Dictionary Learning-Based Cross-Media Retrieval Technology Postprint

Authors: Yudan Qi, Huaxiang Zhang, Liu Yihe

Date: 2018-05-02T00:00:00+00:00

Abstract

Cross-media information retrieval faces challenges posed by the heterogeneity of different modal data. To better overcome this heterogeneity issue and improve retrieval accuracy among multi-modal data, this paper proposes a novel cross-media retrieval technique based on dictionary learning. First, sparse coefficients between two different modal data are learned through dictionary learning methods. Then, a feature mapping scheme projects them into a common feature subspace via two distinct projection matrices. Finally, correlation between different modalities is enhanced by aligning labels within the same class. Experimental results demonstrate that, compared with traditional homogeneous subspace learning methods, the dictionary-based algorithm exhibits superior classification performance, and the proposed method outperforms several state-of-the-art approaches on two datasets.

Full Text

Preamble

Cross-media retrieval technology based on dictionary learning

Qi Yudan, Zhang Huaxiang†, Liu Yihe (School of Information Science & Engineering, Shandong Normal University, Jinan 250358, China)

Abstract: In the study of cross-media retrieval, capturing and correlating heterogeneous features originating from different modalities remains a challenge. To address this problem, this paper presents a novel cross-modal retrieval framework based on coupled dictionary learning. Firstly, it obtains sparse coefficients from different modalities by imposing dictionary learning. Then, it projects the data samples from different modalities into a common feature space. Moreover, it leverages label information to align cross-modal data sample pairs in the common space so as to encourage the inherent correlation across different

modalities. Simulation experimental results show that the method based on dictionary learning algorithm has superior recognition performance in comparison with methods based on traditional mid-level feature subspace. Experiment results on two public datasets demonstrate that our method outperforms several state-of-the-art methods.

Key Words: cross-modal retrieval; dictionary learning; sparse representation; modality-dependent; feature mapping

0 Introduction

Early data retrieval primarily focused on single-modal data, where queries and retrieved data belonged to the same modality. For example, given a text query, single-modal methods directly match it with raw text data on the web rather than consistent images. Typically, these single-modal methods cannot be applied to cross-media retrieval. Cross-media retrieval is a new research field in content-based multimedia retrieval. Due to the heterogeneity among different modalities of data, direct mutual retrieval is difficult to achieve. How to solve the heterogeneity problem between different modalities to enable mutual retrieval of multimedia data has become an important research issue in cross-media retrieval.

In recent years, many new methods have been proposed for cross-media retrieval that achieve mutual retrieval across modalities by mining potential relationships between them. Specifically, the most authoritative Canonical Correlation Analysis (CCA) [**?**] is a classical feature learning method that maximizes the correlation between two groups of features to obtain low-dimensional representations of both features in a subspace with the highest correlation. The proposal of CCA greatly promoted research in cross-media retrieval, and its extended methods have been widely applied in this field. For example, Rasiwasia et al. [**?**] integrated cross-media retrieval problems from both association hypothesis and abstraction hypothesis perspectives. Hwang et al. [**?**] modeled the relative importance of words based on user-provided annotation order to improve cross-modal retrieval accuracy. Ballan et al. [**?**] used Kernel Canonical Correlation Analysis (KCCA) to develop cross-view retrieval methods for establishing correlations between images and text.

In addition to CCA-based methods, many other cross-media retrieval approaches exist. Among them, Partial Least Squares (PLS) [**?**] is a novel multivariate statistical data analysis method first proposed by Wold and Albano in 1983. In recent years, it has rapidly developed in theory, methods, and applications. Chen et al. [**?**] applied PLS to cross-media retrieval, using it to transform visual features into text feature space and learn semantics to measure similarity between two different modalities. Besides these, there are Linear Discriminant Analysis (LDA) and Marginal Fisher Analysis (MFA). Sharma et al. [**?**] extended generalized multiview analysis based on LDA and

MFA into Generalized Multiview LDA (GMLDA) and Generalized Multiview MFA (GMMFA) for cross-media retrieval.

