Pattern Recognition Letters 145 (2021) 247–253

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

CHOP: An orthogonal hashing method for zero-shot cross-modal retrieval☆

Xu Yuan[a,b], Guangze Wang[a], Zhikui Chen[a,b], Fangming Zhong[a,*]

[a] School of Software, Dalian University of Technology, Dalian, China
[b] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, China

ARTICLE INFO

Article history:
Received 13 March 2020
Revised 1 February 2021
Accepted 23 February 2021
Available online 6 March 2021

Keywords:
Zero-shot
Cross-modal retrieval
Orthogonal projection

ABSTRACT

Cross-modal retrieval has recently attracted much attention because it helps users retrieve data across different modalities. However, with the explosive growth of data, a large number of new emerging concepts (unseen classes) that have not been appeared in the training data (seen classes) bring great challenges to the traditional cross-modal retrieval. Nevertheless, most existing approaches mainly focus on improving cross-modal retrieval performance of seen classes, which may fail in the unseen classes. To address the challenge of zero-shot cross-modal retrieval, we propose an orthogonal method in this paper, i.e., Cross-modal Hashing with Orthogonal Projection (CHOP). It projects cross-modal features and class attributes onto a Hamming space, where each projection of cross-modal features is orthogonal to the mismatched class attributes. By so doing, the model can learn a discriminative and binary representation of each modality. In addition, the class attributes build a bridge to transfer knowledge from seen classes to unseen classes. Furthermore, the orthogonal constraint on binary codes can help to mitigate the hubness problem. Extensive experiments on three benchmark datasets show that the proposed CHOP is effective in handling zero-shot cross-modal retrieval.

1. Introduction

With the explosion of multi-modal data on the Internet, the task of information retrieval has become a challenging problem, especially for retrieval from another modality. Due to the cross-modal data (such as images and texts) have significantly different statistical properties (heterogeneous gap) [1], it is impossible to measure the similarity in a direct way. To this end, cross-modal retrieval has been widely investigated recently whose main idea is learning a common representation for various modalities [2–4]. Due to the impressive performance on low storage cost and high retrieval speed, the cross-modal approaches based on hashing have drawn considerable attention [5–8].

It is a remarkable fact that most of the existing cross-modal retrieval approaches work in a closed circumstance. In other words, the testing instances belong to the classes that are used in the training stage. However, with the explosion of data, the new emerging concepts (unseen classes) which have not been appeared in the training data (seen classes) bring great challenges to the traditional cross-modal retrieval. It is infeasible to collect and label sufficient data to retrain the existing cross-modal retrieval models. Therefore, it is necessary to train a model with zero information of unseen classes but it still can handle the cross-modal retrieval for unseen classes. The zero-shot cross-modal retrieval which is conceptually similar to zero-shot learning [9–12] has become an emerging research topic recently.

Over the past few years, only a few works have been proposed to address zero-shot cross-modal retrieval [13–15]. Most of them use the class attributes to build a semantic space. Because the attributes describe some properties of the object, such as color, texture, shape, and geographic information, which can be shared between seen and unseen classes. For instance, Zhong et al. [16] proposed to build a multi-layers network structure to connect the hash codes, data features, class attributes, and class labels. The multi-layers network also builds a bridge for knowledge transfer from seen classes to unseen classes by using the class attributes. However, most of the existing approaches ignore the hubness problem, that is, in the high dimensional space, a few unseen class prototypes will become the nearest neighbors of many points [17,18]. The hubness problem will degrade the performance of cross-modal retrieval. In summary, the zero-shot cross-modal retrieval is still an open problem which is worth further studying.

In this paper, we propose a novel method named Cross-modal Hashing with Orthogonal Projection (CHOP), which aims to tackle

☆ Editor: Jiwen Lu

* Corresponding author.
E-mail address: fmzhong@dlut.edu.cn (F. Zhong).

https://doi.org/10.1016/j.patrec.2021.02.016
0167-8655/

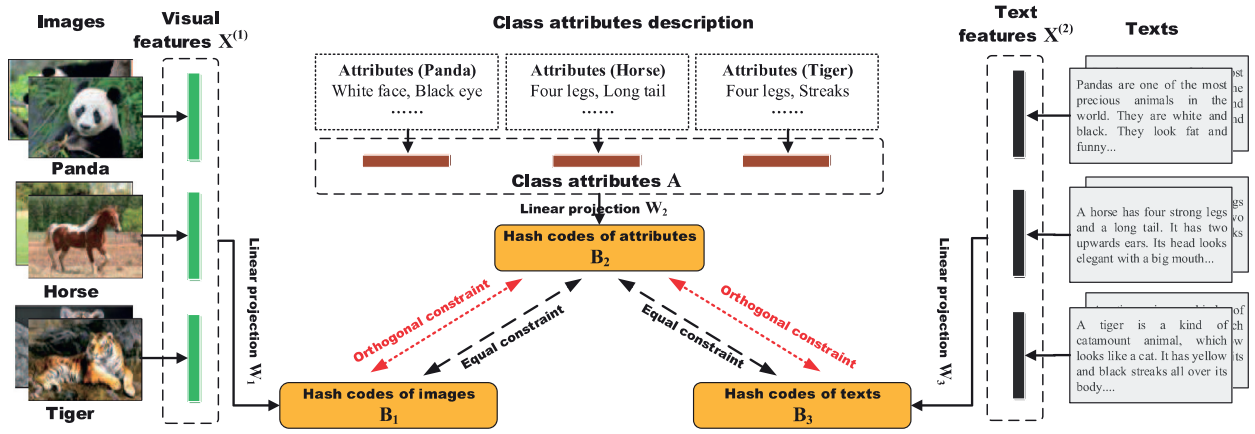

Fig. 1. The framework of the proposed method. $\mathbf{X}^{(1)}$ is the visual features extracted from image modality, $\mathbf{X}^{(2)}$ is the text features of corresponding texts, and $\mathbf{A}$ represents the corresponding class attributes. $\mathbf{X}^{(1)}$, $\mathbf{A}$, and $\mathbf{X}^{(2)}$ are projected respectively to hash codes $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{B}_3$ using linear projections $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{W}_3$. Meanwhile, the orthogonal and equal constraints are imposed on $\mathbf{B}_1$ and $\mathbf{B}_2$, which is similar between $\mathbf{B}_2$ and $\mathbf{B}_3$. Zero-shot cross-modal retrieval is performed on the hash codes through well learned hash functions $\mathbf{W}_1$ and $\mathbf{W}_3$, respectively.

the problem of zero-shot cross-modal retrieval. Motivated by zero-shot hashing and the recently proposed methods [16,19], our CHOP builds a connection between data features and class attributes by imposing constraints on the generated hash codes, by doing this to conduct the knowledge transfer from seen classes to unseen classes. As shown in Fig. 1, our approach projects image modality, text modality, and their corresponding class attributes into a unified Hamming space where the hash codes should be similar if they share the same class labels. Simultaneously, we impose orthogonal constraint on the hash codes. Thus, the hash codes of text modality and image modality should be orthogonal to the hash codes generated by mismatched class attributes, respectively. This makes different categories have the same Hamming distance in the semantic space, so as to alleviate the hubness problem. We validate our method on three non-overlapping cross-modal datasets, and experimental results demonstrate the effectiveness of the proposed CHOP in dealing with zero-shot cross-modal retrieval.

