Principal Affinity based Cross-Modal Retrieval

Jian Liang[1] Dong Cao[1] Ran He[1,2] Zhenan Sun[1,2] Tieniu Tan[1,2]
[1]Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
{jian.liang, dong.cao, rhe, znsun, tnt}@nlpr.ia.ac.cn

Abstract

Multimedia content is increasingly available in multiple modalities. Each modality provides a different representation of the same entity. This paper studies the problem of joint representation of the text and image components of multimedia documents. However, most existing algorithms focus more on inter-modal connection rather than intra-modal feature extraction. In this paper, a ***simple*** *yet* ***effective*** *principal affinity representation (PAR) approach is proposed to exploit the affinity representations of different modalities with local cluster samples. Afterwards, multi-class logistic regression model is adopted to learn the projections from principal affinity representation to semantic labels vectors. Inner product distance is further used to improve cross-modal retrieval performance. Extensive experiments on three benchmark datasets illustrate that our proposed method obtains significant improvements over the state-of-the-art subspace learning based cross-modal methods.*

1. Introduction

With the rapid development of information technology, multi-modal data (e.g., image, text, video or audio) have been widely available on the Internet. For example, an image often co-occurs with text on a web page to describe the same object or event. However, multi-modal data usually span different feature spaces, and this heterogeneous characteristic poses a great challenge to cross-media retrieval tasks. In this work, we mainly address the cross-media retrieval between text and images, i.e., using image (text) to search text documents (images) with the similar semantics.

To address this problem, subspace learning based methods [23, 27] are the most studied approaches. Classical subspace learning algorithms such as the Canonical Correlation Analysis (CCA) [10] and the Partial Least Squares (PLS)[19] have been adopted for learning a common representation for heterogeneous modalities. Taking semantic information into consideration, Gong et al. [9] extended CCA to three-view CCA (CCA-3V) explicitly. Besides, Sharma et al. [20] combined Marginal Fisher Analysis with their supervised multiview counterparts - Generalized Multiview MFA (GMMFA) for the cross-media retrieval problem. Wang et al. [24] proposed to learn coupled feature spaces (LCFS) through low-rank constraints. Moreover, Wang [25] utilized graph regularization to learn common subspace with the between-class covariance and the within-class covariance elegantly combined.

Moreover, probabilistic models are also proposed for cross-modal retrieval. To capture the correlation between images and annotations, Correspondence Latent Dirichlet Allocation (Corr-LDA) [2] was proposed to capture the topic-level relations between images and text annotations. Jia et al. [13] proposed a new probabilistic model to learn a set of shared topics across the modalities through a Markov random field. Besides, deep learning based models [21, 26] also have been proposed to learn a multimodal representation for multi-modal data.

Finally, metric learning approaches directly learn the metric between different modalities. Mignon and Jurie [1] exploited both positive and negative constraints to learn a novel metric while Quadrianto et al. [17] proposed a new metric learning scheme to learn projections from the data in different modalities into a shared feature space, in which the Euclidean distance provides a meaningful intra-modality and inter-modality similarity.

However, Rasiwasia et al. [18] proposed Semantic Correlation Matching (SCM) to match the different modal features through different classifier directly, where the heterogeneous features were transformed through CCA before this. To avoid intra-modal information losses brought by CCA [18], we propose a novel feature representation named Principal Affinity Representation (PAR) for different modalities through anchor points. Then multi-class logistic regression model is adopted to learn projection functions from PAR to semantic label information. Lastly, we evaluate the cross-modal similarity through inner product

978-1-4799-6100-9/15/$31.00 ©2015 IEEE

metric. The main contributions of our proposed method can be summarized as follow:

- We propose a *simple*, yet *effective* principal affinity representation, i.e., affinities with predefined data clusters, for cross-modal retrieval.
- Experimental results on three benchmark datasets demonstrate that our method obtains promising results and performs better than state-of-the-art methods.
- Additionly, we discover that the inner product distance metric can improve the retrieval performance of classification-based algorithms such as SCM and our PACMR.

2. Methodology

In the section, we firstly explain a novel representation named PAR. Then, we will give a brief overview of multi-class Logistic Regression (mcLR) that will be mainly adopted in our cross modal retrieval work. Lastly, we will present to combine PAR and mcLR for cross-modal retrieval.

2.1. Principal Affinity Representation

Affinity information, also known as similarity over pairs of data points, has proven effective in machine learning algorithms such as kernel methods [11] and data clustering [7]. The raw data features need to be explicitly transformed into feature vector representations via a user-specified kernel, i.e., a similarity function over pairs of data points in raw representation. Taking support vector machine [22] for example, kernel methods can obtain much better performance. Frey [7] proposed a classical clustering algorithm named as affinity propagation. Liu et al. [14] discovered that large-scale data affinity graph, constructed through similarities with less anchor points, can obtain comparable even better performance.

