchmark datasets show that the proposed model outperforms current state-of-the-art methods.

1 Introduction

Due to the explosive increase of high-dimensional media data in search engines and social networks, approximate nearest neighbor (ANN) search for large-scale datasets has attracted more and more attention. Among existing ANN techniques, hashing has become the most popular and effective one due to its fast query speed and low memory cost [Deng et al., 2015a; 2015b], which aims to map high-dimensional data into compact binary codes and preserve their original similarities.

Recently, deep hashing methods [Xia et al., 2014; Lai et al., 2015; Cao et al., 2017; Yang et al., 2017; Li et al., 2018] have gained state-of-the-art performance due to their powerful ability of feature learning by using deep network architecture, with which we can build more accurate similarity relationship and then generate more discriminative hash codes. Compared with unsupervised deep hashing methods, supervised ones can achieve better performance with the aid of label information. Even so, how to sufficiently discover the semantic correlation from label information is still a crucial issue to be addressed. In this paper, we mainly focus on extracting abundant semantic correlation from label information with deep neural network.

*Corresponding author

Figure 1: Single-label dataset vs. multi-label dataset.

Actually, existing supervised hashing methods do not rationally exploit label information of samples, almost all of which only simply construct the similarity affinity matrix of sample pairs [Xia et al., 2014; Li et al., 2015; Liu et al., 2016a]. As shown in Fig. 1a, for the ImageNet dataset, each sample is annotated by single label, where the similarity relationship between samples is very sparse, i.e., the number of similar pairs is much smaller than the number of dissimilar pairs, which will result in that the learned hash codes cannot preserve the original similarity relationship effectively. To tackle this problem, HashNet [Cao et al., 2017] alleviates such data imbalance by adjusting the weights of similar pairs. However, the optimal weights cannot be easily obtained, which limits its feasibility to real-world retrieval system. For NUS-WIDE dataset, as shown in Fig. 1b, each sample is annotated with multiple labels, which can provide high level semantic information and complex similarity relationship. Unfortunately, multiple labels in current methods are oversimplified to single-label case, which removes many useful semantic information and cannot maintain the original similarity relationship of sample pairs. Therefore, either single-label or multi-label dataset, we should capture more abundant semantic correlation to indicate the accurate similarity relationship between samples and produce more discriminative hash codes.

In this paper, we propose a novel Deep Joint Semantic-Embedding Hashing method, namely DSEH, in which both LabNet and ImgNet are end-to-end networks containing semantic layers and hash layers. In LabNet, label information are projected into common semantic space and common Hamming space for exploring abundant semantic features and discriminative hash codes, respectively. In ImgNet, an image is embedded into the common semantic space and common
Hamming space. By exploiting the learned semantic correlation and hash codes in LabNet as supervised information and transferring them to ImgNet with the form of two constraints, more accurate semantic correlation can be discovered and thus discriminative hash codes can be generated. Extensive experiments, conducted on three popular datasets including single-label and multi-label ones, demonstrate the proposed DSEH outperforms state-of-the-art hashing approaches.

The main contributions of our DSEH are summarized as follows. 1) We exploit a novel architecture for deep hashing, consisting of LabNet and ImgNet, where common semantic space and common Hamming space are built across the networks. 2) We utilize a couple of constrains to build a relationship between LabNet and ImgNet from semantic feature level and hash code level. 3) We adopt an alternative training strategy to jointly optimize the parameters of these two networks, and produce the optimal hash codes.

2 Related Work

Existing hashing methods can be roughly categorized into unsupervised [Gionis et al., 1999; Weiss et al., 2009; Gong et al., 2013; Liu et al., 2016b] and supervised hashing [Liu et al., 2012; Shen et al., 2015; Deng et al., 2014; 2016; Liu et al., 2016a; 2016a; Deng et al., 2018]. Unsupervised hashing methods learn hash functions from unlabeled data. Locality Sensitive Hashing (LSH) [Gionis et al., 1999] uses random projections as hash function. Graph-based hashing [Liu et al., 2011] learns appropriate hash codes by discovering inherent neighborhood structure. Supervised hashing methods incorporate semantic label or relevance information to improve the quality of hash codes. Binary Reconstruction Embedding (BRE) [Kulis and Darrell, 2009] designs hash functions by minimizing the squared errors between the original distances and the reconstructed distances in Hamming space. Supervised Hashing with Kernels (KSH) [Liu et al., 2012] learns to build compact binary codes by minimizing the Hamming distances on similar pairs and maximizing those on dissimilar pairs.

