Pattern Recognition 79 (2018) 147–161

Contents lists available at ScienceDirect

Pattern Recognition

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

Reconstruction-based supervised hashing

Xin Yuan[a,b,c], Zhixiang Chen[a,b,c], Jiwen Lu[a,b,c,*], Jianjiang Feng[a,b,c], Jie Zhou[a,b,c]

[a] Department of Automation, Tsinghua University, Beijing 100084, China
[b] State Key Lab of Intelligent Technologies and Systems, Beijing 100084, China
[c] Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China

ARTICLE INFO

Article history:
Received 1 September 2017
Revised 29 January 2018
Accepted 2 February 2018
Available online 6 February 2018

ABSTRACT

In the context of large scale similarity search, one promising technique is to encode high dimensional data as compact binary codes to take advantage of the speed and storage efficiencies. Many existing hashing approaches achieve similarity preservation in the Hamming space by preserving similarity relationship between data points. However, most of these methods only consider the relationship between points, which can not capture the data structure comprehensively. In this paper, we propose a reconstruction-based supervised hashing (RSH) method to learn compact binary codes with holistic structure preservation. The proposed method characterizes the similarity structure by the relationship between each data point and the structure generated by the remaining points. The learning objective is set to simultaneously minimize the distance between each point and the structure with the same class label and maximize the distance between each point and the structure with different class labels. In cross-modal retrieval, we propose a reconstruction-based hashing method by distilling the correlation structure in the common latent hamming space. The correlation structure characterizes the semantic correlation by the relationship between data points and structures in the common hamming space. Minimizing the reconstruction error of each single-modal latent model makes hidden layer outputs representative for the input of each modality. Experimental results in both single-modal and cross-modal datasets demonstrate the effectiveness of our methods when compared to several recently proposed approaches.

1. Introduction

With the rapid development of information technology, a large volume of images are generated everyday on the Internet and the amount has been extremely tremendous. In order to mine information from millions or billions images, large scale similarity search, which plays an essential role in a variety of domains(e.g. computer vision [1–10], machine learning [11,12], information retrieval [13–16], and data mining [17,18]), is employed to find the most similar instances to a query and has recently received considerable attention [19–21]. Efforts are made to deal with the scalable problem of similarity search from various aspects. Besides reducing the complexity of indexing with the tree-based like methods [22], the hashing technique is an emerging approach to encode inputs as low-dimensional binary vectors to benefit from the computation and storage efficiencies of Hamming codes.

Most existing hashing methods pursue similarity-preserving binary codes such that similar inputs are mapped to nearby binary codes. Then, the distances between binary codes are regarded as the approximation of the distances between inputs during retrieval. Locality Sensitive Hashing (LSH) [23] and its extension [24], which are based on randomized projections, are the notable examples in the case of no preassigned training dataset. Although such methods can maintain similarity structure in the original data space to some extent, long codes are required to achieve higher precision. This would lead to the problem of low recall and large storage cost. To deal with this dilemma, recently, learning based hashing methods are favored to integrate the power of machine learning and the knowledge of training data to seek binary codes with better similarity preservation. As stated in [3,25], the supervised information of the data labels is usually used to boost the performance of hash learning methods. Supervised learning based hashing methods exploit the knowledge of semantic affinities provided by data labels in different levels. For example, the single-point level semantic attributes are applied to generate the hash functions in [26]. In addition, pairwise supervision, i.e. similar pairs and dissimilar pairs, is utilized to learn binary codes and has been extensively studied in [3,20,21,27] to preserve the relationship between two points. To further exploit the ranking information on top of the pairwise label relationship, the pairwise proximity com-

[*] Corresponding author at: Department of Automation Room 626D Central Main Building, Beijing, China.
E-mail address: lujiwen@tsinghua.edu.cn (J. Lu).

https://doi.org/10.1016/j.patcog.2018.02.003
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

parison among three data points is utilized to learn hashing functions in the form of triplet ranking [28]. Moreover, the listwise information, which indicates the rank order of a set of points with respect to the query point, is considered in the procedure of the binary code learning to preserve the ranking [29]. However, most of these learning based hashing methods only consider the relationship between points, which can not well capture the data structure of samples. On the consideration of the essential role of similarity preservation in the hashing learning procedure, it is fundamentally important to preserve the similarity relationship with comprehensive capture of the data structure.

