ARTICLE IN PRESS

Neurocomputing (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

# Diversity Regularized Latent Semantic Match for Hashing

Yong Chen[a], Hui Zhang[a], Yongxin Tong[a], Ming Lu[b,*]

[a] School of Computer Science and Engineering, Beihang University, Beijing 100191, PR China
[b] IT FLEX, Intel China Research Center, Beijing 100080, PR China

ARTICLE INFO

Communicated by Rongrong Ji

*Keywords:*
Diversity regularizations
Soft orthogonality
Learning to match
Multimodal retrieval
Representation learning
Approximate nearest neighbors

ABSTRACT

Hashing based approximate nearest neighbors (ANN) search has drawn considerable attraction owing to its low-memory storage and hardware-level logical computing which is doomed to be greatly applicable to quantities of large-scale and practical scenarios, such as information retrieval, computer vision and natural language processing. However, most existing hashing methods concentrate either on images only or on pairwise image-texts (labels, short documents) and rarely utilize more common sentences. In this paper, we propose D iversity R egularized L atent S emantic M atch for H ashing (DRLSMH), a new multimodal hashing method that projects images and sentences into a shared latent semantic space with label-supervised semantic constraints to proceed on multimodal retrieval. Notably, soft orthogonality is induced as a novel regularizer to preserve diverse hashing functions for compact and accurate representations; what's more, this kind of regularization also benefits the derivations of closed-form solutions with some proper relaxations under iterative optimization framework. Extensive experiments on two public datasets demonstrate the advantages of our method over some state-of-the-art baselines under cross-modal retrieval both on image-query-image, image-query-text and text-query-image tasks.

## 1. Introduction

Nearest neighbors (NN) search has acted as a fundamental role in lots of important applications, such as machine learning, computer vision, natural language processing and so forth for decades [1–3]; however, recently ever-changing Internet technologies have already pushed forward the big data era to come: high-dimensional, massive, and heterogeneous data throw a huge challenge on NN. Even for the simplest linearly scanning, it would be impractical and unrealistic for real scenarios now. Hashing, as a new approximate nearest neighbors method, embedding data into the binary hamming space which is capable of preserving similarities between objects makes the memory and computing both extremely effective [4,5]; even ordinary PCs can handle large amounts of data.

Hashing methods can be divided into different classifications according to different views. For example, it can be roughly divided into data-independent methods and data-dependent methods by using the data or not, where LSH [6], KLSH [7] and other LSH-like methods [8] are data-independent and ITQ [9], SpH [10], SSH [11] and MLH [12] are data-dependent ones. From another perspective of using supervised information or not, there could be three kinds: unsupervised [6,13], supervised [5,12] and semi-supervised [11,14,15] methods. Here, we would like to divide the hashing methods into traditional image retrieval methods and current multi-view cross-modal retrieval methods.

Methods mentioned above all belong to the former kind. And as regard to the latter one, there are many new methods emerging in the recent years. Inter-media hashing (IMH) [16] introduces inter-media consistency and intra-media consistency to discovery a common hamming space, and uses regularized linear model to learn view specific hash functions. However, IMH needs to construct the similarity matrix for all the data points, which will impede the effectiveness for large-scale datasets. Latent semantic sparse hashing (LSSH) [17] utilizes the sparse coding to capture the salient structures of images and matrix factorizations to learn the latent concepts from text to perform cross-modal similarity search. However, this kind of learning paradigm, especially the sparsity, makes the training stage consume too much time. Collective matrix factorization hashing (CMFH) [18] learns unified hash codes by collective matrix factorization with latent semantic match model from different modes of one instance, while it's too strict to constraint different modalities to identical hash codes. Semantic topic multimodal hashing (STMH) [19] models text as multiple semantic topics and image as latent semantic structures and then learns the relationship of text and image into their latent semantic spaces. Though STMH has obtained superior performances to some state-of-the-art baselines, we find the extension of out-of-sample need to be simplified.

Although there are many multimodal hashing methods and they all have achieved promising performance in multimodal applications [16–

[*] Corresponding author.
*E-mail addresses:* hzhang@buaa.edu.cn (H. Zhang), ming-lu@outlook.com (M. Lu).

http://dx.doi.org/10.1016/j.neucom.2016.11.057
Received 9 April 2016; Received in revised form 5 September 2016; Accepted 30 November 2016

19], there still needs to be more explorations on models (linear/nonlinear, matrix factorization/probabilistic graphical modes, deep neural network or not), algorithms (convex/nonconvex, distributed parallel gradient-based algorithms) and theories (robustness, sparsity, diversity or low rank), or even for some new formalizations of multi-modal data. In this work, we make full use of the self-characterized image-sentences pairwise data, and map them diversely into a shared latent semantic space via match learning with label-supervised semantic regularizations which is able to preserve similarities between images and sentences, and then put forward a novel method Diversity Regularized Latent Semantic Match for Hashing (DRLSMH). The core contributions of our work can be listed as below:

- We incorporate linear projection instead of direct matrix factorization with learning to match framework, which would definitely lead to two advantages: on one hand, it makes the model look simple (more like convex), and more importantly it would greatly benefit the hashing for out-of-samples just through basic matrix-vector multiplications; on the other hand, this kind of formalizations help to the later closed-form solutions.
- Soft orthogonality is introduced as a novel regularizer for diverse hashing functions, which will provide compact and accurate representations with small fixed number of hash bits. Moreover, closed-form solutions can be easily derived with some relaxations on the regularizations under the iterative framework.
- To the best of our knowledge, this is the pioneer exploration to perform learning to hash for cross-modal retrieval tasks on such kind of datasets: pair-wise image-sentences corpus. Extensive experiments on two public datasets highlight the superiority over some of the state-of-the-art methods for image-query-image, image-query-text and text-query-image missions.