- A new hashing method based on orthogonal projection is proposed to address the problem of zero-shot cross-modal retrieval.
- Different from most existing methods, we connect the seen and unseen classes by using orthogonal-constrained and unified hash codes, which can improve the discriminative property of hash codes, and also can mitigate the hubness problem.
- Extensive experiments of zero-shot cross-modal retrieval are conducted on Wiki, Pascal VOC, and LabelMe datasets, and the results indicate that the proposed CHOP achieves better performance on the cross-modal retrieval for data from unseen classes.

The remainder of this paper is organized as follows. The related work on cross-modal retrieval, zero-shot hashing, and zero-shot cross-modal retrieval are introduced in Section 2. Section 3 describes our proposed approach with formulation and optimization. In Section 4, experimental results are presented on three datasets. Finally, Section 5 summarizes our work.

2. Related work

Since our research mainly focuses on the cross-modal retrieval for data from unseen classes, in this section, we review the related work about cross-modal retrieval, zero-shot hashing, and zero-shot cross-modal retrieval.

2.1. Cross-modal retrieval

Cross-modal retrieval aims to find a common semantic space to compute similarity, whose key challenge is to build a bridge to cross the heterogeneous gap between different modalities.

Most of the existing methods can be divided into unsupervised methods and supervised methods. The unsupervised methods learn the common semantic space by investigating the data structure and distribution information. One of the typical unsupervised methods is Collective Matrix Factorization Hashing (CMFH) [20], which uses collective matrix factorization to obtain unified hash codes for each modality. Semantic Topic Multimodal Hashing (STMH) [21] directly learns discrete hash codes in the encoding process to maintain the discrete nature of hashing. While the supervised ones try to utilize the label information. This usually brings better performance than the unsupervised cross-modal hashing methods. For instance, Liu et al. [22] extended CMFH to a supervised manner based on a graph regularization term. The Self-Supervised Adversarial Hashing (SSAH) [23] utilizes two adversarial networks and a self-supervised semantic network to maximize the semantic correlation and explore high-level semantic information in multi-label annotations.

Recently, many cross-modal hashing approaches based on deep learning have been proposed [7,24,25]. For example, Deep Cross-Modal Hashing (DCMH) [26] learns hashing function for each modality by building an end-to-end deep neural network framework.

The existing cross-modal hashing approaches have achieved great progress on cross-modal retrieval for data from seen classes. However, they can easily fail in the data from unseen classes.

2.2. Zero-shot hashing

In recent years, zero-shot learning has attracted increasing interests due to the newly-emerging concepts [27–29]. However, most of the previous zero-shot learning approaches mainly attempt to address the recognition problem. Motivated by zero-shot learning, zero-shot hashing pays more attention to deal with the retrieval problem. Zero-shot hashing via Transferring Supervised Knowledge (TSK) [30] is one of the early zero-shot hashing works, which takes advantage of semantic embedding and projects binary codes into an embedded space, and it constrains the semantic information of binary codes to retain its original semantic. Hashing with Orthogonal Projection (HOP) [19] adds orthogonal constraint

to avoid the disadvantage of max-margin loss and to make binary codes more discriminative. However, these aforementioned methods can only deal with the single-modal retrieval task.

2.3. Zero-shot cross-modal retrieval

The zero-shot cross-modal retrieval that can handle the cross-modal retrieval for data from both seen and unseen classes, has attracted increasing interests in these years. In [13], the method uses a common semantic representation generated by category weight vectors and modal features to perform cross-modal retrieval of unseen classes. Cross-Modal Attribute Hashing (CMAH) [16] constructs a multi-layers network structure to connect the hash codes, data features, class attributes, and class labels. CMAH transfers knowledge from seen classes to unseen classes via the class attributes.

Nevertheless, these methods only consider the preservation of local structure information in each modality, ignoring the discriminative ability of the generated hash codes, which brings the hubness problem.

3. Approach

In this section, we will introduce the proposed CHOP in details.

3.1. Problem definition

Let $\mathbf{X}^{(1)} = \{x_1^{(1)}, \ldots, x_n^{(1)}\}$ and $\mathbf{X}^{(2)} = \{x_1^{(2)}, \ldots, x_n^{(2)}\}$ be n pairs of seen cross-modal data. $\mathbf{X}^{(1)}$ is image modality and $\mathbf{X}^{(2)}$ represents the text modality, where $\mathbf{X}^{(1)} \in \Re^{n\times d_1}$, $\mathbf{X}^{(2)} \in \Re^{n\times d_2}$, d_1 denotes the dimension of image feature, and d_2 represents the dimension of text feature, and usually $d_1 \neq d_2$. Let $\mathbf{A} = \{a_1, \ldots, a_n\}$ be the class attributes corresponding to each image-text pair, where $\mathbf{A} \in \Re^{n\times d_A}$, and d_A denotes the dimension of attribute. In addition, we have the label information $\mathbf{Y} = \{y_1, \ldots, y_s\}$ for seen classes, where s represents the number of seen classes. It is worth noting that the class attributes $\mathbf{A}$ are determined by class labels, and they have the one-to-one relationship. Similarly, the label of unseen classes is stated as $\mathbf{Z} = \{z_1, \ldots, z_u\}$, where u represents the number of unseen classes. For the zero-shot setting, we have $\mathbf{Y} \cap \mathbf{Z} = \emptyset$. The objective of our method is to learn a hash function for each modality $\mathcal{H}^t(x) : x^t \to \{-1, +1\}^\ell$, $t = 1, 2$ from seen classes, which can be generalized to unseen classes, where ℓ is the length of hash codes.

