Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 31(7): 4232–4247.
DOI: 10.3934/era.2023215
Received: 01 April 2023
Revised: 19 May 2023
Accepted: 21 May 2023
Published: 01 June 2023

Research article

Deep quantization network with visual-semantic alignment for zero-shot image retrieval

Huixia Liu* **and Zhihong Qin**

Department of Network Security, Henan Police College, ZhengZhou, Henan 450046, China

* **Correspondence:** Email: lhx@hnp.edu.cn.

Abstract: Approximate nearest neighbor (ANN) search has become an essential paradigm for large-scale image retrieval. Conventional ANN search requires the categories of query images to been seen in the training set. However, facing the rapid evolution of newly-emerging concepts on the web, it is too expensive to retrain the model via collecting labeled data with the new (unseen) concepts. Existing zero-shot hashing methods choose the semantic space or intermediate space as the embedding space, which ignore the inconsistency of visual space and semantic space and suffer from the hubness problem on the zero-shot image retrieval task. In this paper, we present an novel deep quantization network with visual-semantic alignment for efficient zero-shot image retrieval. Specifically, we adopt a multi-task architecture that is capable of 1) learning discriminative and polymeric image representations for facilitating the visual-semantic alignment; 2) learning discriminative semantic embeddings for knowledge transfer; and 3) learning compact binary codes for aligning the visual space and the semantic space. We compare the proposed method with several state-of-the-art methods on several benchmark datasets, and the experimental results validate the superiority of the proposed method.

Keywords: deep quantization; visual-semantic alignment; zero-shot learning; approximate nearest neighbor search

1. Introduction

Content-based image retrieval (CBIR) has been widely studied in the past decade [1]. Due to computational and memory constraints, these methods are unable to deal with large-scale data. In recent years, the large-scale of and ever-growing nature of online image data makes approximate nearest neighbor (ANN) search popular in image semantic retrieval tasks [2–5]. For ANN search, most research efforts have been devoted to developing two promising binarization solutions, such as learning to hash (L2H) [6–13] and learning to quantization (L2Q) [4,5,14–18]. By encoding real-valued images into binary codes, hashing based methods or quantization based methods can achieve efficient storage and retrieval of image data in a large-scale database.

L2H based methods mainly aim to map high-dimensional data into a low-dimensional Hamming space while preserving the data similarities or the semantic information. L2Q based methods mainly aim to approximate feature representation using a quantizer (i.e., sign funciton) [4,5,11,14] or approximate the high-dimensional data with a set of learned quantizers (i.e., different codebooks) [15–18]. Recent studies [5,16–18] indicate that L2Q based methods perform generally better than L2H methods for image semantic retrieval tasks. The reason may be that L2Q methods can control the quantization error until the statistically minimized error is arrived. Therefore, L2Q methods can generate higher quality of binary codes than L2H methods. Generally speaking, the encoding time and retrieval efficiency of quantization methods are slightly more costly than hashing methods [16].

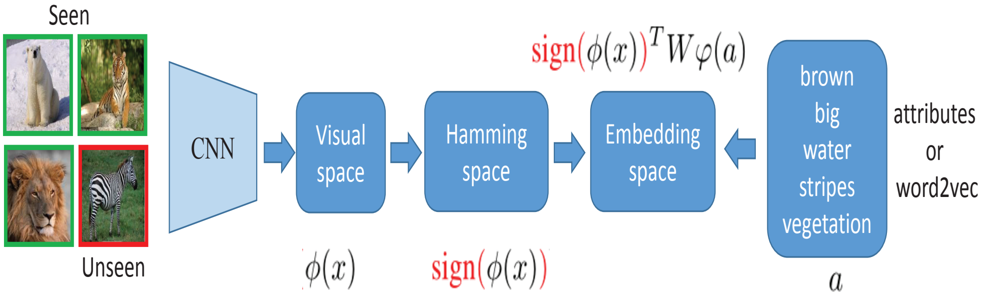

Figure 1. An illustrative diagram of the comparison of the existing zero-shot hashing framework and the proposed one.

It should be noticed that existing ANN search approaches are based on the hypothesis that the concepts of both database samples and query samples are seen at the training stage. However, the hypothesis can be violated with the explosive growth of web data because a fast growing number of images with the new semantic concepts spring up on the web. For the fast growing new concepts, it seems almost impossible to annotate sufficient training data timely, and unrealistic to retrain the model over and over again. Existing ANN search approaches yield poor retrieval performance because they tend to recognize the images of unseen categories as one of the seen categories. Therefore, the generalization ability of the model is essential for solving the retrieval problem of the unseen concepts.

To alleviate the problem mentioned above, zero-shot learning (ZSL) techniques [19–21] assume both seen classes and unseen classes share a common semantic space where all the classes reside. The shared semantic space can be characterized by attributes [22], word2vec [23] or WordNet [24]. In the zero-shot classification task, the image classes in the training set and the test set are referred to as seen classes and unseen classes respectively. During the test phase, the image from the unseen class is assigned to the nearest class embedding vector in the shared space by a simple nearest neighbor search strategy. Although ZSL techniques have achieved progress in zero-shot image classification, zero-shot image retrieval has not yet been well explored.

Recently, zero-shot learning techniques have been introduced into learning to hash to improve the generalization ability of the hashing model [25]. SitNet [25] incorporates a semantic embedding loss and a regularized center loss into a multi-task architecture to capture the semantic structure in the semantic space. To facilitate knowledge transferring and reduce the quantization error in the training process, some quantization based methods [26, 27] propose to simultaneously transfer the semantic information to binary codes and control the quantization error between low-dimensional feature representations and learned binary codes. However, a significant disadvantage of these methods is that the minimization of the quantization error in the training process is still unsatisfactory. Moreover, the inconsistency of the visual space and semantic space has not been considered sufficiently, which can increase the risk of the overfitting the seen classes and reduce the expansibility of the training model to the unseen classes [28]. Last but not least, the works in [26, 27] utilize the semantic space as the embedding space, which means projecting the visual feature vectors or hash codes into the semantic space. This will shrink the variance of the projected data points and thus result in higher hubness (i.e., the projected data points will be closer to each other on average) [20]. In turn, the hubness problem in the semantic space can decrease the semantic transfer ability of the visual feature vectors or hash codes for the zero-shot image retrieval task.

