Neurocomputing 193 (2016) 250–259

Contents lists available at ScienceDirect

Neurocomputing

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

# Semantic consistency hashing for cross-modal retrieval ☆

Tao Yao [a,b], Xiangwei Kong [a,*], Haiyan Fu [a], Qi Tian [c]

[a] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China
[b] Department of Information and Electrical Engineering, LuDong University, Yantai 264025, China
[c] Department of Computer Science, University of Texas at San Antonio, San Antonio 78249, USA

ARTICLE INFO

*Article history:*
Received 19 October 2015
Received in revised form
23 December 2015
Accepted 8 February 2016
Communicated by Tao Mei
Available online 20 February 2016

*Keywords:*
Cross-modal retrieval
Semantic consistency
Hashing
Non-negative matrix factorization
Neighbor preserving

ABSTRACT

The task of cross-modal retrieval is to query similar objects in dataset of multi-modality, such as using text to query images and vice versa. However, most of existing works suffer from high computational complexity and storage cost in large-scale applications. Recently, hashing method mapping the high-dimensional data to compact binary codes has attracted a lot of concerns due to its efficiency and low storage cost over large-scale dataset. In this paper, we propose a Semantic Consistency Hashing (SCH) method for cross-modal retrieval. SCH learns a shared semantic space simultaneously taking both inter-modal and intra-modal semantic correlations into account. In order to preserve the inter-modal semantic consistency, an identical representation is learned using non-negative matrix factorization for the samples with different modalities. Meanwhile, neighbor preserving algorithm is adopted to preserve the semantic consistency in each modality. In addition, an effective optimal algorithm is proposed to reduce the time complexity from traditional $O(N^2)$ or higher to $O(N)$. Extensive experiments on two public datasets demonstrate that the proposed approach significantly outperforms the existing schemes.



## 1. Introduction

With the rapid development of information technology and the Internet, one webpage may contain text, audio, image, video and so on. Although these data are represented by different modalities, they have strong semantic correlation. For example, Fig. 1 displays a number of documents collected from Wikipedia. Each document includes one figure along with surrounding texts. These pairwise images and texts are connected by blue solid line denoting that they have strong semantic correlation. And those connected by blue dotted line mean that the image is relevant to these texts, i.e. they have the same semantic concept, while those connected by red dotted line denote that they are irrelevant to each other. The task of cross-modal retrieval is using one kind of media to retrieve similar samples in dataset of different modalities, and the returned samples are ranked by the correlation. However, with the explosive growth of multimedia on the Internet, storage cost and efficiency are two main challenges in large-scale retrieval.

Hashing method mapping sample from high-dimensional feature space to low-dimensional binary Hamming space has been received much attention due to its efficiency and low memory cost [1–13]. However, most of existing hashing schemes can work only on single modality [1–6]. There have been only a few works addressing multi-modal retrieval so far [7,9–12]. Multi-modal hashing generally can be categorized into two types: multi-modal fusion hashing (MMFH) and cross-modal hashing (CMH). MMFH aims at generating better binary codes by taking advantage of the complementarity of each modality than single modal hashing [7]. While the CMH method is to construct a shared Hamming space to retrieve similar samples over heterogeneous cross-modal dataset [9–12]. In this paper, we focus on CMH method. The key point of CMH is to find the correlation between different modalities in Hamming space. However, how to learn a low-dimensional Hamming space over heterogeneous cross-modal dataset is still a challenging issue.

There have been many recent works focus on this issue. For example, Canonical Correlation Analysis (CCA) hashing method maps the sample from different modalities to a low-dimensional Hamming space by maximizing the correlation between different modalities [14]. Multimodal latent binary embedding (MLBE) employs a binary latent factor with a probabilistic model to learn hashing codes [15]. Co-Regularized Hashing (CRH) is proposed to learn a low-dimensional hamming space by mapping the data far from zero for each bit, and inter-modal similarity is effectively preserved at the same time [8]. Multimodal NN hashing (MM-NNH) proposed in [16] aims at learning a group of hashing functions by preserving intra-modal and inter-modal similarity. However, above cross-modal hashing methods directly learn hashing functions respectively for each modality. It may degrade the

☆Fully documented templates are available in the elsarticle package on CTAN http://www.ctan.org/tex-archive/macros/latex/contrib/elsarticle.