3.2. Cross-modal zero-shot orthogonal projection hashing formulation

So as to obtain the hash codes for each modality, we use the linear projection as hashing functions. In addition, class attributes are imposed to construct the connection between seen classes and unseen classes. Therefore, we also project the class attributes onto the same Hamming space, which can be formulated as,

$$\begin{aligned} &\mathcal{L}_1 = \|\mathbf{X}^{(1)}\mathbf{W}_1 - \mathbf{B}_1\|_F^2 + \|\mathbf{X}^{(2)}\mathbf{W}_3 - \mathbf{B}_3\|_F^2 + \|\mathbf{A}\mathbf{W}_2 - \mathbf{B}_2\|_F^2 \\ &s.t. \quad \mathbf{B}_1,\ \mathbf{B}_2,\ \mathbf{B}_3 \in \{-1,\ +1\}^{n\times\ell}, \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3$ are the binary codes of $\mathbf{X}^{(1)}$, $\mathbf{A}$, and $\mathbf{X}^{(2)}$, respectively. $\mathbf{W}_1 \in \Re^{d_1\times\ell}$, $\mathbf{W}_2 \in \Re^{d_A\times\ell}$, $\mathbf{W}_3 \in \Re^{d_2\times\ell}$ correspond to the projection matrices of image modality, attributes, and text modality.

Motivated by HOP [19], here we impose the orthogonal constraint on the generated hash codes. Since an image or a text and its corresponding class attributes share the same label, the generated binary codes of them should be the same. On the contrary, the hash codes of image or text and its mismatched class attributes should be orthogonal. This can be formulated as follows,

$$\mathcal{L}_2 = \alpha\left\|\mathbf{X}^{(1)}\mathbf{W}_1(\mathbf{A}\mathbf{W}_2)^T - \ell\mathbf{S}\right\|_F^2 + \beta\left\|\mathbf{X}^{(2)}\mathbf{W}_3(\mathbf{A}\mathbf{W}_2)^T - \ell\mathbf{S}\right\|_F^2, \tag{2}$$

where $\mathbf{S}$ represents the similarity matrix. When the image $x_i^{(1)}$ or text $x_i^{(2)}$ and class attributes a_j share the same label, $s_{ij} = 1$, otherwise $s_{ij} = 0$.

In addition, the generated codes of images and texts should be same to the corresponding class attributes, i.e., $\mathbf{B_1} = \mathbf{B_2} = \mathbf{B_3}$. Then we use two Lagrange multipliers to represent the equal constraint, which can be stated as follows,

$$\begin{aligned} &\mathcal{L}_3 = \mu\|\mathbf{B_1} - \mathbf{B_2}\|_F^2 + \theta\|\mathbf{B_3} - \mathbf{B_2}\|_F^2 \\ &s.t. \quad \mathbf{B}_1,\ \mathbf{B}_2,\ \mathbf{B}_3 \in \{-1,\ +1\}^{n\times\ell}. \end{aligned} \tag{3}$$

Finally, combining the above objectives, we can arrive at the overall objective stated as follows,

$$\begin{aligned} &\min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{W}_3,\mathbf{B}_1,\mathbf{B}_2,\mathbf{B}_3} \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \Omega(\mathbf{W}_1,\ \mathbf{W}_2,\ \mathbf{W}_3) \\ &\text{s.t.} \quad \mathbf{B}_1,\ \mathbf{B}_2,\ \mathbf{B}_3 \in \{-1,\ +1\}^{n\times\ell}, \end{aligned} \tag{4}$$

where $\Omega(\mathbf{W}_1,\ \mathbf{W}_2,\ \mathbf{W}_3)$ is the regularization term defined as,

$$\Omega(\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3) = \gamma\|\mathbf{W}_1\|_F^2 + \kappa\|\mathbf{W}_2\|_F^2 + \lambda\|\mathbf{W}_3\|_F^2. \tag{5}$$

3.3. Optimization

To optimize the objective in Eq. (4), we propose an alternative iterative optimization algorithm. We fix other variables while optimizing one variable in each step. The procedure is stated as follows, also shown in Algorithm 1.

Algorithm 1 CHOP.

Input: Class attribute matrix $\mathbf{A}$, seen cross-modal data $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, and parameters α, β, μ, θ, γ, κ, *and* λ.
Output: Hashing functions $\mathbf{W}_1$, $\mathbf{W}_3$, and unified hash codes $\mathbf{B}_1$, $\mathbf{B}_3$.

```
Initialize similarity matrix S and B1, B2, B3, W1, W2, W3.
repeat
    Update B2 by Eq. (7);
    Update B1, B3 by Eqs. (9) and (10);
    Update W1, W3 by Eqs. (13) and (15);
    Update W2 by Eq. (18);
until convergence.
return B1, B3, W1, W3
```

Update $\mathbf{B}_2$. With $\mathbf{B}_1$, $\mathbf{B}_3$, $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}_3$ fixed, the objective function in Eq. (4) can be simplified as,

$$\begin{aligned} &\min_{\mathbf{B}_2}\|\mathbf{A}\mathbf{W}_2 - \mathbf{B}_2\|_F^2 + \mu\|\mathbf{B_1} - \mathbf{B_2}\|_F^2 + \theta\|\mathbf{B_3} - \mathbf{B_2}\|_F^2 \\ &s.t. \quad \mathbf{B}_2 \in \{-1,\ +1\}^{n\times\ell}, \end{aligned} \tag{6}$$

which can be optimized to:

$$\mathbf{B}_2 = sgn(\mathbf{A}\mathbf{W}_2 + \mu\mathbf{B_1} + \theta\mathbf{B_3}). \tag{7}$$

Update $\mathbf{B}_1$, $\mathbf{B}_3$. Fixing the other variables without $\mathbf{B}_1$, the objective can be transformed to,

$$\begin{aligned} &\min_{\mathbf{B}_1}\left\|\mathbf{X}^{(1)}\mathbf{W}_1 - \mathbf{B}_1\right\|_F^2 + \mu\|\mathbf{B_1} - \mathbf{B_2}\|_F^2 \\ &s.t. \quad \mathbf{B}_1 \in \{-1,\ +1\}^{n\times\ell}. \end{aligned} \tag{8}$$

By setting the derivative of Eq. (8) w.r.t $\mathbf{B}_1$ to 0, we can get the following solution,

$$\mathbf{B}_1 = sgn(\mathbf{X}^{(1)}\mathbf{W}_1 + \mu\mathbf{B_2}). \tag{9}$$

Similarly, $\mathbf{B}_3$ can be updated by

$$\mathbf{B}_3 = sgn(\mathbf{X}^{(2)}\mathbf{W}_3 + \theta\mathbf{B_2}). \tag{10}$$

