Pattern Recognition 83 (2018) 64–77

Contents lists available at ScienceDirect

Pattern Recognition

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

Deep Discrete Cross-Modal Hashing for Cross-Media Retrieval

Fangming Zhong[a], Zhikui Chen[a,b,*], Geyong Min[c]

[a] School of Software, Dalian University of Technology, Dalian, China
[b] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, China
[c] College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

ARTICLE INFO

Article history:
Received 30 November 2017
Revised 2 May 2018
Accepted 20 May 2018
Available online 21 May 2018

Keywords:
Cross-modal retrieval
deep learning
discrete hashing
alternative optimization

ABSTRACT

Cross-modal hashing has drawn increasing research interests in multimedia retrieval due to the explosive growth of multimedia big data. It is such a challenging topic due to the heterogeneity gap and high storage cost. However, most of the previous methods based on conventional linear projections and relaxation scheme fail to capture the nonlinear relationship among samples and suffers from large quantization loss, which result in an unsatisfactory performance of cross-modal retrieval. To address these issues, this paper is dedicated to learning discrete nonlinear hash functions by deep learning. A novel framework of cross-modal deep neural networks is proposed to learn binary codes directly. We formulate the similarity preserving in the framework, and also bit-independent as well as binary constraints are imposed on the hash codes. Specifically, we consider intra-modality similarity preserving at each hidden layer of the networks. Inter-modality similarity preserving is formulated by the output of each individual network. By so doing, the cross correlation can be encoded into the network training (i.e. hash functions learning) by back propagation algorithm. The final objective is solved by alternative optimization in an iterative fashion. Experimental results on four datasets i.e. NUS-WIDE, MIR Flickr, Pascal VOC, and LabelMe demonstrate the effectiveness of the proposed method, which is significantly superior to state-of-the-art cross-modal hashing approaches.

1. Introduction

Recently, multimedia retrieval has gained considerable attention, due to the explosive growth of multimedia data on the Internet, especially in the social network with social media big data being produced by users. As a significant component of multimedia retrieval, cross-modal retrieval has drawn increasing interests in many real life applications, such as visual search [1], image captioning [2], and machine translation [3]. However, the heterogeneity gap [4] among various modality data, such as images and texts, makes it difficult to perform Approximate Nearest Neighbor (ANN) search. In addition, the cross-modal retrieval in large-scale and high-dimensional datasets becomes quite challenging due to the high storage cost and computational complexity. To address these challenges, hashing has attracted increasing research interests due to its effectiveness in reducing storage cost and improving retrieval speed. The goal of hashing is to learn hash functions which can map high-dimensional data to a Hamming space in which data are represented by compact binary codes. Thus, the similarities in ANN search can be measured by Hamming distances, which can be obtained efficiently via bit-wise XOR operation in the expected Hamming space [5]. Hence, cross-modal hashing (CMH), which projects different modalities data into a common Hamming space enabling cross-modal retrieval is receiving more and more attentions. It is crucial that learning discriminative hash functions for each modality to boost the cross-modal retrieval.

There is a wide range of cross-modal hashing approaches proposed in recent years [5–15]. Cross-view hashing (CVH) [11] formulated the problem of learning hash functions as a generalized eigenvalue problem. Most recently, Collective Matrix Factorization Hashing (CMFH) [12] uses matrix factorization to learn the latent concepts from each modality which has achieved an impressive result on cross-modal retrieval. Inspired by CMFH, several extensions based on matrix factorization have been proposed to formulate the supervised label information, such as supervised matrix factorization hashing (SMFH [5], SMFCMH [13]), cluster-based joint matrix factorization (C-JMFH) [16], and Supervised CMFH (SCMFH) [17] etc. In particular, SMFCMH [13] integrates the graph regularization into the collective non-negative matric factorization. Furthermore, label information is used to refine the graph regularizer. With the supervised label information being taken in to account, SMFCMH learns more discriminative hash codes. Zhou et al.

* Corresponding author.
E-mail addresses: fmzhong@mail.dlut.edu.cn (F. Zhong), zkchen@dlut.edu.cn (Z. Chen).

https://doi.org/10.1016/j.patcog.2018.05.018
0031-3203/

[18] proposed Latent Semantic Sparse Hashing (LSSH), 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. Although a number of efforts have been made on achieving impressive performance on cross-modal retrieval, there are still several limitations that remain to be exploited in cross-modal hashing.

One fundamental limitation of these methods is that the binary constraints of hash codes are relaxed to real values. In order to learn compact binary codes, the discrete constraints are imposed to most of the existing object functions. However, the object function with mixed-integer optimization results in an NP-hard optimization problem. To simplify the optimization involved in the binary hash functions learning, most of them discard the discrete constraints and then address the problem in a real-valued space with continuous solution. Then, the binary codes are obtained by quantizing the continuous solution. Unfortunately, using such a relaxation scheme will lead to large quantization loss. In such case, the accumulated quantization error, especially when learning long binary codes, is bound to degrade the discrimination capability of binary codes. To this end, it is crucial to learn binary codes directly and design discrete hash functions that can minimize the quantization error.

