Cross-Modal Self-Taught Learning for Image Retrieval

Liang Xie, Peng Pan*, Yansheng Lu, and Sheng Jiang

School of Computer Science and Technology,
Huazhong University of Science and Technology,
Wuhan, China, 430074
whutxl@hotmail.com, {panpeng,lys,jwt}@mail.hust.edu.cn

Abstract. In recent years, cross-modal methods have been extensively studied in the multimedia literature. Many existing cross-modal methods rely on labeled training data which is difficult to collect. In this paper we propose a cross-modal self-taught learning (CMSTL) algorithm which is learned from unlabeled multi-modal data. CMSTL adopts a two-stage self-taught scheme. In the multi-modal topic learning stage, both intra-modal similarity and multi-modal correlation are preserved. And different modalities have different weights to learn the mutli-modal topics. In the projection stage, soft assignment is used to learn projection functions. Experimental results on Wikipedia articles and NUS-WIDE show the effectiveness of CMSTL in both cross-modal retrieval and image hashing.

Keywords: cross-modal retrieval, image retrieval, image hashing, self-taught learning.

1 Introduction

Over last decades, with the advance of computer network and multimedia technologies, we have witnessed a massive explosion of multimedia content on the web. Large amounts of multi-modal data, such as images and texts, are generated, shared and accessed on social websites, e.g., Flickr, Wikipedia and YouTube. Classical uni-modal approach [1] is not able to deal with these multi-modal data.

Cross-modal approach which analyzes the correlation of heterogeneous modalities, has been extensively studied in the multimedia literature [2,3,4,5,6]. They can solve the retrieval of heterogeneous data, e.g., using text query to retrieve images. Since the cross-modal correlation is beneficial for bridging the "semantic gap", the performance of uni-modal retrieval is likely to be improved by cross-modal approach [3,7]. In many previous studies on cross-modal retrieval [3,4,5], class labels are used to construct the cross-modal correlation. If an image and a text belong to the same class, then they are considered as relevant. However, labeling the training data is usually labor intensive and expensive, which makes the

* Corresponding author.

X. He et al. (Eds.): MMM 2015, Part I, LNCS 8935, pp. 257–268, 2015.

labeled data difficult to be collected. Another disadvantage of using class labels for cross-modal learning is that classes may be limited in some domain. For example, 'apple' may not be connected to 'computer' if labeled data is about fruit.

In this paper we propose a novel method: cross-modal self-taught learning (CMSTL) for cross-modal image retrieval. Unlike previous methods which analyze cross-modal correlation according to class labels, CMSTL analyzes the latent correlation according to the co-occurrence of heterogenous data. CMSTL does not require any labeled training data, it only needs multi-modal documents for training. Multi-modal data are usually tend to co-occur in the same documents on many social websites. Therefore, CMSTL is more practical than previous methods in the real world.

CMSTL adopts the two-stage self-taught scheme [8]. Since self-taught scheme have two stages: unsupervised learning and supervised learning, methods based on this scheme can both benefit from effective unsupervised and supervised approaches. In the first stage, an effective hierarchical multi-modal approach is proposed to generate latent topic. Intra-modal topics of each modality are first generated from intra-modal similarity. Then they are all combined to a uniform multi-modal topic space. These two generations are combined in a joint objective function from which the final multi-modal topics can be optimized. In the generation of multi-modal topics, different modalities have different weights. It makes our method more adaptive than traditional methods which treat all modalities equally. In the second stage, a soft supervised projection is used, and all modality are projected to the latent topic space via kernel least square regression (KLSR). We test our method on two real-world datasets: Wikipidia articles [3] and NUS-WIDE [9]. The experimental results show the effectiveness of our method in cross-modal retrieval. We further extend CMSTL to image hashing and the results show that our cross-modal method improves uni-modal image retrieval.

The rest of this paper is organized as follows. In section 2 we discuss the related work. In section 3 we describe the framework of our CMSTL. Section 4 shows the experimental results of cross-modal retrieval and image hashing on two datasets. Finally we conclude in Section 5.