Cross-media hashing embeds different forms of data into a common low-dimensional Hamming space for cross-media retrieval, mapping high-dimensional data objects into concise hash codes such that similar data objects have identical or similar hash codes, and then obtains similarity between original data by measuring similarity between binary hash codes. This approach has attracted extensive attention in recent years. For example, Yu et al. [**?**] proposed a Discriminative Coupled Dictionary Hashing (DCDH) method. Additionally, with breakthrough progress of deep learning in computer vision, some deep learning methods have also been used for cross-media similarity retrieval models, such as Convolutional Neural Networks (CNN), Recursive Neural Networks (RNN), and Auto-encoders. In deep learning-based research, Andrew et al. [**?**] proposed Deep Canonical Correlation Analysis (DCCA), which learns nonlinear projections between different modalities to make the learned data highly linearly correlated. Wang et al. [**?**] proposed a supervised Multi-modal Deep Neural Network (MDCCN). Jiang et al. [**?**] proposed a deep learning-based real-time network cross-media retrieval method that ranks elements in image feature vectors according to their contribution and then eliminates unnecessary features.

Even though these existing methods solve cross-media retrieval problems, most focus only on learning correlations between two modalities through distance in two feature spaces, thereby ignoring different semantic features. Additionally, class label information is not fully utilized. To sufficiently learn heterogeneous features in different feature spaces, sparse dictionary learning has received increasing attention.

1 Overview

This paper combines dictionary learning with modality independence to learn a new cross-media retrieval technique based on modality independence and dictionary learning. Firstly, multi-modal data is transformed into sparse representations through dictionary learning while ensuring the generated representations are uniform. Then, linear regression maps these sparse coefficients, projecting sparse representations from different modalities into two common semantic spaces through two different projection matrices. Figure 2 [Figure 2: see original paper] describes the framework of the proposed method. Figures 2(a) and 2(c) are two linear regression operations, representing the mapping from image and text feature spaces to semantic spaces, respectively. Thus, multi-modal data with the same semantics can be associated in the common latent subspace. Figure 2(b) is a correlation analysis operation that maintains interconnections of multi-modal data in the common space. Combining Figure 2(a) with 2(b) learns projections for I2T; similarly, a different projection for T2I is learned through joint optimization of Figures 2(b) and 2(c).

This paper focuses on multimedia retrieval between images and text (Figure 1 [Figure 1: see original paper]), using images to search text documents or text to search images (I2T and T2I). Figure 1(a) shows that given an image of an airplane, the task is to find text reports related to this image. Figure 1(b) shows that given a text document about two pilots, the task is to find related images.

On the other hand, this paper separates the retrieval tasks of the two modalities, i.e., the modality-independent method. Modality independence [**?**] differs from previous methods that learn a pair of projections; it learns two pairs of mappings that project image-to-text and text-to-image retrieval from their original feature spaces into two common latent subspaces. If two tasks are learned simultaneously, the resulting common subspace is the optimal subspace shared by I2T and T2I, which is typically not optimal for semantic understanding of the retrieval modality. For example, in image-to-text retrieval, accurate representation of the query in the image semantic space is considered more important than the text to be retrieved. If the query semantics are incorrectly judged, it becomes more difficult to retrieve relevant text. If performed separately, when retrieving text from an image, the image can be projected alone into its semantic space without text interference, achieving optimal semantic understanding of the image. After understanding the image semantics, data retrieval becomes more accurate, thereby improving cross-media retrieval precision. Experiments demonstrate that modality independence is also effective compared to other algorithms.

2 Related Work

2.1 Dictionary Learning

Dictionary learning [**?**, **?**, **?**, **?**, **?**] aims to find a special set of sparse codes from training data, where this set of sparse elements can linearly represent features of the original data, thereby representing as much content as possible with as little data as possible. Therefore, dictionaries essentially reduce dimensionality for massive datasets, and sparse representation uses minimal data to represent maximal features to improve retrieval efficiency. Due to this effectiveness, dictionary learning has been widely applied.

The main method for processing heterogeneous data in cross-media retrieval using sparse dictionaries is to convert correlations between original data into relationships between sparse coefficients through dictionary reconstruction. The objective function is expressed as:

$$\min_{D,A} \|X - DA\|_F^2 + \alpha\|A\|_1$$
$$\text{s.t.} \quad \|d_i\|_2 \leq 1, \quad \forall i \in [1:K]$$

where $X \in \mathbb{R}^{P\times N}$ is a dataset with dimension P and sample number N; $D = [d_1, d_2, ..., d_K] \in \mathbb{R}^{P\times K}$ is the learned dictionary; d_i is the i-th atom in the

dictionary; K represents dictionary size; $A \in \mathbb{R}^{K\times N}$ is the sparse coefficient of sample data X obtained according to dictionary D; $\|\cdot\|_F$ denotes the Frobenius norm. The purpose is to make the linear combination of dictionary D and sparse coefficient A as close as possible to data sample X. Additionally, α controls sparsity.