Update $\mathbf{W}_1$, $\mathbf{W}_3$. With $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3$, $\mathbf{W}_2$, and $\mathbf{W}_3$ fixed, we can have,

$$\min_{\mathbf{W}_1} \left\|\mathbf{X}^{(1)}\mathbf{W}_1 - \mathbf{B}_1\right\|_F^2 + \alpha\left\|\mathbf{X}^{(1)}\mathbf{W}_1(\mathbf{A}\mathbf{W}_2)^T - \ell\mathbf{S}\right\|_F^2 + \gamma\|\mathbf{W}_1\|_F^2. \tag{11}$$

We derivative Eq. (11) w.r.t $\mathbf{W}_1$, and set it to 0, then we can have the following formulation,

$$\left(\left(\mathbf{X}^{(1)}\right)^T\mathbf{X}^{(1)} + \gamma\mathbf{I}\right)\mathbf{W}_1 + \left(\alpha\left(\mathbf{X}^{(1)}\right)^T\mathbf{X}^{(1)}\right)\mathbf{W}_1(\mathbf{A}\mathbf{W}_2)^T(\mathbf{A}\mathbf{W}_2) = \left(\mathbf{X}^{(1)}\right)^T\mathbf{B}_1 + \ell\alpha\left(\mathbf{X}^{(1)}\right)^T\mathbf{S}\mathbf{A}\mathbf{W}_2. \tag{12}$$

where $\mathbf{I}$ is the identity matrix.

We define $\hat{\mathbf{A}}_1 = (\alpha(\mathbf{X}^{(1)})^T\mathbf{X}^{(1)})^{-1}((\mathbf{X}^{(1)})^T\mathbf{X}^{(1)} + \gamma\mathbf{I})$, $\hat{\mathbf{B}}_1 = (\mathbf{A}\mathbf{W}_2)^T(\mathbf{A}\mathbf{W}_2)$, $\hat{\mathbf{C}}_1 = (\alpha(\mathbf{X}^{(1)})^T\mathbf{X}^{(1)})^{-1}((\mathbf{X}^{(1)})^T\mathbf{B}_1 + \ell\alpha(\mathbf{X}^{(1)})^T\mathbf{S}\mathbf{A}\mathbf{W}_2)$, then Eq. (12) can be rewritten as:

$$\hat{\mathbf{A}}_1\mathbf{W}_1 + \mathbf{W}_1\hat{\mathbf{B}}_1 = \hat{\mathbf{C}}_1. \tag{13}$$

Eq. (13) is a Sylvester equation which can be solved by using the Sylvester function in MATLAB. Similarly, in terms of $\mathbf{W}_3$, we have,

$$\left(\left(\mathbf{X}^{(2)}\right)^T\mathbf{X}^{(2)} + \gamma\mathbf{I}\right)\mathbf{W}_3 + \left(\beta\left(\mathbf{X}^{(2)}\right)^T\mathbf{X}^{(2)}\right)\mathbf{W}_3(\mathbf{A}\mathbf{W}_2)^T(\mathbf{A}\mathbf{W}_2) = \left(\mathbf{X}^{(2)}\right)^T\mathbf{B}_3 + \ell\beta\left(\mathbf{X}^{(2)}\right)^T\mathbf{S}\mathbf{A}\mathbf{W}_2. \tag{14}$$

Define $\hat{\mathbf{A}}_3 = (\beta(\mathbf{X}^{(2)})^T\mathbf{X}^{(2)})^{-1}((\mathbf{X}^{(2)})^T\mathbf{X}^{(2)} + \gamma\mathbf{I})$, $\hat{\mathbf{B}}_3 = (\mathbf{A}\mathbf{W}_2)^T(\mathbf{A}\mathbf{W}_2)$, $\hat{\mathbf{C}}_3 = (\beta(\mathbf{X}^{(2)})^T\mathbf{X}^{(2)})^{-1}((\mathbf{X}^{(2)})^T\mathbf{B}_3 + \ell\beta(\mathbf{X}^{(2)})^T\mathbf{S}\mathbf{A}\mathbf{W}_2)$, then Eq. (14) can be rewritten as,

$$\hat{\mathbf{A}}_3\mathbf{W}_3 + \mathbf{W}_3\hat{\mathbf{B}}_3 = \hat{\mathbf{C}}_3. \tag{15}$$

Thus, the solution of $\mathbf{W}_3$ can be obtained.

Update $\mathbf{W}_2$. Fix other variables but $\mathbf{W}_2$, then Eq. (4) can be simplified as,

$$\min_{\mathbf{W}_2} \|\mathbf{A}\mathbf{W}_2 - \mathbf{B}_2\|_F^2 + \kappa\|\mathbf{W}_2\|_F^2 + \alpha\|\mathbf{X}^{(1)}\mathbf{W}_1(\mathbf{A}\mathbf{W}_2)^T - \ell\mathbf{S}\|_F^2 + \beta\|\mathbf{X}^{(2)}\mathbf{W}_3(\mathbf{A}\mathbf{W}_2)^T - \ell\mathbf{S}\|_F^2. \tag{16}$$

Setting the derivative of Eq. (16) w.r.t $\mathbf{W}_2$ to 0, we can have the following equation similar to Eq. (13),

$$(\mathbf{A}^T\mathbf{A} + \kappa\mathbf{I})\mathbf{W}_2 + (\mathbf{A}^T\mathbf{A})\mathbf{W}_2\mathbf{Q} = \mathbf{A}^T, \tag{17}$$

where $\mathbf{Q} = \alpha(\mathbf{X}^{(1)}\mathbf{W}_1)^T\mathbf{X}^{(1)}\mathbf{W}_1 + \beta(\mathbf{X}^{(2)}\mathbf{W}_3)^T\mathbf{X}^{(2)}\mathbf{W}_3$.

Similarly, we define $\hat{\mathbf{A}}_2 = (\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{A} + \kappa\mathbf{I})$, $\hat{\mathbf{B}}_2 = \mathbf{Q}$, $\hat{\mathbf{C}}_2 = (\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T)$, then Eq. (17) can be rewritten as:

$$\hat{\mathbf{A}}_2\mathbf{W}_2 + \mathbf{W}_2\hat{\mathbf{B}}_2 = \hat{\mathbf{C}}_2. \tag{18}$$

Thus, the solution of $\mathbf{W}_2$ can be easily calculated.

