Pattern Recognition 113 (2021) 107813

Contents lists available at ScienceDirect

Pattern Recognition

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

Reconstruction regularized low-rank subspace learning for cross-modal retrieval

Jianlong Wu[a,*], Xingxu Xie[b], Liqiang Nie[a,*], Zhouchen Lin[b], Hongbin Zha[b]

[a] School of Computer Science and Technology, Shandong University, China
[b] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, China

ARTICLE INFO

Article history:
Received 11 March 2020
Revised 6 November 2020
Accepted 7 December 2020
Available online 12 January 2021

Keywords:
Cross-modal retrieval
Low-rank subspace learning
Reconstruction regularization

ABSTRACT

With the rapid increase of multi-modal data through the internet, cross-modal matching or retrieval has received much attention recently. It aims to use one type of data as query and retrieve results from the database of another type. For this task, the most popular approach is the latent subspace learning, which learns a shared subspace for multi-modal data, so that we can efficiently measure cross-modal similarity. Instead of adopting traditional regularization terms, we hope that the latent representation could recover the multi-modal information, which works as a reconstruction regularization term. Besides, we assume that different view features for samples of the same category share the same representation in the latent space. Since the number of classes is generally smaller than the number of samples and the feature dimension, therefore the latent feature matrix of training instances should be low-rank. We try to learn the optimal latent representation, and propose a reconstruction based term to recover original multi-modal data and a low-rank term to regularize the learning of subspace. Our method can deal with both supervised and unsupervised cross-modal retrieval tasks. For those situations where the semantic labels are not easy to obtain, our proposed method can also work very well. We propose an efficient algorithm to optimize our framework. To evaluate the performance of our method, we conduct extensive experiments on various datasets. The experimental results show that our proposed method is very efficient and outperforms the state-of-the-art subspace learning approaches.

1. Introduction

Along with the arrival of information age, people get a growing number of multimedia information everyday. These multimedia data are often collected from diverse modalities, such as image, text, video and etc. These heterogeneous data are usually associated with the same entity. There is a great need for users to use data of one modality as query to search relevant documents or files of other modalities. For example, people can use text or keywords to retrieve related images or videos. Therefore, cross-modal matching has received much attention. In this paper, we mainly focus on cross-modal retrieval which has been actively studied in recent years.

For various modalities data of one object, there are heterogeneous properties between them, even though they share the same semantic information. The main challenge of cross-modal retrieval is how to reduce the heterogeneous gap between different modality features and then measure their cross-modal similarity. Towards this issue, a lot of work have been proposed. Among existing methods, latent subspace learning is the most popular approach for cross-modal matching. By learning a shared subspace for multi-modal data, we can reduce the heterogeneous gap and measure the cross-modal relevance in the common space. There are mainly two kinds of subspace learning methods. The first one is the *projection space learning method* which hopes to project multi-modal data into one common space with some regularizations, such as [1–3]. The other one [4] can be referred to as the *intact space learning method* which wants to learn an original space where multi-view data are projected from. Both these two kinds of methods are reasonable and achieve good performance. So the problem is transformed to how to learn an intrinsic subspace for multi-modal data.

Besides, there is another issue that we should take into consideration. There are a lot of situations where the semantic labels are not easy to obtain. So it is necessary to present a framework that can also be applied to unsupervised situation.

In view of all above issues, we propose a reconstruction regularized low-rank subspace learning (RRLSL) method for cross-modal retrieval to learn the essential low-rank representation for multi-

* Corresponding authors.
E-mail addresses: jlwu1992@sdu.edu.cn (J. Wu), xyxie@pku.edu.cn (X. Xie), nieliqiang@gmail.com (L. Nie), zlin@pku.edu.cn (Z. Lin), zha@cis.pku.edu.cn (H. Zha).

https://doi.org/10.1016/j.patcog.2020.107813
0031-3203/

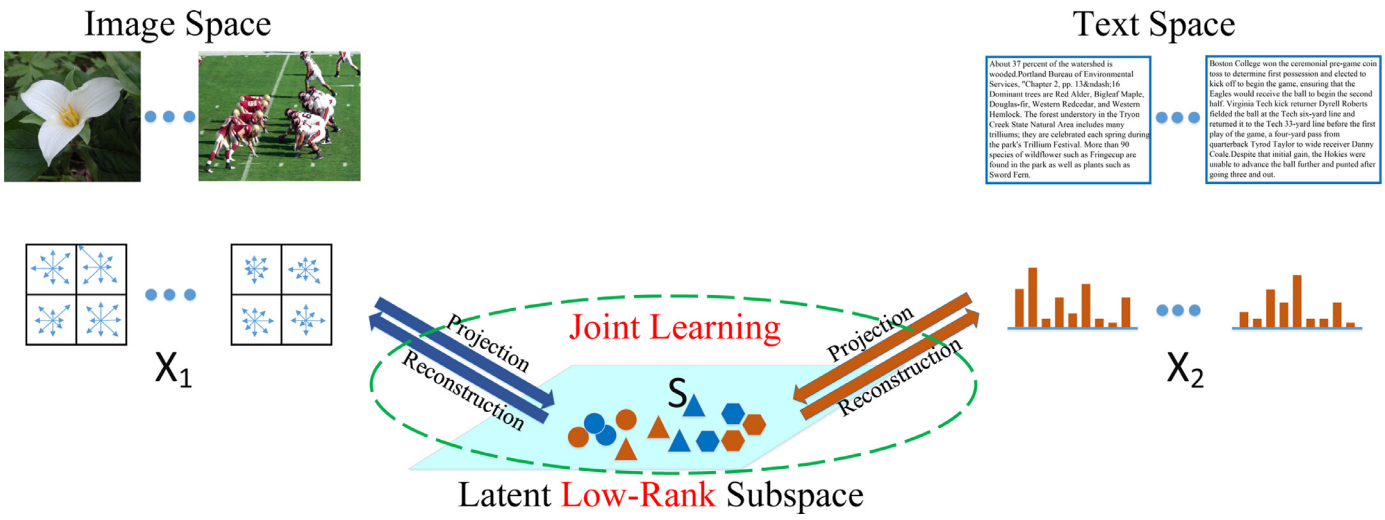


Fig. 1. The brief overview of our proposed reconstruction regularized low-rank subspace learning method. For multi-modal features, we project them into a low-rank subspace with modality-specific projection matrix, and use the relevant weights to reconstruct the original data as a regularization term. For unsupervised learning, the shared low-rank subspace S is jointly learned with all the projection matrices.

modal data. In Fig. 1, we present a brief overview of the proposed method. Specifically, we project multi-modal data into a shared latent subspace with low-rank constraint as we assume that different view features for samples of the same category share the same representation in this latent space. In the meantime, instead of adding traditional regularization terms, we hope the latent representation could recover the multi-modal information, which works as a reconstruction regularization term. Our proposed RRLSL method can be applied to both supervised and unsupervised cross-modal retrieval tasks. For unsupervised situation, the optimal latent subspace is jointly learned with the projection matrices. For supervised task, we directly use the label space as the shared latent subspace. We also propose an efficient algorithm to optimize this problem.

Main contributions of our work are summarized as follows:

(1) We propose a novel reconstruction regularized low-rank subspace learning method for cross-modal retrieval, which can learn the essential low-rank representation shared by multi-modal data. The latent representation can well capture the essential information and reconstruct the original data.

(2) Within this framework, an efficient algorithm is proposed to solve this optimization problem, learning the optimal latent subspace and projection matrices. Both the computation and space complexities of our method are very low. It can be extended to large-scale retrieval applications.

(3) Our proposed method can deal with both supervised and unsupervised cross-modal matching tasks. Experimental results on various datasets with different features show that our method can outperform the start-of-the-art methods for both supervised and unsupervised situations.

The remainder of the paper is organized as follows. In Section 2, we will briefly review the related work of cross-modal retrieval. In Section 3, we will illustrate the details of our RRLSL method and optimization process. Experimental results are presented in Section 4. Finally, Section 5 concludes our paper.

2. Related work

In this section, we briefly review the related studies about latent subspace learning for cross-modal retrieval. Existing methods can be divided into two categories, including the traditional latent subspace learning methods and the deep learning methods.

2.1. Traditional latent subspace learning methods

Latent subspace learning is the most popular method for cross-modal retrieval. By projecting multi-modal data into one common subspace, we can measure their similarity effectively. Canonical Correlation Analysis (CCA) [5], Partial Least Squares (PLS) [6] and Bilinear Model (BLM) [7] are three main classic unsupervised methods. CCA aims to learn a latent space by maximizing the correlation between cross-modal features. There are also many CCA based extensions, such as kernel CCA [8]. PLS tries to learn the optimal projection direction by maximizing their covariance. Sharma et al. [9] applied PLS for cross-modal face recognition. BLM [7] attempts to capture the variation in the latent space and is proposed to separate style and content. Recently, Liang et al. [10,11] tried to learn the unsupervised embedding based on self-paced learning and groupwise correspondences.

