An Exploration of Cross-Modal Retrieval for Unseen Concepts

Fangming Zhong[1(✉)], Zhikui Chen[1,2], and Geyong Min[3]

[1] School of Software, Dalian University of Technology, Dalian, China
fmzhong@mail.dlut.edu.cn
[2] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, China
[3] College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

Abstract. Cross-modal hashing has drawn increasing research interests in cross-modal retrieval due to the explosive growth of multimedia big data. However, most of the existing models are trained and tested in a close-set circumstance, which may easily fail on the newly emerged concepts that are never present in the training stage. In this paper, we propose a novel cross-modal hashing model, named Cross-Modal Attribute Hashing (CMAH), which can handle cross-modal retrieval of unseen categories. Inspired by zero-shot learning, attribute space is employed to transfer knowledge from seen categories to unseen categories. Specifically, the cross-modal hashing functions learning and knowledge transfer are conducted by modeling the relationships among features, attributes, and classes as a dual multi-layer network. In addition, graph regularization and binary constraints are imposed to preserve the local structure information in each modality and to reduce quantization loss, respectively. Extensive experiments are carried out on three datasets, and the results demonstrate the effectiveness of CMAH in handling cross-modal retrieval for both seen and unseen concepts.

Keywords: Cross-modal retrieval · Unseen classes · Zero-shot learning

1 Introduction

Recent years have witnessed the rapidly increasing interests in cross-modal retrieval that is becoming significant and imperative for many real-world applications, such as using image to search the relevant text documents or searching relevant images with given text query [2,6,15,31,32]. Due to the large-scale and high-dimensional properties of multimodal data, cross-modal hashing which has shown fairly impressive performance in reducing storage cost and improving retrieval speed, has been investigated intensively over the last few years [5,11,14,23,33].

The original version of this chapter was revised: An acknowledgement has been added. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-18579-4_46

G. Li et al. (Eds.): DASFAA 2019, LNCS 11447, pp. 20–35, 2019.
https://doi.org/10.1007/978-3-030-18579-4_2

It is worth noting that, most of the current cross-modal hashing models are trained and tested in a close set i.e. the training and test categories are identical. However, with the explosion of newly-emerging concepts, it is infeasible to label data for each class. Additionally, the number of labelled data for these new concepts may be far from sufficient to build high-quality cross-modal hashing model. The existing methods perform well on the seen data, but they may easily fail on the unseen concepts that are never present before in the training stage. This gives rise to an emerging demand to explore the problem of cross-modal retrieval for unseen concepts.

Such learning with no data in unimodal scenario is termed zero-shot learning (ZSL) [12,13,16,26] that has been widely studied in recent years. The fundamental goal of zero-shot learning is recognizing objects from classes that are not seen during training. The key challenge of achieving this goal is to transfer knowledge from the limited seen categories to unseen categories. Most of the previous approaches employ an intermediate semantic space to conduct knowledge transfer as well as to bridge the semantic gap between low-level visual feature and high-level class label. For example, the authors in [26] proposed to learn semantic embedding projection by matrix tri-factorization and manifold regularization. In [13], semantic autoencoder is employed to learn a projection that can generalize better to new unseen classes. In contrast, orthogonal semantic-visual embedding was developed in [16] to inversely use semantic space to infer visual features for unseen classes. However, all the above methods focus only on unimodal classification or recognition scenarios. Only a few works on zero-shot hashing have been proposed. Zero-shot hashing [29] is one of the first works that focus on handling visual indexing by hashing for unseen categories. In [27], a multi-layer hierarchy was proposed for zero-shot image retrieval. However, in real world, users may be not satisfied with image query, but more comfortable to use other types of query such as text and sound. To the best of our knowledge, the zero-shot learning problem in cross-modal retrieval has been rarely investigated in previous works. In [4], the cross-modal retrieval for unseen categories was first explored with external knowledge. It utilizes a weight vector to build the connection between seen and unseen classes for knowledge transfer. However, this method which simply combines the deep networks with dot product operation uses the pre-trained model with ImageNet that actually includes the information of unseen categories. Thus, the experimental results of cross-modal retrieval for unseen classes are not convincing. Ji et al. also noticed the importance of cross-modal retrieval for unseen concepts [10]. However, their work explores the zero-shot cross-modal hashing with images and the class names, which is different from the traditional cross-modal tasks i.e. image to text and text to image. Therefore, this is the first work that explores cross-modal retrieval for unseen concepts using hashing technique.

In this paper, we consider the problem of handling unseen classes in cross-modal hashing. The main focus is on generalizing the cross-modal hashing model from seen classes to unseen classes, which can produce effective hash codes for data from unseen classes. Motivated by zero-short hashing and the recently proposed

approaches [21,27,29], a novel cross-modal attribute hashing (CMAH) model is presented in this work. During the cross-modal hashing functions learning, the knowledge transfer between seen and unseen categories is conducted with the idea of modelling the relationships among features, attributes, and classes as a dual multi-layer network. As shown in Fig. 1, cross-modal data are projected into unified binary codes that are used to construct the relationship between attributes and class labels, as well as build the connection of different modalities. Moreover, graph regularization and binary constraints are imposed to preserve the local structure information in each modality and to reduce quantization loss, respectively. Thus, the learned hashing functions for each modality through seen classes not only can generate discriminative binary codes for seen classes, but also can generalize well to the unseen classes. By conducting experiments on three non-overlapping cross-modal datasets, the effectiveness of our method has been validated. Compared against the existing cross-modal hashing methods, our method can effectively handle the cross-modal retrieval for unseen concepts. In addition, the proposed method also shows superior performance on the cross-modal retrieval of seen classes.