In this paper, we propose a novel deep quantization network with visual-semantic alignment (VSAQ) for efficient zero-shot image retrieval. Specifically, we design a deep quantization network architecture which consists of the following components: 1) an image feature network to generate discriminative and polymeric image representations for facilitating the visual-semantic alignment and guiding the semantic embedding more easily; 2) a semantic embedding network to maximize the compatibility score between the image and semantic vectors for knowledge transfer; 3) a quantization loss layer to control the quantization error of image representation and generate high quality of binary codes for visual-semantic alignment and alleviating the hubness problem. We compare the proposed method with several state-of-the-art methods on several benchmark datasets and the experimental results validate the superiority of the proposed method.

The remainder of this paper is organized as follows: related work is reviewed in Section 2 and we illustrate the proposed method in Section 3. Evaluation on three commonly used benchmark datasets is described in Section 4, followed by conclusions in Section 5.

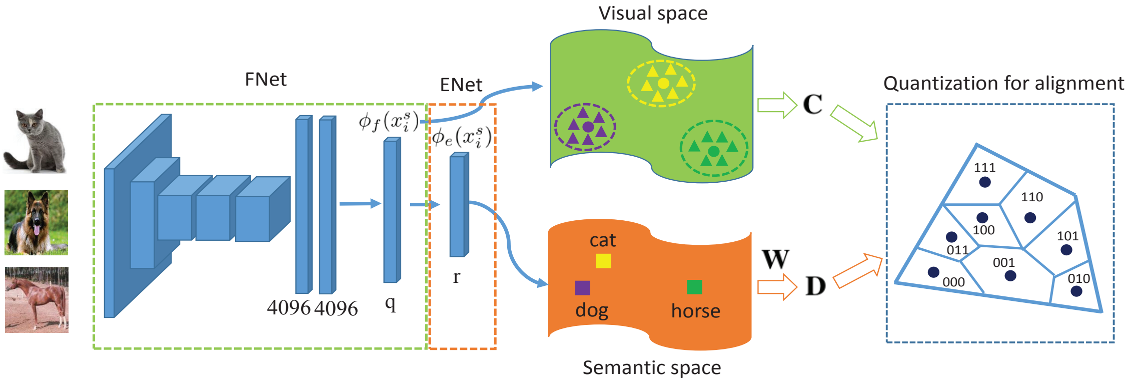

Figure 2. An overall architecture overview of the proposed VSAQ model. We use the label information to learn the image representations with discrimination and polymerization via the image feature network (FNet), and then input the image representations into the image embedding network (ENet) for improving the compatibility between the visual features and the semantic vectors. The semantic vectors are mapped to the visual space and aligned with the corresponding image representations via a collective quantization framework for alleviating the hubness problem.

2. Related work

2.1. Hashing for retrieval

Due to the ever-growing amount of image data on the internet, hashing has become a popular technique for image retrieval. Generally, we can divide existing hashing approaches into two categories: data-independent and data-dependent hashing methods. Data-independent hashing methods map the data points from the original feature space into a binary code space by using random projections as hash functions. Representative data-independent hashing methods include Locality Sensitive Hashing (LSH) [3]. These methods provide theoretical guarantees for mapping the nearby data points into the same hash codes with high probabilities. However, they need long binary codes to achieve high precision. Data-dependent hashing methods learn hash functions and compact binary codes from training data. Typical data-dependent hashing methods include spectral hashing (SH) [6], anchor graph hashing (AGH) [7], supervised hashing with kernels (KSH) [8], supervised discrete hashing (SDH) [9] and column sampling based discrete supervised hashing (COSDISH) [10]. Recently, benefiting from the power of deep convolutional networks, deep hashing methods which integrate feature learning and hash-code learning into the same end-to-end framework have been proposed to further improve the semantic retrieval performance. Typical deep hashing methods include convolutional neural network hashing deep pairwise supervised hashing (DPSH) [5], deep supervised discrete hashing (DSDH) [29], deep supervised hashing (DSH) [13], and deep hashing network (DHN) [12]. Although there has been success in semantic image retrieval, most existing hashing methods fail on zero-shot image retrieval, due to the low generalization ability of learned hashing models for unseen concepts.

2.2. Quantization for retrieval

Quantization-based methods attempt to control the quantization error of the feature representations using a quantizer (i.e., sign funciton) [4, 5, 11, 14, 30] or approximate the high-dimensional data with a set of learned quantizers (i.e., different codebooks) [15–18]. For example, [4, 5, 14] and [30] try to minimize the Euclidean distance and the cosine distance between continuous representations and their signed binary codes respectively. Alternatively, [11] utilizes a sequence of smoothing activation functions to gradually approach the sign function. Although the quantization error can be controlled using a single quantizer, it is not statistically minimized for generating high-quality binary codes. To further reduce the quantization error, [15–18] utilize the vector quantization (VQ) technique [31] to improve the accuracy and efficiency of the quantification process. Benefiting from the power of VQ, the retrieval performance has been improved significantly. However, these methods focus on traditional image retrieval (i.e., the concepts of all samples are seen in the training set), and how to integrate them into zero-shot image retrieval is still an open problem.