Despite the above unsupervised methods, there are many approaches that incorporate label information to improve the retrieval results. Label space is often used as the latent subspace. Regularization terms are well studied in these methods. Sharma et al. [12] presented a general multi-view feature extraction approach called generalized multi-view analysis which extended linear discriminant analysis (GMLDA) and marginal Fisher analysis (GMMFA) to their multi-view cases. Wang et al. [1] proposed a method to learn coupled feature space with ℓ_{21}-norm projection matrix penalty and low-rank constraint on the projected data. Then they employed a multi-modal graph regularization term to preserve the local relationship [2]. Zhai et al. [13] also adopted the graph regularization to learn heterogeneous metric. In [14,15], the authors presented a joint representation learning method to explore the influence of pairwise constraint during latent space regression. Kang et al. [3] added a local group-based priori and an ϵ-dragging term for robust representation. Zhang et al. [16] proposed a metric learning framework to learn multi-ordered discriminative structured subspace. Instead of using the simple label information as the latent subspace. Wu et al. [17] came up with a joint latent subspace learning and regression method to learn the optimal common subspace for projection. Different from the above *projection subspace learning methods*, Xu et al. [4] assumed that multi-modal data are projected from an intact space, and they proposed the multi-view intact space learning method to integrate the encoded complementary information in multiple views.

Besides these above subspace clustering methods, dictionary learning methods [18–20] are also very popular in cross-modal retrieval. Zhuang et al. [21] proposed the coupled dictionary learning with group structures for multi-modal retrieval. Huang and Frank Wang [18] came up with the coupled dictionary and feature space learning method for cross-domain image synthesis and recognition. Xu et al. [19] presented the semi-supervised coupled dictionary learning for cross-modal retrieval. Deng et al. [22] tried to capture the discriminative patterns and presented the discriminative dictionary learning with common label alignment. Liu et al. [23] first combined CNN with dictionary learning to learn sparse representation, and then built a structure-guided multi-modal dictionary learning model to learn the concept-level representation [24].

According to previous work, existing studies mainly focus on the learning of latent subspace and regularization terms. It is also necessary to come up with a framework that can be applied to both supervised and unsupervised situations. So we propose the RRLSL method, where we hope that the representation in the learned subspace could reconstruct the multi-modal data by projecting in some way. And we also investigate the correlation among samples based on the low-rank regularization. This novel reconstruction and low-rank terms work as regularization terms in our framework, which is much different from previous regularization related methods.

2.2. Deep learning based cross-model retrieval

Deep learning based methods receive more and more attention in the past decade. For unsupervised learning, Andrew et al. [25] proposed DCCA that uses the objective function of CCA to guide the training of deep learning. Then Wang et al. [26] combined it with autoencoder. Feng et al. [27] also used the autoencoder but defined the correlation based on distance metric. For supervised cross-modal retrieval, Peng et al. [28] proposed the deep hierarchical learning with multiple deep networks. Hua et al. [29] explored the cross-modal correlation by adaptive hierarchical semantic aggregation. Wang et al. [30] adopted the adversarial learning to investigate the correlation between two modalities. Liong et al. [31] focused on learning the coupled metric for deep cross-modal matching. Liu et al. [32] combined dictionary learning with deep learning to learn better representation. Semedo et al. [33] proposed the scheduled adaptive margin constraints to learn deep subspace. Besides the above cross-modal retrieval, there are also many work focus on cross-modal hashing [34–38].

We mainly focus on traditional latent subspace learning in this paper. But we can also use the pre-trained model to extract deep features as the input for our subspace learning method to improve the performance of cross-modal retrieval.

3. Reconstruction regularized low-rank subspace learning

3.1. Model formulation

Suppose that we have a collection of data from M different modalities, $X_i = (x_1^i, x_2^i, \ldots, x_n^i), i = 1, \ldots, M$, where features in X_i are in d_i dimensions, and n is the total number of samples. Multi-modal features $x_j^1, x_j^2, \ldots, x_j^M$ of the jth object share the same semantic label. For traditional latent subspace learning methods, they learn a projection matrix $W_i \in \mathbb{R}^{c\times d_i}$ for each modality to map each modality features $X_i \in \mathbb{R}^{d_i\times n}$ into the shared latent space $S \in \mathbb{R}^{c\times n}$, where c is the dimension in the latent space. By adding related regularization on projection matrix W and subspace S, the general objective function for latent space learning can be formulated as:

$$\min_{W,S} \sum_{i=1}^{M} \|W_i X_i - S\|_F^2 + \alpha\Omega(S) + \beta\Psi(W_i, \ldots, W_M), \tag{1}$$

where α and β are balance parameters, $\Omega(S)$ and $\psi(W)$ are regularizations on W and S, respectively.

For the regularization term $\Omega(W)$, the commonly used forms include sparse [15], low rank [1], and the graph Laplacian [2,17] regularizations. Different from them and inspired by the intact space learning method [4], we hope that the representation in the latent subspace S should contain all essential information of this object, so that we can reconstruct the original multi-modal data based on S. Besides, for the latent representation, samples belong to same category should have the same representation. As the number of classes is much smaller than the number of training samples and the dimension of the latent subspace, the representation matrix of training samples S should be low-rank. Since it is NP-hard to optimize the problem of rank(S) minimization, it is a commonly used strategy to relax it to the nuclear norm $\|S\|_*$.

By adding the low-rank constraint and the reconstruction term as the regularization on subspace S, we can get the primary objective function for our RRLSL method:

$$\min_{W,S} \sum_{i=1}^{M} \left(\|W_i X_i - S\|_F^2 + \alpha\|W_i^* S - X_i\|_F^2\right) + \gamma\|S\|_*, \tag{2}$$

where α and γ are constants to balance the contributions of different terms, W_i^* denotes the projection matrix to reconstruct the ith modality data, and the nuclear norm $\|\cdot\|_*$ is defined as the sum of singular values for low-rank constraint.

Similar to that in deep belief networks [39,40], we adopt the tied weights to simplify the framework in Eq. (2). That is:

$$W_i^* = W_i^T, \quad i = 1, 2, \ldots, M. \tag{3}$$

For the nuclear norm, it is also called the trace norm according to its definition, and we have:

$$\|S\|_* = \mathrm{tr}\left((S^T S)^{1/2}\right) = \mathrm{tr}\left(S^T (SS^T)^{-1/2} S\right). \tag{4}$$

The above transformation can benefit the optimization. Moreover, we add a Frobenius norm regularization term to penalize the projection matrix to avoid trivial solution. Then the final objective function for our proposed RRLSL can be written as:

$$\min_{W,S,H} \sum_{i=1}^{M} \left(\|W_i X_i - S\|_F^2 + \alpha\|W_i^T S - X_i\|_F^2 + \beta\|W_i\|_F^2\right) + \gamma\,\mathrm{tr}\left(S^T H S\right), \tag{5}$$

where $H = (SS^T + \epsilon I)^{-1/2}$, ϵ is a small positive constant, α, β, and γ are balance parameters, and γ should decay with the increasing of iterations for convergence reason.

The overview of our work is shown in Fig. 1. We project multi-modal data into a common low-rank subspace, and use the relevant weights in Eq. (3) to reconstruct the original data based on the representation in the latent space. The reconstruction part can be regarded as a regularization term. For unsupervised situation, the latent low-rank subspace S is jointly learned with all the projection matrices. After we get the optimal weights and latent subspace based on the training samples, we project the testing samples into the latent subspace and retrieve related cross-modal samples.

3.2. Optimization

For the problem in Eq. (5), there are three variables to optimize, and it is hard to solve it jointly. However, with other variables fixed, it is convex to optimize the specific variable and we can

compute the close-form solution. We adopt the alternating minimization to solve this problem.

W sub-problem:

We first fix S and H. According to the norm properties $\|X\|_F^2 = \|X^T\|_F^2$, the problem to optimize W can be transformed into:

$$\min_W \sum_{i=1}^{M} \left(\|W_i X_i - S\|_F^2 + \alpha \|S^T W_i - X_i^T\|_F^2 + \beta \|W_i\|_F^2 \right). \tag{6}$$

We can optimize W_i ($i \in \{1, 2, \ldots, M\}$) for each view, respectively. By setting the derivative of the objective function in Eq. (6) with respect to W_i to zero, we can get:

$$(\alpha SS^T + \beta I) W_i + W_i X_i X_i^T = (1+\alpha) S X_i^T, \quad i = 1, \ldots, M. \tag{7}$$

By denoting $A = \alpha SS^T + \beta I$, $B_i = X_i X_i^T$, and $C_i = (1+\alpha) S X_i^T$, optimization in Eq. (7) can be reformulated as:

$$AW_i + W_i B_i = C_i, \quad i = 1, 2, \ldots, M. \tag{8}$$

It is obvious that the above equation is the well-known Sylvester equation, and it can be solved efficiently by the Bartels-Stewart algorithm [41]. Please note that $A = \alpha SS^T + \beta I$ should be positive definite in Sylvester equation. As the initialization and update of S cannot guarantee positive definiteness of SS^T, so we incorporate the regularization term on W to add term βI to guarantee positive definiteness of A.

H sub-problem:

H can be simply computed by $H = (SS^T + \epsilon I)^{-1/2}$.

S sub-problem:

With W and H fixed, S can be solved by minimizing the following problem:

$$\min_S \sum_{i=1}^{M} \left(\|W_i X_i - S\|_F^2 + \alpha \|W_i^T S - X_i\|_F^2 \right) + \gamma \operatorname{tr}\left(S^T H S\right). \tag{9}$$

The above problem in Eq. (9) is convex with respect to S. By setting the derivative of the objective function in Eq. (5) with respect to S to zero, we can get:

$$\sum_{i=1}^{M} \left(-2(W_i X_i - S) + 2\alpha W_i (W_i^T S - X_i) \right) + \gamma (H + H^T) S = 0. \tag{10}$$

Then we can get:

$$S = \left(2M \cdot I + 2\alpha \sum_{i=1}^{M} W_i W_i^T + \gamma (H + H^T) \right)^{-1} \left(2(1+\alpha) \sum_{i=1}^{M} W_i X_i \right). \tag{11}$$

We summarize the optimization process of RRLSL in Algorithm 1.