The main contributions of this paper are:

- A novel cross-modal attribute hashing model is proposed to explore the problem of cross-modal retrieval in zero-shot scenario. To our best knowledge, this is one of the first works which explores the cross-modal hashing for handling unseen classes.
- A cross-modal multi-layer network is developed for simultaneously connecting features, binary codes, attributes, and classes and building relationship among different modalities. Furthermore, local structure information in each modality has been preserved in the expected Hamming space.
- Experiments on cross-modal retrieval for both unseen query and seen query are conducted to evaluate the effectiveness of the proposed method. We find that the proposed method shows superior performance on both the seen query and unseen query.

The rest of this paper is organized as follows. The previous works on cross-modal hashing and zero-shot learning are reviewed in Sect. 2. The proposed approach is presented in Sect. 3. Section 4 presents the experimental results. Finally, we conclude our work in Sect. 5.

2 Related Work

Since our work mainly concerns handling the unseen classes in cross-modal retrieval based on hashing, this section reviews the previous works from two aspects, i.e. conventional cross-modal hashing and unimodal zero-shot learning.

2.1 Cross-Modal Hashing

Motivated by hashing, a number of methods have been proposed to conduct cross-modal retrieval. For example, in Collective Matrix Factorization Hashing

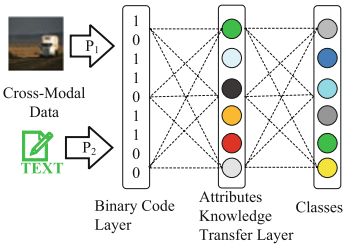

Fig. 1. The framework of the proposed method.

(CMFH) [5], matrix factorization is utilized to learn the latent concepts from each modality, which has achieved an impressive result on cross-modal retrieval. In [33], Latent Semantic Sparse Hashing (LSSH) was presented which learns the semantic concepts of images and text by sparse coding and matrix factorization respectively. The learned latent semantic features from images and text are then mapped to a common abstraction space in which the unified hash codes are generated by quantization. Inspired by CMFH, several supervised extensions [14,23] have been proposed to formulate the label information for boosting retrieval performance. A unified linear regression model with dragging technique based on semi-supervised learning for cross-modal retrieval was proposed in [30]. Most of these methods adopt hand-crafted features as input. Recently, deep neural networks such as convolutional neural networks (CNN) have drawn considerable attention in cross-modal retrieval [11,28]. Due to the high-level abstract of original data, the CNN based methods perform better than those based on low-level hand-crafted features. However, these methods suffer from high time complexity of training CNN. Most importantly, none of hand-crafted or CNN based methods have considered the cross-modal retrieval for unseen concepts.

2.2 Zero-Shot Learning

Zero-shot learning has become an active topic in recent years due to the rapid evolution of newly-emerging concepts. Most of the existing methods aim at solving the recognition task of unseen categories. A promising solution is to find an intermediate representation which can bridge the semantic gap between visual features and class labels, and can also transfer knowledge from seen classes to unseen classes. For instance, the methods including [12,16] project both the images and class labels into an attribute space, where a simple nearest neighbor classifier can be adopted to recognize the instances from unseen categories. However, these methods focus only on classification or recognition. Few works on zero-shot hashing for retrieval have been presented. In [18], the authors investigated the hashing in the zero shot scenario for image retrieval, in which hashing function is learned based on the combination of similarity preserving and unsupervised domain adaptation. In [8], a zero-shot hashing based on CNN is proposed, which considers the similarity transfer, discriminability, and discrete

hashing comprehensively. However, the existing zero-shot hashing approaches can only deal with single modality, the circumstance of multiple modalities has not been explored. Therefore, a novel cross-modal hashing framework that can handle cross-modal retrieval for unseen classes draws a significant need.

Overall, the cross-modal retrieval for unseen concepts has not yet been investigated well. The works in [4] and [10] have noticed the importance of this problem. However, due to the utilization of pre-trained model for feature extraction and retrieval based on class name, their results are not convincing enough. To our best knowledge, this paper is one of the first works that explore cross-modal retrieval for unseen concepts using hashing technique.

3 Approach

In this section, the proposed CMAH for tackling cross-media retrieval of unseen classes is described in detail followed by the optimization algorithm.

3.1 Problem Definition

The definition of zero-shot cross-modal hashing follows [29]. Given n pairs of "seen" cross-modal data $\mathbf{X}^{(1)} = \{x_1^{(1)}, \cdots, x_n^{(1)}\}$ and $\mathbf{X}^{(2)} = \{x_1^{(2)}, \cdots, x_n^{(2)}\}$, such as images and the associated text, where $\mathbf{X}^{(1)} \in \Re^{d_1 \times n}$, $\mathbf{X}^{(2)} \in \Re^{d_2 \times n}$, d_1 represents the dimensionality of image feature, d_2 denotes the dimensionality of text feature (usually $d_1 \neq d_2$). The semantic label of the given data is $\mathbf{Y} \in \{0,1\}^{n \times c}$, where c is the size of "seen" classes. Different from the conventional cross-modal hashing setting in which the training data and testing data are from the seen classes, we assume that some testing data are from unseen classes which are never present during training. The goal of our proposed method is to learn cross-modal hashing model via seen classes and then generalize it to unseen classes for generating high-quality discriminative binary codes.

