Joint Latent Subspace Learning and Regression for Cross-Modal Retrieval

Jianlong Wu, Zhouchen Lin, Hongbin Zha
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, P. R. China
Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, P. R. China
{jlwu1992,zlin}@pku.edu.cn,zha@cis.pku.edu.cn

ABSTRACT

Cross-modal retrieval has received much attention in recent years. It is a commonly used method to project multi-modality data into a common subspace and then retrieve. However, nearly all existing methods directly adopt the space defined by the binary class label information without learning as the shared subspace for regression. In this paper, we first adopt the spectral regression method to learn the optimal latent space shared by data of all modalities based on the orthogonal constraints. Then we construct a graph model to project the multi-modality data into the latent space. Finally, we combine these two processes together to jointly learn the latent space and regress. We conduct extensive experiments on multiple benchmark datasets and our proposed method outperforms the state-of-the-art approaches.

KEYWORDS

Cross-modal retrieval; Latent subspace learning; Regression

1 INTRODUCTION

With the rapid increase of multimedia data on the internet, the cross-modal retrieval topic has received much attention. As we can describe one common thing with multiple views information such as texts, images and videos, the cross-modal retrieval aims to retrieval one modal data from others directly [8]. For different modal data, they share the same underlying content, but there are semantic gaps and heterogeneous properties.

The main challenge of cross-modal retrieval is how to reduce the heterogeneous gap between different modal features and then measure their cross-modal similarity. Towards this issue, a lot of works have been proposed and the most popular solution is the latent subspace learning [3, 5, 9, 13, 15, 16, 21, 22]. For this kind of methods, they learn view-specific projection directions to project different features into a common latent space under which we can measure the similarity between learned features of different modalities.

There are mainly two basic issues for cross-modal retrieval that we should take into consideration. The first one is how to learn a common latent subspace for multi-modal data. The other one is how to preserve the local structure during regression and projection. The existing methods mainly focus on the second basic issue. Instead of learning the common latent space, most subspace learning based methods [4, 8, 17, 18] directly use the binary label space as the shared space. Although [17] tried to learn the latent space, they still used the binary class label matrix as the common space. We need to mention that we aim to learn a shared space that is good for matching instead of classification directly, so the binary label space might not be the optimal common space.

To solve the above problem, we propose a novel unified framework to simultaneously learn the common latent space and projection with correlation preserving. On the one hand, inspired by CCA, we construct a graph model based on the label information to learn the latent space with orthogonal constraints. On the other hand, we learn the view-specific projection matrix with regularization and correlation preserving items. Instead of simply regressing after latent space learning, we propose to combine these two aspects together to form a unified framework. We give a close form solution to optimize this unified problem.

For cross-modal retrieval, based on the training set, assume there are c categories in total, we hope to learn an underlying space which consists of c orthogonal basis vectors. The latent space learned by our proposed framework satisfies the following three basic conditions: (1) These latent basis vectors are orthogonal to each other; (2) Each latent basis vector is regularized to length one; (3) The learned latent space is related to both the label information and multiple modalities features. However, for the binary label space, it only meets the first requirement. It is not reasonable for retrieval to set the length of all projected vectors in the latent space to one as the binary label matrix does. Besides, while CCA does not utilize label information and other related latent space regression methods directly use the binary label matrix, our proposed methods learn the orthogonal latent space by incorporating both the label information and features of different modalities. The orthogonal latent space learned by our method is more suitable for cross-modal retrieval. Experimental results also demonstrate the superiority of our proposed method.

Main contributions of our work are summarized as follows:

1) We propose a novel unified framework to simultaneously learn the latent space and regress based on correlations preserving for cross-modal retrieval.

2) Instead of using the binary label information as the latent space, we adopt the spectral regression with orthogonal constraints to learn the optimal low dimension embedding.

