Joint Feature Selection and Latent Subspace Regression for Cross-Media Retrieval Postprint

Authors: Liu Yun, Yu Zhi Building, Fu Qiang

Date: 2018-08-13T00:00:00+00:00

Abstract

Due to the massive existence of multimodal data, cross-modal retrieval has recently attracted considerable attention and typically involves two fundamental problems: similarity measurement and feature selection. Most existing methods focus solely on addressing the first problem: projecting multimodal data into a common subspace, measuring similarity between different data modalities, and then performing retrieval. To address the second problem, an _{2,1}-norm penalty term is imposed on the projection matrix to select relevant and discriminative features from the feature space. Simultaneously, spectral regression is employed to learn the optimal latent space with orthogonal constraints shared across all modalities. Subsequently, a graph model is constructed to project multimodal data into the latent space while preserving intra-modal similarity relationships. Extensive experiments conducted on two datasets demonstrate the effectiveness of the proposed method for cross-modal retrieval tasks.

Full Text

Preamble

Joint Feature Selection and Latent Subspace Regression for Cross-Media Retrieval

Liu Yun[1], Yu Zhilou[2]†, Fu Qiang[3]
(School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China)

Abstract: Cross-modal retrieval has recently drawn much attention due to the widespread existence of multi-modality data, which generally involves two fundamental problems: relevance measurement and coupled feature selection. However, most existing methods focus only on solving the first problem by projecting multi-modality data into a common subspace where similarity between different modalities can be measured. To address the second problem, we impose

, -norm penalties on the projection matrices to select relevant and discriminative features from different feature spaces. Simultaneously, we employ spectral regression to learn an optimal latent space shared by all modalities under orthogonal constraints. We then construct a graph model to project multi-modality data into this latent space while preserving intra-modality similarity relationships. Extensive experiments on two datasets demonstrate the effectiveness of our proposed method for cross-modal retrieval tasks.

Keywords: cross-media retrieval; feature selection; subspace learning; spectral regression

---

0 Introduction

With the rapid development of Internet technology, multi-modal data (such as images, text, video, and audio) have become ubiquitous online. Cross-media retrieval aims to use one data type as a query to retrieve relevant data pairs of another type. For instance, users can retrieve relevant images using text queries (Figure 1 [Figure 1: see original paper]) or search for relevant textual descriptions by submitting an interesting image as a query (Figure 2 [Figure 2: see original paper]). Cross-modal retrieval enables users to retrieve various modalities of data using any content form as a query, yielding more comprehensive results than single-modality retrieval.

A major challenge in cross-media retrieval tasks is the heterogeneity gap between different modality features, as multi-modal data typically reside in different feature spaces. The most direct approach to address this issue is to map data from different modalities into a shared space where similarity between modalities can be directly measured. Canonical Correlation Analysis (CCA) [1] is the most popular method, which learns a latent subspace by finding optimal basis vectors that maximize correlation between two sets of variables. CCA can be formulated as follows:

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

where W_1 and W_2 represent the mapping matrices for each modality' s features.

Building upon CCA, other algorithms have been proposed to handle different modality problems, such as Partial Least Squares (PLS) [2] and Bilinear Models (BLM) [3], which also attempt to learn subspaces for cross-modal retrieval. Beyond CCA, PLS, and BLM, additional methods address cross-modal issues. For example, Mahadevan et al. [4] proposed Maximum Covariance Unfolding, a manifold learning algorithm for dimensionality reduction of data from different input modalities. Mao et al. [5] introduced a parallel field alignment method for cross-media retrieval that integrates a manifold alignment framework from

the perspective of vector fields. Lin et al. [6] proposed a Coupled Discriminant Feature Extraction (CDFE) method to learn a common feature subspace where within-class scatter is minimized and between-class scatter is maximized. Sharma [7] extended Linear Discriminant Analysis (LDA) and Marginal Fisher Analysis (MFA) to their multi-view versions, namely Generalized Multi-view LDA (GMLDA) and Generalized Multi-view MFA (GMMFA), for cross-media retrieval problems. GMLDA and GMMFA consider semantic categories and have achieved promising results.

Furthermore, Zhai et al. [8] proposed Joint Representation Learning (JRL), which incorporates pairwise correlation and semantic information into a unified optimization framework. Zhuang et al. [9] presented a supervised coupled dictionary learning algorithm for cross-media retrieval. Zhai et al. [10] also proposed a heterogeneous metric learning approach capable of measuring content similarity between different media types.