2.3. Zero-shot learning

Zero-shot learning recognizes unseen or novel classes that did not appear in the training stage [19–21]. The zero-shot learning framework learns a compatible visual-semantic embedding space and utilizes the learned embedding space as an intermediate to accomplish the zero-shot image classification task. The method in [20] utilizes a latent space as the visual-semantic embedding space and introduces the least square loss between the embedded visual features and the embedded semantic vectors to cope with the hubness problem. The method in [21] utilizes the semantic space as the visual-semantic embedding space and introduce an image feature structure constraint and a semantic embedding structure constraint to learn structure-preserving image features and improve the generalization ability of the learned embedding space respectively. Recently, some works [25–27] attempt to utilize the zero-shot learning for solving the zero-shot image retrieval problem. The method in [26] projects the binary codes to the semantic space with the ridge regression formulation, which can exacerbate the hubness problem. However, the quantization error is not statistically minimized and the inconsistency of the visual space and semantic space has not been considered sufficiently.

3. Deep quantization network with visual-semantic alignment for zero-shot image retrieval

3.1. Problem definition

We follow the definition of zero-shot image retrieval in [25, 26]. The training set is defined as $\mathcal{S} \equiv \{x_i^s, y_i^s, a_i^s\}_{i=1}^{n_s}$. Each image $x_i^s \in \mathcal{X}_\mathcal{S}$ is associated with a corresponding class label $y_i^s \in \mathcal{Y}_\mathcal{S}$. Similarly, the test set is defined as $\mathcal{U} \equiv \{x_j^u, y_j^u, a_j^u\}_{i=j}^{n_u}$. Each image $x_j^u \in \mathcal{X}_\mathcal{U}$ is associated with a corresponding class label $y_j^u \in \mathcal{Y}_\mathcal{U}$. The side information matrix $\mathbf{A} \in \mathbb{R}^{r\times(|\mathcal{Y}_\mathcal{S}|+|\mathcal{Y}_\mathcal{U}|)}$ is obtained from the user-defined attributes or word2vec to transfer knowledge across concepts. The side information of image x_i^s can be denoted as $a_i^s = \mathbf{A}_{y_i^s}$, which corresponds to the y_i^s-th column of $\mathbf{A}$. According to the setting of zero-shot learning, $\mathcal{Y}_\mathcal{S} \cap \mathcal{Y}_\mathcal{U} = \emptyset$, i.e., the seen classes are disjoint from the unseen classes. The goal of zero-shot hashing is to predict the binary codes of images from both seen classes and unseen classes.

3.2. Network architecture

As illustrated in Figure 2, the proposed architecture mainly consists of three different components: 1) the image feature network (FNet) for learning discriminative and polymeric image representations; 2) the embedding network (Enet) for learning an embedding space to associate the visual information with the semantic information; and 3) the quantization loss layer for controlling coding quality, aligning the visual and semantic information and alleviating the hubness problem.

3.2.1. The image feature network (FNet)

The image feature network (FNet) aims to learn the semantic image representations with discrimination and polymerization. We adopt AlexNet [32] as the base network using the layers from conv1 to fc7 and replace fc8 with a q-dimensional fully-connected layer (4096-128). In addition, the $tanh(\cdot)$ activation function and an L2 Normalization Layer are added to enhance the nonlinear representation ability and constrain the range of the output features. Inspired by [33], a variant of the softmax loss is utilized to increase the discrimination of inter-class features and the compactness of intra-class features as follows:

$$\mathcal{L}_f = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp(\gamma_1 \langle \phi_f(x_i^s), \hat{\mathbf{c}}_k \rangle)}{\sum_{j=1}^{|\mathcal{Y}_\mathcal{S}|} \exp(\gamma_1 \langle \phi_f(x_i^s), \hat{\mathbf{c}}_j \rangle)} \tag{3.1}$$

where $\hat{\mathbf{c}}_j$ denotes the centroid of the features associated with the j-th class, and γ_1 is set to 10 in all experiments. The $\phi_f(x)$ refers to the output of the FNet. Under the guidance of the label information $\mathcal{Y}_\mathcal{S}$ of the seen classes, the FNet can learn semantic-preserving image representations. In addition, the following image embedding network can learn the visual-semantic embedding space more easily with the help of such semantic-preserving image representations. Finally, it can assist the visual-semantic alignment more easily.

3.2.2. The semantic embedding network (ENet)

The embedding network (ENet) aims to learn an embedding space to associate the visual information with the semantic information. According to most of previous ZSL methods, we utilize the semantic space of $\mathcal{A}$ as the visual-semantic embedding space, i.e., projecting the outputs of FNet into the semantic space. Therefore, the ENet is constructed by an r-dimensional fully-connected layer (128-d) followed by the $tanh(\cdot)$ activation function and an L2 Normalization Layer, where r denotes the length of the semantic vectors. We use the following inner product to define the compatibility score between the visual embedding $\phi_e(x)$ and the semantic vector a^y. Similar to traditional image classification tasks, we replace the classification score with the compatibility score in the following softmax loss:

$$\mathcal{L}_e = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp(\gamma_2 \langle \phi_e(x_i^s), \hat{\mathbf{a}}_k^s \rangle)}{\sum_{j=1}^{|\mathcal{A}_S|} \exp(\gamma_2 \langle \phi_e(x_i^s), \hat{\mathbf{a}}_j^s \rangle)} \tag{3.2}$$

where $\phi_e(x_i^s)$ denotes the output of the ENet, $\hat{\mathbf{a}}_j^s$ denotes the L2-normalized side information (attribute or word2vec) associated with the j-th class and γ_2 is set to 10 in all experiments.

3.2.3. Deep quantization with visual-semantic alignment

The acquirement of the semantic information is independent of visual samples. Therefore, the class structures between the visual space and semantic space are usually inconsistent. For example, the concepts of ‘cat’ and ‘dog’ locate quite close to each other in the semantic space, while the appearance features of ‘cat’ and ‘dog’ are far away from each other in the visual space. If we only use the semantic space as the visual-semantic embedding space, the mapped visual embeddings can be collapsed to hubs [34], i.e., nearest neighbours to many other projected visual feature representation vectors. To alleviate the hubness problem, we map the semantic information to the visual space and align the projected semantic vectors with the visual features in the visual space using a collective quantization framework.