3.2 Cross-Modal Attribute Hashing Formulation

Motivated by the recently proposed zero-shot learning approaches [21], the knowledge transfer is conducted in the intermediate attribute space. As shown in Fig. 1, we formulate the relationship of cross-modal data, binary codes, attribute, and class labels as the following loss function:

$$\begin{aligned} \mathcal{L}_1 = & \|\mathbf{Y} - \mathbf{BVS}\|_F^2 + \alpha \left\|\mathbf{B} - (\mathbf{X}^{(1)})^T \mathbf{P}_1\right\|_F^2 \\ & + \beta \left\|\mathbf{B} - (\mathbf{X}^{(2)})^T \mathbf{P}_2\right\|_F^2 \\ s.t. \quad & \mathbf{B} \in \{-1, +1\}^{n \times k} \end{aligned} \tag{1}$$

where $\mathbf{B}$ is the unified binary codes of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, k is the length of binary codes, $\mathbf{P}_1 \in \Re^{d_1 \times k}$ and $\mathbf{P}_2 \in \Re^{d_2 \times k}$ are cross-modal hashing functions projecting

images and text to hash codes, and $\mathbf{V} \in \Re^{k \times a}$ is the mapping matrix from binary codes to attributes, $\mathbf{S} \in \Re^{a \times c}$ is the mapping from attributes to semantic class labels, where a is the number of attributes. Here, we use the word vector representation of class name as $\mathbf{S}$ following the idea of [3].

In addition, the local structure information preserving in each modality is explored during learning cross-modal hashing functions. Thus, the similar items in original feature space will share similar binary codes in the Hamming space, which can further enhance the discriminative capability of learned hashing functions. Laplacian Eigenmaps (LE) [1] is utilized to formulate the structure preservation based on manipulations on an undirected weight graph which indicates the neighborhood relationship of pairwise data. The objective with respect to $\mathbf{X}^{(1)}$ can be stated as:

$$\min_{\mathbf{P}_1} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \mathbf{P}_1^T x_i^{(1)} - \mathbf{P}_1^T x_j^{(1)} \right\|^2 w_{ij}^{(1)} \tag{2}$$

where $w_{ij}^{(1)}$ is the similarity of $x_i^{(1)}$ and $x_j^{(1)}$. It usually can be defined according to the neighborhood relationship as below:

$$w_{ij}^{(1)} = \begin{cases} \exp(-\frac{\|x_i^{(1)} - x_j^{(1)}\|^2}{2\sigma^2}), & \text{if } x_i^{(1)} \in \mathcal{N}_k(x_j^{(1)}) \text{ or } x_j^{(1)} \in \mathcal{N}_k(x_i^{(1)}), \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

where $\mathcal{N}_k(x_j^{(1)})$ is the k-nearest neighbors of $x_j^{(1)}$. The Euclidean distance between samples $x_i^{(1)}$ and $x_j^{(1)}$ is used for finding nearest neighbors. σ is the bandwidth parameter which is set to $\sigma = 1$ in our experiments.

Through algebraic calculation, the objective function in Eq. (2) can be reformulated as:

$$\min_{\mathbf{P}_1} tr(\mathbf{P}_1^T \mathbf{X}^{(1)} \mathbf{L}_1 (\mathbf{X}^{(1)})^T \mathbf{P}_1) \tag{4}$$

where $\mathbf{L}_1$ is the Laplacian matrix, $\mathbf{L}_1 = \mathbf{D}_1 - \mathbf{W}^{(1)}$, $\mathbf{D}_1$ is a diagonal matrix, $\mathbf{D}_1(i,i) = \sum_j w_{ij}^{(1)}$. The elements of $\mathbf{W}^{(1)}$ are $w_{ij}^{(1)}$. $tr(\cdot)$ is the trace operator. Similarly, for modality $\mathbf{X}^{(2)}$, we can have:

$$\min_{\mathbf{P}_2} tr(\mathbf{P}_2^T \mathbf{X}^{(2)} \mathbf{L}_2 (\mathbf{X}^{(2)})^T \mathbf{P}_2) \tag{5}$$

where $\mathbf{L}_2$ is the Laplacian matrix of $\mathbf{X}^{(2)}$. Finally, combining the relationship modeling and local structure information preserving, the overall objective can be stated as follows:

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{V},\mathbf{P}_1,\mathbf{P}_2} \mathcal{L}_1 + \mathcal{L}_2 + \Omega(\mathbf{B},\mathbf{V},\mathbf{S},\mathbf{P}_1,\mathbf{P}_2) \\ &s.t. \quad \mathbf{B} \in \{-1,+1\}^{n \times k} \end{aligned} \tag{6}$$

where

$$\mathcal{L}_2 = \lambda_1 tr(\mathbf{P}_1^T \mathbf{X}^{(1)} \mathbf{L}_1 (\mathbf{X}^{(1)})^T \mathbf{P}_1) + \lambda_2 tr(\mathbf{P}_2^T \mathbf{X}^{(2)} \mathbf{L}_2 (\mathbf{X}^{(2)})^T \mathbf{P}_2) \tag{7}$$

where λ_1, λ_2 are balancing parameters. Inspired by [21], a regularization term is also integrated which is defined as:

$$\begin{aligned}&\Omega(\mathbf{B},\mathbf{V},\mathbf{S},\mathbf{P}_1,\mathbf{P}_2)\\&=\gamma\left\|\mathbf{V}\mathbf{S}\right\|_F^2+\mu\left\|\mathbf{B}\mathbf{S}\right\|_F^2+\gamma\mu\left\|\mathbf{V}\right\|_F^2+\xi_1\left\|\mathbf{P}_1\right\|_F^2+\xi_2\left\|\mathbf{P}_2\right\|_F^2\end{aligned}\tag{8}$$

where γ, μ, ξ_1, and ξ_2 are trade-off parameters.