The remainder of this paper is organized as follows. In Section 2, we introduce related work about diversity regularizations, learning to match and deep learning for representations. In Section 3, we define our problem and give necessary notations. In Section 4, we propose our method DRLSMH and present an approximate learning process for match learning and then derive the optimization algorithms. We conduct experiments on three kinds of tasks to evaluate the proposed models in Section 5, and finally draw conclusions in Section 6.

## 2. Related work

### 2.1. Diversity regularizations

Very recently, it's quite interesting that there seems a more and more growing attention on the diversity regularizations explored in various aspects of data mining and machine learning, such as ensemble methods, self-paced learning, metric learning, multi-view clustering and so on, without any prior consolations. And lots of superior performances are mined out with the utilizations of diversity constraints in different formalizations. For example, [20] proposed the diversity regularized machine to construct an ensemble of diverse SVMs which lead to an effective reduction on its hypothesis space complexity and better generation ability verified both in theoretical analysis and experiments; [21] threw focus on the preferences both easy and diverse samples into a general non-convex regularizer which would greatly contribute to the self-pace learning; [22] discussed about the tasks of keeping a small number of latent factors meanwhile making them as effective as a large set of factors for the sake of computational efficiency and put forward an diversity constraints with the mean and variance of latent factors, and then learned compact and effective distance metrics for retrieval, clustering and classifications; last but not the least, [23] utilized the Hilbert Schmidt Independence Criterion as a diversity term to explore the complementarity of multi-view representations that could explicitly enforce the learned subspace to be novel with each other for better clustering.

Definitely, diversity is an intuitive and effective idea to be taken advantage of for its compact and effective information presentations in large scale data. However, it's still an open research problem both in wide varieties of tasks, formalizations, algorithms and its theoretical analysis. Here, soft orthogonal constraints are induced on projection matrices as a novel diversity regularizer to obtain diverse hash functions (another perspective different from the former works) for compact representations of both image and text data in our paper. Soft orthogonality not only can achieve comparable effects with a small number of hash functions as that of large sets of hash functions, but also can be made use of for relaxations to derive closed-form solutions which are all of great benefits to multi-modal retrieval.

### 2.2. Learning to match

Relevance has always been considerably important in search and will always be, and match is a key factor for similarity, especially in the contemporary heterogeneous, multi-view, associated big data era. Learning to match (match learning) [24–27] is a sharp sword in such scenarios including question answering, recommender systems, machine translation, cross-language information retrieval, online advertising, image annotation, drug design and couple pairing. In recent research, [28] leveraged both clicks and content to learn to match heterogeneous objects via shared latent structures for web search. Likewise, image annotations [29], recommendation systems [30], and Cross-modal Search [17] all mapped different modals or views (i.e. keywords v.s. images, users v.s. products, images v.s. texts etc.) into a shared latent high-level semantic space with low dimensions and bridged them each other for better and effective relevance.

However, in this paper, the datasets explored are formed with images and sentences pair-wisely; therefore we can naturally connect them into a common latent semantic space from two distinct image and sentence spaces with the assumptions that they both describe the same object/thing with just different languages.

### 2.3. Deep learning for representations

Deep learning (deep machine learning, or deep neural network learning, or hierarchical learning, or sometimes DL) is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers with complex structures, or otherwise composed of multiple non-linear transformations [31–33]. A large amount of exploration and research on AutoEncoder, CNN, LSTM, and other types of DNNs have brought unprecedented changes in fields such as image understanding and recognition, speech recognition, and distributed representations and language processing in recent few years since 2006, when Hinton and Salakhutdinov gave a second birth to the traditional neural network [34]. In this paper, we would like to focus on two well-used DL tools VGG-16 [35] and Sentence2vec [36,37] for image and text representations respectively, which will be prepared for the next match learning parts illustrated in the middle of Fig. 1.

## 3. Problem statement

Suppose that $O = \{o_s\}s = 1^N$ is a set of multimodal instances, which consists of an image and its corresponding texts (sentences), i.e. $o_s = (D_s^i, D_s^t)$, where $D_s^i \in R^{M_1}$ is an $M_1$-dimensional image descriptor extracted from VGG-16[1] and PCA, and $D_s^t \in R^{M_2}$ is an $M_2$-dimensional text feature obtained from Sentence2vec[2] (usually $M_1 \neq M_2$). Given the bits length $K$, the purpose of DRLSMH is to learn an integrated binary

[1] http://www.robots.ox.ac.uk/vgg/research/very_deep/

[2] https://github.com/klb3713/sentence2vec