2 Related Work

In recent years, many efforts have been devoted to the cross-modal multimedia retrieval. Most cross-modal methods focus on learning an uniform space where different modalities are correlated. One type of methods analyze latent correlation, which is based on the co-occurrence of multi-modal data, to construct the uniform space. In [2], multimedia correlation space (MMCS) is constructed from multi-modal data. However, its main limitation is the lack of out-of-sample generalization. New queries must be first mapped to their nearest neighbors in the training set. [10] uses canonical correlation analysis (CCA) and cross-modal factor analysis (CFA) in the context of audio-image retrieval. Both CCA and CFA learn a latent space where two modalities are correlated. Kernel CCA (KCCA)

is proposed in [11] to extract translation invariant semantics of text documents written in multiple languages. [12] also uses KCCA to model correlation between web images and corresponding text captions.

Another type of cross-modal methods learn the cross-modal correlation space from class labels. In [3], both images and texts are represented by the posteriors of class labels. Logistic regression is used to project documents into the probabilistic space of class. In [4], a semantic generation model is proposed for cross-modal retrieval, the correlation of different modalities is described by their generation from semantic labels. In [13], a joint graph regularized heterogeneous metric learning (JGRHML) algorithm is proposed to improve the semantic metric, which is learned through label propagation. [5] proposes the latent semantic cross-modal ranking (LSCMR) to discriminatively learn a latent low-rank embedding space by structural large margin learning. LSCMR is trained from supervised ranking examples, which are determined by the class labels shared among documents. These methods have the limitation that their performance depends on the class labels. If the class labels are not sufficient to describe the cross-modal correlation well, their performance may be affected.

Recently some methods are proposed to solve the problem of large-scale multimedia retrieval, including cross-modal retrieval. Spectral hashing [14] and self-taught hashing [8] are two representative methods for large-scale uni-modal retrieval. Spectral Hashing uses a subset of thresholded eigen-vectors of the graph Laplacian as hashing codes. Self-taught hashing adopts the self-taught scheme which is similar to our methods, the difference is that it uses hard assignment to obtain the binary codes, which may cause the loss of semantic information. Cross-modality similarity-sensitive hashing (CMSSH) [15] uses the supervised similarity learning method boosting to embed data into the hamming space. Multimodal latent binary embedding (MLBE) learns the hashing codes in a probabilistic framework. Both CMSSH and MLBE rely on training data labeled by classes, and they cannot work while the labels are missing. Cross-view hashing (CVH) [16] requires predefined cross-modal similarities of the training data. But if the cross-modal similarity matrix is set to identity matrix, CVH can be learned from unlabeled multi-modal data.

3 The Description of Our Framework

In this section we describe the framework of cross-modal self-taught learning (CMSTL), which contains two learning stages: multi-modal topic learning (MMTL) and projection function learning. MMTL learns latent topics from multi-modal training data, then projection function is learned to represent test data from each modality by these topics.

3.1 Learning Multi-modal Latent Topics

MMTL learns latent topics which can correlate different modalities, from unlabeled multi-modal training data. Suppose there are N multi-modal documents

$D_1, \ldots, D_N$. Each document D_n contains M modalities, and $D_n = \left\{x_n^1, \ldots, x_n^M\right\}$, where x_n^m is the feature of m-th modality. In our experiments, we only consider two modalities: image and text, thus $M = 2$ and D_n is an image-text pair in fact. But our methods can also be applied to more modalities ($M > 2$).

For each modality, we construct its intra-modal similarity graph A_m, which is defined as:

$$A_{ij}^m = \begin{cases} sim_m(x_i^m, x_j^m), & x_i^m \text{ and } x_j^m \text{ are } k \text{ nearest neighbors} \\ 0, & otherwise \end{cases} \tag{1}$$

where $sim_m(x_i^m, x_j^m)$ is the similarity of x_i^m and x_j^m.

To learn the latent topics from multi-modal documents, we first generate intra-modal topic matrix $F_m \in \mathbb{R}^{N \times T_m}$ for each modality. T_m is the number of intra-modal topics. F_m can preserve the latent information of the m-th modal features, it is obtained by minimizing the following Graph Laplacian regularizer:

$$\sum_{k=1}^{T_m} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}^m \left(\frac{f_{ik}^m}{d_{ii}^m} - \frac{f_{jk}^m}{d_{jj}^m} \right) = Tr\left(F_m^T L_m F_m\right) \tag{2}$$

where f_{ik}^m is an element of F_m, and d_{ii}^m is the sum of i-th row of A_m. $L_m = I - D_m^{-1/2} A_m D_m^{-1/2}$, I is the identity matrix, D_m is the diagonal matrix and its diagonal element is d_{ii}^m. $Tr(\cdot)$ denotes the trace operator.