3) Within this framework, we propose an efficient algorithm to optimize this problem. Experimental results show that our method can outperform the start-of-the-art methods.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

ACM ISBN 978-1-4503-5022-8/17/08…$15.00
https://doi.org/http://dx.doi.org/10.1145/3077136.3080678

2 RELATED WORK

Canonical Correlation Analysis (CCA) [3] is one of the most popular approaches for cross-modal retrieval. It aims to learn a latent space by maximizing the correlation between the projected features of two modalities. Denoting X_1 and X_2 as extracted features of two modalities, CCA [3] can be formulated as follows:

$$\begin{aligned} &\max_{W_1, W_2} \ tr(W_1^T X_1 X_2^T W_2) \\ &s.t.\ W_1^T X_1 X_1^T W_1 = \mathbf{I},\ W_2^T X_2 X_2^T W_2 = \mathbf{I}, \end{aligned} \tag{1}$$

where W_1 and W_2 represent the learned projection matrix for each modality features. Then, Rupnic et al. [14] generalized CCA for multi-view situation. Gong et al. [2] proposed a framework of three-view CCA and its kernel extension. Similar to CCA, Partial Least Squares (PLS) [13] and Bilinear Model (BLM) [16] are two classical methods that also try to learn subspace for cross-modal retrieval. However, these above methods only use the pairwise relationship and do not take the label information into consideration.

For discriminative analysis, Lin et al. [10] proposed common discriminant feature extraction (CDFE) method and Sharma et al. [15] came up with generalized multiview analysis (GMA) which extended linear discriminant analysis and marginal Fisher analysis (MFA) to their multiview cases. Kan et al. [7] proposed a multi-view discriminant analysis approach that jointly learns multiple view-specific linear transforms. For cross-modal hashing, Yu et al. [23] exploited coupled dictionary learning method, and Zhou et al. [24] adopted matrix factorization method.

Recently, researchers pay much attention to joint common space representation and feature selection. Wang et al. [18] presented a method to learn coupled feature spaces (LCFS). Wang et al. [17] extended LCFS by adding a Laplacian term to preserve the neighbourhood relationships during regression. He et al. [4] adopted pairwise constraints during projection. Kang et al. [8] used local group prior for better representation. By incorporating label information, these common space representation methods achieve good performance. However, all these methods directly used the binary label space as the common space without learning.

3 THE PROPOSED METHOD

3.1 Latent Subspace Learning

According to Eq. 1, we can see that CCA tries to project cross-modal features into an orthogonal space to maximize the correlation. Inspired by this, we hope to learn a common space with orthogonal constraint instead of directly using the binary label space. As spectral regression (SR) [1] enjoys very good performance in feature learning and graph embedding method can well characterize the local relationships, we adopt SR to learn the latent space. We first construct a graph to capture the local relationship between samples. For supervised retrieval tasks, based on the label information, the weight matrix W is defined as follows:

$$W_{ij} = \begin{cases} 1/N_t, & \text{if both the } i\text{-th and the } j\text{-th samples} \\ & \text{belong to the same class } t; \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where N_t is the number of samples in the t-th class. In the learned latent subspace, we hope that the neighbourhood relationships should be preserved and samples belong to same class should share the same representation. Let y_i denote the representation of the i-th sample in the learned latent space. Then the objective function for latent space learning is:

$$\min_Y \frac{1}{2} \sum_{i,j} \|y_i - y_j\|_2^2 W_{ij} = tr(Y^T L Y) \quad s.t.\ Y^T Y = \mathbf{I}, \tag{3}$$

where $L = D - W$ is the graph Laplacian matrix, D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$, and $Y = [y_1, y_2, \cdots, y_n]^T$. It is obvious that the binary label space is not the solution to this problem as it does not satisfy the orthogonal constraints. The problem in Eq. 3 can be easily solved by eigenvalue decomposition.

3.2 Latent Subspace Regression

Suppose we have M sets of features, $X_i = (x_1^i, x_2^i, \cdots, x_n^i), i = 1, \cdots, M$, from M modalities, where features in X_i are in d_i dimensions. n is the total number of samples. In general, M is often defined as 2 for the text image retrieval tasks. Given the latent space $Y \in \mathbb{R}^{n \times c}$, we regress each sample to its low dimensional embedding. For each modality features $X_i \in \mathbb{R}^{d_i \times n}$, we want to learn a projection matrix $U_i \in \mathbb{R}^{d_i \times c}$ to map each modality features into the common space. The objective function for latent space regression can be formulated as:

$$\min_U \sum_{i=1}^{M} \left(\|U_i^T X_i - Y^T\|_F^2 + \beta \|U_i\|_F^2 + \gamma tr(U_i^T X_i L X_i^T U_i) \right), \tag{4}$$

where β and γ are balance parameters. The regression problem in Eq. 4 can be regarded as an extended regularized least-squares problem. In Eq. 4, we simply use the Frobenius norm to regularize the projection matrix, which can also be regularized by other norms, such as ℓ_1 norm and $\ell_{2,1}$ norm, to achieve other desired characteristics. For the third term in Eq. 4, as we have:

$$\sum_{i,j} W_{ij} \|U^T x_i - U^T x_j\|_F^2 = tr(U^T X L X^T U), \tag{5}$$

the graph Laplacian is to preserve the structure of the original data. Here, we use the same weight matrix W as that in Eq. 2 to define the neighbourhood relationship.