Specifically, we use a matrix $\mathbf{W} \in \mathbb{R}^{r \times q}$ to map the L2-normalized semantic vectors to the visual space. The semantic image representations $\phi_f(x_i)$ and the corresponding mapped semantic vectors $\mathbf{W}^T\hat{\mathbf{a}}_j$ are quantized using two codebooks $\mathbf{C} = [\mathbf{C}_1, \cdots, \mathbf{C}_M]$ and $\mathbf{D} = [\mathbf{D}_1, \cdots, \mathbf{D}_M]$ respectively. Each sub-codebook $\mathbf{C}_m$ (or $\mathbf{D}_m$) consists of K codewords $\mathbf{C}_m = [\mathbf{C}_{m1}, \cdots, \mathbf{C}_{mK}]$ where the k-th codeword $\mathbf{C}_{mk}$ corresponds to a q-dimensional vector. The basic idea for visual-semantic alignment is to learn two codebooks to quantize the visual features and the corresponding mapped semantic vectors into binary codes and enforce the binary codes to be the same between them. The loss function can be written as:

$$\begin{aligned} \mathcal{L}_q =& \frac{1}{n}\sum_{i=1}^{n} \|\phi_f(x_i^s) - \sum_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mi}\|^2 + \\ & \frac{1}{n}\sum_{i=1}^{n} (\|\mathbf{W}^T \hat{\mathbf{A}}_{y_i^s} - \sum_{m=1}^{M} \mathbf{D}_m \mathbf{b}_{mi}\|^2 + \lambda \|\mathbf{W}\|^2), \\ s.t.,\quad & \|\mathbf{b}_{mi}\|_0 = 1,\ \mathbf{b}_{mi} \in \{0,1\}^K, \end{aligned} \tag{3.3}$$

where $\lambda > 0$ is a balancing parameter, and $\hat{\mathbf{A}}_{y_i^s}$ is the L2-normalized semantic vector of i-th image. $\|\cdot\|_0$ refers to the ℓ_0-norm which returns the number of the vector's non-zero values. The constraint indicates that $\{\mathbf{b}_{mi}\}_{m=1}^{K}$ are the one-of-K encodings which means only one of the codeword per sub-codebook in codebooks $\mathbf{C}$ and $\mathbf{D}$ can be activated to approximate the semantic image representations $\phi_f(x)$ and the corresponding mapped semantic vectors $\mathbf{W}^T\hat{\mathbf{a}}_j$. Each one-of-$K$ encodings $\{\mathbf{b}_{mi}\}_{m=1}^{M}$ can be compressed in $\log_2 K$ bits. We can obtain compact binary codes with $B = M\log_2 K$ bits by concatenating all M compressed encodings. The one-of-K encodings $\{\mathbf{b}_{mi}\}_{m=1}^{M}$ play the key role to align the visual space and the semantic space, thus the consistency of the class structures can be guaranteed in the two spaces.

The final objective function for training the whole network is constructed by aggregating all the loss functions as follows:

$$\mathcal{L} = \mathcal{L}_f + \alpha\mathcal{L}_e + \beta\mathcal{L}_q \tag{3.4}$$

where α and β are two hyperparameters to balance the influence of different terms.

Approximate nearest neighbor search with the inner product distance is a powerful tool for quantization techniques. Given an unseen image query x_q and the binary codes of database points $\{\mathbf{b}_n = [\mathbf{b}_{1n}; \cdots; \mathbf{b}_{Mn}]\}_{n=1}^{N}$, we first use the trained image feature network to obtain the image representations. Following the asymmetric search method in [16–18], we adopt the asymmetric quantizer distance (AQD) to compute the inner-product similarity between the unseen query x_q^u and database point x_n as follows:

$$AQD(x_q^u, x_n) = \sum_{m=1}^{M} \phi_f(x_q^u)^T (\sum_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mn}) \tag{3.5}$$

where $\sum_{m=1}^{M} \mathbf{C}_m\mathbf{b}_{mn}$ is used to approximate the image representation of the database point x_n. Given an unseen query x_q^u, the inner-products between $\phi_f(x_q^u)$ and all M codebooks $\{\mathbf{C}_m\}_{m=1}^{M}$ and all K possible values of $\mathbf{b}_{mn}$ can be pre-computed and stored in a $M \times K$ lookup table. Therefore, the computation of AQD between the unseen query and all database points can be speed up. Considering computational complexity, it is slightly more costly than the Hamming distance, since M table lookups and additions are involved.

4. Learning algorithm

The optimization problem contains four sets of variables including the network parameters Θ, the centroid of the features $\widehat{\mathbf{C}} = \{\hat{\mathbf{c}}_1, \cdots, \hat{\mathbf{c}}_{|\mathcal{Y}_S|}\}$, the projection matrix $\mathbf{W}$, the codebooks $\mathbf{C}$ and $\mathbf{D}$, and the binary codes $\mathbf{B} = [\mathbf{b}_1, \cdots, \mathbf{b}_n]$. In the following optimization process, we adopt an alternating optimization strategy that updates one variable while holding fixed all other variables iteratively.

Updating Θ. We adopt the standard back-propagation algorithm with automatic differentiation techniques in Pytorch [35] to update the network parameters Θ.

Updating $\widehat{\mathbf{C}}$. We can update $\{\hat{\mathbf{c}}_i\}_{i=1}^{|\mathcal{Y}_S|}$ as follows:

$$\hat{\mathbf{c}}_j = \frac{1}{|\{y_i^s \in j\}_{i=1}^{n_s}|} \sum_{y_i^s \in j} \phi_f(x_i^s) \tag{4.1}$$

where $\{y_i^s \in j\}_{i=1}^{n_s}$ denotes the set of samples from class j.