After we obtain intra-modal latent semantic matrices $F_m|_{m=1}^M$, we use them to generate the multi-modal latent topic matrices $F \in \mathbb{R}^{N \times T}$. T is the number of multi-modal topics. F combines the latent information of all modalities, and each F_m generates the final F by minimizing the following function:

$$\|F - F_m W_m\|_F^2 \tag{3}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm. $W_m \in \mathbb{R}^{T_m \times T}$ is the weight matrix for the generation of F.

In sum, the multi-modal topic matrix F is hierarchically generated. At first, intra-modal topic matrices $F_m|_{m=1}^M$ are generated by intra-modal similarity. Then multi-modal topic matrix F is generated from all $F_m|_{m=1}^M$.

We optimize the hierarchical generation in a joint framework. By combining (2) and (3), we arrive at the following objective function:

$$\begin{aligned} \min_{F} \sum_{m=1}^{M} & \left(Tr\left(F_m^T L_m F_m\right) + \alpha_m^2 \left\|F - F_m W_m\right\|_F^2 \right) \\ s.t. \quad & F_m^T F_m = I, \ m = 1, \ldots, M \\ & F^T F = I \\ & \sum_{m=1}^{M} \alpha_m = 1 \end{aligned} \tag{4}$$

where α_m is the weight parameter, it represents the importance of m-th modality for the generation of F, and we can easily find that (4) is convex respect to α_m.

By setting the derivative of (4) w.r.t. W_m to zero, we have:

$$W_m = F_m^T F \tag{5}$$

Substituting W_m in (4), the objective function becomes:

$$\sum_{m=1}^{M} \left(Tr\left(F_m^T L_m F_m\right) + \alpha_m^2 Tr\left(I - F^T F_m F_m^T F\right)\right) \tag{6}$$

We adopt an alternating optimization to solve (6). More specifically, we alternatively update F, F_m and α_m to optimize the objective function.

1) Optimizing F: We fix F_m and α_m, then (6) can be reformulated as:

$$\begin{aligned} &\max_F Tr\left(F^T \sum_{m=1}^{M} \left(\alpha_m^2 F_m F_m^T\right) F\right) \\ &s.t. \qquad\qquad F^T F = I \end{aligned} \tag{7}$$

It is obviously that (7) is an eigenvalue problem, and we can obtain F by eigen-decomposition of $\sum_{m=1}^{M} \left(\alpha_m^2 F_m F_m^T\right)$.

2) Optimizing F_m: We fix F and α_m. According to the trace property: $Tr\left(F^T F_m F_m^T F\right) = Tr\left(F_m^T F F^T F_m\right)$, (6) can be transformed to:

$$\begin{aligned} &\min_{F_m} Tr\left(F_m^T C_m F_m\right) \\ &s.t. \quad F_m^T F_m = I \end{aligned} \tag{8}$$

where

$$C_m = L_m - \alpha_m^2 F^T F \tag{9}$$

We can also find that F_m is learned by solving the eigenvalue problem of (8),

3) Optimizing α_m: F and F_m are fixed, by using Lagrange multiplier, we can obtain:

$$\alpha_m = \frac{1/Tr\left(I - F^T F_m F_m^T F\right)}{\sum_{i=1}^{M} 1/Tr\left(I - F^T F_i F_i^T F\right)} \tag{10}$$

The whole alternating optimization process is illustrated in Algorithm 1. In the implementation of this algorithm, we initialize F_m by solving the eigenvalue problem of (2), and $\alpha_m|_{m=1}^{M}$ are set to the same. Since the objective function is lower bounded by 0 and it will keep decreasing in each step, its convergence is guaranteed. One advantage of our topic learning is that the importance of different modality for generating the latent topic is different, while previous cross-modal methods such as CCA treat all modalities equally. Thus our topic learning methods is more adaptive. Another advantage is that the hierarchical generation preserves not only intra-modal similarity, but also multi-modal correlation which is seldom considered in previous methods.