3.3 Joint Learning and Regression

We hope to combine the latent space learning and regression together to find the optimal common space that can minimize the projection error. By combining the objective functions in Eq. 3 and Eq. 4, we get the unified objective function for our joint latent subspace learning and regression (JLSLR) method:

$$\begin{aligned} \mathcal{L}(U, Y) = \arg \min_{U, Y^T Y = I_c} & tr(Y^T L Y) + \alpha \sum_{i=1}^{M} \left(\|U_i^T X_i - Y^T\|_F^2 \right. \\ & \left. + \beta \|U_i\|_F^2 + \gamma tr(U_i^T X_i L X_i^T U_i) \right), \end{aligned} \tag{6}$$

where α, β and γ are balance parameters.

For the above problem, by fixing Y(or U), we can compute U(or Y) directly. We will give a close-form solution to the joint optimization problem in the following.

Fix Y, the problem in Eq. 6 is convex with respect to U. By setting the derivative of the objective function in Eq. 6 with respect to U_i to zero, we can get:

$$\frac{\partial \mathcal{L}(U,Y)}{\partial U_i} = 2X_iX_i^TU_i - 2X_iY + 2\beta U_i + 2\gamma X_iLX_i^TU_i = 0. \quad (7)$$

Then we can calculate the corresponding projection matrix by

$$U_i = (X_iX_i^T + \beta I + \gamma X_iLX_i^T)^{-1}X_iY, \quad i = 1, \cdots, M. \quad (8)$$

By substituting the above U_i into Eq. 6, the second part of Eq. 6 can be replaced with:

$$\begin{aligned}
&\alpha \sum_{i=1}^{M} \left(\|U_i^TX_i - Y^T\|_F^2 + \beta \|U_i\|_F^2 + \gamma tr(U_i^TX_iLX_i^TU_i) \right) \\
&= \alpha \sum_{i=1}^{M} \Big(tr(U_i^TX_iX_i^TU_i) - 2tr(U_i^TX_iY) + tr(Y^TY) \\
&\qquad + \beta \|U_i\|_F^2 + \gamma tr(U_i^TX_iLX_i^TU_i) \Big) \quad (9) \\
&= \alpha \sum_{i=1}^{M} \left(-tr\left(U_i^T(X_iX_i^T + \beta I + \gamma X_iLX_i^T)U_i \right) + tr(Y^TY) \right) \\
&= tr\left(Y^T \left(\alpha I_n - \alpha \sum_{i=1}^{M} X_i^T(X_iX_i^T + \beta I + \gamma X_iLX_i^T)^{-1}X_i \right) Y \right).
\end{aligned}$$

By denoting $Q_i = X_iX_i^T + \beta I + \gamma X_iLX_i^T$, the optimization problem in Eq. 6 with respect to Y can be reformulated as:

$$\min_{Y^TY=I_c} tr\left(Y^T(L + \alpha I_n - \alpha \sum_{i=1}^{M} X_i^TQ_i^{-1}X_i)Y \right). \quad (10)$$

The above optimization problem for Y has a closed form solution which can be well solved by the eigen-decomposition of matrix $L + \alpha I_{n\times n} - \alpha \sum_{i=1}^{M} X_i^TQ_i^{-1}X_i$. We pick up the eigenvectors corresponding to the c smallest eigenvalues.

For the above algorithm, the main computational complexity lies on the eigen-decomposition of solving the problem in Eq. 10 and inversion of matrix Q in Eq. 8 and Eq. 10, all of which can be solved conveniently.

In summary, we can efficiently solve the proposed JLSLR model with the close-form solutions. For latent space learning, we can easily see that the orthogonal space learned in Eq. 10 can well preserve the correlation with label information based graph and is highly related to the multi-modalities features. For latent space regression, projection matrices are well regularized and local relationship can also be preserved during regression to the common space.