Updating W. We can update the projection matrix **W** by optimizing the following subproblem

$$\min_{\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T \hat{\mathbf{A}}_{y_i^s} - \sum_{m=1}^{M} \mathbf{D}_m \mathbf{b}_{mi}||^2 + \lambda ||\mathbf{W}||^2. \tag{4.2}$$

We can obtain an analytic solution for this unconstrained quadratic problem as follows:

$$\mathbf{W} = (\hat{\mathbf{A}}\mathbf{Y}^{\mathcal{S}}\mathbf{Y}^{\mathcal{S}^T}\hat{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}\hat{\mathbf{A}}\mathbf{Y}^{\mathcal{S}}\mathbf{B}^T\mathbf{D}^T \tag{4.3}$$

where $\mathbf{Y}^{\mathcal{S}} = [y_1^s, \cdots, y_n^s] \in \{0,1\}^{(|\mathcal{Y}_{\mathcal{S}}|+|\mathcal{Y}_{\mathcal{U}}|)\times n}$ is the label matrix of training images with each column corresponding to a one-hot vector and **I** is an identity matrix.

Updating C. We rewrite the optimization problem w.r.t. the dictionary **C** in matrix formulation as follows:

$$\min_{\mathbf{C}} ||\Phi_f - \mathbf{C}\mathbf{B}||^2 \tag{4.4}$$

where $\Phi_f = [\phi_f(x_1^s), \cdots, \phi_f(x_n^s)]$. We can update **C** with the following analytic solution

$$\mathbf{C} = \Phi_f \mathbf{B}^T (\mathbf{B}\mathbf{B}^T)^{-1}. \tag{4.5}$$

Updating D. Similarly to the update method for **C**, we can update **D** with the following analytic solution

$$\mathbf{D} = \mathbf{W}^T \hat{\mathbf{A}} \mathbf{Y}^{\mathcal{S}} \mathbf{B}^T (\mathbf{B}\mathbf{B}^T)^{-1}. \tag{4.6}$$

Updating B. We can decompose the optimization problem for **B** into n subproblems, since $\{\mathbf{b}_i\}_{i=1}^n$ are independent of each other. For $\mathbf{b}_i$, the subproblem can be written as

$$\min_{\mathbf{b}_i} ||\phi_f(x_i^s) - \sum_{m=1}^{M} \mathbf{C}_m \mathbf{b}_{mi}||^2 + ||\mathbf{W}^T \hat{\mathbf{A}}_{y_i^s} - \sum_{m=1}^{M} \mathbf{D}_m \mathbf{b}_{mi}||^2 \tag{4.7}$$

which can be further simplified as

$$\min_{\mathbf{b}_i} || \begin{bmatrix} \phi_f(x_i^s) \\ \mathbf{W}^T \hat{\mathbf{A}}_{y_i^s} \end{bmatrix} - \sum_{m=1}^{M} \begin{bmatrix} \mathbf{C}_m \\ \mathbf{D}_m \end{bmatrix} \mathbf{b}_{mi} ||^2. \tag{4.8}$$

Generally, the above optimization problem is NP-hard. We adopt the iterated conditional modes (ICM) algorithm [36] to solve M indicators $\{\mathbf{b}_{mi}\}_{m=1}^M$ alternatively. Specifically, fixing $\{\mathbf{b}_{m'i}\}_{m'\neq m}$, we check all the elements in $\begin{bmatrix} \mathbf{C}_m \\ \mathbf{D}_m \end{bmatrix}$ exhaustively and find the element such that the obective function is minimized. Then, the corresponding entry of $\mathbf{b}_{mi}$ is updated to 1 and the rest is updated to 0. The ICM algorithm is guaranteed to converge until the maximum iterations reached. The algorithm is summarized in Algorithm 1.

Algorithm 1 VSAQ algorithm

```
Input:
    Training set $\mathcal{S} \equiv \{x_i^s, y_i^s, a_i^s\}_{i=1}^{n_s}$;
Output:
    Parameter $\Theta$ of the deep neural networks.
Initialization:
    Initialize network parameter $\Theta$, mini-batch size $M$, the iteration number $T$;
for $epoch = 1, 2, \ldots, T$ do
    Update $\mathbf{W}$ according to Eq (4.2);
    Update $\mathbf{C}$ according to Eq (4.5);
    Update $\mathbf{D}$ according to Eq (4.6);
    Update $\mathbf{B}$ according to Eq (4.8);
    Update the parameter $\Theta$ by using backpropagation;
end for
```

5. Experiments

We evaluate and compare the proposed method with state-of-the-art baselines on several benchmark datasets. The proposed method is implemented with the open-source deep learning toolbox Pytorch [35]. All the experiments are carried out on a server with an Intel(R) Xeon(R) E5-2620 v4@2.10GHz CPU, 128GB RAM and two GeForce TITAN X GPUs with 24GB memory.

5.1. Datasets

Three widely used datasets including Animals with Attributes [37], CIFAR10 [32] and ImageNet [38] are adopted to evaluate the proposed method and other baselines.

Animals with Attributes: contains 30,475 images from 50 animal categories. Each class is provided with 85 semantic attributes.

CIFAR-10: consists of 60,000 color images. The image size is 32 × 32 pixels. Each image is associated with one of the ten classes with each class containing 6000 images.

ImageNet: consists of 1.2 million images labeled with 1000 categories/synsets for the Large Scale Visual Recognition Challenge 2012 (ILSVRC2012).