Algorithm 1. The learning process of MMLSA

Input:
$A_m|_{m=1}^M$
Output:
F
1: Compute $L_m|_{m=1}^M$;
2: Initialize $F_m|_{m=1}^M$ and $\alpha_m|_{m=1}^M$;
3: **while** $t < T$ **do**
4: Update F by solving the eigenvalue problem of (7);
5: Update $F_m|_{m=1}^M$ by solving the eigenvalue problem of (8);
6: Update $\alpha_m|_{m=1}^M$ according to (10);
7: t=t+1;
8: **end while**

3.2 Learning Projection Functions

Heterogenous data in training documents are correlated by multi-modal topics. However, for new data out of the training set, we still have to learn the explicit projection from each modality to the topic space. In the previous self-taught scheme [8], F is converted into binary codes via thresholding, and then binary topics can be treated as class labels. At last, classifiers which project documents into the latent topic space, are trained via some classification methods, such as support vector machine (SVM). However, using binary codes may lose some semantic information. Generally, it's better to use a soft assignment for latent topics.

In this paper we use a probabilistic soft assignment to learn the projection functions. Previous methods [3,17] have shown the effectiveness of image retrieval in probabilistic space. Therefore, the probabilistic representation for latent topics should be better than binary codes. To obtain a probabilistic space, we use Gaussian mixture on F to learn S probabilistic topics, and predict $Y = \left[y_1^T, \ldots, y_N^T\right]^T$. y_n is the posterior vector of f_n and it is predicted by the Gaussian mixture model.

Y is the final topic representation for training documents, we can use supervised methods to learn the function to project new documents into this topic space. Since Y is not binary, classification methods cannot work in this condition, and we use kernel least square regression (KLSR) instead. The objective function of KLSR is described as follows:

$$\min_{P_m} \|K_m P_m - Y\|_F^2 + \lambda Tr\left(P_m^T K_m P_m\right) \tag{11}$$

where K_m is the kernel function of the m-th modal training features. P_m is the projection weight for the m-th modality. $\lambda > 0$ is the regularization parameter. By setting the derivative of (11) w.r.t P_m to zero, we can easily get the weight matrix:

$$P_m = (K_m + \lambda I)^{-1} Y \tag{12}$$

In order to effectively compute the inverse. We perform singular singular value decomposition (SVD) to obtain a pseudo-inverse of $K_m + \lambda I = U_m \Lambda {V_m}^T$, which is computed by $(K_m + \lambda I)^{-1} = V_m \bar{\Lambda} U_m^T$, where $\bar{\Lambda}$ is defined as:

$$\bar{\Lambda}_{ii} = \begin{cases} 0, & \text{if } \Lambda_{ii} < \epsilon \\ \Lambda_{ii}^{-1}, & \text{otherwise} \end{cases} \tag{13}$$

where $\epsilon = 1$ is a threshold. Then the weight matrix is computed by:

$$P_m = V_m \bar{\Lambda} U_m^T Y \tag{14}$$

Algorithm 2. The process of cross-modal self-taught learning

Input:
$A_m|_{m=1}^M, K_m|_{m=1}^M$
Output:
$P_m|_{m=1}^M$
1: Compute F according Algorithm (1);
2: Train the Gaussian Mixture Model and predict the posterior matrix Y on F;
3: **for** each modality m **do**
4: Do the SVD: $K_m + \lambda I = U_m \Lambda {V_m}^T$;
5: Compute $\bar{\Lambda}$ according to (13);
6: Compute P_m according to (14) ;
7: **end for**

The self-taught learning process is described in Algorithm 2. Given a new document from the m-th modality, we first compute its kernel vector k, then we can compute its topic representation according to $y = kP_m$. Note that some elements in y may not between 0 and 1, thus we have to normalize y to make it be a probabilistic vector. Softmax function is used to normalize y, for each element y_i in y, it is normalized by:

$$\bar{y}_i = \frac{\exp(y_i)}{\sum_{j=1}^{S} \exp(y_j)} \tag{15}$$

Finally, we use the normalized vector $\bar{y} = [\bar{y}_1, \ldots, \bar{y}_T]$ to represent new document for retrieval.

4 Experiments

4.1 Datasets and Features

In this paper, two real world multi-modal image datasets: Wikipedia articles [3] and NUS-WIDE [9] are used for evaluation. These two datasets are both split

to training set and test set. All methods are only learned from training set, and their performance is evaluated on test set.