Deep hashing methods have been presented recently, which achieve promising performance due to the powerful arbitrary nonlinear representation of deep neural network. With the help of this structure, CNNH [Xia et al., 2014] learns approximate hash codes from the pairwise similarity regularization first, then tries to learn feature representation and hash function based on the hash codes in the first stage. DNNH [Lai et al., 2015] and DPSH [Li et al., 2015] integrate feature learning and hashing learning into a unified end-to-end network to improve the discrimination of hash codes. DSH [Liu et al., 2016a] groups training data into similar pairs and dissimilar pairs to generate similarity correlation and controls the quantization error. One further study, HashNet [Cao et al., 2017] uncovers the inherent problem caused by data imbalance of some single-label dataset and alleviates this drawback by adjusting the weights of semantic correlation matrix. However, the data imbalance remains a challenge and almost all of these methods do not or little exploits semantic information to generate semantic correlation from label information directly.

3 Proposed DSEH

Fig. 2 shows the flowchart of the proposed method, which mainly consists of two parts: LabNet and ImgNet. LabNet is an end-to-end fully connected deep neural network, where a semantic layer and a hash layer are built to generate semantic features and hash codes from label information. Meanwhile, ImgNet consists of a convolution neural network with a semantic layer and a hash layer, which is used to learn hash codes of the input images.

3.1 Problem Formulation

In similarity retrieval scenario, given a dataset \( O = \{ \alpha_i \}_{i=1}^n \), \( \alpha_i = (v_i, l_i) \), where \( v_i \in \mathbb{R}^{1 \times d_v} \) is a feature vector of the \( i \) th sample, which could be hand-crafted feature, deep feature, or raw pixels of an image. \( l_i = [l_{i1}, \ldots, l_{ic}] \) is the label annotations assigned to \( \alpha_i \), where \( c \) is the number of classes. \( \alpha_i \) and \( \alpha_j \) are associated with similarity label \( s_{ij} \), where \( s_{ij} = 1 \) implies \( \alpha_i \) and \( \alpha_j \) are similar, or otherwise \( s_{ij} = 0 \). In our setting, we define \( s_{ij} = 1 \) if \( \alpha_i \) and \( \alpha_j \) share at least one label, and \( s_{ij} = 0 \) if \( \alpha_i \) and \( \alpha_j \) have no common label. The goal of deep hashing is to learn nonlinear hash function, \( i.e., f: O \rightarrow h \in \{-1, 1\}^K \), to encode each sample \( O \) into compact \( K \)-bit hash code \( h \), such that the original similarity between sample pairs can be well preserved.

For two binary hash codes \( h_i \) and \( h_j \), their Hamming distance \( dis_H(h_i, h_j) \) and inner product \( \langle h_i, h_j \rangle \) can be formu-
lated as:
\[
    \text{dis}_H(h_i, h_j) = \frac{1}{2} (K - \langle h_i, h_j \rangle).
\]
(1)

If the inner product of two binary codes is small, their Hamming distance should be large, and vice versa. Given the hash codes \( h_i \) and \( h_j \), the similarity probability between \( o_i \) and \( o_j \) is defined as a likelihood function:
\[
    p(s_{ij} | h_i, h_j) = \begin{cases} 
        \sigma(\langle f_i^T f_j \rangle), & s_{ij} = 1 \\
        1 - \sigma(\langle f_i^T f_j \rangle), & s_{ij} = 0 
    \end{cases}
\]
(2)

where \( \sigma(x) = \frac{1}{1 + e^{-x}} \) is the sigmoid function. Similar to logistic regression, we can see that the smaller Hamming distance \( \text{dis}_H(h_i, h_j) \) is, the larger their inner product \( \langle h_i, h_j \rangle \) is. A smaller condition probability \( P(0|h_i, h_j) \) implies \( h_i \) and \( h_j \) should be similar; otherwise, a larger condition probability \( P(1|h_i, h_j) \) means \( h_i \) and \( h_j \) should be dissimilar. Thus, quantifying the similarity relationship between hash codes in Hamming space can be transformed into calculating the inner product of original hash codes.