4. Experiments

In this section, we conduct extensive experiments on three widely used datasets to verify the performance of the proposed method on zero-shot cross-modal retrieval. We design Text-to-Image (T2I) and Image-to-Text (I2T) two kinds of tasks, which are the typical tasks in the cross-modal retrieval.

4.1. Datasets

Wiki [31] dataset involves 10 semantic categories, that contains a total of 2866 pairs of image-text data, including 2173 pairs of training data and 693 pairs of test data. The 128-dimensional bag-of-visual words SIFT feature vectors are used to represent the images, while the latent Dirichlet allocation (LDA) model generates 10-dimensional topic vectors to describe text modality.

Pascal VOC [32] dataset can be divided into 20 categories and it contains 9963 pairs of picture-label data. In this dataset, we only choose single-label data. The publicly available features are used for the experiment, where the image modality is represented with 512-dimensional GIST features, and the text modality is described with 399-dimensional word frequency features.

LabelMe [33] dataset includes 2688 outdoor scenes, which can be divided into 8 categories. After we delete the words that appear less than three times, the remaining 366 unique words form a 366-dimensional textual representation. The image representations are 512-dimensional GIST features. In addition, the data set contains 2686 pairs of image-text data after we discard some unlabeled samples.

For each dataset, we use the word vectors of class names as the attributes of each class i.e., $\mathbf{A}$. 300-dimensional real-valued vectors are extracted from Glove [34] as the way in Zhong et al. [16].

4.2. Settings

Firstly, we need to construct the zero-shot scenarios for each dataset. Following the settings in Zhong et al. [16], for Wiki and LabelMe, each experiment randomly selects two classes as unseen categories, while for Pascal VOC, four classes are selected as unseen classes, and the rest are trained as seen classes. Ten repeated experiments are conducted to achieve the average as the final result. We follow the generalized zero-shot setting described in Zhong et al. [16], which means that training data are selected from seen classes, the query data in the testing phase are randomly selected from unseen classes, and the remaining data from unseen classes and all the samples from seen classes consist the retrieval set. There are two measurement metrics that can be used to measure the performance of cross-modal retrieval, i.e., Mean Average Precision (MAP) and the mean precision within Hamming radius 2 (PH2) [38].

Due to the particularity of dataset, the parameters are set by different values for each dataset to ensure the optimal performance. For the Wiki dataset, α and β are 1 and 10^{-3}, μ and θ are all 1, then γ, κ, and λ are all set as 10^4. For Pascal VOC, α and β are 10^{-1} and 10^{-2}, μ and θ are all 1, thenγ, κ, and λ are all 10^3. For LabelMe dataset, α and β are 10^{-1} and 10^{-2}, μ and θ are 1 and 10^{-1}, and γ, κ, and λ are all set as 10^4. We limit the iterative number as 10. The results presented in this paper are the average of 10 repeated experiments.

4.3. Baseline methods

We compare the proposed CHOP with three cross-modal hashing approaches, two zero-shot hashing approaches, and one typical zero-shot cross-modal retrieval approach. CMFH [20], Latent semantic sparse hashing (LSSH) [35], and Supervised matrix factorization hashing (SMFH) [36] are cross-modal hashing approaches. While the zero-shot hashing approaches are TSK [30] and Attribute Hashing (AH) [37]. The typical zero-shot cross-modal method is CMAH [16]. We build the baseline with the parameters which are recommended by the original papers.

4.4. Experimental results

4.4.1. Cross-modal retrieval of unseen classes

Our method mainly focuses on cross-modal retrieval performance of unseen classes, i.e., the queries are from unseen classes. Next, we will evaluate the performance from two aspects, i.e., MAP and PH2.

Fig. 2 shows the MAP results of cross-modal retrieval for unseen queries, i.e., cross-modal retrieval performance for unseen classes. The experimental results show that the CHOP performs