3.3 Optimization

It is intractable to directly minimize the objective in Eq. (6) because of the non-convexity with four matrix variables $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{V}$, and $\mathbf{B}$. Fortunately, it is convex with respect to any of them when the others are fixed. Therefore, we employ an alternative optimization in an iterative manner to address the optimization problem until convergence. The detailed optimization steps are listed as follows:

Update $\mathbf{P}_1, \mathbf{P}_2$. Fix other variables but $\mathbf{P}_1$, then the objective function shown in Eq. (6) can be simplified as:

$$\begin{aligned}&\min_{\mathbf{P}_1}\lambda_1 tr(\mathbf{P}_1^T\mathbf{X}^{(1)}\mathbf{L}_1(\mathbf{X}^{(1)})^T\mathbf{P}_1)\\&+\alpha\left\|\mathbf{B}-(\mathbf{X}^{(1)})^T\mathbf{P}_1\right\|_F^2+\xi_1\left\|\mathbf{P}_1\right\|_F^2\end{aligned}\tag{9}$$

By setting its derivative w.r.t $\mathbf{P}_1$ to 0, we can have the closed-form solution stated as follows:

$$\mathbf{P}_1=(\alpha\mathbf{X}^{(1)}(\mathbf{X}^{(1)})^T+\lambda_1\mathbf{X}^{(1)}\mathbf{L}_1(\mathbf{X}^{(1)})^T+\xi_1\mathbf{I})^{-1}\alpha\mathbf{X}^{(1)}\mathbf{B}\tag{10}$$

Similarly, $\mathbf{P}_2$ can be updated by:

$$\mathbf{P}_2=(\beta\mathbf{X}^{(2)}(\mathbf{X}^{(2)})^T+\lambda_2\mathbf{X}^{(2)}\mathbf{L}_2(\mathbf{X}^{(2)})^T+\xi_2\mathbf{I})^{-1}\beta\mathbf{X}^{(2)}\mathbf{B}\tag{11}$$

Update V. The objective can be transformed to the following when fixing the other variables but $\mathbf{V}$:

$$\min_{\mathbf{V}}\ \left\|\mathbf{Y}-\mathbf{B}\mathbf{V}\mathbf{S}\right\|_F^2+\gamma\left\|\mathbf{V}\mathbf{S}\right\|_F^2+\mu\left\|\mathbf{B}\mathbf{V}\right\|_F^2+\gamma\mu\left\|\mathbf{V}\right\|_F^2\tag{12}$$

By setting its derivative w.r.t $\mathbf{V}$ to 0, we can have the closed-form solution stated as follows:

$$\mathbf{V}=(\mathbf{B}^T\mathbf{B}+\gamma\mathbf{I})^{-1}\mathbf{B}^T\mathbf{Y}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T+\mu\mathbf{I})^{-1}\tag{13}$$

Update B. Fixing other variables but $\mathbf{B}$, we can learn the unified binary codes of image and text directly without relaxation by solving the reformulated optimization stated as:

$$\begin{aligned}&\min_{\mathbf{B}}\ \left\|\mathbf{Y}-\mathbf{B}\mathbf{V}\mathbf{S}\right\|_F^2+\alpha\left\|\mathbf{B}-(\mathbf{X}^{(1)})^T\mathbf{P}_1\right\|_F^2\\&\quad+\beta\left\|\mathbf{B}-(\mathbf{X}^{(2)})^T\mathbf{P}_2\right\|_F^2+\mu\left\|\mathbf{B}\mathbf{V}\right\|_F^2\\&s.t.\quad\mathbf{B}\in\{-1,+1\}^{n\times k}\end{aligned}\tag{14}$$

The optimization defined in Eq. (14) under binary constraint can be easily solved by using discrete cyclic coordinate descent (DCC) method [22].

The proposed model is summarized in Algorithm 1. Through alternative optimization, the objective is minimized in each iterative step, and it will converge in the end.

4 Experiments

Extensive experiments are carried out in this section to evaluate the effectiveness of the proposed method in cross-modal retrieval, where some classes may not have been seen during training. Two tasks i.e. text to image (T2I) and image to text (I2T), are designed to validate the proposed approach in handling "seen" and "unseen" cross-modal retrieval. In our experiments, an image and a text are considered to be relevant if they share the same semantic label.

4.1 Datasets

Wiki [20] dataset consists of 2866 image-text documents. These documents can be grouped into 10 semantic categories. The images are described in 128-dimensional bag-of-visual words SIFT feature vectors, while text is represented by 10-dimensional topic vectors generated by the latent Dirichlet allocation (LDA) model.

Pascal VOC [7] dataset contains 9963 testing image-tag pairs, which can be classified into 20 categories. Since several image-tag pairs are multi-labeled, we select the pairs with only one label as the way in [24]. The image modality is represented by 512-dimensional GIST features [9], and the representations of text modality are 399-dimensional word frequency features.

LabelMe [17] dataset contains 2688 outdoor scenes from 8 different classes. We discard the words that occur in less than 3 times, resulting in 366 unique words. Thus, the representation of text is a 366-dimensional word frequency. The images are represented by 512-dimensional GIST features. Additionally, we delete the samples without tags, which results in a dataset with 2686 image-text pairs.