4 EXPERIMENTAL RESULTS

4.1 Experimental settings

We evaluate the performance of our proposed method on two commonly used datasets, the Wiki [12] image-text dataset and the Pascal VOC [6] dataset. We mainly consider two cross-modal retrieval tasks: (1) Image query vs. Text database and (2) Text query vs. Image database. We compare the proposed JLSLR method with several related state-of-the-art methods, such as the PLS [13], BLM [16], CCA [3], CDFE [10], CCA-3V [2], GMLDA [15], GMMFA [15], LCFS [18], JFSSL [17], and LGCFL [8]. In this paper, we cite the results of other methods in [17, 19] except the LGCFL [8], which originally adopted different dataset setting on the Wiki dataset.

Table 1: MAP Comparison of different methods on the Wiki dataset with SIFT image features and LDA text features.

Methods	Image query	Text query	Average
PLS [13]	0.2402	0.1633	0.2032
BLM [16]	0.2562	0.2023	0.2293
CCA [3]	0.2549	0.1846	0.2198
CDFE [10]	0.2655	0.2059	0.2357
GMMFA [15]	0.2750	0.2139	0.2445
GMLDA [15]	0.2751	0.2098	0.2425
CCA-3V [2]	0.2752	0.2242	0.2497
LCFS [18]	0.2798	0.2141	0.2470
JFSSL [17]	0.3063	0.2275	0.2669
LGCFL [8]	0.3009	**0.2377**	0.2693
JLSLR	**0.3168**	0.2346	**0.2757**

The mean average precision (MAP) is a classical performance evaluation criterion for cross-modal retrieval. For details of the MAP computation, please refer to [12]. Higher MAP scores represent better performance.

For our proposed JLSLR method, we fine-tune the parameters α, β and γ in Eq. 6 by searching the gird of $\{10^{-2}, 10^{-1}, \cdots, 10^{2}, 10^{3}\}$ based on cross validation.

4.2 Results on Wiki dataset

The Wiki dataset [12] contains 2866 image-text pairs from 10 semantic classes. We use the same dataset setting as that in [17, 18], which splits 2866 pairs into a training set of 1300 pairs (130 pairs per class) and a testing set of 1566 pairs. For text, latent Dirichlet allocation (LDA) is adopted to extract 10 dimensions representation. 128 dimensional SIFT descriptor histograms [11] are used to represent the images.

In Table 1, we present the MAP scores of different approaches on the Wiki dataset. We can see that the performance of PLS [13], BLM [16], and CCA[3] is much lower than that of other methods, which can be attributed to the fact that these three methods don't take the label semantic information into consideration. More importantly, by learning the optimal latent space, our proposed JLSLR method can achieve an average MAP score of 0.2757, which is better than the results of all other state-of-the-art methods. Compared with the second best result achieved by LGCFL [8], our performance improves 2.4%.

Besides the given 128-dimensional SIFT image features of Wiki, we also adopt the convolutional neural network (CNN) to extract the 4096-dimensional features [20] for image presentation to evaluate the effect of different features. Table 2 shows the results based on the CNN image features and LDA text features on the Wiki dataset. Compared with the results in Table 1, CNN based image features lead to a significant improvement of the MAP scores, which can be attributed to the discriminative characteristics of CNN. Our method also outperforms all other start-of-the-art methods. JLSLR achieves nearly 3% performance gain compared to the second best method JFSSL [17].

Table 2: MAP Comparison of different methods on the Wiki dataset with CNN image features and LDA text features.

Methods	Image query	Text query	Average
GMLDA [15]	0.4084	0.3693	0.3889
CCA-3V [2]	0.4049	0.3651	0.3850
LCFS [18]	0.4132	0.3845	0.3989
JFSSL [17]	0.4279	**0.3957**	0.4118
LGCFL [8]	0.4347	0.3849	0.4098
JLSLR	**0.4579**	0.3901	**0.4240**

4.3 Results on Pascal VOC dataset

The Pascal VOC dataset [6] consists of 5011/4952 (training/testing) image-tag pairs from 20 different classes. The images corresponding to only one object are selected in the experiment, which result in 2808 pairs for the training set and 2841 pairs for the test set.