Inspired by recent advances in deep learning, Ngiam et al. [11] applied deep networks to learn features from multiple modalities, focusing on learning representations of speech audio combined with lip video. Deep Restricted Boltzmann Machines [12] successfully learn joint representations of multi-modal data by first using separate modality-friendly latent models to learn low-dimensional representations for each modality, then fusing them into higher-dimensional deep architectures for joint representation. Building on deep representation learning, Andrew et al. [13] introduced Deep Canonical Correlation Analysis (DCCA), a deep learning method that learns complex non-linear projections of data from different modalities to produce highly linearly correlated representations.

However, most existing methods primarily focus on correlation measurement, while coupled feature selection remains inadequately addressed. Since real-world data often have high dimensionality with redundant and irrelevant features, selecting discriminative features for different modalities is crucial for cross-media retrieval.

1.1 Latent Space Learning

We first briefly introduce the notation used in this paper. For matrix X, x_i represents the i-th row, and x^j represents the j-th column. The $\ell_{2,1}$-norm of matrix X is defined as:

$$\|X\|_{2,1} = \sum_{i=1}^{n} \|x_i\|_2 = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} X_{ij}^2}$$

which represents the sum of ℓ_2 norms of all rows of matrix X.

As shown in Equation (1), CCA attempts to project features from different modalities into orthogonal spaces to maximize correlation between modalities.

In this regard, we aim to learn a common space through orthogonal constraints rather than directly using binary label space. Since Spectral Regression (SR) [14] demonstrates excellent performance in feature learning and graph embedding methods can effectively characterize local relationships, we adopt SR to learn the latent space. We first construct a graph to capture intra-modality relationships. For supervised retrieval tasks, the weight matrix W is defined based on label information as follows:

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

where N_t represents the number of samples in class t.

The objective function for latent space learning is:

$$\begin{aligned} \min_{Y} \quad & \|y - yW\|_2^2 = \mathrm{tr}(YLY^T) \\ \text{s.t.} \quad & YY^T = I \end{aligned}$$

where $Y = [y_1, y_2, ..., y_n] \in \mathbb{R}^{d \times n}$ represents the representations of samples in the learned latent space, and $L = D - W$ is the graph Laplacian matrix with D being a diagonal matrix where $D_{ii} = \sum_j W_{ij}$. In the learned latent subspace, we hope to preserve neighborhood relationships such that samples belonging to the same class share similar representations. Equation (3) can be solved through eigenvalue decomposition.

1.2.1 Feature Selection

Feature selection aims to identify an optimal set of features using selection criteria. By retaining some original features, it preserves their physical meaning and provides better readability and interpretability for the model. It is an important technique widely used in pattern analysis that reduces data variance by eliminating irrelevant and redundant features, decreases storage and computational costs, improves learning accuracy, and helps better understand the learning model or data. Therefore, feature selection is considered an effective dimensionality reduction method.

1.2.2 Latent Space Regression and Feature Selection

Assume we are given M groups of features from M modalities, denoted as $\{X^p \in \mathbb{R}^{d_p \times n}\}_{p=1}^{M}$, where features in X^p are d_p-dimensional and n is the total number of samples. Typically, in cross-media retrieval tasks, M is set to 2, representing image and text modalities. Given a latent space $Y \in \mathbb{R}^{d \times n}$, we

regress each sample to its low-dimensional embedding. For features X^p of each modality, we aim to learn mapping matrices $U^p \in \mathbb{R}^{d_p \times d}$ that project each modality' s features to the common space. The regression objective function can be expressed as:

$$\min_{U^p,Y} \sum_{p=1}^{M} \left(\|U^p X^p - Y\|_F^2 + \beta \|U^p\|_{2,1} + \gamma \mathrm{tr}(U^p X^p L X^{pT} U^{pT}) \right)$$

where β and γ are balancing parameters. The regression problem in Equation (4) can be viewed as an extended regularized least squares problem.

The second term in Equation (4) performs feature selection. Following the analysis of the $\ell_{2,1}$-norm in literature [15], we define:

$$\|U^p\|_{2,1} = \sum_{i=1}^{d_p} \|u_i^p\|_2 = \sum_{i=1}^{d_p} r_i^p \|u_i^p\|_2^2$$

where r_i^p is defined as:

$$r_i^p = \frac{1}{2\|u_i^p\|_2 + \varepsilon}$$

Here, ε is a smoothing term typically set to a very small value. It can be shown that r_i^p satisfies all conditions in Equation (5).