All the datasets are completely mutually exclusive, i.e. no overlapping samples between classes. In terms of attribute mapping $\mathbf{S}$, the word vectors of class names extracted from GloVe [19] are used in our experiments.

4.2 Settings

We construct the zero-shot scenario for three datasets as follows. For Wiki and LabelMe, we randomly select 2 classes as the unseen concepts each time, and the rest as seen classes. For Pascal dataset 4 classes are randomly selected as unseen classes each time. We report the average result of 10 experiments with randomly selected unseen concepts.

Algorithm 1. CMAH

Input: Seen cross-modal data $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, label $\mathbf{Y}$, attribute mapping $\mathbf{S}$, the length of hash codes k, and parameters α, β, λ_1, λ_2, ξ_1, ξ_2, μ, and γ.
Output: Unified hash codes $\mathbf{B}$, hashing functions $\mathbf{P}_1, \mathbf{P}_2$.

Compute Laplacian matrix $\mathbf{L}_1, \mathbf{L}_2$
Initialize $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{B}$, $\mathbf{V}$.
repeat
    Update $\mathbf{P}_1$, $\mathbf{P}_2$ by Eqs. (10) and (11);
    Update $\mathbf{V}$ by Eq. (13);
    Update $\mathbf{B}$ by solving Eq. (14);
until convergence
return $\mathbf{B}$, $\mathbf{P}_1$, $\mathbf{P}_2$

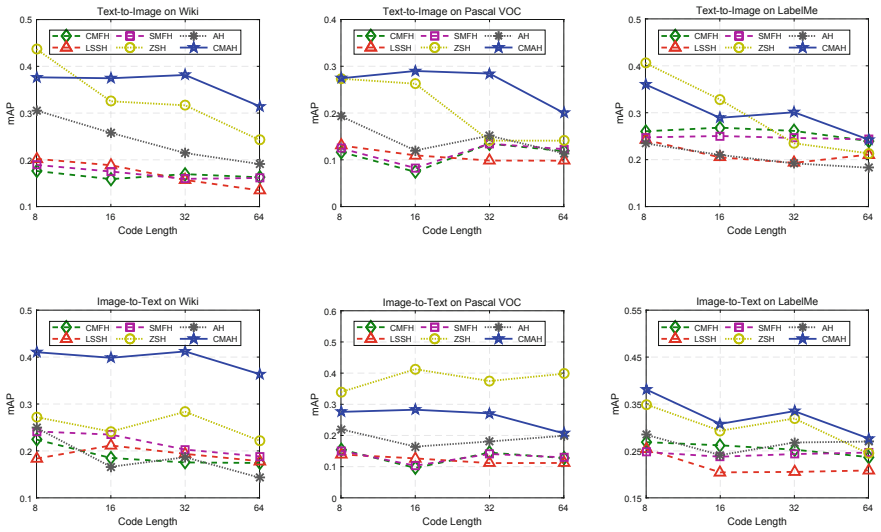

Fig. 2. MAP on unseen query of various approaches with varied hash code lengths on three datasets.

In order to evaluate the performance on handling unseen classes cross-modal retrieval. We use the testing data of unseen classes as query set, and we construct the retrieval set by merging the retrieval set of both seen and unseen classes, which follows the generalized zero-shot setting in [25]. In addition, the proposed new model should still have the comparable or even better performance on the seen classes. To this end, extra experiments of cross-modal retrieval are designed on the seen classes. The testing data of seen classes are chosen as query set, and the training data are regarded as retrieval set.

Two widely used metrics are employed to evaluate the performance of cross-modal retrieval. One is the mean Average Precision (mAP) based on Hamming ranking of all the retrieval set. The other is the mean precision within Hamming distance radius 2 (PH2) based on lookup table.

In our experiments, we empirically set α and β to 0.1. For balancing parameters λ_1 and λ_2, we set them to 0.1. The trade-off parameters γ, μ, ξ_1, and ξ_2 are set to 10^{-3}, 10^{-3}, 10, and 10, respectively. The Laplacian matrix is constructed within the 5 nearest neighbors. For the optimization procedure, we restrain the iteration number to 10.

4.3 Baselines

Since this is the first work of cross-modal hashing which considers the zero-shot scenario, we compare the proposed CMAH against five state-of-the-art methods including three conventional cross-modal hashing methods and two zero-shot hashing methods. The former three are CMFH [5], LSSH [33], and Supervised Matrix Factorization Hashing (SMFH) [23], respectively. The latter two are Attribute Hashing (AH) [27] and Zero-Shot Hashing (ZSH) [29]. In particular, for unimodal methods AH and ZSH, we only fix the length of hash codes such as 8 bits. Then hashing functions are trained independently on image and text modality. In the testing phase, the binary codes of text and image for query and retrieval are generated by the learned hashing functions respectively. The parameters of the baselines are set according to the suggestion of their original papers.

4.4 Experimental Results

Results on Unseen Query. Firstly, the performance of handling cross-modal retrieval for unseen concepts is evaluated from two aspects.

First, the mAP results that indicate the overall performance of cross-modal retrieval for unseen query are shown in Fig. 2. Compared against the conventional cross-modal hashing approaches CMFH, LSSH, and SMFH, the proposed method CMAH outperforms them with significant margins in most cases. This is because they are trained in a close-set circumstance, which makes it limited to generalize the hashing functions to newly emerged concepts that have never been present. We also notice that AH and ZSH perform better than CMFH, LSSH, and SMFH. Since they are excellent zero-shot learning methods, they can handle the knowledge transfer from seen to unseen classes. However, AH and ZSH are unimodal methods that ignore the correlation of different modalities. Thus, our CMAH outperforms AH and ZSH in most cases. It demonstrates the effectiveness of our proposed CMAH in handling cross-modal retrieval for unseen concepts.