Similar to hash learning, replacing two features \( f_i \) and \( f_j \) in Eq. (2), the similarity between two features can also be calculated. The larger \( \langle f_i, f_j \rangle \) is, the greater the similarity of them is, and vice versa. The similarity probability of \( f_i \) and \( f_j \) can be expressed as likelihood function:
\[
    p(s_{ij} | f_i, f_j) = \begin{cases} 
        \sigma(\langle f_i^T f_j \rangle), & s_{ij} = 1 \\
        1 - \sigma(\langle f_i^T f_j \rangle), & s_{ij} = 0 
    \end{cases}
\]
(3)

### 3.2 LabNet Learning

For discovering the abundant semantic correlation from label information, our LabNet is constrained in both semantic space and Hamming space. Pairwise correlation loss in these two spaces should be concerned. Let \( f(l_i; \theta^l) \) denote embedding labels for point \( i \), and \( \theta^l \) is the parameter of LabNet.

Different from generating supervised information only in the Hamming space in most existing methods, a new semantic space is constructed in our method, within which similarity relationship can be well preserved at semantic level. For all the instances in semantic space, given features \( F^l = \{f_i^l\}_{i=1}^n \) and pairwise similarity labels \( S = \{s_{ij}\} \), the logarithm Maximum a Posterior (MAP) estimation of semantic features \( F^l = \{f_i^l, \cdots, f_N^l\} \) can be expressed as:
\[
    \log p(F^l | S) \propto \log p(S | F^l) p(F^l) = \sum_{s_{ij} \in S} \log p(s_{ij} | f_i^l, f_j^l) p(f_i^l, f_j^l)
\]
(4)

where \( p(S | F^l) \) is the likelihood function, and \( p(F^l) \) is the prior distribution. By taking the negative log-likelihood of the observed pairwise labels in \( S \), we can frame the following optimization problem as:
\[
    \min_{F^l, \theta^l} J_1 = - \log p(S | F^l) = - \sum_{s_{ij} \in S} (s_{ij} f_i^T f_j^l - \log(1 + \exp(f_i^T f_j^l)))
\]
(5)

It is easy to find that the above optimization problem can make semantic features \( F^l \) to preserve the original similarity relationship in semantic space.

Then, semantic features are embedded into Hamming space to produce compact binary codes which also need to keep the original similarities. The MAP estimation of hash codes \( H^l = [h_1^l, \cdots, h_N^l] \) can be represented as:
\[
    \log p(H^l | S) \propto \log p(S | H^l) p(H^l) = \sum_{s_{ij} \in S} \log p(s_{ij} | h_i^l, h_j^l) p(h_i^l, h_j^l).
\]
(6)

When substituting Eq. (2) into MAP estimation in Eq. (6), the problem can be formulated as:
\[
    \min_{H^l, \theta^l} J_2 = - \sum_{s_{ij} \in S} \left( s_{ij} h_i^T h_j^l - \log(1 + \exp(h_i^T h_j^l)) \right)
\]
(7)

Furthermore, in order to promote the hash value discretization, binary regularization should be considered additionally, which can be formulated as follow:
\[
    \min_{H^l, \theta^l} J_3 = \sum_{s_{ij} \in S} (\|h_i^l - 1\|_1 + \|h_j^l - 1\|_1)
\]
(8)

where \( 1 \in \mathbb{R}^K \) is the vector of ones, and \( \| \cdot \|_1 \) denotes the \( \ell_1 \) norm of a vector.

Finally, to maintain the semantic information during the training of LabNet, the achieved hash codes from Hamming space is mapped to original label. Therefore, the output of LabNet can be written as:
\[
    \hat{Y}^l = W^T H^l + b
\]
(9)

where \( \hat{Y}^l \) is the predicted label of output, and \( W \) is the mapping weight. To minimize the distance between the predict label \( y_i^l \) and ground truth \( y_i \), the least squares loss is adopted as follows:
\[
    \min_{Y^l, \theta^l} J_4 = \sum_{i=1}^N \|y_i - y_i^l\|_2^2 = \sum_{i=1}^N \|y_i - W^T h_i^l - b\|_2^2
\]
(10)

where \( \| \cdot \|_2 \) is the \( l_2 \) norm of a vector.

The overall objective function for LabNet can be written as follows:
\[
    \min_{F^l, H^l, \theta^l} J_Lab = J_1 + \alpha J_2 + \beta J_3 + \gamma J_4
\]
(11)

where \( \alpha, \beta, \gamma \) are the hyper-parameters corresponding to the loss function, respectively.