2.2 Retrieval Task Description

Based on the overview of the proposed optimization framework, the two retrieval tasks are described in detail below.

2.2.1 Image Retrieval Text This section discusses retrieving consistent text in cross-media retrieval. The I2T linear regression term is a regression operation from image space to semantic space. Let $X_V = [v_1, v_2, ..., v_n] \in \mathbb{R}^{p\times n}$ be the image dataset with dimension p and sample number n; $X_T = [t_1, t_2, ..., t_n] \in \mathbb{R}^{q\times n}$ be the text dataset with dimension q and sample number n; $D_V \in \mathbb{R}^{p\times k}$ be the dictionary for learning images; $D_T \in \mathbb{R}^{q\times k}$ be the dictionary for learning text; $A_V \in \mathbb{R}^{k\times n}$ be sparse coefficients for images; $A_T \in \mathbb{R}^{k\times n}$ be sparse coefficients for text, where image sparse coefficients A_V and text sparse coefficients A_T are obtained based on learned dictionaries.

Let $Y(i) = [y_1, y_2, ..., y_n] \in \mathbb{R}^{c\times n}$ be the keyword matrix, i.e., the common semantic subspace. The projection matrix for images is $W_V^1 \in \mathbb{R}^{c\times k}$ and for text is $W_T^1 \in \mathbb{R}^{c\times k}$. The purpose of dictionary learning is to learn two projection matrices for images and text separately. The framework is formulated as follows:

$$\begin{aligned}\min_{D_V,D_T,A_V,A_T,W_V^1,W_T^1} &\|X_V - D_V A_V\|_F^2 + \|X_T - D_T A_T\|_F^2 \\ &+ \|W_V^1 A_V - Y\|_F^2 + \|W_T^1 A_T - Y\|_F^2 + \alpha_1\|A_V\|_1 + \alpha_1\|A_T\|_1 \\ &+ \alpha_2(\|W_V^1\|_F^2 + \|W_T^1\|_F^2) + \alpha_3\|W_V^1 A_V - W_T^1 A_T\|_F^2\end{aligned}$$

where: - A_V and A_T are sparse coefficients obtained through dictionary learning for image and text modalities respectively - $W_V^1 A_V$ and $W_T^1 A_T$ are linear regression terms that project sparse coefficient matrices into semantic space, gathering multimedia data with same semantics together - Parameters α_1, α_2, α_3 are balance parameters: α_1 controls sparsity, α_2 controls projection matrix complexity to avoid overfitting - $\|W_V^1 A_V - W_T^1 A_T\|_F^2$ is the correlation analysis term that makes data of the same class closer and enhances correlation between different modalities

2.2.2 Text Retrieval Image This section discusses retrieving consistent images in cross-media retrieval. The T2I linear regression term is a regression operation from text space to semantic space, similar to image retrieval. Let $X_V = [v_1, v_2, ..., v_n] \in \mathbb{R}^{p\times n}$ be the image dataset; $X_T = [t_1, t_2, ..., t_n] \in \mathbb{R}^{q\times n}$ be the text dataset; $A_V \in \mathbb{R}^{k\times n}$ be image sparse coefficients; $A_T \in \mathbb{R}^{k\times n}$ be text

sparse coefficients; $D_V \in \mathbb{R}^{p\times k}$ be the image dictionary; $D_T \in \mathbb{R}^{q\times k}$ be the text dictionary.

Let $Y(t) = [y_1, y_2, ..., y_n] \in \mathbb{R}^{c\times n}$ be the common semantic subspace. Two different projection matrices from image retrieval text are set up: $W_V^2 \in \mathbb{R}^{c\times k}$ and $W_T^2 \in \mathbb{R}^{c\times k}$. Using projection matrices, sparse representations A_V and A_T from both modalities are projected into a common feature space. The framework is described as follows:

$$\begin{aligned}\min_{D_V,D_T,A_V,A_T,W_V^2,W_T^2} &\|X_V - D_V A_V\|_F^2 + \|X_T - D_T A_T\|_F^2 \\ &+ \|W_V^2 A_V - Y(t)\|_F^2 + \|W_T^2 A_T - Y(t)\|_F^2 + \alpha_1\|A_V\|_1 + \alpha_1\|A_T\|_1 \\ &+ \alpha_2(\|W_V^2\|_F^2 + \|W_T^2\|_F^2) + \alpha_3\|W_V^2 A_V - W_T^2 A_T\|_F^2\end{aligned}$$

where $W_V^2 A_V$ and $W_T^2 A_T$ project sparse coefficient matrices into the keyword subspace, gathering multimedia data with same semantics together; α_2 controls complexity to avoid overfitting; and the correlation analysis term makes data of the same class closer, improving their correlation. Similarly, in this model, correlations between different modalities are represented.