Wikipedia dataset was assembled from the "Wikipedia feature articles". It contains 2,866 multi-modal documents (image-text pairs), and each of them is labeled with exactly one of 10 semantic concepts which can be used as the ground truth. Documents share the same concept are regarded as relevant. 2,173 of the image-text pairs in Wikipedia dataset are chosen as training set, and the rest 693 are used as test set. 10-D LDA features are extracted for texts, and images are represented by 128-D SIFT BoVWs [1].

NUS-WIDE dataset contains 269,648 multi-modal documents, each multi-modal documents is also an image-text pair and text in NUS-WIDE refers to the associated social tags. Each image-text pairs are labeled by 81 concepts that can be used for evaluation. We prune the original NUS-WIDE to form a new dataset consisting of 203,597 image-text pairs by keeping the images that have at least one tag and one concept. Then this dataset is split to 5,090 training set and 198,507 test set. 1000-D binary features are used for tags, and 500-D SIFT BoVWs are used for images [2].

4.2 Results of Cross-Modal Retrieval

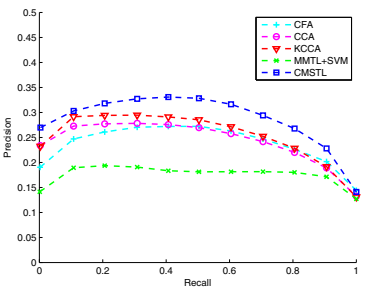
In this subsection we compare our method with several representative methods for cross-modal retrieval. CFA[10], CCA[18], and KCCA[12] are used for comparison. We also show the performance of a baseline method MMLT+SVM, where the latent topics is first learned by Algorithm 1, and then the self-taught scheme described in [8] is used to learn projection functions for each modality respectively. In our methods, the topic dimension of $F_m|_{m=1}^M$, F, and Y are all set to the same, they are set to 8 on Wikipedia dataset and 32 on NUS-WIDE dataset. In the computing of intra-modal similarity matrices $A_m|_{m=1}^M$, histogram intersection distance is used for both image and text features. The nearest neighbors for $A_m|_{m=1}^M$ is set to 500 on Wikipedia dataset, and 1000 on NUS-WIDE dataset. In all kernel based methods, the histogram intersection kernel which is same to histogram intersection distance, is used for both image and text features. The dimension of latent space in all methods are set to the same, 8 in Wikipedia dataset and 32 in NUS-WIDE dataset. Normalize correlation (NC) distance [3] are used for all methods in cross-modal retrieval.

We adopt the non-interpolated mean average precision (MAP) to evaluate the performance of cross-modal retrieval. Given a query and the rank list of retrieval set, the average precision (AP) is defined as:

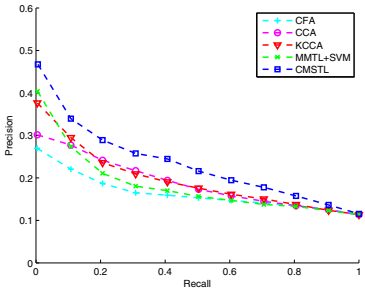
$$AP = \frac{1}{R}\sum_{i=1}^{N} pre(i)rel(i) \tag{16}$$

[1] All features can be downloaded from `http://www.svcl.ucsd.edu/projects/crossmodal/`

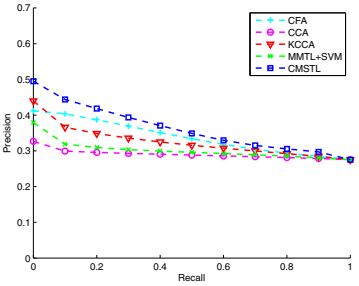
[2] All features can be downloaded from `http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm`



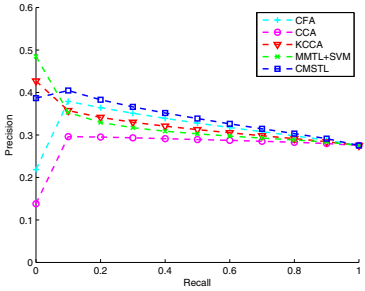
(a) Image query (Wikipedia)



(b) Text query (Wikipedia)



(c) Image query (NUS-WIDE)



(d) Text query (NUS-WIDE)

Fig. 1. The PR curves of cross-modal retrieval on two datasets

where N is the size of retrieval set. R is the number of relevant documents in the retrieved set, $pre(i)$ is the precision of top i retrieved documents. $rel(i) = 1$ if the i-th retrieved documents is relevant to query, otherwise $rel(i) = 0$. The MAP score is the mean of AP scores from all the queries. Besides MAP, the retrieval performance is also measure by Precision-Recall (PR) curve.