### 3.3 ImgNet Learning

ImgNet is supervised by LabNet from semantic features as well as hash codes. Let \( g(v_i; \theta^v) \) be the learned image feature for the \( i \)th samples, where \( \theta^v \) is the network parameter of ImgNet.

...
In the common semantic space between LabNet and ImgNet, if the sample pairs $v_i$ and $v_j$ are similar, their corresponding features $f_i^v$ and $f_j^v$ should also be similar. Supervised by the semantic feature of LabNet, the semantic feature $F^v$ of ImgNet can be depicted as:

$$
\min_{F^v, \theta^v} J_1 = - \log p(S|F^v)
$$

$$
= - \sum_{s_{ij} \in S} (s_{ij} f_i^v f_j^v - \log(1 + \exp(f_i^v f_j^v)))
$$

(12)

where $f_i^v$ is the semantic feature generated by ImgNet, and $f_j^v$ is semantic feature from LabNet.

In common Hamming space, different from the traditional methods that employ pairwise similarity and iterative search hash codes, we guide the hash codes learning in ImgNet by utilizing the learned hash codes in LabNet. The hash layer of ImgNet is constrained to approach precise binary code \(\{0, 1\}^K\) by utilizing sigmoid function with cross-entropy loss. Since the activation function of hash layer in LabNet is \(\tanh(\cdot)\), the hash codes of LabNet need to adjust from $h_i^l \in \{-1, 1\}^K$ to $h_i^l' \in \{0, 1\}^K$ to match the \(\text{sigmoid}(\cdot)\) activation function in ImgNet. The loss of hash codes in common Hamming space is defined as:

$$
\min_{H^v, \theta^v} J_2 = - \sum_{i=1}^{N} \left[ h_i^l \log \sigma(y_i^v) + (1 - h_i^l) \log(1 - \sigma(y_i^v)) \right]
$$

(13)

where $y_i^v$ is the output of ImgNet.

Therefore, the whole objective function of ImgNet is denoted as follow:

$$
\min_{F^v, H^v, \theta^v} \mathcal{L}_{img} = J_1 + \eta J_2
$$

(14)

where $\eta$ is the hyper-parameter to balance the two loss function terms.

3.4 Training Strategy

LabNet takes advantage of all label information to generate semantic features and hash codes. However, the learned semantic features and hash codes in LabNet may not match well with the corresponding semantic features and hash codes to be learned in ImgNet at the beginning. Therefore, we should exploit the strategy of alternative training to reconstruct the optimal semantic features and hash codes in semantic space and Hamming space, respectively.

Specifically, we first randomly initialize LabNet and train it until $\mathcal{L}_{lab}$ reaches convergence. Then, utilizing the obtained semantic features and hash codes in LabNet, we supervise the ImgNet training in semantic space and Hamming space, respectively. Next, we initialize the semantic features and hash codes of LabNet with the resulting semantic feature and hash codes in ImgNet generated from the second step. Finally, repeating such training procedure for LabNet and ImgNet until convergence.

Algorithm 1 outlines the whole leaning algorithm in detail. It is noted that we learn all network parameters by utilizing stochastic gradient descent (SGD) with back-propagation (BP) algorithm, which is also widely used in existing deep learning methods.

Algorithm 1 The learning algorithm for our DSEH

| Input: | Image set $X$, Label set $L$ |
| Output: | Parameters $\theta^v$ of ImgNet, Optimal code matrix $H$ |
| **Initialization** | Initialize network parameters $\theta^l, \theta^v$; hyper-parameters: $\alpha, \beta, \gamma$, and $\eta$; learning rate: $\mu$; mini-batch size: $N^l = 32$, $N^v = 128$ |
| **repeat** | for $t^v$ iteration do |
| | Update $\theta^l$ by BP algorithm: $\theta^l \leftarrow \theta^l - \mu \cdot \nabla_{\theta^l} J_1 (\mathcal{C}_{lab})$ |
| | end for |
| | Update the parameter $h_i^l$ by $h_i^l = \text{sign}(h_i^l)$ |
| | Update the parameter $h_i^l$ by adjusting $h_i^l \in \{-1, 1\}^K$ to $\{0, 1\}^K$ for $t^v$ iteration do |
| | Update $\theta^v$ by BP algorithm: $\theta^v \leftarrow \theta^v - \mu \cdot \nabla_{\theta^v} \left( J_1 + \eta J_2 \right)$ |
| | end for |
| | Update the parameter $h_i^l$, $h_i^l$ by $h_i^v = \text{sign}(y_i^v)$, $h_i^l = \text{sign}(y_i^l)$ |
| | Update the parameter $B$ by $B = H^v$ |
| **until** | convergence |

4 Experiments

4.1 Datasets and Settings

The experiments are conducted on three benchmark image retrieval datasets: NUS-WIDE [Chua et al., 2009], ImageNet [Russakovsky et al., 2015], and MS-COCO [Lin et al., 2014].