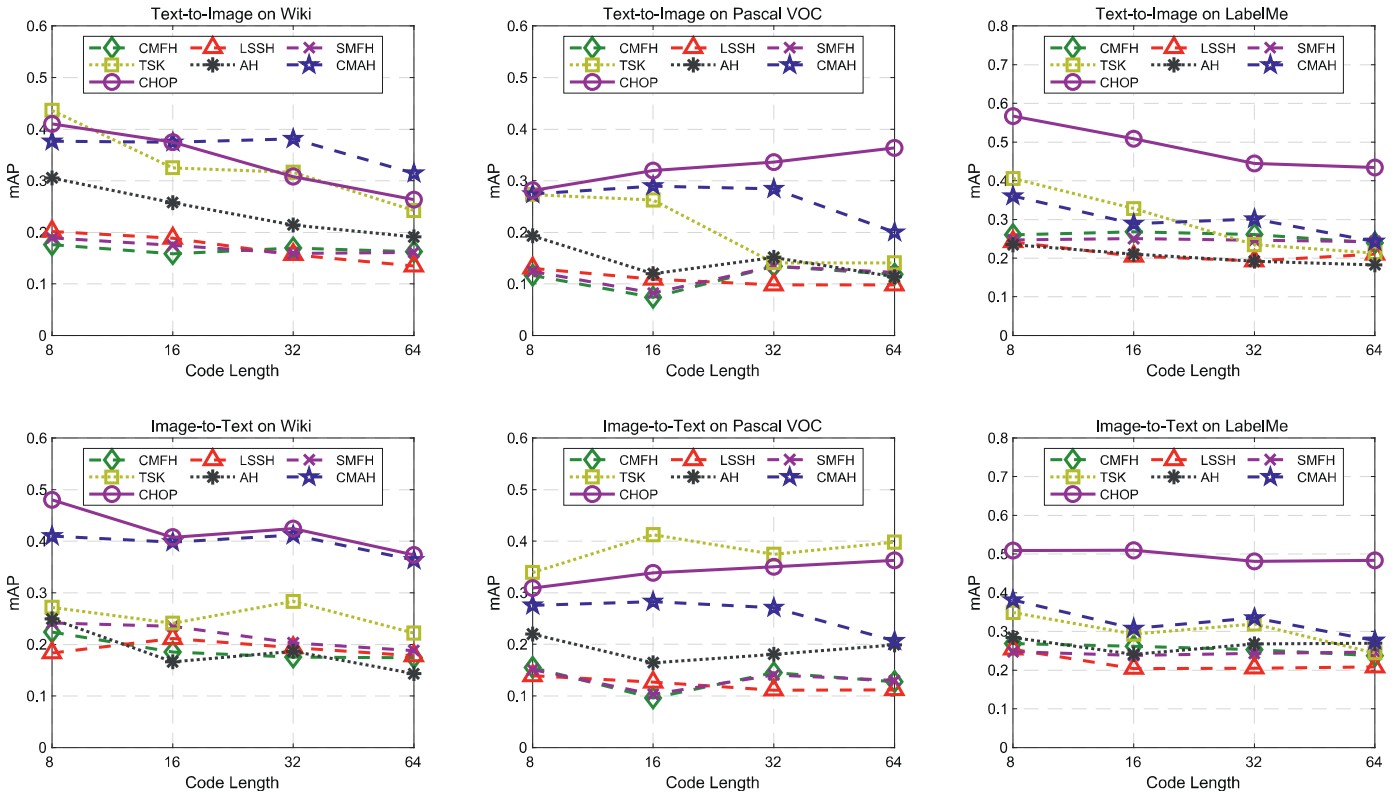

Fig. 2. The MAP results of cross-modal retrieval (Text-to-Image and Image-to-Text) for unseen classes of all methods on Wiki, Pascal VOC, and LabelMe datasets with different hash code lengths.

Table 1
The PH2 results of cross-modal retrieval (Text-to-Image and Image-to-Text) for unseen queries of all methods on Wiki, Pascal VOC, and LabelMe datasets with different hash code lengths.

Task	Method	WIKI_UNSEEN				PASCAL_UNSEEN				LABELME_UNSEEN			
		8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits
T2I	CMFH [20]	0.1384	0.0436	0.0027	0	0.0847	0.0225	0	0	0.2071	0.2533	0.0061	0
	LSSH [35]	0.1604	0.1151	0.0342	0	0.0847	0.0674	0.0008	0	0.1683	0.1472	0.0195	0
	SMFH [36]	0.1432	0.1387	0.0475	0.0106	0.0838	0.0828	0.0347	0	0.2144	0.4050	0.2907	0.2358
	TSK [30]	0.0978	0.0189	0	0	0.0776	0.0623	0.0042	0	0.1282	0.0343	0.0176	0
	AH [37]	0.0940	0.0424	0.0047	0	0.0240	0.0015	0	0	0.0420	0.0015	0	0
	CMAH [16]	0.1443	0.1581	0.1588	**0.1661**	0.0999	0.1803	0.1661	**0.1194**	0.1826	0.1913	0.2310	0.1471
	Ours	**0.2287**	**0.3920**	**0.3079**	0.0448	**0.1865**	**0.3158**	**0.2578**	0.0123	**0.2890**	**0.4856**	**0.5604**	**0.4462**
I2T	CMFH [20]	0.1087	0.0582	0.0053	0.0011	0.0886	0.0364	0.0003	0	0.1738	0.1678	0.0026	0
	LSSH [35]	0.1212	0.0745	0.0187	0.0009	0.0838	0.0450	0.0004	0	0.1555	0.0895	0.0033	0
	SMFH [36]	0.1021	0.1206	0.0472	0.0082	0.0896	0.0618	0.0080	0	0.2200	0.3279	0.2594	0.1831
	TSK [30]	0.1204	0.0951	0.0002	0	0.0648	0.0312	0.0005	0	0.1335	0.0222	0.0084	0
	AH [37]	0.1116	0.0900	0.0400	0.0084	0.0348	0.0128	0	0	0.0720	0.0197	0	0
	CMAH [16]	0.1297	0.1372	0.1393	**0.1532**	0.0930	0.1802	0.1925	**0.2148**	0.1725	0.1918	0.2195	0.1469
	Ours	**0.3053**	**0.3712**	**0.1923**	0.0201	**0.2049**	**0.3584**	**0.2426**	0.0093	**0.2202**	**0.3958**	**0.3687**	**0.1904**

significantly better than the three cross-modal hashing methods. The reason is that traditional cross-modal hashing approaches can only handle seen classes retrieval. It is easy to observe that, the results of the two single-modal zero-shot approaches TSK and AH are better than the traditional cross-modal hashing approaches. This is because that they can learn and transfer the knowledge from seen classes to unseen classes. However, since TSK and AH cannot close the heterogeneous gap from different modalities, the proposed CHOP outperforms TSK and AH in most cases. As for CMAH, our method is superior to it in most cases. This is because CMAH does not consider the orthogonal constraint between hash codes, which also certificates the effectiveness of CHOP for handling the zero-shot cross-modal retrieval. Though CMAH can get better performance for Text-to-Image unseen categories retrieval on Wiki datasets with 32 bits and 64 bits, it cannot perform well on the seen categories retrieval. Similarly, TSK performs well for Image-to-Text unseen categories retrieval on Pascal datasets, but it fails in seen categories retrieval. The proposed CHOP tries to balance the results between the unseen and seen query, and it has better performance overall.

The PH2 results of cross-modal retrieval for unseen classes of various methods are reported in Table 1. It shows that CHOP is mostly superior to the other six methods. Since the PH2 indicates the local distribution of the hash codes, results indicate that CHOP can achieve better and discriminative hash codes in the Hamming space. We also can find that the PH2s of TSK and AH are lower than CMFH, LSSH, and SMFH. Owing to the constraints on the hash codes, our CHOP performs better than CMAH at varied code lengths. In addition, Table 1 shows that the results increase gradually when the length of hash codes is less than 32 bits and

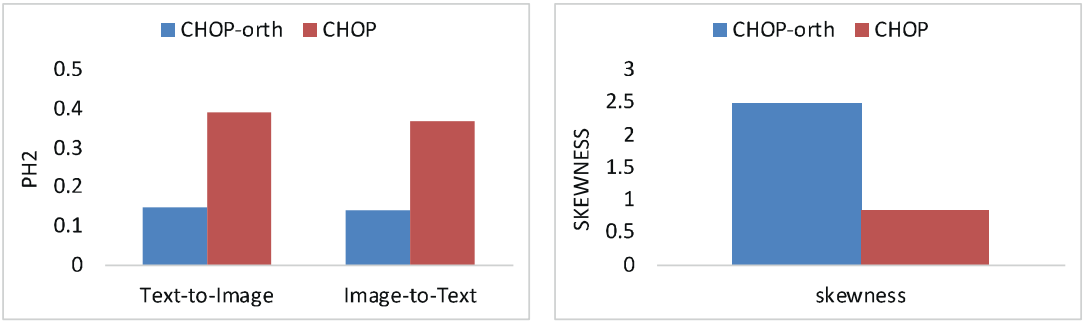

Fig. 3. Ablation study based on PH2 and N_k_*skewness* results on Wiki.