For the above problem, we can directly compute U^p (or Y) by fixing Y (or U^p). We provide a closed-form solution for the joint optimization problem below.

1.3 Unified Framework for Latent Space Learning, Regression, and Feature Selection

By combining the objective functions in Equations (3) and (7), we obtain the unified objective function:

$$\min_{U^p,Y} \mathrm{tr}(YLY^T) + \sum_{p=1}^{M} \left(\alpha \|U^p X^p - Y\|_F^2 + \beta \mathrm{tr}(U^p R^p U^{pT}) + \gamma \mathrm{tr}(U^p X^p L X^{pT} U^{pT}) \right)$$

where α, β, and γ are balancing parameters, and $R^p = \mathrm{Diag}(r_1^p, r_2^p, ..., r_{d_p}^p)$ with r_i^p defined as in Equation (8).

According to Lemma 1, the objective function in Equation (4) can be reformulated as:

$$\min_{U^p,Y} \sum_{p=1}^{M} (\|U^p X^p - Y\|_F^2 + \beta \mathrm{tr}(U^p R^p U^{pT}) + \gamma \mathrm{tr}(U^p X^p L X^{pT} U^{pT}))$$

Lemma 1 Let ϕ be a function satisfying all conditions in Equation (5). For any fixed $x \in \mathbb{R}$, there exists a dual potential function ψ such that:

$$\phi(x) = \inf_{s \in \mathbb{R}} \left(\frac{sx^2}{2} + \psi(s) \right)$$

where s is minimized by the function $\delta(x)$.

By fixing Y, Equation (9) is convex with respect to U^p. Taking the derivative of the objective function with respect to U^p and setting it to zero yields:

$$2X^p X^{pT} U^p - 2X^p Y^T + \beta R^p U^p + \gamma X^p L X^{pT} U^p = 0$$

The corresponding projection matrix can be computed as:

$$U^p = \left(X^p X^{pT} + \beta R^p + \gamma X^p L X^{pT}\right)^{-1} X^p Y^T, \quad p = 1, ..., M$$

Substituting U^p from Equation (11) into Equation (9), the second part of Equation (9) can be reformulated as an optimization problem with respect to Y:

$$\min_Y \mathrm{tr} \left(Y \left(L + \alpha I - \alpha \sum_{p=1}^{M} X^{pT} Q^p X^p \right) Y^T \right)$$

where $Q^p = \left(X^p X^{pT} + \beta R^p + \gamma X^p L X^{pT}\right)^{-1}$.

Y can be obtained through eigenvalue decomposition of matrix $\left(L + \alpha I - \alpha \sum_{p=1}^{M} X^{pT} Q^p X^p\right)$. To solve this, we select the eigenvectors corresponding to the d smallest eigenvalues.

For the third term of the equation, the Laplacian graph preserves the structure of the original data. Here we use the same weight matrix W as defined in Equation (2) to define neighborhood relationships.

In summary, we can efficiently solve the approximate solution of the model. For latent space learning, it is evident that the orthogonal space obtained in Equation (13) well preserves correlation based on graph label information and is closely related to multi-modal features. For latent space regression and feature selection, the projection matrices are well-regularized, enabling feature selection during the projection process to select effective features while preserving local relationships when regressing to the common space.

2.1 Experimental Setup

We evaluate our proposed method on two commonly used datasets: the Wiki image-text dataset [17] and the Pascal VOC dataset [18]. We consider two cross-modal retrieval tasks: image querying text database and text querying image database. Our method is compared with several state-of-the-art methods, including PLS [1], BLM [3], CCA [1], CDFE [6], CCA-3V [19], GMLDA [7], GMMFA [7], LCFS [20], SM [17], and SCM [17].

- **PLS, BLM, CCA**: Three classical methods that use pairwise information to learn a common latent subspace from multi-modal data, where similarity between different data modalities can be measured.
- **CDFE**: Learns a common feature subspace where within-class scatter is minimized and between-class scatter is maximized.
- **CCA-3V**: Three-view Canonical Correlation Analysis.
- **GMLDA**: Finds a set of projection matrices that bring samples from the same class closer while separating samples from different classes.
- **GMMFA**: A supervised extension of CCA that considers both CCA constraints and semantic information.
- **LCFS**: Integrates coupled linear regression, , -norm, and trace norm into a unified minimization formula, enabling simultaneous subspace learning and coupled feature selection.
- **SM**: Analyzes representations of images and text at a high-level abstract layer, specifically using multi-class logistic regression for classification.
- **SCM**: A combination of CCA and SM. SCM first obtains feature representations using CCA, then uses these representations to construct a semantic space, which improves retrieval performance over CCA and SM.