Algorithm 1 Alternating Minimization for Unsupervised RRLSL.

Input: Multi-view data $X_i \in \mathbb{R}^{d_i \times n}$, $i = 1, 2, \ldots, M$.

1: Set $k = 1$, and initialize the subspace S^0 by K-means clustering.
2: **while** not converged **do**
3: **for** $i = 1, 2, \ldots, M$ **do**
4: Compute $A = \alpha S^{k-1}(S^{k-1})^T + \beta I$, $B_i = X_i X_i^T$, and $C_i = (1+\alpha) S^{k-1} X_i^T$.
5: Update W_i^k by solving the Sylvester equation in Eq.~(8).
6: **end for**
7: Update H^k by $H^k = (S^{k-1}(S^{k-1})^T + \epsilon I)^{-1/2}$.
8: Update S^k by Eq.~(11).
9: Update γ_k by $\gamma_k = \frac{1}{2}\gamma_{k-1}$.
10: $k = k + 1$.
11: **end while**

Output: Projection matrices $W_i \in \mathbb{R}^{c \times d_i}$, $i = 1, \ldots, M$.

3.3. Convergence and complexity

We briefly discuss the convergence and the computational complexity of our method.

We solve the problem of unsupervised situation by alternating minimization. At each iteration, the subproblem to optimize each variable with others fixed is convex, so we can compute the close-form solution to update the variables. As we alternatively optimize W and S, the objective function is monotonically decreasing. According to Tseng [42], our algorithm will converge to a stationary point. For detailed convergence analysis, please refer to the appendix.

For the update of W_i, the complexity to compute the intermediate matrices A, B, and C is $\mathcal{O}(nc^2 + nd_i^2 + ncd_i)$, where d_i is the dimension of ith modality feature and c is the dimension in the latent subspace. According to Bartels and Stewart [41], the complexity to solve the Sylvester equation in Eq. (8) is $\mathcal{O}(d_i^3 + c^3)$, which is irrelevant to the sample number n and only depends on the feature dimension. Let $d = \max(d_1, d_2, \ldots, d_M)$, then the complexity to optimize W is $\mathcal{O}(nc^2 + nd^2 + d^3 + c^3)$. For the update of S and H, the computational complexity is $\mathcal{O}(c^3 + dc^2 + ncd)$. Denote the iteration number as k, which is a small constant in practice as our algorithm converges in a few steps. Then the total time complexity of our algorithm is $\mathcal{O}(k(nc^2 + nd^2 + d^3 + c^3))$, which is linear to the number of samples n. In general, c is much smaller than the largest feature dimension d, and d is smaller than the number of samples n. So the total time complexity of unsupervised RRLSL is $\mathcal{O}(knd^2)$, which is very efficient. It can be easily applied to large scale datasets.

3.4. Extension to supervised situation

Our method can be directly extended to supervised version. Just as most existing supervised methods do, we use the label space $Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{c_g \times n}$ to define the latent subspace. Here, c_g denotes the number of classes. Then the objective function is transformed to:

$$\min_W \sum_{i=1}^{M} \left(\|W_i X_i - Y\|_F^2 + \alpha \|W_i^T Y - X_i\|_F^2 + \beta \|W_i\|_F^2 \right). \tag{12}$$

Then we set the derivative of Eq. (12) with respect to each W_i to zero. The above problem in Eq. (12) can be optimized by solving the following Sylvester equation:

$$(\alpha YY^T + \beta I) W_i + W_i X_i X_i^T = (1+\alpha) Y X_i^T, \quad i = 1, \ldots, M. \tag{13}$$

As introduced in the last complexity subsection, the above equation can be solved efficiently. It can be applied to large scale datasets.

3.5. Relations to existing methods

Compared with existing subspace learning methods, the main difference of RRLSL method lies in the regularization term. As traditional methods add regularization terms such as structured sparsity [43], low rankness [1], group-based priori [3], and graph Laplacian [2], we propose a novel reconstruction based regularization, which helps the latent representation preserve the essential information. And our low-rank constraint can better explore the correlation among samples. Besides, our RRLSL model can deal with both supervised and unsupervised situations, while previous studies only focus on one situation.

For unsupervised learning, our model looks like the autoencoder as we both try to learn the latent embedding. However, there are still much difference. Correspondence autoencoder [27,44] and DCCAE [26] are two main autoencoder related work. For cross-modal problems, correspondence autoen-

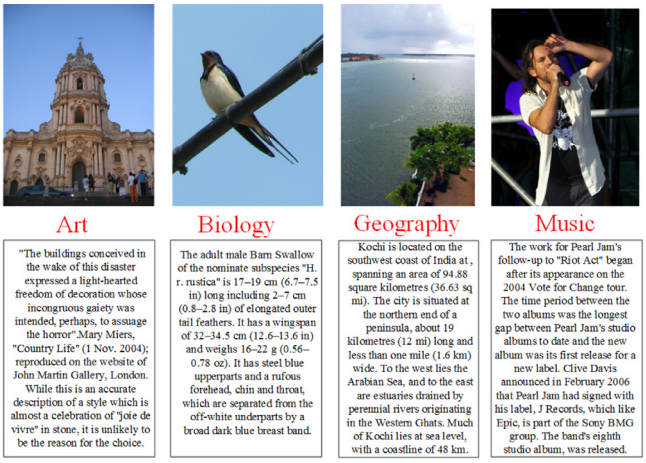


Fig. 2. Some example image-text pairs of the Wiki dataset (a) and the Pascal VOC dataset (b).

coder in [27,44] used the distance between cross-modal embeddings and DCCAE [26] adopted the CCA in the latent layer as the regularization to optimize the latent embedding. Instead of constraining the distance or correlation, we learn a specific low-rank latent subspace S which is shared by all modalities. Besides, the optimization algorithms are different. We propose an efficient iterative algorithm to compute the optimal the projection matrices, while other autoencoder related work uses the back propagation method to update the weights based on the loss function. In addition, our methods can be easily extended to multi-view cases, while the above two methods can only deal with the two-view situation. Moreover, when we use the label space as the latent subspace for supervised mode in our RRLSL method, it can be regarded as the label-aware autoencoder, which will benefit the retrieval. And previous autoencoder methods mainly work on the unsupervised situation.

4. Experiments

4.1. Datasets

For cross-modal retrieval, the main task is the matching between image and text. There are mainly three commonly used datasets, including the Wiki dataset [45], the Pascal VOC dataset [46], and the NUS-WIDE dataset [47]. Some sample image-text pairs of Wiki and Pascal VOC are shown in Fig. 2.

The Wiki dataset [45] contains 2,866 image-text pairs, which are generated from the featured article of Wikipedia. There are 10 semantic categories in total. For each pair, the text is a long article describing the label related information, and the image is highly correlated to the content of the article. We adopt the same setting as that in [1,2], which splits 2,866 pairs into a training set of 1,300 pairs (130 pairs per class) and a testing set of 1,566 pairs. For text features, latent Dirichlet allocation (LDA) [48] is adopted to extract 10 dimensional representations. For images, we extract the 128 dimensional SIFT [49] descriptor histograms.

The Pascal VOC dataset [46] contains 5,011 training and 4,952 testing image-tag pairs collected from 20 different classes. As image-tags pairs that belong to only one object of the 20 concepts are selected in the experiment, there are 2,808 pairs for the training set and 2,841 pairs for the testing set. For feature extraction, 512 dimensional GIST [50] features are used to represent the images. 399 dimensional word frequency features are used to represent texts.

The NUS-WIDE database [47] is a real-world web image database which is created by NUS's lab for media search. It contains 269,648 images and the associated tags from Flickr, with a total of 5,018 unique tags. There are 81 concepts in total, and some pairs may belong to more than one concepts. Like many previous work, we select image-tags pairs that exclusively belong to only one of the 21 largest concepts, which results in 97,079 pairs in total for our experiments. For this task, some work also select 10 largest concepts, which is much easier. We use the same setting as that in [3], one third of the samples in each class are randomly selected to form the testing set, and the remaining samples are used as the training set. For this database, six different types of low-level image features are provided by the authors. We directly adopt the 500-dimensional bag of words feature vectors based on SIFT descriptions and 1000-dimensional bag-of-words feature vectors to represent images and textual tags, respectively.

4.2. Evaluation metrics and parameters setting

We mainly consider two cross-modal retrieval tasks: (1) Image query vs. Text database and (2) Text query vs. Image database. Based on the training dataset, we learn the latent space as well as the projection matrices. Then we map the multi-modal features of test dataset into the common subspace based on the learned view-dependent projection matrices. During testing phase, one modality data of the testing set serves as the query set and the other modality data of the testing set serves as the database. We use the query to retrieval relevant objects of the other modality from the database. We adopt the cosine distance to measure the similarity between cross-modal features.

To evaluate the performance of our proposed RRLSL method, the mean average precision (MAP) is used in the experiments. The average precision (AP) of N retrieved results is defined by $AP = \frac{1}{T}\sum_{i=1}^{N} P(i)\delta(i)$, where T is the actual number of same category objects in the database, $P(i)$ represents the precision of the top i retrieved results, and $\delta(i) = 1$ only when the ith retrieved result belongs to the same class with the query. Then we can compute the MAP by averaging the AP of all queries in the query set. Higher MAP scores represent better result. Besides the MAP, we also adopt the precision-recall curve to thoroughly evaluate the effectiveness of different methods.

Table 1
MAP scores comparison between the usupervised RRLSL and other methods on the Wiki, the Pascal VOC, and the NUS-WIDE datasets.