Table 2
The MAP results of cross-modal retrieval (Text-to-Image and Image-to-Text) for seen queries of all methods on Wiki, Pascal VOC, and LabelMe datasets with different hash code lengths.

Task	Method	WIKI_SEEN				PASCAL_SEEN				LABLEME_SEEN			
		8bits	16bits	32bits	64bits	8bits	16bits	32bits	64bits	8bits	16bits	32bits	64bits
T2I	CMFH [20]	0.4982	0.5281	0.5549	0.5779	0.4125	0.5885	0.5558	0.5314	0.5856	0.5306	0.4879	0.4801
	LSSH [35]	0.4698	0.5345	0.5688	0.5926	0.5013	0.5891	0.6362	0.6372	0.6192	0.6513	0.6707	0.7116
	SMFH [36]	**0.6195**	**0.6653**	0.6677	0.6874	0.4231	0.7216	0.7553	0.8277	0.6438	0.7329	0.6951	0.6291
	TSK [30]	0.1418	0.1435	0.1522	0.1484	0.1165	0.1122	0.1161	0.1139	0.2305	0.2643	0.2657	0.2847
	AH [37]	0.2294	0.2295	0.2240	0.2074	0.1312	0.1283	0.1226	0.1319	0.2441	0.2386	0.2579	0.2656
	CMAH [16]	0.5819	0.6256	0.6420	0.6282	**0.7208**	**0.8417**	0.8564	0.8713	**0.9151**	**0.9260**	**0.9252**	0.9222
	Ours	0.5460	0.6160	**0.6693**	**0.7040**	0.5330	0.7612	**0.8730**	**0.9073**	0.7125	0.8350	0.9032	**0.9339**
I2T	CMFH [20]	0.2565	0.2579	0.2783	0.2817	0.1848	0.2235	0.2078	0.1969	0.4540	0.4266	0.3927	0.3879
	LSSH [35]	0.2292	0.2557	0.2708	0.2673	**0.2650**	**0.2987**	0.3094	0.3294	0.6217	0.6705	0.6878	0.7256
	SMFH [36]	**0.3089**	**0.3240**	0.3414	0.3402	0.2266	0.2828	0.2959	0.3245	0.5504	0.6221	0.6140	0.5607
	TSK [30]	0.1777	0.1803	0.1689	0.1597	0.1105	0.1059	0.1071	0.1055	0.2376	0.2798	0.2566	0.2704
	AH [37]	0.2107	0.1924	0.1694	0.1736	0.1119	0.1103	0.1155	0.1120	0.1777	0.1796	0.1861	0.1998
	CMAH [16]	0.2342	0.2329	0.2365	0.2503	0.1919	0.2020	0.1900	0.2351	**0.7694**	**0.8017**	**0.8136**	**0.8239**
	Ours	0.2423	0.2979	**0.3424**	**0.3660**	0.2395	0.2941	**0.3408**	**0.3497**	0.5488	0.6465	0.7278	0.7690

reach the peak at 32 bits in most cases. This indicates that the hash codes with too few bits cannot contain enough information while too long bits will impose more noise or irrelevant information. Consequently, the appropriate length of hash codes should be chosen in different tasks to ensure the performance.

4.4.2. Cross-modal retrieval of seen classes

We also evaluate the performance of the cross-modal retrieval for seen classes which uses the conventional setting of cross-modal retrieval methods. In such case, the queries are from seen classes. Table 2 shows the MAP results for seen classes of all methods, some of which are cited from [16]. The best MAP results are in bold and the second-best MAP results are underlined. Table 2 indicates that TSK and AH perform rather poorly, while the methods CMFH, LSSH, SMFH perform better. This is because CMFH, LSSH, SMFH are traditional cross-modal hashing methods and are trained with data from seen classes. However, they cannot tackle cross-modal retrieval for unseen classes well. We can further find that CMAH performs better than the traditional cross-modal hashing method. This is because CMAH although mainly focuses on unseen classes, it still considers balancing the unseen and seen query results. Comprehensively, the proposed CHOP is competitive for seen classes.

4.5. Ablation study

One of the contributions of our work is that the proposed orthogonal constraint can mitigate the hubness problem. In order to validate that, we need to define the measure of hubness. According to the definition of the hubness problem in Radovanovic et al. [17], the authors in Zhang et al. [18], Shigeto et al. [39] proposed to measure the hubness using the skewness score of N_k distribution, which is the distribution of the number of times $N_k(i)$ each prototype i is found in the top k of the ranking for test samples (i.e. their k-nearest neighbors). This is because the skewness of N_k distribution reflects the existence of target objects that frequently appear in the k-nearest neighbors lists of source objects, i.e., the emergence of hubs. Therefore, we also employ the N_k skewness value to measure the degree of hubness problem. For more details, please refer to [18,39].

Following the way of [18,39], the skewness of N_k distribution is defined as follows,

$$N_k\_skewness = \frac{\sum_{i=1}^{n}(N_k(i) - E[N_k])^3}{n(Var[N_k])^{\frac{3}{2}}}, \tag{19}$$

where $N_k(i)$ is the ithth prototype with the top k nearest neighbors in the test samples, and n represents the total number of test prototypes. In the experiment, k is set to 1.

Then, we can compute the value of N_k_*skewness* to measure the degree of hubness. The larger value of N_k_*skewness* leads to the severer hubness problem. Here, we design an experiment to validate the importance of orthogonal constraint in mitigating the hubness problem. We define the CHOP-orth as the CHOP method without orthogonal constraint. Fig. 3 shows the PH2 and N_k_*skewness* values of CHOP-orth and CHOP on the Wiki dataset at 16 bits. Obviously, the CHOP achieves higher PH2 and lower N_k_*skewness* score. This validates that the orthogonal constraint of our CHOP can alleviate the hubness problem.

5. Conclusion

To address the problem of zero-shot cross-modal retrieval, we have proposed an orthogonal hashing method i.e., Cross-modal Hashing with Orthogonal Projection (CHOP). CHOP maps different

modalities and the class attributes onto a unified Hamming space, in which the orthogonal and equal constraints are imposed on the hash codes. The equal constraint is imposed on the paired image (text) and class attributes. While the orthogonal constraint is imposed on image (text) with mismatched class attributes. Owing to this, our method can have a more discriminative Hamming space and can alleviate the hubness problem, which results in better performance of the proposed approach in zero-shot cross-modal retrieval. Extensive experiments and results demonstrate the effectiveness of the proposed CHOP in handling the cross-modal retrieval of data from unseen classes. In addition, we also evaluated the performance in the traditional cross-modal retrieval of the proposed CHOP. The results demonstrated that our CHOP can achieve a balanced results between the zero-shot cross-modal retrieval and the traditional cross-modal retrieval. Since the deep neural networks perform well in feature extracting, we plan to extend our method to a deep end-to-end framework for future research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China [Grant number 2018YFC0831305] and National Natural Science Foundation of China [Grant numbers 62006035, 62076047].