- **NUS-WIDE** dataset is a multi-label image dataset, which contains 269,648 images with 81 ground truth concepts. We follow similar experimental protocols as DPSH [Li et al., 2015] and use the subset of 195,834 images that are associated with the 21 most frequent concepts, where each concept contains at least 5,000 images. We randomly select 100 images per class as the query set, and 500 images per class as the training set.

- **ImageNet** dataset is a benchmark image dataset for Large Scale Visual Recognition Challenge (ILSVRC 2015), containing over 1.2M images. It is a single-label dataset, where each image is labeled by one of 1,000 categories. We randomly select 100 categories, and randomly select 50 images per class as the query set, 100 images per class as the training set.

- **MS-COCO** dataset is an image recognition, segmentation, and caption dataset which contains 82,783 training images and 40,504 validation images. It is a multi-label dataset labeled by 80 categories. After pruning images without category information, we obtain 122,218 images and randomly sample 5,000 images as queries, 10,000 images as training points.

We evaluate the retrieval quality using three evaluation metrics: Mean Average Precision (MAP), Precision-Recall curves, and Precision curves with respect to the number of top returned results. With the same training and test set, all methods were tested under the same conditions. Given a query, the ground truth is defined as: if a result shares at least one common concept with the query, it is relevant; otherwise it is irrelevant.

We compare our method with ten classical or state-of-art hashing methods, including unsupervised methods.
LSH [Gionis et al., 1999], SH [Weiss et al., 2009], ITQ [Gong et al., 2013], supervised shallow methods KSH [Liu et al., 2012], ITQ-CCA [Gong et al., 2013], SDH [Shen et al., 2015], and deep supervised methods CNNH [Xia et al., 2014], DPSH [Li et al., 2015], DHN [Zhu et al., 2016], HashNet [Cao et al., 2017].

For fair comparison, we extract 4,096-dimensional deep features by CNN-F [Chatfield et al., 2014] model which is re-trained on ImageNet dataset. We construct ImageNet to reserve first seven layers same with those in CNN-F followed fc8 with 512 nodes for semantic layer and $K$ nodes for hash layer, i.e., $(I \rightarrow \text{CNNF} \rightarrow 512 \rightarrow K)$. LabNet is initialized randomly and constructed as $(L \rightarrow 4096 \rightarrow 512 \rightarrow K \rightarrow c)$, which contains $c$ nodes for total class labels.

Since the semantic layer and hash layer are trained from scratch, we set its learning rate 10 times of the ones for the other layers. The learning rate is chosen from $10^{-2}$ to $10^{-6}$ with a validation set. The batch size of LabNet and ImageNet are set to 32 and 128 respectively. Since the semantic correlation of ImageNet is sparse, we set the values in similarity matrix as $S \in \{0, 5\}$. For the hyper-parameters in LabNet, we conduct cross-validation to search $\alpha$ and $\gamma$ from $10^{-3}$ to $10^{5}$, and search $\beta$ from $10^{-6}$ to $10^{-1}$. We find that the optimal result can be obtained when $\alpha = \gamma = 1$, and $\beta = 0.005$. Then we search from $10^{-3}$ to $10^{-2}$ and discover $\eta = 1$ is the best for ImageNet. It is noted that the parameter searching operations are performed with the searching step set to 5. Our model is implemented on TensorFlow [Abadi et al., 2016] on a server with two NVIDIA TITAN X GPUs.

### 4.2 Results and Discussions

Table 1 shows the results of different hashing methods on three benchmark datasets when the code length is 16, 32, 48, and 64 bits respectively. Fig. 3 and Fig. 4 show the Precision-Recall curves and Precision curves respectively for different methods on the code length of 32 bits.