Datasets and Methods	Wiki			Pascal VOC			NUS-WIDE		
	Image query	Text query	Average	Image query	Text query	Average	Image query	Text query	Average
CCA	0.2541	0.2058	0.2299	0.3022	0.2298	0.2660	0.2463	0.2415	0.2434
BLM	0.2565	0.2029	0.2297	0.3178	0.2329	0.2754	0.2621	0.2542	0.2581
PLS	0.2615	0.2190	0.2403	0.3273	0.2578	0.2925	0.2796	0.2684	0.2740
SCSM	0.2740	0.2170	0.2450	0.3435	0.2787	0.3111	–	–	–
SPGCM	0.2847	0.2229	0.2538	0.3512	0.2770	0.3141	0.2907	0.2734	0.2821
URRLSL	**0.2949**	**0.2267**	**0.2608**	**0.3663**	**0.2835**	**0.3249**	**0.3003**	**0.2892**	**0.2947**

Table 2
P-value of statistic significance test between the results of related methods on all three datasets.

Methods and Datasets	Unspervised		Supervised	
	URRCSL VS SCSM	URRCSL VS SPGCM	RRCSL VS JRL	RRCSL VS LGCFL
Wiki	$1.00 * 10^{-16}$	$7.33 * 10^{-9}$	$2.69 * 10^{-11}$	$2.31 * 10^{-24}$
Pascal VOC	$1.26 * 10^{-16}$	$1.32 * 10^{-5}$	$1.90 * 10^{-10}$	$4.00 * 10^{-16}$
NUS-WIDE	–	$7.83 * 10^{-30}$	$3.55 * 10^{-8}$	$4.72 * 10^{-15}$

For the parameters in our proposed RRLSL method, we fine-tune the parameters α, β and γ in Eq. (5) by searching the gird of $\{10^{-2}, 10^{-1}, \ldots, 10^{2}, 10^{3}\}$ based on cross validation. The dimension of latent subspace is fine-tuned in the range of $[1 * c_g, 10 * c_g]$.

4.3. Compared methods

For unsupervised cross-modal retrieval, we compare the unsupervised RRLSL (URRLSL) with five unsupervised algorithms, including the CCA [5], PLS [7], BLM [7], self-paced cross-modal subspace matching (SCSM) [10], and simultaneous pairwise and groupwise correspondences maximization (SPGCM) [11].

For supervised cross-modal retrieval, supervised RRLSL (SRRLSL) is compared with seven supervised algorithms, including the GMMFA [12], GMLDA [12], learning coupled feature spaces (LCFS) [1], joint representation learning (JRL) [14], and the local group based consistent feature learning (LGCFL) [3]. In order to compare fairly, the kernel method used in [14] to improve the MAP scores is not adopted for all the methods in this paper.

4.4. Experimental results of unsupervised situation

We test the performance of unsupervised RRLSL and other methods on the Wiki, the Pascal VOC, and the NUS-WIDE datasets. The MAP scores are presented in Table 1. The value in bold represents the best result. SCSM fails to work on the large NUS-WIDE dataset since it has very high computation complexity. We can see that our unsupervised RRLSL method outperforms all other unsupervised methods in MAP scores for both tasks of image and text query. On the Wiki dataset, the average MAP score of our method is 0.2608, which is relatively $\frac{0.2608}{0.2538} - 1 = 2.76\%$ higher than the second best result 0.2538 achieved by the SPGCM. On the Pascal VOC dataset, our MAP score 0.3249 is 3.44% higher than the SPGCM's result 0.3141.

On the NUS-WIDE dataset, compared with the second best result 0.2821 achieved by the SPGCM, the average MAP score of our method 0.2947 is 4.47% higher. Compared with the results of supervised methods shown in Table 3, our results on the Wiki and the Pascal VOC datasets are much better than that of GMMFA, GMLDA, and LCFS. Even if compared with LGCFL, the gap is also acceptable. In Table 2, we test the statistic significance [51] between the results of SCSM, SPGCM and URRLSL, and present the *p*-values. We can see that the *p*-values are less than 0.01 on all datasets, which shows that there is significant difference between the results of SCSM, SPGCM and our URRLSL method.

Figs. 3 and 4 present the precision-recall curves on the Wiki and the Pascal VOC datasets. Our method also outperforms other methods.

4.5. Experimental results of supervised situation

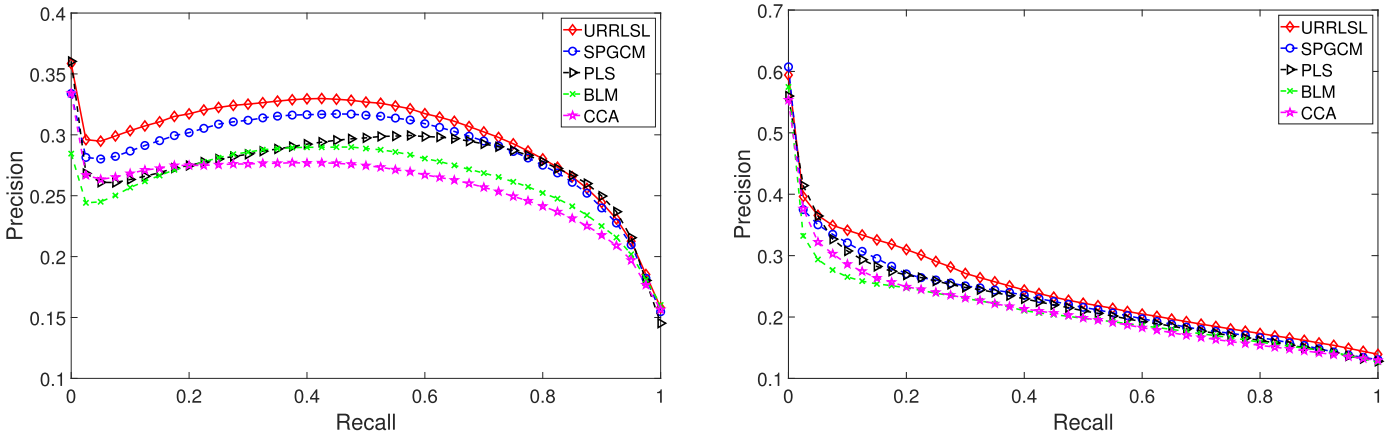
For the supervised cross-modal retrieval, we test the performance of these supervised methods on all three datasets, including the Wiki, Pascal VOC, and NUS-WIDE datasets. Table 3 shows the MAP scores of all these methods. We can observe that the performance of our SRRLSL method surpasses the results of all other methods in tasks of both image query and text query. Compared with the second best results, our results achieve an average 2% improvement on these three datasets, which further verifies the effectiveness of the RRLSL method. Thanks to the authors of LGCFL and JRL to share their codes, we present the *p*-value of statistic significance test between results of these related methods in Table 2. As *p*-values on all datasets are less than 0.01, there is also significant difference between the results of these supervised methods.

Figs. 5 and 6 show the precision-recall curves for both tasks on the Wiki and the Pascal VOC datasets. We can also see that for both tasks on two datasets, our results are better than other methods.

4.6. Performance with different types of features

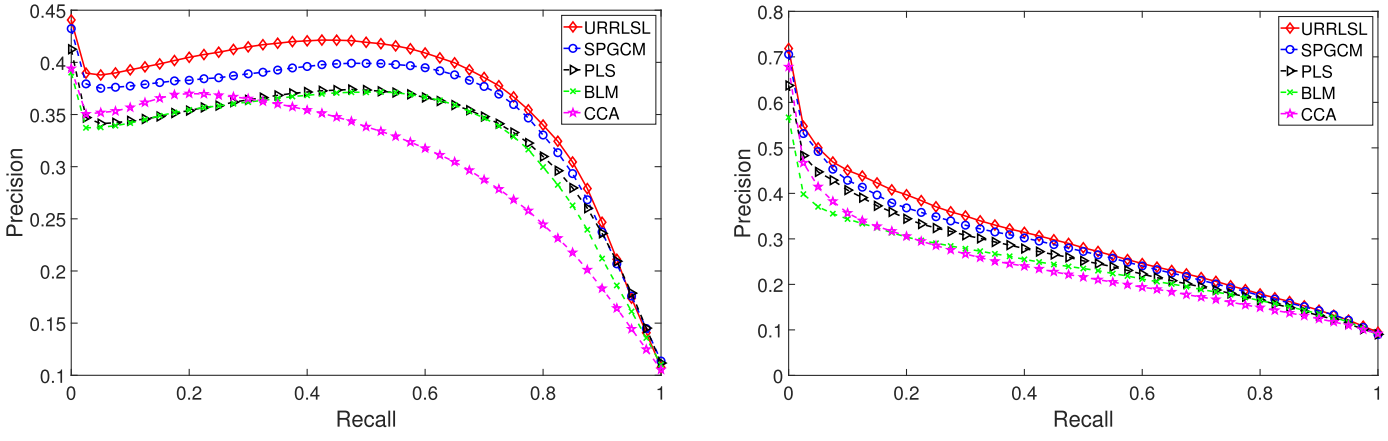
We also test the influence of different types of features of both images and texts to the cross-modal retrieval performance on the Wiki dataset. Besides the SIFT image features and LDA text features given by the Wiki dataset itself, deep image features and another type of text features are also extracted. The convolutional neural network (CNN) is adopted to extract the 4096-dimensional features [52] for image presentation. For text features, 5000-dimensional BOW feature vectors are learned based on the TF-IDF. PCA [53] is adopted to reduce the dimension. In Table 4, we present the MAP scores of GMLDA, LCFS, LGCFL, JRL and our supervised RRLSL with different types of features on the Wiki dataset. We can see that with CNN features, all methods achieve much better results, which can be attributed to the powerful representation ability of deep neural networks. Compared with the result of LDA features, there is no obvious improvement for the TF-IDF features. Compared with all other methods, our SRRLSL method achieves the best result with all these different kinds of features.