5.2. Experimental settings

Following the settings in [25, 26], we construct the zero-shot scenario by splitting the benchmark datasets into seen classes and unseen classes. Specifically, for the Animals with Attributes (AwA) dataset, we randomly split the 50 animal categories into five groups with each group containing ten categories. In turn, we use one group as the unseen classes and the remaining groups as the seen classes. Therefore, we can obtain 5 different seen-unseen splits. We utilize 85-dim attribute vectors as the semantic vector. For the CIFAR10 dataset, we use one category as the unseen class and the remaining categories as the seen classes. Consequently, we can obtain 10 different seen-unseen splits. The 300-dimensional semantic vector is extracted from class names using the word2vec tool. For the ImageNet dataset, we randomly select a subset of ImageNet with 100 categories, which gives us about

130,000 images for evaluation. The 100 selected categories have the semantic vector from word2vec. We use 10 categories as seen classes and the remaining 90 categories as unseen classes, and thus we can obtain 10 different seen-unseen splits. Similar to CIFAR10, we use the word2vec tool to extract 300-dimensional semantic vectors from class names. For all three datasets, we randomly take 1000 images from the unseen categories as the query set. The remaining images from the remaining unseen categories images and all the seen categories images are treated as the retrieval database. For training, we randomly select 10,000 images from the seen categories as the training set.

We use the widely used mean Average Precision (mAP) based on Hamming ranking as the evaluation metric. The final experimental results are averaged over the different seen-unseen splits for all datasets.

5.3. Baselines

We compare the proposed method with the following state-of-the-art hashing methods. These methods fall into two categories: 1) Hashing methods for traditional image retrieval: Iterative Quantization (ITQ) [4], supervised discrete hashing (SDH) [9], deep pairwise supervised hashing (DPSH) [5], deep supervised discrete hashing (DSDH) [29]; 2) zero-shot hashing methods: zero-shot hashing via transferring supervised knowledge (TSK) [26] and zero-shot hashing with discrete similarity transfer network (SitNet) [25]. We implement SitNet with Pytorch by ourselves. For the other compared methods, we adopt the public codes and suggested parameters from the their papers. For the non-CNN hashing methods, we adopt the pre-trained AlexNet model for extracting the 4096-dimensional CNN features as image representations for fair comparison.

5.4. Implementation details

We implement the VSAQ model via Pytorch. For the Animals with Attributes and CIFAR-10 datasets, the initial learning rate was set to 0.001. For the ImageNet dataset, the initial learning rate was set to 0.01. As the last fully connected layers in FNet and ENet are training from scratch, the learning rates of these layers are set to 10 times the other layers. We set the batch size to 128 and train the model for 10 epochs. The dimension of image representations q is set to 128 following [17]. The hyperparameters are set as $\alpha = 1, \beta = 10, \lambda = 0.01$ across all the following experiments.

5.5. Experimental results

5.5.1. Results on AwA

The zero-shot image retrieval performances on AwA in terms of MAP with respect to different code lengths (i.e., $\{8, 16, 32, 48\}$) are shown in Table 1. We find that our VSAQ method outperforms all other baseline methods by a large margin in terms of MAP, especially from 8 to 32 bits. In addition, we find that the unsupervised hashing method ITQ achieves comparable results with some supervised hashing method SDH. This demonstrates that the generalization ability of existing supervised hashing is limited for unseen concepts. The existing state-of-the-art deep hashing methods, including DPSH and DSDH, perform poorly on the zero-shot retrieval task over the AwA dataset. The main reason can be that the trained CNN compatible with the label information can fall into the risk of overfitting the seen classes, which reduces the expansibility of the training model to the unseen classes. We also find that TSK performs worse, especially at the lower bits (e.g., 8 and 16 bits). The main reason is that the hubness problem is exacerbated by projecting the binary codes to the semantic space with the ridge regression formulation [20], which will decrease the semantic transfer ability of hash codes in turn. To alleviate such a problem, the proposed VSAQ model utilizes the visual space as the embedding space for learning compact binary codes. In addition, we adopt a collective quantization technique for visual-semantic alignment which can improve the generalization ability of the proposed model.

Table 1. The comparisons of mAP on zero-shot image retrieval over AwA dataset from 12 to 48 bits.

Method	AwA 8 bits	16 bits	32 bits	48 bits
ITQ	0.0886	0.1359	0.1723	0.2024
SDH	0.0966	0.1370	0.1835	0.2122
DPSH	0.0726	0.1080	0.1435	0.1525
DSDH	0.0808	0.1081	0.1320	0.1469
TSK	0.0349	0.0591	0.1320	0.1617
SitNet	0.1036	0.1651	0.1870	0.2121
VSAQ	**0.1948**	**0.2099**	**0.2187**	**0.2218**

5.5.2. Results on CIFAR-10

The performances of the proposed VSAQ and other baselines on CIFAR-10 with different code length are illustrated in Table 2. From Table 2, we can find that VSAQ consistently outperforms other baselines at all bits by a large margin. For example, VSAQ surpasses SitNet with the second best performance by 3 to 4 percent. Even though the code length is short, VSAQ still achieves superior retrieval performance compared to the baselines with longer code length. It can be attributed to the lower quantization error controlled by the quantization technique. The deep hashing methods DPSH and DSDH perform better than the non-deep hashing methods ITQ and SDH, which demonstrates that CNNs can utilize the proper supervision to discover the complicated semantic similarity structure. VSAQ utilizes the label information to learn the semantic image representations with discriminative and polymeric structure, which can assist the visual-semantic alignment more easily. The unsupervised hashing mehtod ITQ achieves comparable performance with TSK which demonstrates that the generalization ability degenerates due to the existing hubness problem.

5.5.3. Results on ImageNet

The performances of the proposed VSAQ and other baselines on ImageNet with different code length are demonstrated in Table 3. As we can see, the proposed VSAQ model outperforms the baseline approaches by significant margins. For example, VSAQ surpasses SitNet with the second best performance by 2 to 9 percent. It clearly demonstrates that the VSAQ model generalizes better for unseen concepts compared with other state-of-the-art methods, which validates the effectiveness of the proposed method for zero-shot image retrieval.

Table 2. The comparisons of mAP on zero-shot image retrieval over CIFAR-10 dataset from 12 to 48 bits.