We evaluate two types of cross-modal retrieval on test set. In image query, images in test set are used as queries and texts in test set form retrieval set. In text query, test texts are queries and test images form retrieval set. Table 1 shows the MAP scores of cross-modal retrieval. It should be noted that the results of SCM have been reported in [3], thus its results on NUS-WIDE is null. We can find our CMSTL performs best on both two datasets. It performs even better than SCM which is learned from labeled training data on Wikipedia dataset. In usual methods learned from the training data labeled by ground truth should obtain better performance. CMSTL performs better than KCCA, CCA and CFA which treat all modalities equally, which confirms that it is worth weighting different modalities for the learning of latent topics. CMSTL also obtains higher MAP scores than MMTL+SVM, which demonstrates the advantage of our soft

Table 1. The comparison of MAP scores for cross-modal retrieval. The best results are marked in bold.

Datasets	Wikipedia			NUS-WIDE		
Methods	Image Query	Text Query	Mean	Image Query	Text Query	Mean
SCM[3]	0.277	0.226	0.252	-	-	-
CFA	0.245	0.166	0.210	0.313	0.315	0.314
CCA	0.249	0.193	0.221	0.291	0.292	0.292
KCCA	0.262	0.196	0.229	0.326	0.320	0.323
MMLT+SVM	0.181	0.181	0.181	0.305	0.313	0.309
CMSTL	**0.295**	**0.234**	**0.265**	**0.366**	**0.344**	**0.355**

projection scheme. The PR curves on two datasets are shown in Figure 1, we can see they are consistent with the MAP scores.

4.3 Results of Image Hashing

We have shown the superiority of CMSTL in cross-modal retrieval. In this subsection we show its performance on image hashing for large-scale retrieval. CMSTL can be easily extended for image hashing by thresholding the probabilistic latent vector $\bar{y}$. For the i-th topic, we obtain its threshhold θ_i by computing the mean of all values of this topic in the retrieval set. If $\bar{y}_i > \theta_i$, we set it to 1, otherwise we set it to 0. Generally cross-modal correlation is benefit to image retrieval. Thus CMSTL should be effective in image hashing.

We compare CMSTL based hashing (CMSTLH) to several representative hashing methods, including laplacian co-hashing (LCH)[19], spectral hashing (SH) [14] [3], self-taught hashing (STH) [8] [4], cross-view hashing (CVH) [16], where CVH is a cross-modal hashing method. For all hashing methods, their code length is set to 16 on Wikipedia and 32 on NUS-WDE, which are optimal on two datasets respectively. We use MAP50 to measure the performance, MAP50 is similar to MAP described in section 4.2, the difference is that MAP50 is compute by the top 50 relevant documents, it can be compute by setting $N = 50$ in (16). The performance is also evaluated on test set. In each retrieval, an image is used as query and other images in test set form retrieval set.

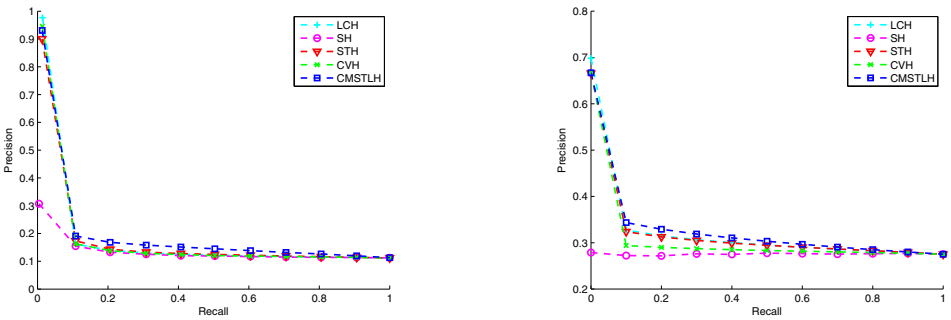
Table 2 shows the MAP scores on Wikipedia and NUS-WIDE. We can observe that CMSTL obtains the highest MAP scores on both two datasets. Although CVH is a cross-modal method, it does not always performs better than uni-modal methods. These results confirm that our method is also suit to image hashing. Since our methods is not specially designed for hashing, the improvement of CMSTLH is not large. Figure 2 shows the PR curve of image hashing, the results are consistent with Table 2.