References

[1] Y. Cao, M. Long, J. Wang, Q. Yang, P.S. Yu, Deep visual-semantic hashing for cross-modal retrieval, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1445–1454.
[2] H. Liu, F. Wang, X. Zhang, F. Sun, Weakly-paired deep dictionary learning for cross-modal retrieval, Pattern Recognit. Lett. 130 (2020) 199–206.
[3] L. Wang, L. Zhu, E. Yu, J. Sun, H. Zhang, Task-dependent and query-dependent subspace learning for cross-modal retrieval, IEEE Access 6 (2018) 27091–27102.
[4] Y. Jia, L. Bai, S. Liu, P. Wang, J. Guo, Y. Xie, Semantically-enhanced kernel canonical correlation analysis: a multi-label cross-modal retrieval, Multimed. Tools Appl. 78 (10) (2019) 13169–13188.
[5] Z. Chen, F. Zhong, G. Min, Y. Leng, Y. Ying, Supervised intra-and inter-modality similarity preserving hashing for cross-modal retrieval, IEEE Access 6 (2018) 27796–27808.
[6] F. Zhong, Z. Chen, G. Min, Deep discrete cross-modal hashing for cross-media retrieval, Pattern Recognit. 83 (2018) 64–77.
[7] H. Peng, J. He, S. Chen, Y. Wang, Y. Qiao, Dual-supervised attention network for deep cross-modal hashing, Pattern Recognit. Lett. 128 (2019) 333–339.
[8] Y. Cao, M. Long, B. Liu, J. Wang, Deep cauchy hashing for hamming space retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1229–1237.
[9] T. Long, X. Xu, F. Shen, L. Liu, N. Xie, Y. Yang, Zero-shot learning via discriminative representation extraction, Pattern Recognit. Lett. 109 (2018) 27–34.
[10] Z. Ye, F. Lyu, L. Li, Q. Fu, J. Ren, F. Hu, SR-GAN: semantic rectifying generative adversarial network for zero-shot learning, in: IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 85–90.
[11] X. Li, M. Fang, H. Li, J. Wu, Zero shot learning based on class visual prototypes and semantic consistency, Pattern Recognit. Lett. 135 (2020) 368–374.
[12] H. Zhang, J. Liu, Y. Yao, Y. Long, Pseudo distribution on unseen classes for generalized zero shot learning, Pattern Recognit. Lett. 135 (2020) 451–458.
[13] J. Chi, X. Huang, Y. Peng, Zero-shot cross-media retrieval with external knowledge, in: International Conference on Internet Multimedia Computing and Service, Springer, 2017, pp. 200–211.
[14] Z. Ji, Y. Sun, Y. Yu, Y. Pang, J. Han, Attribute-guided network for cross-modal zero-shot hashing, IEEE Trans. Neural Netw. Learn. Syst. 31 (1) (2019) 321–330.
[15] T. Dutta, S. Biswas, Cross-modal retrieval in challenging scenarios using attributes, Pattern Recognit. Lett. 125 (2019) 618–624.
[16] F. Zhong, Z. Chen, G. Min, An exploration of cross-modal retrieval for unseen concepts, in: International Conference on Database Systems for Advanced Applications, Springer, 2019, pp. 20–35.
[17] M. Radovanovic, A. Nanopoulos, M. Ivanovic, Hubs in space: popular nearest neighbors in high-dimensional data, J. Mach. Learn. Res. 11 (sept) (2010) 2487–2531.
[18] L. Zhang, T. Xiang, S. Gong, Learning a deep embedding model for zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3010–3019.
[19] H. Zhang, Y. Long, L. Shao, Zero-shot hashing with orthogonal projection for image retrieval, Pattern Recognit. Lett. 117 (2019) 201–209.
[20] G. Ding, Y. Guo, J. Zhou, Collective matrix factorization hashing for multimodal data, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2075–2082.
[21] D. Wang, X. Gao, X. Wang, L. He, Semantic topic multimodal hashing for cross–media retrieval, in: International Joint Conference on Artificial Intelligence, 2015, pp. 3890–3896.
[22] H. Liu, R. Ji, Y. Wu, G. Hua, Supervised matrix factorization for cross-modality hashing, in: International Joint Conference on Artificial Intelligence, 2016, pp. 1767–1773.
[23] C. Li, C. Deng, N. Li, W. Liu, X. Gao, D. Tao, Self-supervised adversarial hashing networks for cross-modal retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4242–4251.
[24] X. Gong, L. Huang, F. Wang, Deep semantic correlation learning based hashing for multimedia cross-modal retrieval, in: IEEE International Conference on Data Mining (ICDM), IEEE, 2018, pp. 117–126.
[25] X. Xu, L. He, H. Lu, L. Gao, Y. Ji, Deep adversarial metric learning for cross–modal retrieval, World Wide Web 22 (2) (2019) 657–672.
[26] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.
[27] H. Huang, C. Wang, P.S. Yu, C.-D. Wang, Generative dual adversarial network for generalized zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 801–810.
[28] Y. Yu, Z. Ji, J. Guo, Z. Zhang, Zero-shot learning via latent space encoding, IEEE Trans. Cybern. 49 (10) (2019) 3755–3766.
[29] B. Zhao, X. Sun, X. Hong, Y. Yao, Y. Wang, Zero-shot learning via recurrent knowledge transfer, in: IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1308–1317.
[30] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, H.T. Shen, Zero-shot hashing via transferring supervised knowledge, in: ACM International Conference on Multimedia, 2016, pp. 1286–1295.
[31] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 251–260.
[32] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[33] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.
[34] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.
[35] 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 & Development in Information Retrieval, ACM, 2014, pp. 415–424.
[36] J. Tang, K. Wang, L. Shao, Supervised matrix factorization hashing for cross–modal retrieval, IEEE Trans. Image Process. 25 (7) (2016) 3157–3166.
[37] 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, pp. 133–138.
[38] X. Liu, J. He, C. Deng, B. Lang, Collaborative hashing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2147–2154.
[39] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, Y. Matsumoto, Ridge regression, hubness, and zero-shot learning, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015, pp. 135–151.