Method	CIFAR-10 8 bits	16 bits	32 bits	48 bits
ITQ	0.1507	0.1736	0.1871	0.1972
SDH	0.1226	0.1331	0.1553	0.2068
DPSH	0.2176	0.2205	0.2280	0.2261
DSDH	-	-	-	-
TSK	0.1507	0.1759	0.1740	0.2132
SitNet	0.2208	0.2303	0.2351	0.2471
VSAQ	**0.2615**	**0.2682**	**0.2670**	**0.2867**

Table 3. The comparisons of mAP on zero-shot image retrieval over ImageNet dataset from 12 to 48 bits.

Method	ImageNet 8 bits	16 bits	32 bits	48 bits
ITQ	0.0507	0.0732	0.1123	0.1357
SDH	0.0400	0.0727	0.1107	0.1312
DPSH	0.0409	0.0524	0.0712	0.0881
DSDH	-	-	-	-
TSK	0.0162	0.0206	0.0247	0.0609
SitNet	-	-	-	-
VSAQ	**0.1472**	**0.1516**	**0.1579**	**0.1614**

5.6. Effectiveness of the proposed framework

The proposed VSAQ model consists of three components: an image feature loss layer $\mathcal{L}_f$ for learning discriminative and polymeric image representations, a semantic embedding loss layer $\mathcal{L}_e$ for maximizing the compatibility score between the image and semantic vectors for knowledge transfer, and a quantization loss layer $\mathcal{L}_q$ for visual-semantic alignment. The quantization loss layer $\mathcal{L}_q$ is an essential part of generating binary codes. To study the contribution of different components for the zero-shot image retrieval performance, we compare the proposed method with the following submodels: 1) $\mathcal{L}_f + \mathcal{L}_q$ (VSQA-1); 2) $\mathcal{L}_e + \mathcal{L}_q$ (VSQA-2); 3) $\mathcal{L}_f + \mathcal{L}_e + \mathcal{L}_q^1$ (VSQA-3), where $\mathcal{L}_q^1$ refers to the first term in Eq (3.3), i.e., only considering the visual features for quantization. Table 4 illustrates the experimental results of different submodels. From Table 4, we can see that the combination of the image feature loss, the semantic embedding loss and the quantization loss achieves the best performance. The results demonstrate that the proposed framework improves the zero-shot image retrieval performance indeed. Comparing the performance of VSQA-3 and VSQA, we can find that the visual-semantic alignment will help the knowledge transfer from the seen concepts to the unseen concepts. Though the comparisons of VSQA-2 and VSQA, we can find that the discriminative and polymeric image representations will improve the performance a lot, which means that it can assist the visual-semantic alignment and semantic embedding more easily. Comparing the performance of VSQA-1 and VSQA, we can find that the knowledge transfer ability can be significantly improved by the semantic embedding.

Table 4. The impact of different submodels of our VSAQ on mAP for AwA, CIFAR-10 and ImageNet datasets from 12 to 48 bits.

Method	AwA				CIFAR-10				ImageNet			
	12 bits	24 bits	32 bits	48 bits	12 bits	24 bits	32 bits	48 bits	12 bits	24 bits	32 bits	48 bits
VSAQ	**0.1948**	**0.2099**	**0.2187**	**0.2218**	**0.2615**	**0.2682**	**0.2670**	**0.2867**	**0.1472**	**0.1516**	**0.1579**	**0.1614**
VSAQ-1	0.1830	0.1849	0.1923	0.2012	0.2360	0.2487	0.2538	0.2539	0.1288	0.1319	0.1386	0.1497
VSAQ-2	0.1911	0.1956	0.2026	0.2089	0.2412	0.2533	0.2613	0.2665	0.1296	0.1365	0.1463	0.1495
VSAQ-3	0.1816	0.1736	0.1825	0.1998	0.2278	0.2324	0.2405	0.2487	0.1053	0.1150	0.1194	0.1256

6. Conclusions

In this paper, we propose a novel deep quantization network with visual-semantic alignment for efficient zero-shot image retrieval. In the proposed deep architecture, we use the label information and the sematic vector to supervise the image feature extraction and improve the compatibility between the image representations and the semantic vectors respectively. The semantic vectors are mapped to the visual space and aligned with the corresponding image representations via a collective quantization framework for alleviating the hubness problem. The experimental results on three datasets show that the proposed model outperforms the state-of-the-art methods on zero-shot image retrieval tasks. In the future work, we will investigate the zero-shot multi-label image (i.e., an image is assigned with multiple categories) retrieval task.

Acknowledgments

I would like to thank all anonymous reviewers for their constructive comments through each stage of the process.

Conflict of interest

All authors declare that they have no conflicts of interest.

References

1. W. Zhou, H. Li, Q. Tian, Recent advance in content-based image retrieval: a literature survey, preprint, arXiv:1706.06064.
2. J. H. Friedman, J. L. Bentley, R. A. Finkel, An algorithm for finding best matches in logarithmic expected time, *ACM Trans. Math. Software*, **3** (1977), 209–226. https://doi.org/10.1145/355744.355745
3. A. Gionis, P. Indyk, R. Motwani, Similarity search in high dimensions via hashing, in *International Conference on Very Large Data Bases*, **99** (1999), 518–529. Available from: https://www.cs.princeton.edu/courses/archive/spring13/cos598C/Gionis.pdf.

4. Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin, Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.*, **35** (2012), 2916–2929. https://doi.org/10.1109/TPAMI.2012.193

5. W. J. Li, S. Wang, W. C. Kang, Feature learning based deep supervised hashing with pairwise labels, preprint, arXiv:1511.03855.

6. Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in *Advances in Neural Information Processing Systems*, **21** (2008), 1753–1760. Available from: https://proceedings.neurips.cc/paper_files/paper/2008/file/d58072be2820e8682c0a27c0518e805e-Paper.pdf.