In similarity search involving cross-modal data, the main challenge is how to preserve the cross-modal semantic correlation in the common hamming space. Many existing cross-modal methods [30–37] project data across different modalities into a common Hamming space and discover the correlations for fast cross-modal similarity search. Similarly, cross-modal hashing methods can be categorized into unsupervised methods and supervised methods according the utilized information. Unsupervised methods are more general and can be trained without labels. For example, co-occurrence information in cross-modal data is utilized to learn common representations. To obtain more discriminative representations, supervised methods explore the semantic labels to keep samples with the same class close, while to push samples with different label faraway. For example, some existing methods utilize similar pairs (or dissimilar pairs) to learn common representations. For the rank based methods, which study the cross-modal retrieval as a ranking problem, rank lists are utilized to learn common representations. However, most existing methods only consider to distill the relationship between data points in the common space, which can't exploit complex correlation across different modalities with holistic similarity preservation. Hence it remains a problem how to distill the correlation for cross-modal data.

In this paper, we propose a reconstruction-based supervised hashing method (RSH) to learn discriminative binary codes for large scale similarity search. To capture the data structure comprehensively, we characterize the similarity structure by the relationships between each data point and the set of the remaining samples. As shown in Fig. 1(a), the objective of the proposed hashing learning problem is set as the goal that binary codes of similar data points are closer to each other and binary codes of dissimilar data points are far away from each other. Additionally, the supervised information of class labels is further exploited to enhance the discriminative power of the learned binary codes. Specifically, the learning objective is set to minimize the reconstruction loss function, which simultaneously minimizes the distance between each point and the set of points in the same class and maximizes the distances between each point and the sets of points of different class labels. To deal with the cross-modal data, we explore both feature representation and semantic correlation across different modalities simultaneously. We develop a reconstruction-based method for our cross-modal retrieval system. The method characterizes the semantic correlation by the relationship between each data point and generated structure in the common latent space. As illustrated in Fig. 1(b), a reconstruction-based objective is designed to preserve the correlation structure between the hidden representations in different modalities. In addition, minimizing the reconstruction error of each single-modal latent model makes the outputs in the common hamming space representative for the input of each modality. To evaluate the effectiveness of the proposed hashing method, we conduct extensive experiments on both single-modal and cross-modal datasets. Experimental results demonstrate the effectiveness of our proposed methods show when compared to prior representative hashing methods.

2. Related work

In this section, we give a brief review of the related work in two categories. In Section 2.1, we review some existing single-modal hashing learning methods. In Section 2.2, we review several cross-modal hashing methods.

2.1. Single-modal hash learning

Recent years have witnessed the success of single modality hashing techniques in various real-world tasks. Existing single modality hashing methods can be classified into three categories: unsupervised, supervised and semi-supervised. The representatives of each category have achieved promising performance in many applications.

The unsupervised hashing methods learn from the data attributes like distributions and structures. For example, spectral hashing [38] formulates a graph partitioning problem and approximately solved the problem with the assumption of the uniform data distribution. A graph-based hashing method [39] is proposed to approximate the neighborhood using the anchor graph. KMH [40] is proposed to simultaneously perform k-means clustering and learn the binary indices of the quantized cells. ITQ [13] is an iterative quantization method to minimize quantization loss by seeking a rotation matrix. Inductive manifold method in [41] is proposed to learn binary embeddings from the cluster centers into a low-dimension manifold.

To enhance the discriminative power of the learning method, Many supervised hashing methods have been proposed in recent years. For example, the kernalized supervised hashing [3] is proposed to utilize the equivalence between code inner products and Hamming distances, which aims to keep the inner product of hash codes consistent with the pairwise supervision. Fast supervised hashing [42] which is based on boosted decision trees adopts an iteratively alternative optimization on a subset of the binary codes. Supervised discrete hashing [43] reformulates the discrete optimization objective by introducing an auxiliary variable and uses a existing kernel based hashing function to learn binary codes. The supervised version of deep hashing in [44] takes supervision information to seek nonlinear transformations from the continuous output to the compact binary codes.

Since the labels can be noisy and sparse, supervised methods may suffer from the overfitting problem. Semi-supervised hashing is developed, which aims to avoid the overfitting problem to some extent. For example, SSH [20] is a semi-supervised hashing method to minimize the empirical error for pairwise labeled training data and maximize the variances of all the training data. Binary reconstructive embedding [21] aims to minimize the reconstructive error between the learned Hamming distance and the Euclidean distance. Minimal loss hashing [45] formulates the hashing problem using a hinge-like loss function to minimize the loss between the Hamming distance and the quantization error. However, most existing semi-supervised methods only use the pairwise supervision, which may not capture the local structure of the data points.

2.2. Cross-modal hash learning

In cross-modal hashing, it takes one type of data as the query to retrieve relevant data of another type. For example, given a textual description, one would like to retrieve some relevant pictures or videos. To deal with the data across different modalities, cross-modal hashing has attracted more research attention recently. Existing cross-modal hashing methods can be roughly categorized into unsupervised methods and supervised methods. Cross-View Hashing (CVH) [30] extends spectral hashing to the cross-modal