To further demonstrate the effectiveness of our proposed method, we conduct experiments on the NUS-WIDE dataset and



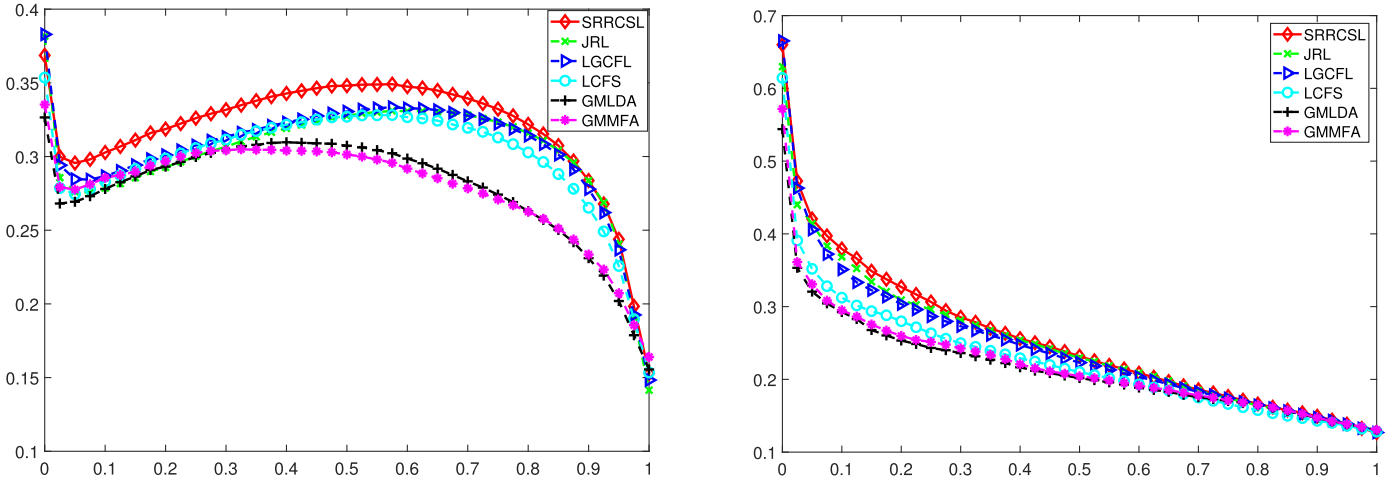
(a) Image query (b) Text query

Fig. 3. Performance of different unsupervised methods on the Wiki dataset, based on precision-recall curve. Best view on screen!



(a) Image query (b) Text query

Fig. 4. Performance of different unsupervised methods on the Pascal VOC dataset, based on precision-recall curve. Best view on screen!

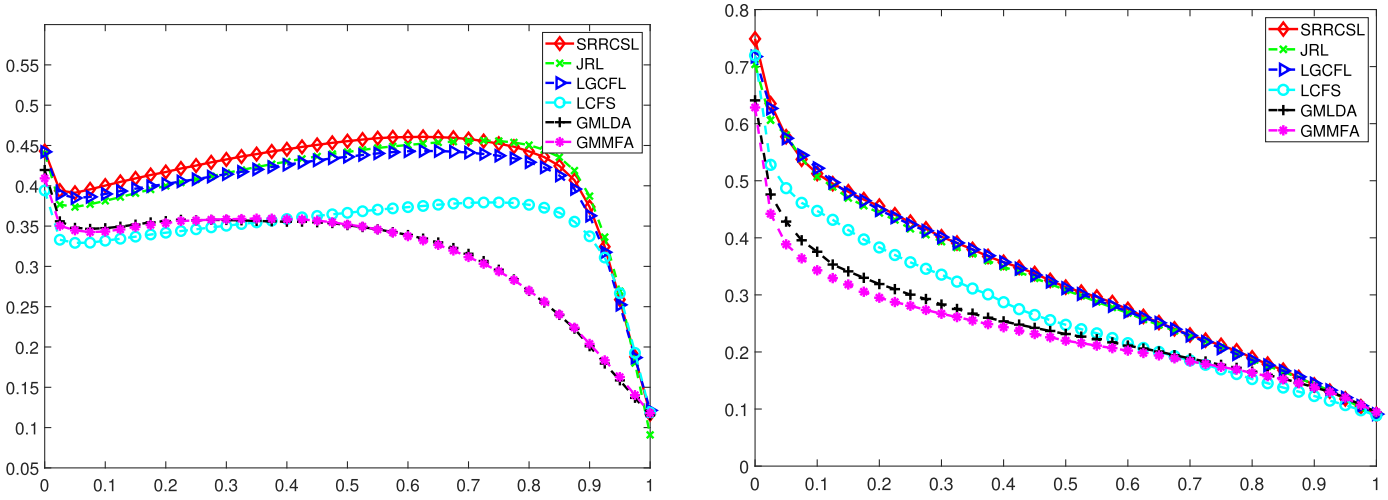


(a) Image query (b) Text query

Fig. 5. Performance of different supervised methods on the Wiki dataset, based on precision-recall curve. Best view on screen!

Table 3
MAP scores comparision between the supervised RRLSL and other methods on the Wiki, the Pascal VOC, and the NUS-WIDE datasets.

Datasets and Methods	Wiki			Pascal VOC			NUS-WIDE		
	Image query	Text query	Average	Image query	Text query	Average	Image query	Text query	Average
GMLDA	0.2751	0.2098	0.2425	0.3094	0.2448	0.2771	0.3243	0.3076	0.3159
GMMFA	0.2750	0.2139	0.2445	0.3090	0.2308	0.2699	0.2983	0.2939	0.2961
LCFS	0.2798	0.2141	0.2470	0.3438	0.2674	0.3056	0.3830	0.3460	0.3645
LGCFL	0.3009	0.2377	0.2693	0.3988	0.3212	0.3600	0.3780	0.3290	0.3535
JRL	0.2979	0.2413	0.2696	0.4044	0.3166	0.3605	0.3818	0.3360	0.3589
SRRLSL	**0.3146**	**0.2466**	**0.2806**	**0.4134**	**0.3228**	**0.3681**	**0.4120**	**0.3666**	**0.3893**



(a) Image query (b) Text query

Fig. 6. Performance of different supervised methods on the Pascal VOC dataset, based on precision-recall curve. Best view on screen!

Table 4
MAP comparison with different features on the Wiki dataset.

Query	Image Query					Text Query				
Image/Text Features	GMLDA	LCFS	LGCFL	JRL	SRRLSL	GMLDA	LCFS	LGCFL	JRL	SRRLSL
SIFT/LDA	0.2751	0.2798	0.3009	0.2979	**0.3146**	0.2098	0.2141	0.2377	0.2413	**0.2466**
CNN/LDA	0.4084	0.4132	0.4532	0.4208	**0.4633**	0.3693	0.3845	0.3887	0.3854	**0.4028**
SIFT/TF-IDF	0.2782	0.2978	0.3157	0.3023	**0.3188**	0.1925	0.2134	0.2461	0.2395	**0.2496**
CNN/TF-IDF	0.4455	0.4553	0.4535	0.4412	**0.4681**	0.3661	0.3978	0.4033	0.3900	**0.4169**

Table 5
MAP comparison with deep learning based methods on NUS-WIDE-10.

	Corr-AE [27]	DCCA [25]	CMDN [28]	CCL [54]	SAM [33]	SRRLSL
Image query	0.391	0.475	0.643	0.671	**0.701**	0.695
Text query	0.429	0.500	0.626	0.676	0.707	**0.719**
Average	0.410	0.488	0.635	0.674	0.704	**0.707**

compare the results with deep learning based cross-modal retrieval methods. For fair comparison, we use the same setting as [33,54], where only the largest 10 categories are selected to construct NUS-WIDE-10 with more than 60,000 instances. For our SRRCSL, we use a pre-trained VGG-19 convolutional network to extract the 4,096-dimensional image features, and we adopt the 1,000-dimensional bag-of-words vector for the text feature, while other deep learning methods train the network on the training instances. The results are shown in Table 5. We directly copy the results of other methods from Peng et al. [28]. We can also see that our method achieves the best average MAP. Based on the pre-trained deep features without fine-tune, our method can even outperform the state-of-the-art deep learning based methods.