On two multi-label datasets NUS-WIDE and MS-COCO, DSEH substantially outperforms all the compared baseline methods. Besides, almost all deep hashing methods outper-
form the traditional hashing baselines, which highlights the benefit of feature learning by deep networks that more discriminative representation can be obtained. Compared with other deep methods which utilize similarity pairs, DSEH achieves a substantial increase in average MAP at different code lengths. All the results shown in Table 1, Fig. 3 and Fig. 4 illustrate the superiority of our method. One reason may be that instead of utilizing similarity pairs information roughly, DESH exploring label information to generate semantic feature is very effective to generate more sufficient semantic information and thus produce more discriminative hash codes. Another reason is that sufficient semantic information obtained from LabNet can be retained completely and thus supervise ImgNet effectively when training ImgNet with the supervised information on the semantic level and hash codes level.

On ImageNet dataset which is annotated with single label, DHN, DPSh, and CNNH achieve under-performing results compared with the shallow baseline SDH, which demonstrates that network learning capacity can be dropped on single-label dataset because of the imbalance of pairs similarity. CNNH generates undiscriminating hash codes only under the supervision of pairwise similarity matrix. By adjusting the weight of similarity correlation, HashNet outperforms other baselines, which shows that adjusting weight can only alleviate influence of the data imbalance. The proposed DSEH significantly outperforms all other baselines. Compared with the state-of-the-art HashNet, we achieve about 34.50% increase in average MAP at different code lengths on this imbalanced dataset. It means that the proposed semantic feature learning and supervision to hashing learning can solve the issue of data imbalance in single-label dataset and thus hash codes can be generated more discriminative.

### 4.3 Empirical Analysis

Two different experiment settings are designed additionally to analyse the proposed method.

**Visualization of Semantic Features:** We visualize the semantic features generated by LabNet and ImgNet on NUS-WIDE at 32 bits in Fig. 5 (for convenience, 100 points are sampled and encapsulated by PCA [Wold et al., 1987]). We observe that the semantic features of LabNet are abundant, indicating that the semantic information of labels is effectively exploited. Furthermore, the semantic features of ImgNet are similar to those in LabNet, inferring that ImgNet is well supervised in the common semantic space.

**Ablation Study:** We investigate the variants of DSEH on the three datasets. DSEH-S denotes that ImgNet without supervision on semantic layer from LabNet. DSEH-SS refers to that both LabNet and ImgNet without semantic supervision. DSEH-L denotes that LabNet drops direct label supervision. DSEH-A refers to that LabNet and ImgNet are trained only once without alternating manner.

<table>
<thead>
<tr>
<th>Method</th>
<th>NUS-WIDE</th>
<th>ImageNet</th>
<th>MS-COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n_h$</td>
<td>$\text{map}_l$</td>
<td>$\text{map}_i$</td>
</tr>
<tr>
<td>DSEH</td>
<td>2008</td>
<td>0.9664</td>
<td>0.7312</td>
</tr>
<tr>
<td>DSEH-S</td>
<td>1959</td>
<td>0.9633</td>
<td>0.7208</td>
</tr>
<tr>
<td>DSEH-SS</td>
<td>1164</td>
<td>0.9322</td>
<td>0.7013</td>
</tr>
<tr>
<td>DSEH-L</td>
<td>1036</td>
<td>0.9607</td>
<td>0.7251</td>
</tr>
<tr>
<td>DSEH-A</td>
<td>1684</td>
<td>0.9558</td>
<td>0.7234</td>
</tr>
</tbody>
</table>

**Figure 5:** The visualization of semantic features.

**Table 2:** The results of ablation study @ 32bits of our DSEH.

5 Conclusion

In this paper, we proposed a novel deep hashing method, namely DSEH, for image retrieval, which consists of LabNet and ImgNet. The LabNet is explored to discover abundant semantic correlation and generate accurate hash codes. Meanwhile, the ImgNet is jointly constrained with the supervision information from common semantic space and common Hamming space for generating similarity-preserving yet discriminative hash codes. Extensive experiments conducted on three widely-used datasets demonstrate that our proposed method significantly outperforms many state-of-the-art hashing approaches, including both traditional and deep learning-based ones.

**Acknowledgments**

This work was supported in part by the National Natural Science Foundation of China under Grant 61572388 and Grant 61703327, and in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02 and Grant 2017ZDCXL-GY-05-04-02.
References