3 Optimization

The optimization problems for I2T and T2I are unconstrained matrix optimization problems. Therefore, equations (2) and (3) are non-convex with many local optimal solutions. To solve this problem, an algorithm is designed to find fixed points. It can be noted that when fixing the other two terms, equation (2) is convex with respect to the remaining term. Similarly, equation (3) can be convex when fixing the other two terms. Minimization of each term is completed by fixing two of D_V (D_T), A_V (A_T), or W_V^1 (W_T^2) and iteratively updating the third. The specific optimization strategy is as follows:

Input: Image feature matrix X_V, text feature matrix X_T, and consistent semantics Y for images and text.

Initialization: Initialize dictionaries D_V, D_T and sparse coefficients A_V, A_T. Relying on FDDL [**?**], set W_V^1, W_T^1 as identity matrices.

Iteration: Repeat until convergence:

1. Update dictionaries D_V, D_T using equations (4) and (5), fixing sparse coefficients A_V, A_T and projection matrices W_V^1, W_T^1.
2. Update sparse coefficients A_V, A_T using equation (6), fixing dictionaries D_V, D_T and projection matrices W_V^1, W_T^1.
3. Update projection matrices W_V^1, W_T^1 using equation (7), fixing dictionaries D_V, D_T and coefficients A_V, A_T.

Output: Dictionaries D_V, D_T and projection matrices W_V^1, W_T^1.

First, updating dictionary D_V with fixed sparse coefficients A_V and projection matrix W_V^1 yields:

$$\min_{D_V} \|X_V - D_V A_V\|_F^2 \quad \text{s.t.} \quad \|d_i\|_2 \leq 1, \quad \forall i \in [1:K] \tag{4}$$

This is a Quadratically Constrained Quadratic Program (QCQP) that can be solved through Lagrangian dual techniques [**?**].

Similarly, dictionary D_T can be solved as:

$$\min_{D_T} \|X_T - D_T A_T\|_F^2 \quad \text{s.t.} \quad \|d_i\|_2 \leq 1, \quad \forall i \in [1:K] \tag{5}$$

Then, with dictionaries D_V and projection matrices W_V^1 fixed, sparse coefficients are solved. From equation (2):

$$\min_{A_V} \|X_V - D_V A_V\|_F^2 + \|W_V^1 A_V - Y\|_F^2 + \alpha_1 \|A_V\|_1 \tag{6}$$

Through analysis and partial derivative calculation:

$$\frac{\partial}{\partial A_V} = -2D_V^T X_V + 2D_V^T D_V A_V + 2(W_V^1)^T W_V^1 A_V - 2(W_V^1)^T Y + \alpha_1 \text{sign}(A_V) = 0$$

Finally, with dictionary D_V and sparse coefficients A_V fixed, projection matrix W_V^1 is updated:

$$\min_{W_V^1} \|W_V^1 A_V - Y\|_F^2 + \alpha_2 \|W_V^1\|_F^2 + \alpha_3 \|W_V^1 A_V - W_T^1 A_T\|_F^2 \tag{7}$$

Similarly, the solution can be obtained as:

$$W_V^1 = (2YA_V^T + 2\alpha_3 W_T^1 A_T A_V^T)(2A_V A_V^T + 2\alpha_2 I + 2\alpha_3 A_V A_V^T)^{-1}$$

In summary, each part of the objective function designed in this paper is convex, thus having an optimal solution. To obtain the final result, the above steps need to be repeated continuously until final convergence. This process is summarized in the algorithm below. A similar method can be applied to text retrieval image.

Algorithm 1: Alternating Iterative Optimization Process for I2T

4 Experiments

To verify the performance of the proposed cross-media retrieval method, the following experiments were conducted: first, experimental settings and evaluation metrics are described, then the proposed method is compared with several other models.

4.1 Experimental Design

The method is evaluated on two public image-text datasets: Wikipedia text-image dataset [**?**] and Pascal Sentence dataset [**?**].