4.7. Complexity and processing time

In Table 6, we compare the time complexity and space complexity of these state-of-the-art methods. We also list the training computation time of all these methods on the Wiki, the Pascal VOC, and the NUS-WIDE datasets. All these methods are achieved with matlab code and tested on a 3.4GHZ PC with 64GB RAM. In Table 6, top three methods are unsupervised and last seven methods are supervised. In the results, d is the max dimension of features, n is the number of samples, k is the number of iterations until convergence, s denotes seconds, t is the number of groups during unsupervised clustering, c is the group number for LGCFL. In general, d is much smaller than n. For example, on the NUS-WIDE dataset, d is equal to 1000, while n is 97,079. We can see

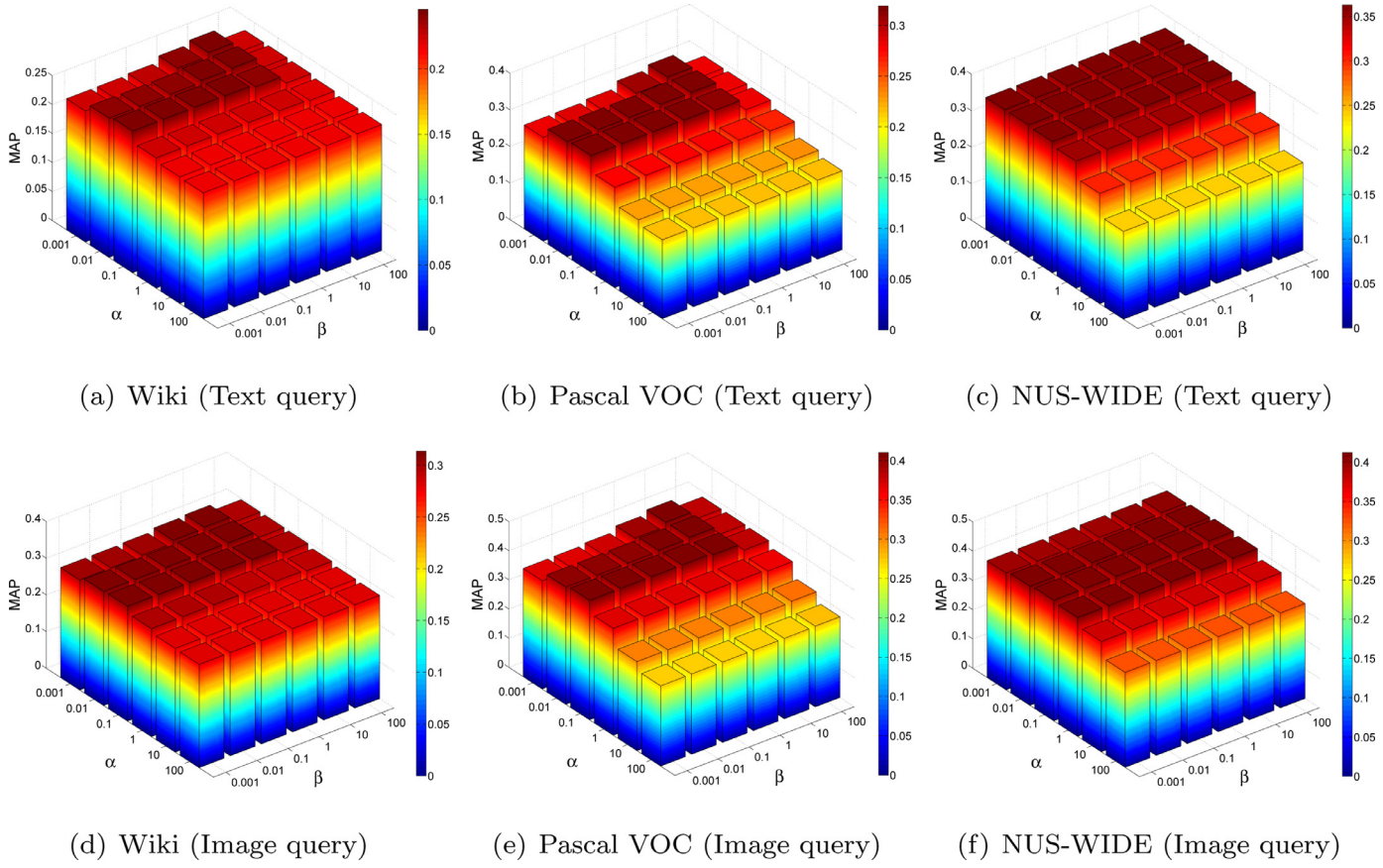


Fig. 7. Performance variations of SRRLSL for the Text query vs. Image database and Image query vs. Text database tasks with respect to α and β on the Wiki, the Pascal VOC, and the NUS-WIDE datasets, respectively.

Table 6
Complexity and processing time of all related methods on the Wiki, the Pascal VOC, and the NUS-WIDE datasets, where d is the max dimension of features, n is the number of samples, k is the number of iterations, s denotes seconds, t is the number of groups during unsupervised clustering, c is the group number for LGCFL.

Methods	Time complexity	Space complexity	Time on Wiki	Time on Pascal VOC	Time on NUS-WIDE
SCSM	$\mathcal{O}(ktn^3)$	$\mathcal{O}(n^2)$	96.56s	1015.35s	–
SPGCM	$\mathcal{O}(kd^3 + ktn^2)$	$\mathcal{O}(nd)$	0.15s	1.52s	32.73s
URRLSL	$\mathcal{O}(knd^2)$	$\mathcal{O}(nd)$	0.07s	1.31s	21.25s
GMLDA	$\mathcal{O}(dn^2)$	$\mathcal{O}(n^2)$	0.15s	1.13s	495s
GMMFA	$\mathcal{O}(dn^2)$	$\mathcal{O}(n^2)$	0.22s	2.29s	790s
LCFS	$\mathcal{O}(k(d^3 + n^{2.376}))$	$\mathcal{O}(n^2)$	1.54s	21.28s	21000s
LGCFL	$\mathcal{O}(kdnc_g)$	$\mathcal{O}(nd)$	0.02s	0.38s	8.04s
JRL	$\mathcal{O}(kdn^2)$	$\mathcal{O}(n^2)$	2.28s	21.23s	8000s
SRRLSL	$\mathcal{O}(nd^2)$	$\mathcal{O}(nd)$	0.01s	0.23s	2.59s

our methods under both supervised and unsupervised modes have the lowest complexities in both time and space aspects. The processing time is also the shortest among all these methods. For unsupervised methods, SCSM has the highest computation complexity $\mathcal{O}(ktn^3)$. Even on the Pascal VOC dataset, it needs more than 1,000 s, while our URRLSL only need 1.31 s. For the NUS-WIDE dataset where n is very large, SCSM fails to work. SPGCM has the similar complexity with our method. For supervised methods, even on the largest NUS-WIDE dataset, our method can finish training within 3 seconds, which is very fast. In comparison, JRL needs 8,000 s to finish training on this dataset, which is about 3,000 times more than that of our method. In this case, our algorithm is very efficient and it can be easily extended to large scale applications.

4.8. Parameter sensitivity analysis

Our method is relatively stable. Due to the page limit, we mainly analyze the supervised situation, which is similar to unsupervised condition. There are two parameters α and β in supervised mode. We tune these two parameters in the range of $\{0.001, 0.01, 0.1, 1, 10, 100\}$. In Fig. 7, we present the results on all three datasets. We can observe that the experimental results are very stable with various β on all three datasets. So the proposed method is insensitive to β. When α is between 0.01 and 1, our method can achieve the best performance. It is also insensitive to α.

Based on all above experimental results, we can get the following conclusions. First of all, the result of supervised methods is better than that of unsupervised methods. By incorporating the

label information, we can get more discriminative separation between classes in the latent subspace, which benefits the cross-modal retrieval. Secondly, our RRLSL method achieves very good performance in both unsupervised and supervised tasks, which can be attributed to the effectiveness of reconstruction regularization. With the regularization to recover the original data, the embedding in the latent space of our method can preserve more essential information. And then, good features will benefit the retrieval. CNN features show great improvement on the results. Finally, the time and space complexity of our method is very low, so it can be scaled to very large datasets.

5. Conclusion

In this paper, we have proposed a novel RRLSL method that takes both the low-rank subspace learning and original data reconstruction into consideration to jointly learn the latent subspace and the projection matrices, which can be applied to both unsupervised and supervised cross-modal retrieval. An efficient algorithm is presented for optimization. Extensive experiments on several related datasets demonstrate the superiority of the proposed method over other state-of-the-art methods, especially on the supervised situation. Compared with existing methods, our main strengths lie in the lower computational complexity, better performance, and wider application for different situations. For the weakness, its application to the semi-supervised situation could be further investigated. For the future work, we would like to extend the proposed RRLSL to the semi-supervised situation, where only a few samples are labeled.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Jianlong Wu is supported by the National Natural Science Foundation (NSF) of China (grant no. 62006140), the NSF of Shandong Province (grant no. ZR2020QF106), the Future Talents Research Funds of Shandong University, and the Fundamental Research Funds of Shandong University. Liqiang Nie is supported by the NSF of China (grant nos. 61772310, 61702300, 61702302, 61802231, and U1836216), the Project of Thousand Youth Talents 2016, the Shandong Provincial Natural Science and Foundation (grant no.s ZR2019JQ23 and ZR2019QF001). Zhouchen Lin is supported by the NSF of China (grant nos. 61625301 and 61731018), Major Scientific Research Project of Zhejiang Lab (grant no.s 2019KB0AC01 and 2019KB0AB02), Beijing Academy of Artificial Intelligence, and Qualcomm. Hongbin Zha is supported by the National Key Research and Development Program of China (grant no. 2017YFB1002601) and NSF of China (grant nos. 61632003 and 61771026).

Appendix

In the following, we provide the detailed convergence analysis of the proposed algorithm in unsupervised situation.

Recall that our objective function is:

$$\min_{W,S,H} L(W,S,H) = \sum_{i=1}^{M}\left(\|W_iX_i - S\|_F^2 + \alpha\|W_i^TS - X_i\|_F^2 + \beta\|W_i\|_F^2\right) + \gamma_k \mathrm{tr}\left(S^THS\right). \tag{A.1}$$

Since we have three variables, we divide the convergence analysis into the following four parts.