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2016.02.016
0925-2312/

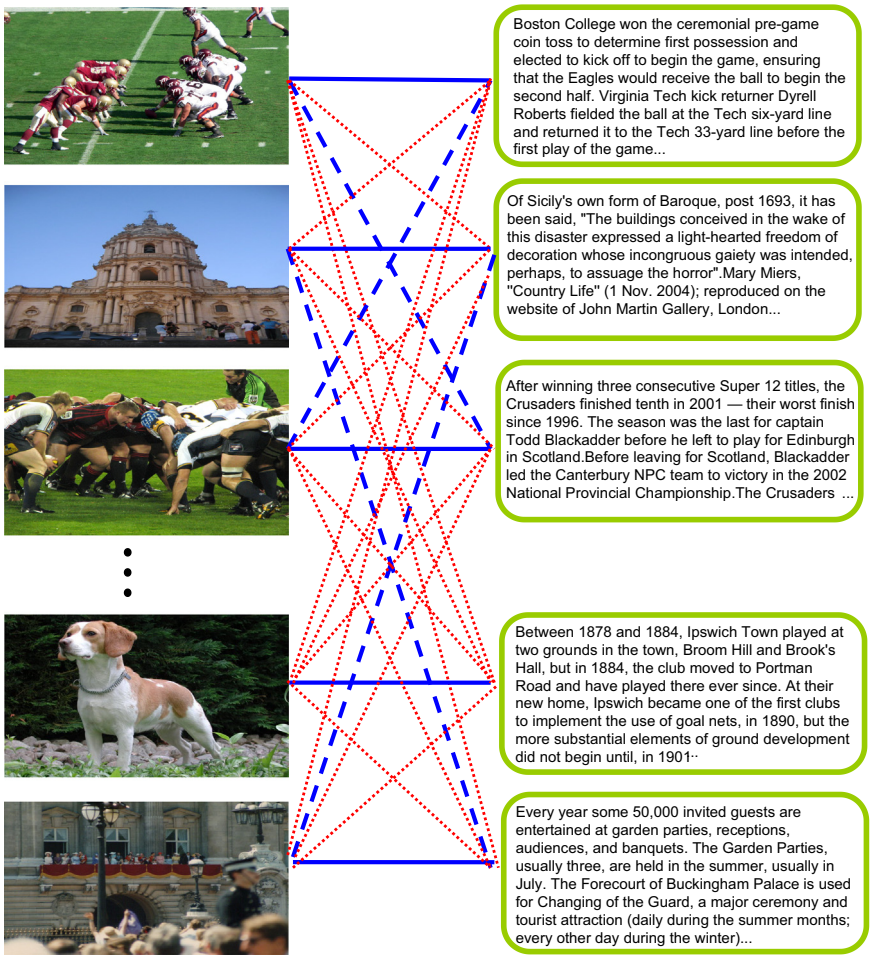


**Fig. 1.** A number of examples are collected from Wikipedia. The correlations between images and texts are denoted by different lines and colors. The blue solid line represents that the image and texts have strong semantic correlation with each other. The blue dashed line represents that they are relevant to each other. The red dotted line represents that they are irrelevant to each other. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

performance because the learned Hamming space is not semantically distinguishing.

To address above issue, a supervised scheme (SliM2) [17] is proposed to embed heterogeneous data into a semantic space by dictionary learning and sparse coding. In [18], Latent Semantic Sparse Hashing (LSSH) algorithm is proposed to learn a semantic space by sparse coding and matrix factorization. Sparse coding is used to capture the salient structure of text, and matrix factorization is used to learn the latent concept for image. At last, learning a linear mapping matrix bridges the semantic space between the text and the image. Collective Matrix Factorization Hashing (CMFH) [19] intends to project sample to a common semantic space by collective matrix factorization, thus inter-modal semantic similarity is preserved effectively. The results of above methods prove that learning a semantic space is helpful to cross-modal retrieval. However, those methods only consider to preserve inter-modal semantic consistency, but ignore to preserve intra-modal semantic consistency. Inter-modal semantic consistency aims at preserving the global similarity structure, while intra-modal semantic consistency aims at preserving the local similarity structure for each modality in the learned low-dimensional semantic space. Moreover, recent studies have proved that samples from high-dimensional space actually lie on a low-dimensional manifold in real-world [20,21]. Hence it will be beneficial for introducing the intra-modal semantic consistency to cross-modal retrieval framework.

In this paper, we put forward a semantic consistency hashing method for cross-modal retrieval. We aim to efficiently learn binary codes for different modalities by jointing intra-modal and inter-modal semantic consistency into a framework. In order to preserve inter-modal semantic consistency, an identical representation is learned by non-negative matrix factorization (NMF) for the samples with different modalities. The main advantages of NMF are as follows: (1) Nonnegative representation is consistent with the cognition of human brain [22,23]. (2) The constraint of non-negative brings sparse, and relatively sparse representation can resist noise to a certain extent [24], which will be beneficial for learning the shared semantic space with the noisy labels. In order to preserve intra-modal semantic consistency, neighbor preserving algorithm is utilized to preserve the local similarity structure. This allows to exploit richer information existing in data to learn a better shared semantic space.

Our main contributions are as follows:

1. We propose a semantic consistency hashing method to effectively find the semantic correlation between different modalities in the shared semantic space. Not only the inter-modal semantic consistency is preserved by NMF, but also the intra-modal semantic consistency is preserved by neighbor preserving algorithm in the shared semantic space.
2. We propose an efficient and iterative optimization framework. In experiments, we find that satisfactory performance can be achieved in about 10–20 iterations. Meanwhile, the training