Second, the precision within Hamming radius 2 (PH2) that indicates the local distribution performance which reveals how far the relevant instances is from the query item is plotted in Fig. 3. It can be seen that our method outperforms others in most cases except SMFH on the Labelme dataset. This outlier may be caused by the weak correlations between unseen and seen classes in LabelMe. Different from the mAP results, the PH2 of AH and ZSH are inferior than CMFH, LSSH, and SMFH. This is because the encoding of cross-modal correlation enhances

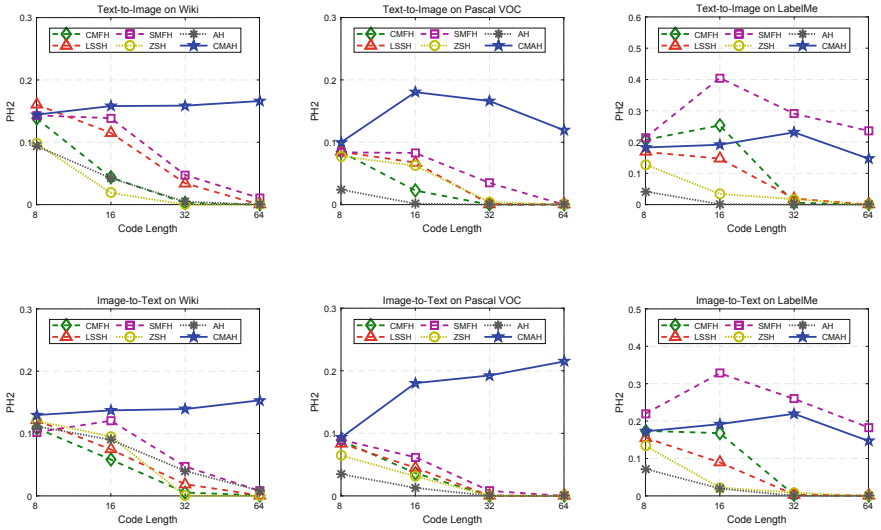

Fig. 3. Precision on unseen query of various approaches with varied hash code lengths on three datasets.

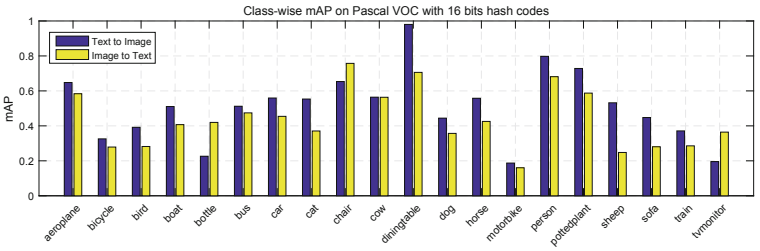

Fig. 4. Class-wise mAP of unseen query on Pascal VOC with 16 bits hash codes.

the performance of cross-modal methods, which is why the unimodal zero-shot methods AH and ZSH obtain higher mAP but lower precision in most cases.

Moreover, from Fig. 2, we can find that all the methods show a slightly trend toward degradation. The trend of PH2 of all baselines is similar to mAP but with more rapid decreasing speed. It is because the longer binary codes can carry more discriminant information but also introduces noise into the codes. In contrast, our method generally has a rising trend of PH2 from 8 bits to 32 bits. It demonstrates that the overall performance decreases slightly as the code length increases, but more relevant instances are distributed around the query item. It further depicts our cross-modal multi-layer network can enhance the robustness to noise. The proposed CMAH is thus able to generate hash codes for unseen query with high discriminative capability.

An additional observation to the hash code length is that as it increases, the mAP shows a slightly trend toward degradation, while the precision has a

rising trend from 8 bits to 16 bits and then decreases rapidly from 16 bits to 64 bits. This phenomenon indicates that an appropriate hash code length which can balance the information encoding and noise is significant for cross-modal retrieval for unseen concepts.

We also investigate the correlation between unseen and seen classes by conducting additional experiments on Pascal VOC. One class is selected as the unseen concept and the rest as seen concepts each time, and thus we have 20 different splits. The mAP results of cross-modal retrieval for the 20 unseen classes are shown in Fig. 4, respectively. Generally, the T2I task performs better than I2T task. More importantly, we find that the unseen class that shares more similar attributes with seen classes will lead to a better performance when conducting unseen query. For example, 'dining table' is quit close to four-leg 'chair', and 'cow' is similar to 'horse' and 'sheep'. Therefore, the selection of unseen classes will affect the performance of cross-modal retrieval for unseen concepts. For this reason, we present the average performance of repeated experiments with randomly selected unseen classes.

Results on Seen Query. Then, we still evaluate the performance of our method on seen query i.e. the same test setting with conventional cross-modal retrieval methods. The mAP results of all approaches are reported in Table 1. Similarly, the T2I task outperforms I2T tasks. This is because the representation of text feature is closer to the object semantic than the visual feature. Different from unseen query, it can be seen that AH and ZSH perform rather poorly, while CMFH, LSSH, SMFH perform well in the cross-modal retrieval for seen classes. This is because CMFH, SMFH and LSSH are trained in a close-set circumstance, which makes them limited to generalize the hashing functions to unseen concepts. We can see that the results of our method with varied code lengths are superior to AH and ZSH with large margin. Compared to the best cross-modal hashing method, our proposed CMAH performs comparable or even better on the three datasets. The reason is that our CMAH strives to achieve a balance between unseen and seen query. CMAH mainly focus on the knowledge transfer for unseen classes, which degrades slightly the discriminant of generated binary codes of seen classes. However, due to the utilization of attributes the binary codes can carry additional discrimination information, which results in superior performance in some cases such as on LabelMe and Pascal datasets. In addition, a rising trend is observed on seen query as the code length increases.