(1) With S^k and H^k fixed, the problem to optimize W can be formulated as:

$$\min_W L(W,S^k,H^k) = \sum_{i=1}^{M}\left(\|W_iX_i - S^k\|_F^2 + \alpha\|(S^k)^TW_i - X_i^T\|_F^2 + \beta\|W_i\|_F^2\right), \tag{A.2}$$

which is strongly convex with respect to W. By setting the derivative of the objective function in Eq. (6) with respect to W_i to zero, we can get:

$$(\alpha S^k(S^k)^T + \beta I)W_i + W_iX_iX_i^T = (1+\alpha)S^kX_i^T, \quad i = 1,\ldots,M. \tag{A.3}$$

By denoting $A = \alpha S^k(S^k)^T + \beta I$, $B_i = X_iX_i^T$, and $C_i = (1+\alpha)S^kX_i^T$, optimization in Eq. (7) can be reformulated as:

$$AW_i + W_iB_i = C_i, \quad i = 1,2,\ldots,M, \tag{A.4}$$

which is equivalent to the following problem:

$$\left(I_m \otimes A + B_i^T \otimes I_n\right)\mathrm{vec}(W_i) = \mathrm{vec}(C_i), \quad i = 1,2,\ldots,M. \tag{A.5}$$

Since $A \succeq \beta I$, $B_i \succeq \mathbf{0}$, and $S^kS^{kT} \succeq \mathbf{0}$, we have $\left(I_m \otimes A + B_i^T \otimes I_n\right) \succeq \beta I$. Then:

$$\begin{aligned} L(W^k,S^k,H^k) - L(W^{k+1},S^k,H^k) &\geq \frac{\beta}{2}\left\|\mathrm{vec}(W_i^{k+1}) - \mathrm{vec}(W_i^k)\right\|_2^2 \\ &= \frac{\beta}{2}\left\|W_i^{k+1} - W_i^k\right\|_F^2. \end{aligned} \tag{A.6}$$

(2) With S^k fixed, $H^{k+1} = (S^k(S^k)^T + \epsilon I)^{-1/2}$. Denote $\|X\|_F = \max_i \|X_i\|_F$ and $\|W^k\|_F = \max_i \|W_i^k\|_F$, then according to Eq. (11), we have:

$$\begin{aligned} \left\|S^k\right\|_F &\leq \frac{1}{2M} * (2(1+\alpha))\left\|\sum_{i=1}^{M} W_i^kX_i\right\|_F \\ &\leq \frac{1+\alpha}{M} * M * \|X\|_F * \left\|W^k\right\|_F \\ &= (1+\alpha)\|X\|_F * \left\|W^k\right\|_F. \end{aligned} \tag{A.7}$$

Recall that we also add a regularization term on W, so W^k is bounded. Therefore, S^k is also bounded. For convenience, let $\|S^k\|_F \leq B_s$ for all k. Denote $U = [S^{K-1}, \epsilon I]$ and $V = [S^K, \epsilon I]$, then $\epsilon \leq \|U\|_F \leq B_s + \epsilon$ and $\epsilon \leq \|V\|_F \leq B_s + \epsilon$, where $\epsilon > 0$ is a small positive constant. We have:

$$\begin{aligned} &\left|L(W^{k+1},S^k,H^k) - L(W^{k+1},S^k,H^{k+1})\right| \\ &= \left|\gamma_k\mathrm{tr}\left((S^k)^TH^kS^k\right) - \gamma_k\mathrm{tr}\left((S^k)^TH^{k+1}S^k\right)\right| \\ &\leq \gamma_kB_s^2\left\|H^k - H^{k+1}\right\|_F \\ &\leq \gamma_kB_s^2\left\|(UU^T)^{-1/2} - (VV^T)^{-1/2}\right)\right\|_F. \end{aligned} \tag{A.8}$$

Since

$$\begin{aligned} &\left((UU^T)^{-1/2} - (VV^T)^{-1/2}\right)(UU^T)^{-1/2} \\ &\quad + (VV^T)^{-1/2}\left((UU^T)^{-1/2} - (VV^T)^{-1/2}\right) \\ &= (UU^T)^{-1} - (VV^T)^{-1} = -(UU^T)^{-1}(UU^T - VV^T)(VV^T)^{-1}, \end{aligned} \tag{A.9}$$

then we can get:

$$\begin{aligned} &\left|L(W^{k+1},S^k,H^k) - L(W^{k+1},S^k,H^{k+1})\right| \\ &\leq \gamma_kB_s^2\left\|(UU^T)^{-1/2} - (VV^T)^{-1/2}\right)\right\|_F \\ &\leq \gamma_kB_s^2\left\|\left((UU^T)^{-1/2} \otimes I + I \otimes (VV^T)^{-1/2}\right)^{-1}\right\| \\ &\quad \times \left\|UU^T - VV^T\right\|_F\left\|UU^T\right\|^{-1}\left\|VV^T\right\|^{-1} \\ &\leq \gamma_kB_s^2 * \frac{B_s+\epsilon}{2} * \frac{1}{\epsilon^4} * \left\|UU^T - VV^T\right\|_F \end{aligned}$$

$$= \gamma_k \frac{B_s^2(B_s+\epsilon)}{2\epsilon^4} * \left\| U(U-V)^T + (U-V)V^T \right\|_F$$
$$\leq \gamma_k \frac{B_s^2(B_s+\epsilon)^2}{\epsilon^4} \|U-V\|_F$$
$$= \gamma_k \frac{B_s^2(B_s+\epsilon)^2}{\epsilon^4} \left\| S^{k-1} - S^k \right\|_F. \quad (A.10)$$

By setting $\gamma_k \leq \frac{\epsilon^4}{(B_s+\epsilon)^4} \left\| S^{k-1} - S^k \right\|_F$, we can get:

$$\left| L(W^{k+1}, S^k, H^k) - L(W^{k+1}, S^k, H^{k+1}) \right| \leq \left\| S^{k-1} - S^k \right\|_F^2. \quad (A.11)$$

Therefore,

$$L(W^{k+1}, S^k, H^k) - L(W^{k+1}, S^k, H^{k+1}) \geq -\left\| S^{k-1} - S^k \right\|_F^2. \quad (A.12)$$

(3) Similarly, with W^{k+1} and H^{k+1} fixed, S can be solved by minimizing the following problem:

$$S^{k+1} = \underset{S}{\operatorname{argmin}} L(W^{k+1}, S, H^{k+1})$$
$$= \underset{S}{\operatorname{argmin}} \sum_{i=1}^{M} \left(\|W_i^{k+1} X_i - S\|_F^2 + \alpha \|(W_i^{k+1})^T S - X_i\|_F^2 \right)$$
$$+ \gamma_k \operatorname{tr}\left(S^T H^{k+1} S\right). \quad (A.13)$$

The above problem has the following close-form solution:

$$S^{k+1} = \left(2M \cdot I + 2\alpha \sum_{i=1}^{M} W_i^{k+1} (W_i^{k+1})^T + \gamma_k (H^{k+1} + (H^{k+1})^T) \right)^{-1}$$
$$\times \left(2(1+\alpha) \sum_{i=1}^{M} W_i^{k+1} X_i \right). \quad (A.14)$$

Since $H^{k+1} = (S^k (S^k)^T)^{-1/2} \succeq \mathbf{0}$, then $H^{k+1} + (H^{k+1})^T \succeq \mathbf{0}$. We also have $W_i^{k+1}(W_i^{k+1})^T \succeq \mathbf{0}$. So $L(W^{k+1}, S, H^{k+1})$ is $2M$-strongly convex. Therefore,

$$L(W^{k+1}, S^k, H^{k+1}) - L(W^{k+1}, S^{k+1}, H^{k+1}) \geq 2M \|S^{k+1} - S^k\|_F^2. \quad (A.15)$$

(4) By combining Eqs. (A.6), (A.12) and (A.16), we can get:

$$L(W^k, S^k, H^k) - L(W^{k+1}, S^{k+1}, H^{k+1})$$
$$\geq (2M-1)\|S^{k+1} - S^k\|_F^2 + \frac{\beta}{2} \left\| W_i^{k+1} - W_i^k \right\|_F^2. \quad (A.16)$$

In summary, our objective function will monotonically decrease and it has lower bound, so it will converge to a stationary point.