To evaluate performance, we conduct both image-to-text and text-to-image cross-modal retrieval tasks. Mean Average Precision (MAP) [17] is a classic performance metric for cross-modal retrieval. Specifically, given a set of queries, the Average Precision (AP) for each query is calculated as:

$$\mathrm{AP} = \frac{1}{T} \sum_{r=1}^{T} P(r) \cdot \delta(r)$$

where T is the number of relevant documents in the retrieval set, $P(r)$ denotes precision at the top r retrieved documents, and $\delta(r) = 1$ if the r-th retrieved document is relevant (i.e., belongs to the query's class), otherwise $\delta(r) = 0$. The MAP is then computed by averaging AP values across all queries in the query set. Higher MAP values indicate better cross-modal retrieval performance. In addition to MAP, we use precision-recall curves to evaluate the effectiveness of different methods.

2.2 Results on Wiki Dataset

The Wiki dataset [17] contains 2,866 image-text pairs from 10 semantic categories. We use 2,173 image-text pairs for training and 693 pairs for testing. For text, we employ Latent Dirichlet Allocation (LDA) to extract 10-dimensional representations. For images, we use 128-dimensional SIFT descriptor histograms [21] as features. We set $\alpha = 0.001$, $r = 14$, and $\beta = 4$.

Table 1 shows the MAP scores of our method compared with other algorithms. Figures 3 and 4 [Figure 3: see original paper][Figure 4: see original paper] display the recall rates for cross-modal retrieval. Our method achieves a MAP score of 0.2871 for image queries and 0.2232 for text queries, outperforming previous algorithms. Due to the incorporation of semantic information, CDFE, GMMFA, GMLDA, CCA-3V, and our method perform better than PLS, BLM, and CCA.

2.3 Results on Pascal VOC Dataset

The Pascal VOC dataset [18] consists of 5,011/4,952 (training/test) image-tag pairs from 20 different categories. In our experiments, we select images corresponding to only a single object, resulting in 2,808 pairs for training and 2,841 pairs for testing. We use 512-dimensional GIST features to represent images and 399-dimensional term frequency features to represent text. We set $\alpha = 0.01$, $r = 3$, and $\beta = 4$.

Table 2 shows the MAP scores of our method compared with other algorithms. Figures 5 and 6 [Figure 5: see original paper][Figure 6: see original paper] display the recall rates for cross-modal retrieval. Our method achieves a MAP score of 0.2871 for image queries and 0.2232 for text queries, outperforming previous algorithms.

2.4 Performance with Different Feature Types

We also test cross-modal retrieval performance using different types of features for images and text in the Wiki dataset. In addition to the features provided with the Wiki dataset, we extract 4,096-dimensional CNN features for images using Caffe and 100-dimensional text features using LDA. Table 3 shows the MAP scores of GMMFA, GMLDA, SM, and SCM on the Wiki dataset with these different feature types.

---

3 Conclusion

In this paper, we propose a novel joint learning framework to address cross-modal retrieval problems. The framework includes latent space learning for different modalities, , -norm for feature selection, latent space regression, and graph modeling. Under this framework, different projection matrices are learned to map multi-modal data into a common subspace, where relevant and discriminative features for different modalities are selected during the projection process, and local relationships are characterized using graph models. Experimental results on the Wikipedia dataset and Pascal VOC dataset demonstrate that our proposed method improves retrieval efficiency across modalities.

In future work, we can incorporate correlations between modalities to preserve inter-modal relationships in the common mapping space, or combine multiple views to find optimal representations and learn common feature spaces from multi-view spaces.

References

[1] Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods [J]. Neural Computation, 2004, 16(12): 2639-2664.

[2] Rosipal R, Krämer N. Overview and recent advances in partial least squares [C]//Proc of International Conference on Subspace, Latent Structure and Feature Selection. Springer-Verlag, 2005: 34-51.

[3] Tenenbaum J B, Freeman W T. Separating style and content with bilinear models [J]. Neural Computation, 2000, 12(6): 1247-1283.

[4] Mahadevan V, Pereira J C, Vasconcelos N, et al. Maximum covariance unfolding: manifold learning for bimodal data [C]//Advances in Neural Information Processing Systems. 2011: 918-926.