Another limitation of the recent methods is that they fail to formulate the nonlinear relationship of instances in each modality. Most of the existing cross-modal hashing methods [5,12,13,16,17] impose linear transformations as hash functions that project data from the original high-dimensional space into a lower dimensional Hamming space. Although impressive performances are achieved, however, in many applications, the data are linearly inseparable and thus the nonlinear manifold structure cannot be well captured by such simple linear projections. Therefore, to encourage the learned hash functions for encoding the nonlinear relationship of samples, it is an urgent required to design innovative nonlinear hash functions.

To tackle the nonlinear hashing challenge, several promising approaches introduced deep learning to cross-modal retrieval, [19–23]. However, few work explored the similarity preservation including intra-modality and inter-modality similarities in the learning phase of hash functions. Generally, the instances from different modalities while describing the same semantic object should share similar hash codes. In contrast, the samples representing different semantic objects are supposed to be pushed far away with dissimilar binary codes.

To address the above challenges, we propose a discrete cross-modal hashing based on deep neural network which is termed Deep Discrete Cross-Modal Hashing (DDCMH). We learn the compact binary codes directly by formulating it as a quantization optimization problem incorporated into the final object function. Owing to the successful application in feature extraction in computer vision of deep learning, it is reliable to learn high level nonlinear hash functions by using deep neural network. Therefore, we develop a new framework of cross-modal deep neural networks to seek multiple hierarchical nonlinear transformations to jointly learn compact binary codes and nonlinear hash functions. By so doing, the nonlinear relationship of instances in each modality would be well captured. Furthermore, we consider the similarity preservation in the learning of hash functions. We embed the intra-modality similarity preservation into every hidden layer of each modality. Additionally, the inter-modality preservation is formulated between the outputs of two deep networks at the top layers. Thus, the cross-modal correlation will be incorporated into the updating of deep neural networks by back propagation algorithm, which encourages the learned networks i.e. hash functions, to be more discriminative. Moreover, inspired by [24,25] that learn binary codes using classifiers, we integrate a linear classification with expected binary codes as input and label information as output. The framework of DDCMH is illustrated in Fig. 1. We employ the visual and text features extracted from images and texts as input of our method. In the testing phase, a query is also firstly transformed to visual or text representations followed by hash codes learning. Finally, cross-modal retrieval can be conducted based on the generated binary codes.

Motivated by nonlinear discrete hashing (NDH) [25], our proposed DDCMH is also optimized under four constraints: 1) quantization loss of learned binary codes and the output of cross-modal deep neural networks, 2) intra-modality similarity preservation in each hidden layer of each modality and inter-modality similarity preservation between the output layers of two networks, 3) independent bits in the learned binary codes, and 4) minimized the classification loss between expected binary codes and supervised label information. Comparing against the previous proposed methods, the main contributions of our work are summarized as follows:

- We develop a novel framework of cross-modal deep neural networks. Two deep neural networks are trained as hash functions to learn binary codes for image and text modality, respectively. Different from linear transformations, such cross-modal deep neural networks can well capture the nonlinear relationship in each modality.
- Binary codes of training data are learned directly without any relaxation. Furthermore, a quantization loss between deep neural networks and to-be-learned binary codes are imposed to minimize the quantization error. By so doing, we can jointly learn binary codes and hash functions with low quantization loss, which is especially significant for out-of-sample instances.
- Additionally, intra-modality and inter-modality similarity preservation is considered in the learning of nonlinear hash functions and compact binary codes. We introduced a graph regularization in each layer of the deep network to preserve intra-modality similarity. Similarly, a graph regularization based on semantic label information is imposed into the output layer of deep neural network for preserving inter-modality similarity. Moreover, the inter-modal similarity preservation will contribute to the whole network updating by back propagation algorithm. Hence, the learned binary codes will possess more discriminative power.

The rest of this paper is organized as follows. The previous work on cross-modal hashing is reviewed and analyzed in Section 2. Section 3 presents the detailed design of our proposed method. In Section 4, extensive experimental details and results are described in comparison with the state-of-the-art methods on four benchmark datasets. Finally, this work is concluded in Section 5.

2. Related Work

Due to the efficiency of retrieval and low storage cost, hashing methods are widely investigated in both unimodal hashing and cross-modal hashing. In the last decade, a number of cross-modal hashing approaches have been proposed due to the increasing interest attracted by cross-modal retrieval in various applications. In this section, we will review and analyze the main differences of these methods from three aspects: 1) inter-modality and intra-modality similarity preserving, 2) using of relaxation scheme, and 3) nonlinear relationship capturing.