7. W. Liu, J. Wang, S. Kumar, S. F. Chang, Hashing with graphs, in *Proceedings of the 28 th International Conference on Machine Learning*, (2011), 1–8. Available from: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/37599.pdf.

8. W. Liu, J. Wang, R. Ji, Y. G. Jiang, S. F. Chang, Supervised hashing with kernels, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (2012), 2074–2081. https://doi.org/10.1109/CVPR.2012.6247912

9. F. Shen, C. Shen, W. Liu, H. T. Shen, Supervised discrete hashing, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 37–45.

10. W. C. Kang, W. J. Li, Z. H. Zhou, Column sampling based discrete supervised hashing, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **30** (2016), 1230–1236. https://doi.org/10.1609/aaai.v30i1.10176

11. Z. Cao, M. Long, J. Wang, P. S. Yu, Hashnet: deep learning to hash by continuation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 5608–5617.

12. H. Zhu, M. Long, J. Wang, Y. Cao, Deep hashing network for efficient similarity retrieval, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **30** (2016), 2415–2421. https://doi.org/10.1609/aaai.v30i1.10235

13. H. Liu, R. Wang, S. Shan, X. Chen, Deep supervised hashing for fast image retrieval, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2064–2072.

14. G. Irie, H. Arai, Y. Taniguchi, Alternating co-quantization for cross-modal hashing, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2015), 1886–1894.

15. M. Long, Y. Cao, J. Wang, P. S. Yu, Composite correlation quantization for efficient multimodal retrieval, in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, (2016), 579–588. https://doi.org/10.1145/2911451.2911493

16. Y. Cao, M. Long, J. Wang, S. Liu, Deep visual-semantic quantization for efficient image retrieval, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1328–1337.

17. Y. Cao, M. Long, J. Wang, S. Liu, Collective deep quantization for efficient cross-modal retrieval, in *Thirty-First AAAI Conference on Artificial Intelligence*, **31** (2017), 3974–3980. https://doi.org/10.1609/aaai.v31i1.11218

18. E. Yang, C. Deng, C. Li, W. Liu, J. Li, D. Tao, Shared predictive cross-modal deep quantization, *IEEE Trans. Neural Networks Learn. Syst.*, **29** (2018), 5292–5303. https://doi.org/10.1109/TNNLS.2018.2793863

19. Y. Fu, T. Xiang, Y. Jiang, X. Xue, L. Sigal, S. Gong, Recent advances in zero-shot recognition: toward data-efficient understanding of visual content, *IEEE Signal Process Mag.*, **35** (2017), 112–125. https://doi.org/10.1109/MSP.2017.2763441

20. L. Zhang, T. Xiang, S. Gong, Learning a deep embedding model for zero-shot learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2021–2030.

21. Y. Li, Z. Jia, J. Zhang, K. Huang, T. Tan, Deep semantic structural constraints for zero-shot learning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **32** (2018), 7049–7056. https://doi.org/10.1609/aaai.v32i1.12244

22. A. Farhadi, I. Endres, D. Hoiem, D. A. Forsyth, Describing objects by their attributes, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 1778–1785. https://doi.org/10.1109/CVPR.2009.5206772

23. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, preprint, arXiv:1301.3781.

24. G. A. Miller, Wordnet: a lexical database for English, *Commun. ACM*, **38** (1995), 39–41. https://doi.org/10.1145/219717.219748

25. Y. Guo, G. Ding, J. Han, Y. Gao, Sitnet: discrete similarity transfer network for zero-shot hashing, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, (2017), 1767–1773. Available from: https://www.ijcai.org/proceedings/2017/0245.pdf.

26. Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, H. T. Shen, Zero-shot hashing via transferring supervised knowledge, in *Proceedings of the 24th ACM International Conference on Multimedia*, (2016), 1286–1295. https://doi.org/10.1145/2964284.2964319

27. Y. Xu, Y. Yang, F. Shen, X. Xu, Y. Zhou, H. T. Shen, Attribute hashing for zero-shot image retrieval, in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, (2017), 133–138. https://doi.org/10.1109/ICME.2017.8019425

28. H. Jiang, R. Wang, S. Shan, X. Chen, Learning class prototypes via structure alignment for zero-shot recognition, in *Computer Vision – ECCV 2018*, (2018), 121–138. https://doi.org/10.1007/978-3-030-01249-6_8

29. Q. Li, Z. Sun, R. He, T. Tan, Deep supervised discrete hashing, in *Advances in Neural Information Processing Systems*, **30** (2017), 2479–2488. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/e94f63f579e05cb49c05c2d050ead9c0-Paper.pdf.

30. Y. Cao, M. Long, J. Wang, Correlation hashing network for efficient cross-modal retrieval, preprint, arXiv:1602.06697.

31. T. Ge, K. He, Q. Ke, J. Sun, Optimized product quantization, *IEEE Trans. Pattern Anal. Mach. Intell.*, **36** (2013), 744–755. https://doi.org/10.1109/TPAMI.2013.240

32. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Commun. ACM*, (2017), 84–90. https://doi.org/10.1145/3065386

33. Y. Liu, H. Li, X. Wang, Rethinking feature discrimination and polymerization for large-scale recognition, preprint, arXiv:1710.00870.

34. A. Lazaridou, G. Dinu, M. Baroni, Hubness and pollution: delving into cross-space mapping for zero-shot learning, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, **1** (2015), 270–280. https://doi.org/10.3115/v1/P15-1027

35. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, et al., Automatic differentiation in pytorch, 2017. Available from: https://openreview.net/forum?id=BJJsrmfCZ.

36. J. Besag, On the statistical analysis of dirty pictures, *J. R. Stat. Soc.*, **48** (1986), 48–259. https://doi.org/10.1111/j.2517-6161.1986.tb01412.x

37. C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, *IEEE Trans. Pattern Anal. Mach. Intell.*, **36** (2013), 453–465. https://doi.org/10.1109/TPAMI.2013.140

38. J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: a large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848