[5] Mao X, Lin B, Cai D, et al. Parallel field alignment for cross-media retrieval [C]//Proc of ACM International Conference on Multimedia. ACM Press, 2013: 897-906.

[6] Lin D, Tang X. Inter-modality face recognition [C]//Proc of European Conference on Computer Vision. Springer, 2006: 13-26.

[7] Sharma A. Generalized multiview analysis: a discriminative latent space [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2012: 2160-2167.

[8] Zhai X, Peng Y, Xiao J. Learning cross-media joint representation with sparse and semisupervised regularization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(6): 965-978.

[9] Zhuang Y, Wang Y, Wu F, et al. Supervised coupled dictionary learning with group structures for multi-modal retrieval [C]//Proc of AAAI Conference on Artificial Intelligence. 2013.

[10] Zhai X, Peng Y, Xiao J. Heterogeneous metric learning with joint graph regularization for cross-media retrieval [C]//Proc of the 27th AAAI Conference on Artificial Intelligence. AAAI Press, 2013: 1198-1204.

[11] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning [C]//Proc of International Conference on Machine Learning. 2011: 689-696.

[12] Srivastava N, Salakhutdinov R. Multimodal learning with deep boltzmann machines [C]//Proc of International Conference on Neural Information Processing Systems. Curran Associates Inc, 2012: 2222-2230.

[13] Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis [C]//Proc of International Conference on Machine Learning. 2013: 1247-1255.

[14] Cai D, He X, Han J. Spectral regression for efficient regularized subspace learning [C]//Proc of IEEE International Conference on Computer Vision. IEEE Press, 2007: 1-8.

[15] He R, Tan T, Wang L, et al. , regularized correntropy for robust feature selection [C]//Proc of IEEE International Conference on Computer Vision. IEEE, 2012: 2504-2511.

[16] Nikolova M, Ng M K. Analysis of half-quadratic minimization methods for signal and image recovery [J]. SIAM Journal on Imaging Sciences, 2005.

[17] Rasiwasia N, Pereira J C, Coviello E, et al. A new approach to cross-modal multimedia retrieval [C]//Proc of International Conference on Multimedia. ACM, 2010: 251-260.

[18] Hwang S J, Grauman K. Reading between the lines: object localization using implicit cues from image tags [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(6): 1145-1158.

[19] Gong Y, Ke Q, Isard M, et al. A multi-view embedding space for modeling internet images, tags, and their semantics [J]. International Journal of Computer Vision, 2014, 106(2): 210-233.

[20] Wang K, He R, Wang W, et al. Learning coupled feature spaces for cross-modal matching [C]//Proc of IEEE International Conference on Computer Vision. IEEE Computer Society, 2013: 2088-2095.

[21] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[22] Wu J, Lin Z, Zha H. Joint latent subspace learning and regression for cross-modal retrieval [C]//Proc of International ACM SIGIR Conference. ACM Press, 2017: 917-920.

[23] Wang K, He R, Wang L, et al. Joint feature selection and subspace learning for cross-modal retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2010-2023.

[24] Peng Y, Zhai X, Zhao Y, et al. Semi-supervised cross-media feature learning with unified patch graph regularization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(3): 583-596.

[25] Wang K, Yin Q, Wang W, et al. A comprehensive survey on cross-modal retrieval [J]. 2016.

[26] Wei Y, Zhao Y, Zhu Z, et al. Modality-dependent cross-media retrieval [J]. ACM Transactions on Intelligent Systems and Technology, 2016, 7(4): 57.

[27] Zhou J, Ding G, Guo Y. Latent semantic sparse hashing for cross-modal similarity search [M]. ACM Press, 2014.

[28] Yu Z, Wu F, Yang Y, et al. Discriminative coupled dictionary hashing for fast cross-media retrieval [C]//Proc of International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM Press, 2014: 395-404.

[29] Wei Y, Zhao Y, Lu C, et al. Cross-modal retrieval with CNN visual features: a new baseline [J]. IEEE Transactions on Cybernetics, 2017, 47(2): 449-460.

[30] Kang C, Xiang S, Liao S, et al. Learning consistent feature representation for cross-modal multimedia retrieval [J]. IEEE Transactions on Multimedia, 2015, 17(3): 370-381.

[31] Yan J, Zhang H, Sun J, et al. Joint graph regularization based modality-dependent cross-media retrieval [J]. Multimedia Tools and Applications, 2017(6): 1-19.

Note: Figure translations are in progress. See original paper for figures.