[3] `http://www.cs.huji.ac.il/~yweiss/SpectralHashing/`

[4] `http://www.dcs.bbk.ac.uk/~dell/publications/dellzhang_sigir2010_suppl.html`

Table 2. The comparison of MAP50 scores for image hashing. The best results are marked in bold.

Datasets	LCH	SH	STH	CVH	CMSTLH
Wikipedia	0.382	0.263	0.374	0.385	**0.397**
NUS-WIDE	0.545	0.353	0.542	0.511	**0.569**



(a) Wikipedia (b) NUS-WIDE

Fig. 2. The PR curves of image hashing on two datasets

5 Conclusion

In this paper we propose a cross-modal self-taught topic learning (CMSTL) algorithm which contains two stages: multi-modal topic learning (MMTL) and projection function learning. In MMTL, hierarchial generation is used to obtain multi-modal topics. Different modalities have different weights in the generation process. Then MMTL optimizes the intra-modal similarity and multi-modal correlation jointly. In projection learning stage, the soft assignment for topics is used. Topic matrix is converted to a probabilistic form and KLSR is used to learn the projections function. Experimental results on two real world image dataset demonstrate the effectiveness of CMSTL in cross-modal retrieval. We further extend CMSTL to image hashing and find that CMSTL can also improve the hashing performance.

References

1. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR) 40(2), 5 (2008)
2. Yang, Y., Xu, D., Nie, F., Luo, J., Zhuang, Y.: Ranking with local regression and global alignment for cross media retrieval. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 175–184. ACM (2009)
3. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the International Conference on Multimedia, pp. 251–260. ACM (2010)

4. Xie, L., Pan, P., Lu, Y.: A semantic model for cross-modal and multi-modal retrieval. In: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, pp. 175–182. ACM (2013)
5. Lu, X., Wu, F., Tang, S., Zhang, Z., He, X., Zhuang, Y.: A low rank structural large margin method for cross-modal ranking. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 433–442. ACM (2013)
6. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the 2013 International Conference on Management of Data, pp. 785–796. ACM (2013)
7. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. International Journal of Computer Vision 100(2), 134–153 (2012)
8. Zhang, D., Wang, J., Cai, D., Lu, J.: Self-taught hashing for fast similarity search. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 18–25. ACM (2010)
9. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from National University of Singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval, p. 48. ACM (2009)
10. Li, D., Dimitrova, N., Li, M., Sethi, I.K.: Multimedia content processing through cross-modal association. In: Proceedings of the Eleventh ACM International Conference on Multimedia, pp. 604–611. ACM (2003)
11. Vinokourov, A., Cristianini, N., Shawe-Taylor, J.S.: Inferring a semantic representation of text via cross-language correlation analysis. In: Advances in Neural Information Processing Systems, pp. 1473–1480 (2002)
12. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
13. Zhai, X., Peng, Y., Xiao, J.: Heterogeneous Metric Learning with Joint Graph Regularization for Cross-Media Retrieval. In: AAAI (2013)
14. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: Advances in Neural Information Processing Systems, pp. 1753–1760 (2009)
15. Bronstein, M.M., Bronstein, A.M., Michel, F., Paragios, N.: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: IEEE Conference on Cumputer Vision and Pattern Recognition (CVPR), pp. 3594–3601. IEEE (2010)
16. Kumar, S., Udupa, R.: Learning hash functions for cross-view similarity search. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22(1), p. 1360 (2011)
17. Rasiwasia, N., Moreno, P., Vasconcelos, N.: Bridging the gap: Query by semantic example. IEEE Transactions on Multimedia 9(5), 923–938 (2007)
18. Hotelling, H.: Relations between two sets of variates. Biometrika, 321–377 (1936)
19. Zhang, D., Wang, J., Cai, D., Lu, J.: Laplacian co-hashing of terms and documents. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 577–580. Springer, Heidelberg (2010)