Wikipedia Dataset: Contains 2,866 image-text pairs from 10 classes. The dataset is randomly divided into 2,173 training pairs and 693 test pairs.

Pascal Sentence Dataset: Contains 1,000 image-text pairs annotated with 20 semantic categories (50 pairs per category). For each category, 30 image-text pairs are randomly selected as training set and the rest as test set.

For both datasets, ground-truth labels of each image-text pair are used to construct semantic vectors (10-dimensional for Wikipedia dataset and 20-dimensional for Pascal Sentence dataset) for semantic representation. Specifically, 4,096-dimensional CNN features are used to represent images and 100-dimensional LDA features to represent text.

Experiments focus on two retrieval tasks: a) text query in image database; b) image query in text database. Normalized Correlation (NC) is used to measure similarity between features of different media objects in the transformed subspace. Retrieval performance is evaluated through Precision-Recall (PR) curves and mean Average Precision (mAP). mAP is the average of Average Precision (AP) for each query. AP is defined as $\mathrm{AP} = \frac{1}{T}\sum_{r=1}^{N} P(r)\delta(r)$, where T is the number of retrieved data belonging to the same category; $P(r)$ denotes precision at the r-th retrieved data; $\delta(r) = 1$ if the r-th retrieved data shares the same label as the query, otherwise $\delta(r) = 0$. In experiments, $N = 50$. The mAP value is obtained by averaging AP values over all queries, where larger mAP indicates higher algorithm accuracy.

4.2 Performance Comparison

To objectively evaluate the proposed method, it is compared with several major algorithms: Canonical Correlation Analysis (CCA) [**?**], Deep Canonical Correlation Analysis (DCCA) [**?**], Semantic Matching (SM) [**?**], Semantic Correlation Matching (SCM) [**?**], Three-view CCA (TV CCA) [**?**], Generalized Multiview Linear Discriminant Analysis (GMLDA) [**?**], Generalized Multiview Marginal Fisher Analysis (GMMFA) [**?**], and Modality-Dependent Cross-Media Retrieval (MDCR) [**?**]. All comparison methods use the same features and training set in our experiments.

On Wikipedia Dataset: After testing different parameter settings, $\mu = 0.02$,

$\epsilon = 10^{-2}$ are determined. To further verify experimental efficiency, 4,096-dimensional CNN image features and 100-dimensional LDA text features are used. Parameters are set as $\alpha_1 = 0.1$, $\alpha_2 = 0.5$, $\alpha_3 = 0.5$ for optimizing I2T and T2I. Comparison results are shown in Table 1 , demonstrating that the proposed method improves mAP by 1.9% to 18.4% on average. Precision-recall curves for image-to-text and text-to-image tasks are shown in Figure 3 [Figure 3: see original paper], with scope being the number of top retrieved data. It can be observed that our method achieves better results and outperforms several state-of-the-art methods.

On Pascal Sentence Dataset: Parameters are set as $\mu = 0.02$, $\epsilon = 10^{-4}$, $\alpha_1 = 0.01$, $\alpha_2 = 0.5$, $\alpha_3 = 0.5$ for optimizing I2T and T2I. Comparison results in Table 2 show the proposed method improves mAP by 2.4% to 17.7% on average. Precision-recall curves for both tasks are shown in Figure 4 [Figure 4: see original paper]. The proposed method achieves better results for both tasks.

In experiments, μ is the step size in the alternating update process and ϵ is the convergence condition, with their values ranging between 0 and 1. Smaller values yield more accurate alternating update results. Parameters on the test set are determined based on cross-validation results on the training set rather than arbitrary selection.

5 Conclusion

This paper designs an effective cross-media retrieval model that generates sparse coefficients through dictionary learning and projects different forms of data into a common subspace. Label alignment is used to enhance correlation between different modalities, allowing intrinsic relationships between modalities to be well exploited in this space. Additionally, this paper separates image-to-text and text-to-image retrieval tasks, training them separately to learn two pairs of projections that fully exploit their respective feature advantages. Extensive experiments on Wikipedia and Pascal Sentence datasets demonstrate that the proposed method not only improves retrieval efficiency between multi-modal data but is also effective for single-modal data recognition. It expands sparse representation for dictionary learning and proposes an effective iterative algorithm for solving minimization problems. Experimental results show that the proposed method is effective.

References

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

[2] Rasiwasia N, Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval[C]// Proc of the 18th ACM International Conference on Multimedia. 2010: 251-260.