Overall Results. Finally, we analyze the average result of cross-modal retrieval for seen and unseen classes. The average results on three datasets are plotted in Fig. 5. Generally, it can be observed that our proposed CMAH is superior to the others, which demonstrates the effectiveness of our method. More over, the cross-modal methods CMFH, LSSH, SMFH, and our CMAH perform better than the unimodal method AH and ZSH. In the view of mAP, our method outperforms others in most cases except ZSH on Pascal with Image to Text task. The results of cross-modal methods increase steadily as the code length varies from 8 to

Table 1. MAP results on seen query of all methods with varied hash code lengths on three datasets.

Task	Method	Wiki				Pascal VOC				LabelMe			
		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	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	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	**0.6195**	**0.6653**	**0.6677**	**0.6874**	0.4231	0.7216	0.7553	0.8277	0.6438	0.7329	0.6951	0.6291
	ZSH	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	0.2294	0.2295	0.224	0.2074	0.1312	0.1283	0.1226	0.1319	0.2441	0.2386	0.2579	0.2656
	CMAH	0.5819	0.6256	0.6420	0.6282	**0.7208**	**0.8417**	**0.8564**	**0.8713**	**0.9151**	**0.926**	**0.9252**	**0.9222**
I2T	CMFH	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	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	**0.3089**	**0.3240**	**0.3414**	**0.3402**	0.2266	0.2828	0.2959	0.3245	0.5504	0.6221	0.6140	0.5607
	ZSH	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	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	0.2342	0.2329	0.2365	0.2503	0.1919	0.2020	0.1900	0.2351	**0.7694**	**0.8017**	**0.8136**	**0.8239**

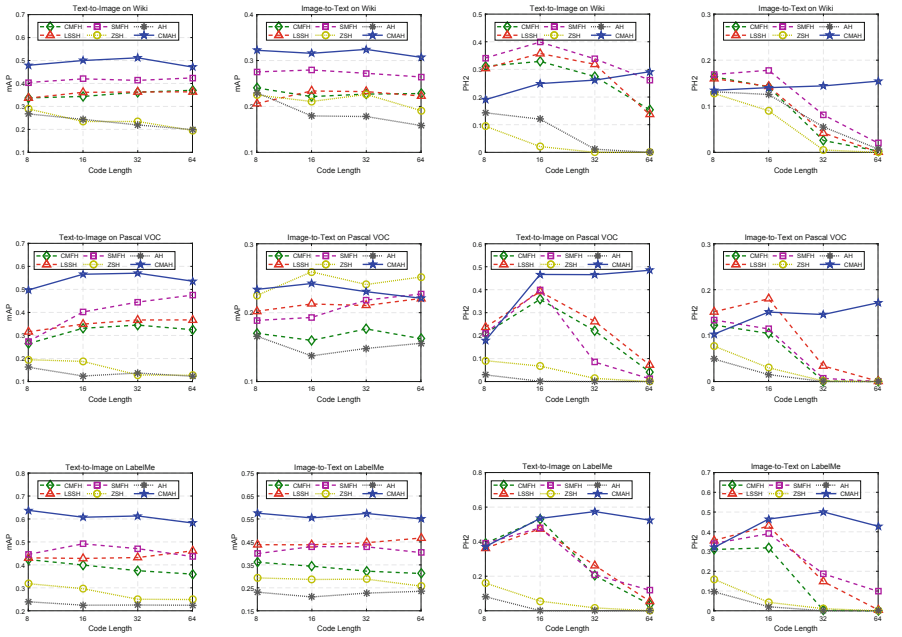

Fig. 5. Average results of seen and unseen query of various approaches with varied hash code lengths on three datasets.

64 bits. Whereas AH and ZSH present a slightly trend toward degradation. In terms of PH2, all the baselines reach their peaks at 16 bits and then decrease dramatically. The overall performance of our proposed CMAH shows a rising trend on three datasets.

Therefore, we can conclude that our novel model is effective and the competitive in cross-modal retrieval for both seen and unseen concepts.

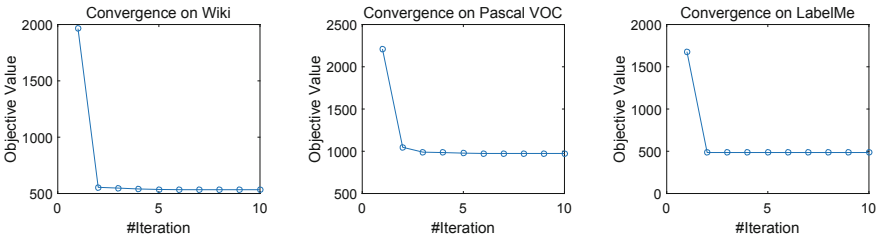

Fig. 6. Convergence analysis.

4.5 Convergence Analysis

The convergence of our method is evaluated via empirical experiments on Wiki, Pascal VOC, and LabelMe when hash code length is 16 bits. As shown in Fig. 6, our method can swiftly converge within 5 iterations, which demonstrates its efficiency in real-life applications.