References

[1] K. Wang, R. He, W. Wang, L. Wang, T. Tan, Learning coupled feature spaces for cross-modal matching, in: IEEE International Conference on Computer Vision, 2013, pp. 2088–2095.
[2] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Trans. Pattern Anal. Mach.Intell. 38 (10) (2016) 2010–2023.
[3] C. Kang, S. Xiang, S. Liao, C. Xu, C. Pan, Learning consistent feature representation for cross-modal multimedia retrieval, IEEE Trans. Multimed. 17 (3) (2015) 370–381.
[4] C. Xu, D. Tao, C. Xu, Multi-view intact space learning, IEEE Trans. Pattern Anal. Mach.Intell. 37 (12) (2015) 2531–2544.
[5] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[6] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, Lect. Notes Comput. Sci. 3940 (2006) 34.
[7] J.B. Tenenbaum, W.T. Freeman, Separating style and content with bilinear models, Neural Comput. 12 (6) (2000) 1247–1283.
[8] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, Int. J. Comput. Vis. 106 (2) (2014) 210–233.
[9] A. Sharma, D.W. Jacobs, Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 593–600.
[10] J. Liang, Z. Li, D. Cao, R. He, J. Wang, Self-paced cross-modal subspace matching, in: ACM SIGIR Conference on Research and Development in Information Retrieval, 2016, pp. 569–578.
[11] J. Liang, R. He, Z. Sun, T. Tan, Group-invariant cross-modal subspace learning, in: International Joint Conference on Artificial Intelligence, 2016, pp. 1739–1745.
[12] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2160–2167.
[13] X. Zhai, Y. Peng, J. Xiao, Heterogeneous metric learning with joint graph regularization for cross-media retrieval, in: AAAI Conference on Artificial Intelligence, 2013, pp. 1198–1204.
[14] X. Zhai, Y. Peng, J. Xiao, Learning cross-media joint representation with sparse and semisupervised regularization, IEEE Trans. Circuits Syst.Video Technol. 24 (6) (2014) 965–978.
[15] R. He, M. Zhang, L. Wang, Y. Ji, Q. Yin, Cross-modal subspace learning via pairwise constraints, IEEE Trans. Image Process. 24 (12) (2015) 5543–5556.
[16] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Cross-modal retrieval using multi-ordered discriminative structured subspace learning, IEEE Trans. Multimed. 19 (6) (2017) 1220–1233.
[17] J. Wu, Z. Lin, H. Zha, Joint latent subspace learning and regression for cross–modal retrieval, in: ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 917–920.
[18] D.-A. Huang, Y.-C. Frank Wang, Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition, in: IEEE International Conference on Computer Vision, 2013, pp. 2496–2503.
[19] X. Xu, Y. Yang, A. Shimada, R.-i. Taniguchi, L. He, Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts, in: ACM Conference on Multimedia, 2015, pp. 847–850.
[20] J. Wu, Z. Lin, H. Zha, Joint dictionary learning and semantic constrained latent subspace projection for cross-modal retrieval, in: ACM Conference on Information and Knowledge Management, 2018, pp. 1663–1666.
[21] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, W. Lu, Supervised coupled dictionary learning with group structures for multi-modal retrieval, in: AAAI Conference on Artificial Intelligence, 2013, pp. 1070–1076.
[22] C. Deng, X. Tang, J. Yan, W. Liu, X. Gao, Discriminative dictionary learning with common label alignment for cross-modal retrieval, IEEE Trans. Multimed. 18 (2) (2016) 208–218.
[23] M. Liu, L. Nie, M. Wang, B. Chen, Towards micro-video understanding by joint sequential-sparse modeling, in: ACM Conference on Multimedia, 2017, pp. 970–978.
[24] M. Liu, L. Nie, X. Wang, Q. Tian, B. Chen, Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning, IEEE Trans. Image Process. 28 (3) (2018) 1235–1247.
[25] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International Conference on Machine Learning, 2013, pp. 1247–1255.
[26] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: International Conference on Machine Learning, 2015, pp. 1083–1092.
[27] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: ACM Conference on Multimedia, 2014, pp. 7–16.
[28] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, in: International Joint Conference on Artificial Intelligence, 2016, pp. 3846–3853.
[29] Y. Hua, S. Wang, S. Liu, A. Cai, Q. Huang, Cross-modal correlation learning by adaptive hierarchical semantic aggregation, IEEE Trans. Multimed. 18 (6) (2016) 1201–1216.
[30] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: ACM Conference on Multimedia, 2017, pp. 154–162.
[31] V.E. Liong, J. Lu, Y.-P. Tan, J. Zhou, Deep coupled metric learning for cross–modal matching, IEEE Trans. Multimed. 19 (6) (2017) 1234–1244.
[32] H. Liu, F. Wang, X. Zhang, F. Sun, Weakly-paired deep dictionary learning for cross-modal retrieval, Pattern Recognit. Lett. (2018).
[33] D. Semedo, J. Magalhães, Cross-modal subspace learning with scheduled adaptive margin constraints, in: ACM Conference on Multimedia, 2019, pp. 75–83.
[34] X. Shen, F. Shen, Q.-S. Sun, Y. Yang, Y.-H. Yuan, H.T. Shen, Semi-paired discrete hashing: learning latent hash codes for semi-paired cross-view retrieval, IEEE Trans. Cybern. (2016).
[35] X. Xu, F. Shen, Y. Yang, H.T. Shen, X. Li, Learning discriminative binary codes for large-scale cross-modal retrieval, IEEE Trans. Image Process. 26 (5) (2017) 2494–2507.
[36] L. Liu, Z. Lin, L. Shao, F. Shen, G. Ding, J. Han, Sequential discrete hashing for scalable cross-modality similarity retrieval, IEEE Trans. Image Process. 26 (1) (2017) 107–118.
[37] F. Zhong, Z. Chen, G. Min, Deep discrete cross-modal hashing for cross-media retrieval, Pattern Recognit. 83 (2018) 64–77.
[38] V.E. Liong, J. Lu, Y.-P. Tan, Cross-modal discrete hashing, Pattern Recognit. 79 (2018) 114–129.
[39] M. Ranzato, Y.-L. Boureau, Y. LeCun, Sparse feature learning for deep belief networks, in: Advances in Neural Information Processing Systems, 2008, pp. 1185–1192.
[40] E. Kodirov, T. Xiang, S. Gong, Semantic autoencoder for zero-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[41] R.H. Bartels, G.W. Stewart, Solution of the matrix equation ax+ xb= c [f4], Commun. ACM 15 (9) (1972) 820–826.

[42] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (3) (2001) 475–494.
[43] H. Wang, F. Nie, H. Huang, Multi-view clustering and feature learning via structured sparsity., in: International Conference on Machine Learning, 2013, pp. 352–360.
[44] F. Feng, X. Wang, R. Li, I. Ahmad, Correspondence autoencoders for cross-modal retrieval, ACM Trans. Multimed. Comput.Commun. Appl. 12 (1s) (2015) 26.
[45] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM Conference on Multimedia, 2010, pp. 251–260.
[46] S.J. Hwang, K. Grauman, Reading between the lines: object localization using implicit cues from image tags, IEEE Trans. Pattern Analy. Mach.Intell. 34 (6) (2012) 1145–1158.
[47] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng, NUS-WIDE: a real-world web image database from National University of Singapore, in: ACM Conference on Image and Video Retrieval, 2009.
[48] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[49] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[50] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.
[51] M.D. Smucker, J. Allan, B. Carterette, A comparison of statistical significance tests for information retrieval evaluation, in: ACM International Conference on Information and Knowledge Management, 2007, pp. 623–632.
[52] Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, S. Yan, Modality-dependent cross-media retrieval, ACM Trans. Intell. Syst.Technol. 7 (4) (2016) 57.
[53] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[54] Y. Peng, J. Qi, X. Huang, Y. Yuan, CCL: cross-modal correlation learning with multigrained fusion by hierarchical network, IEEE Trans. Multimed. 20 (2) (2017) 405–420.

Jianlong Wu received his B.Eng. and Ph.D. degree from Huazhong University of Science and Technology in 2014 and Peking University in 2019, respectively. He is currently an assistant professor with the School of Computer Science and Technology, Shandong University. His research interests lie primarily in computer vision and machine learning, especially the cross-modal retrieval, unsupervised and semi-supervised learning. He has published more than 20 research papers in top journals and conferences, such as TIP, ICML, NeurIPS, and ICCV. He serves as a Senior Program Committee Member of IJCAI 2021, an area chair of ICPR 2020, and a reviewer for many top journals and conferences, including TPAMI, IJCV, ICML, and CVPR.

Xingyu Xie received his Bachelor and Master degree in Automation from Nanjing University of Aeronautics and Astronautics in 2016 and 2019, respectively. Now he is a Ph.D. candidate in the School of Electronics Engineering and Computer Science, Peking University. His research interests include machine learning and optimization.

Liqiang Nie received his B.Eng. and Ph.D. degree from Xi'an Jiaotong University in 2009 and National University of Singapore (NUS) in 2013, respectively. After PhD, Dr. Nie continued his research in NUS as a research follow for three and half years. He is currently a professor with the School of Computer Science and Technology, Shandong University. Meanwhile, he is the adjunct dean with the Shandong AI institute. His research interests lie primarily in multimedia computing and information retrieval. Dr. Nie has co-/authored more than 150 papers, like SIGIR, ACM MM, TOIS, TIP, received more than 6,000 Google Scholar citations. He is an AE of Information Science, and an area chair of ACM MM 2018/2019. He was granted several awards like SIGMM emerging leaders in 2018 and SIGIR 2019 best paper honorable mention.

Zhouchen Lin received the Ph.D. degree in applied mathematics from Peking University in 2000. He is currently a Professor with the Key Laboratory of Machine Perception (MOE), School of Electronics Engineering and Computer Science, Peking University. His research interests include computer vision, image processing, machine learning, pattern recognition, and numerical optimization. He is an Area Chair of CVPR 2014/2016/2019/2020/2021, ICCV 2015, NIPS/NeurIPS 2015/2018/2019/2020, AAAI 2019/2020, IJCAI 2020/2021, ICLR 2021, and ICML 2020/2021. He was an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and currently is an associate editor of the International Journal of Computer Vision. He is a Fellow of the IAPR and the IEEE.

Hongbin Zha received the M.S. and Ph.D. degrees in electrical engineering from Kyushu University, Fukuoka, Japan, in 1987 and 1990, respectively. In 1991, he joined Kyushu University as an Associate Professor. He was a Research Associate with the Kyushu Institute of Technology. He was also a Visiting Professor with the Center for Vision, Speech, and Signal Processing, Surrey University, U.K., in 1999. Since 2000, he has been a Professor with the Key Laboratory of Machine Perception, Peking University, Beijing, China. He has authored more than 300 technical publications in journals, books, and international conference proceedings. His research interests include computer vision, digital geometry processing, and robotics. He is a member of the IEEE Computer Society. He received the Franklin V. Taylor Award from the IEEE Systems, Man, and Cybernetics Society in 1999.