[3] Hwang S, Grauman K. Learning the relative importance of objects from tagged images for retrieval and cross-modal search[J]. International Journal of Computer Vision, 2012, 100: 134-153.

[4] Ballan L, Uricchio T, Seidenari L, et al. A cross-media model for automatic image annotation[C]// Proc of International Conference on Multimedia Retrieval. Glasgow, United Kingdom: ACM Press. 2014: 73.

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

[6] Chen Y, Wang L, Wang W, et al. Continuum regression for cross-modal multimedia retrieval[C]// Proc of IEEE International Conference on Image Processing. 2013: 1949-1952.

[7] Sharma A, Kumar A, Daume H, et al. Generalized multiview analysis: a discriminative latent space[C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2012: 2160–2167.

[8] 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. Gold Coast, Queensland, ACM Press, 2014: 395-404.

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

[10] Wang W, Yang X, Ooi B C, et al. Effective deep learningbased multi-modal retrieval[J]. The VLDB Journal, 2016, 25(1): 79–101.

[11] Jiang B, Yang J, Lv Z, et al. Internet cross-media retrieval based on deep learning[J]. Journal of Visual Communication & Image Representation, 2017, 48: 356-366.

[12] Pan Q, Liang Y, Zhang L, et al. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis[C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE Computer Society, 2012: 2216-2223.

[13] Zhuang Y, Wang Y F, Wu F, et al. Supervised coupled dictionary learning with group structures for multi-modal retrieval[C]// Proc of the 27th AAAI Conference on Artificial Intelligence. 2013: 1070-1076.

[14] Huang Dean, Wang Y C F. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition[C]// Proc of IEEE International Conference on Computer Vision. 2013: 2496-2503.

[15] Xu X, Yang Y, Shimada A, et al. Semi-supervised coupled dictionary learning for cross-modal retrieval in Internet images and texts[C]// Proc of ACM International Conference on Multimedia. 2015: 847-850.

[16] Xu X S. Dictionary Learning Based Hashing for Cross-Modal Retrieval[C]// Proc of ACM on Multimedia Conference. 2016: 177-181.

[17] Wei Y, Zhao Y, Zhu Z, et al. Modality-dependent cross-media retrieval[J]. ACM Trans on Intelligent Systems and Technology, 2016, 17(4): 57.

[18] Wang Kaiye, He Ran, Wang Wei, et al. Learning coupled feature spaces for cross-modal matching[C]// Proc of IEEE International Conference on Computer Vision. 2013: 2088-2095.

[19] Putthividhy D, Attias H T, Nagarajan S S. Topic regression multi-modal Latent Dirichlet Allocation for image annotation[C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2010.

[20] Schölkopf B, Platt J, Hofmann T. Efficient sparse coding algorithms[C]// Advances in Neural Information Processing Systems. 2006: 801-808.

[21] Wu Fei, Han Yahong, Liu Xiang, et al. The heterogeneous feature selection with structural sparsity for multimedia annotation and hashing: a survey[J]. International Journal of Multimedia Information Retrieval, 2012, 1(1): 3-18.

[22] Yang M, Zhang D, Feng X. Fisher discrimination dictionary learning for sparse representation[C]// Proc of IEEE International Conference on Computer Vision. 2011: 543-550.

[23] Rasiwasia N, Mahajan D, Mahadevan V, et al. Cluster canonical correlation analysis[C]// Proc of the 17th International Conference on Artificial Intelligence and Statistics. 2014: 823-831.

[24] Wang Yanfei, Wu Fei, Song Jun, et al. Multi-modal mutual topic reinforce modeling for cross-media retrieval[C]// Proc of the 22nd ACM International Conference on Multimedia. 2014: 307-316.

[25] 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.

[26] Cao Y, Long M, Wang J, et al. Deep visual-semantic hashing for cross-modal retrieval[C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California: ACM Press, 2016: 1445-1454.

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

[28] Shang X, Zhang H, Chua T S. Deep learning generic features for cross-media retrieval[M]// MultiMedia Modeling. Miami, FL: Springer International Publishing, 2016: 264-275.

[29] Feng F, Wang X, Li R. Cross-modal Retrieval with Correspondence Autoencoder[C]// Proc of International Conference on Multimedia. Orlando, Florida: ACM Press, 2014: 7-16.

[30] Cao Yue, Long Mingsheng, Wang Jiamin, et al. Correlation hashing network for efficient cross-modal retrieval[C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2016.

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