512 dimensional GIST features are used to represent the images. 399 dimensional word frequency features are used to represent texts. Table 3 displays the MAP scores of various methods for cross-modal retrieval. We can also observe that the proposed JLSLR method achieves the best performance with the average MAP of 0.3635. On both the image-to-text retrieval and the text-to-image retrieval tasks, our results are better than others.

From the experimental results on these two datasets, we can have the following three observations. First of all, the label information is very helpful for classification. As PLS [13], BLM [16] and CCA [3] only utilize the pairwise closeness, their performance is much lower than others that use the class information. Secondly, powerful representation can benefit the cross-modal retrieval performance. Compared with the score of the SIFT image features on Wiki dataset, the average performance of CNN features is about 50% higher. Finally, for the common subspace regression based methods, the binary label matrix based space is not the optimal one. As our JLSLR method learns the optimal latent space based on the label information and the minimized projection error, the performance of JLSLR is better than that of other methods.

5 CONCLUSIONS

In this paper, we proposed a novel framework to joint learn the latent space and regress for cross-modal retrieval. An optimal latent space is learned by jointly using label information to preserve structure and minimizing the total regression error for different modality features. Experiments on two commonly used datasets demonstrate the superiority of our method over the existing methods. One of our future works is to systematically investigate the influence of different norm regularization and multi-modal graph regularization terms.

6 ACKNOWLEDGEMENTS

Zhouchen Lin is supported by National Basic Research Program of China (973 Program) (Grant no. 2015CB352502), National Natural Science Foundation (NSF) of China (Grant nos. 61625301 and 61231002), and Qualcomm. Hongbin Zha is supported by Beijing Municipal Natural Science Foundation (Grant no. 4152006).

Table 3: MAP Comparison of different methods on the Pascal VOC dataset.

Methods	Image query	Text query	Average
PLS [13]	0.2757	0.1997	0.2377
BLM [16]	0.2667	0.2408	0.2538
CCA [3]	0.2655	0.2215	0.2435
CDFE [10]	0.2928	0.2211	0.2569
GMMFA [15]	0.3090	0.2308	0.2699
GMLDA [15]	0.3094	0.2448	0.2771
CCA-3V [2]	0.3146	0.2562	0.2854
LCFS [18]	0.3438	0.2674	0.3056
JFSSL [17]	0.3607	0.2801	0.3204
LGCFL [8]	0.4010	0.3200	0.3600
JLSLR	**0.4020**	**0.3250**	**0.3635**

REFERENCES

[1] D. Cai, X. He, and J. Han. Spectral regression for efficient regularized subspace learning. In *ICCV*, 2007.
[2] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.
[3] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
[4] R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin. Cross-modal subspace learning via pairwise constraints. *IEEE TIP*, 24(12):5543–5556, 2015.
[5] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE TCYB*, 44(6):793–804, 2014.
[6] S. J. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *IEEE TPAMI*, 34(6):1145–1158, 2012.
[7] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *IEEE TPAMI*, 38(1):188–194, 2016.
[8] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE TMM*, 17(3):370–381, 2015.
[9] J. Liang, Z. Li, D. Cao, R. He, and J. Wang. Self-paced cross-modal subspace matching. In *SIGIR*, 2016.
[10] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, 2006.
[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
[12] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010.
[13] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*. Springer, 2006.
[14] J. Rupnik and J. Shawe-Taylor. Multi-view canonical correlation analysis. In *Slovenian KDD Conference on Data Mining and Data Warehouses*, 2010.
[15] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012.
[16] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.
[17] K. Wang, R. He, L. Wang, W. Wang, and T. Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE TPAMI*, 38(10):2010–2023, 2016.
[18] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, 2013.
[19] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval. *ArXiv*, 2016.
[20] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE TCYB*, 47(2):449–460, 2017.
[21] Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, and S. Yan. Modality-dependent cross-media retrieval. *ACM TIST*, 7(4):57, 2016.
[22] J. Wu, Z. Lin, and H. Zha. Multi-view common space learning for emotion recognition in the wild. In *ICMI*, 2016.
[23] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *SIGIR*, 2014.
[24] J. Zhou, G. Ding, and Y. Guo. Latent semantic sparse hashing for cross-modal similarity search. In *SIGIR*, 2014.