5 Conclusion

In this paper, an exploration on the zero-shot problem in cross-modal retrieval is conducted. We proposed a novel cross-modal attribute hashing model that can generalize the hashing functions to newly emerged concepts. A dual multi-layer network is developed, where attribute plays a crucial role in not only helping to well transfer knowledge from seen to unseen concepts, but also narrowing down the semantic gap across visual features, text features, and class labels. Experiments demonstrated the effectiveness of our model in cross-modal retrieval for both seen and unseen concepts. With this initial exploration, many problems are still worthy of further investigation, such as the selection of unseen classes and the balance between seen query and unseen query.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China (2018YFC0831305) and in part by the Nature Science Foundation of China (61672123).

References

1. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
2. Cao, Y., Long, M., Wang, J., Liu, S.: Collective deep quantization for efficient cross-modal retrieval. In: AAAI, pp. 3974–3980 (2017)
3. Changpinyo, S., Chao, W.L., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. In: CVPR, pp. 5327–5336 (2016)
4. Chi, J., Huang, X., Peng, Y.: Zero-shot cross-media retrieval with external knowledge. In: Huet, B., Nie, L., Hong, R. (eds.) ICIMCS 2017. CCIS, vol. 819, pp. 200–211. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-8530-7_20

5. Ding, G., Guo, Y., Zhou, J.: Collective matrix factorization hashing for multimodal data. In: CVPR, pp. 2075–2082 (2014)
6. Ding, K., Fan, B., Huo, C., Xiang, S., Pan, C.: Cross-modal hashing via rank-order preserving. IEEE Trans. Multimedia **19**(3), 571–585 (2017). https://doi.org/10.1109/TMM.2016.2625747
7. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
8. Guo, Y., Ding, G., Han, J., Gao, Y.: SitNet: discrete similarity transfer network for zero-shot hashing. In: IJCAI, pp. 1767–1773 (2017)
9. Hwang, S.J., Grauman, K.: Reading between the lines: object localization using implicit cues from image tags. IEEE Trans. Pattern Anal. Mach. Intell. **34**(6), 1145–1158 (2012)
10. Ji, Z., Sun, Y., Yu, Y., Pang, Y., Han, J.: Attribute-guided network for cross-modal zero-shot hashing. arXiv preprint arXiv:1802.01943 (2018)
11. Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. In: CVPR, pp. 3270–3278 (2017)
12. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Unsupervised domain adaptation for zero-shot learning. In: CVPR, pp. 2452–2460 (2015)
13. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: CVPR, pp. 3174–3183 (2017)
14. Liu, H., Ji, R., Wu, Y., Hua, G.: Supervised matrix factorization for cross-modality hashing. In: IJCAI, pp. 1767–1773 (2016)
15. Liu, L., Lin, Z., Shao, L., Shen, F., Ding, G., Han, J.: Sequential discrete hashing for scalable cross-modality similarity retrieval. IEEE Trans. Image Process. **26**(1), 107–118 (2017)
16. Long, Y., Liu, L., Shao, L.: Towards fine-grained open zero-shot learning: inferring unseen visual features from attributes. In: IEEE Winter Conference on Applications of Computer Vision, pp. 944–952 (2017)
17. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
18. Pachori, S., Deshpande, A., Raman, S.: Hashing in the zero shot framework with domain adaptation. Neurocomputing **275**, 2137–2149 (2018)
19. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014)
20. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 251–260 (2010)
21. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning, pp. 2152–2161 (2015)
22. Shen, F., Shen, C., Liu, W., Tao Shen, H.: Supervised discrete hashing. In: CVPR, pp. 37–45 (2015)
23. Tang, J., Wang, K., Shao, L.: Supervised matrix factorization hashing for cross-modal retrieval. IEEE Trans. Image Process. **25**(7), 3157–3166 (2016)
24. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2010–2023 (2016)
25. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: CVPR, pp. 4582–4591 (2017)
26. Xu, X., Shen, F., Yang, Y., Zhang, D., Shen, H.T., Song, J.: Matrix tri-factorization with manifold regularizations for zero-shot learning. In: CVPR (2017)

27. Xu, Y., Yang, Y., Shen, F., Xu, X., Zhou, Y., Shen, H.T.: Attribute hashing for zero-shot image retrieval. In: IEEE International Conference on Multimedia and Expo, pp. 133–138 (2017)
28. Yang, E., Deng, C., Liu, W., Liu, X., Tao, D., Gao, X.: Pairwise relationship guided deep hashing for cross-modal retrieval. In: AAAI, pp. 1618–1625 (2017)
29. Yang, Y., Luo, Y., Chen, W., Shen, F., Shao, J., Shen, H.T.: Zero-shot hashing via transferring supervised knowledge. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 1286–1295 (2016)
30. Zhang, L., Ma, B., He, J., Li, G., Huang, Q., Tian, Q.: Adaptively unified semi-supervised learning for cross-modal retrieval. In: AAAI, pp. 3406–3412 (2017)
31. Zhang, L., Ma, B., Li, G., Huang, Q., Tian, Q.: Generalized semi-supervised and structured subspace learning for cross-modal retrieval. IEEE Trans. Multimedia **20**(1), 128–141 (2018)
32. Zhong, F., Chen, Z., Min, G.: Deep discrete cross-modal hashing for cross-media retrieval. Pattern Recogn. **83**, 64–77 (2018)
33. Zhou, J., Ding, G., Guo, Y.: Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th ACM International Conference on Research and Development in Information Retrieval, pp. 415–424 (2014)