Here we adopt similar ideas with [14], where affinities within pairs of data points are measured through Gaussian kernel $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$. Instead of calculating merely the affinities with nearby anchor points in [14], affinities with all anchor points are exploited without normalization to avoid the information loss. Given m anchor points $\{u_1, u_2, \ldots, u_m\}$, the transformed m-dimension representations

$$f(x_i) = [K(x_i, u_1), K(x_i, u_2), \ldots, K(x_i, u_m)] \quad (1)$$

are treated as PAR which can preserve the local and global structure. Nearby raw data points or intra-class data points can own similar PAR while distant raw data or inter-class data points can own a large margin between PAR. Moreover, with increasing amount of anchor points, even local ordinal similarity in source space can be preserved.

To keep the source space structure and guarantee the reasonable memory and computation cost, we adopt Litekmeans[1] to obtain anchor points from features concatenated by different modalities effectively and efficiently. Noting that the anchor points are obtained from both modalities simultaneously to push the PAR in different modalities more consistent.

2.2. Multi-class Logistic Regression

Logistic Regression (LR) [8] is a direct probability model that was developed by statistician D. R. Cox in 1958. Logistic regression, directly exploiting condition probability $p(y|\mathbf{x})$ instead of derived version from joint probability $p(y, \mathbf{x})$, is a discriminative approach. Due to the binary variables, LR adopts Bernoulli distribution below as condition distribution:

$$p(y|\mathbf{x}, \mathbf{w}) = Ber(y|sigm(\mathbf{w^T x})) \quad (2)$$

Besides, the estimated probabilities are restricted to $[0, 1]$ through the following form: $f(z) = \frac{e^{\alpha+\beta z}}{1+e^{\alpha+\beta z}}$. Let $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$be the training dataset of n examples with $\{x_1, x_2, \ldots, x_n\}$ being n input variables and $\{y_1, y_2, \ldots, y_n\}$ being corresponding target variables where $y_i \in \{0, 1\}^L$ with each element y_{ij}, $j \in [1, L]$ denoting the correct label for that sample. Following these assumptions above, mcLR attempts to minimize the negative log-likelihood below:

$$NLL(W) = \sum_{i=1}^{n} log(1 + e^{-y_i^T W x_i}) \quad (3)$$

Taking the max-margin model into consideration, this original model can be extended to a more preferred method called L2-regularized Logistic Regression [6], which solves the following unconstrained optimization problem:

$$\min_W \frac{1}{2}\|W\|_F^2 + C\sum_{i=1}^{n} log(1 + e^{-y_i^T W x_i}) \quad (4)$$

where the first term aims to maximize the classification margin, the last term is negative log-likelihood of traditional logistic regression.

The dual optimization function for L2-regularized Logistic Regression takes the following form:

$$\begin{aligned} &\min_\alpha \frac{1}{2}\alpha^T Q\alpha + \sum_{i:\alpha_i>0} \alpha log\alpha + \\ &\sum_{i:\alpha_i<C} (C-\alpha)log(C-\alpha) - \sum_{i=1}^{l} ClogC \\ &subject\ to \quad 0 \leqslant \alpha_i \leqslant C,\ \forall i \in [1:l]. \end{aligned} \quad (5)$$

[1] http://www.zjucadcg.cn/dengcai/Data/Clustering.html

where $Q_{ij} = y_i^T y_j x_i^T x_j$, α are the dual variables, the ith element is given as α_i; C is the regularization parameter; l is the number of training vectors.

2.3. Cross-modal Retrieval

Almost all subspace based methods for cross-modal retrieval attempt to learn a common space where projected features from different modal can be measured with some ordinary metrics such as Euclidean distance, cosine distance, etc. However, optimizing these objective functions of learning methods is challenging and the running time is rarely well-pleasing. Different from previous subspace learning approaches [24, 20], our method adopts a similar classification approach with [18, 5] which can be seen as a particular form of subspace based method. Our method considers the label space as common space, i.e., the provided ground truth label vectors $\{y_i\}_{i=1}^n$, $y_i \in R^k$, where k is the amount of labels or classes, are fixed as common space vectors. Apparently, these external labels are consistent for two or multiple modalities in every metric space. Then different kinds algorithms can be adopted to learn the projection of PAR in different modal to common label space. Here we choose primal L2-regularized Logistic Regression [6] for efficiency, whose objective function is shown in Eq.(4).

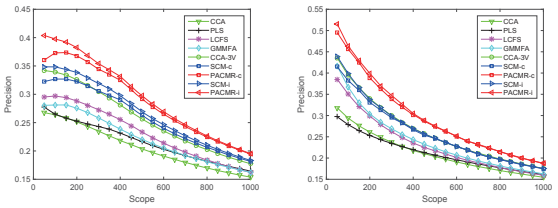
It is worth noting that, we take PAR for classification algorithms while [18] acquires either dimensionality reduction through Principal Component Analysis (PCA) or raw data representation for classification task in each modalities. Here we do not consider nosiy conditions shown in [3]. Furthermore, even various metrics in common space has been thoroughly studied in [5], we discover that extra straightforward inner product distance in our common space for retrieval, which can be seen as cosine distance with L1 normalization, gains the best performance over all three standard public datasets, shown apparently in Table 1 and Figures 1, 2, 3.

3. Experiments and Results

In this section, our approach and existing methods are compared on three publicly available datasets - Pascal VOC2007 [12] dataset, NUS-WIDE [4] dataset and Wiki [18] image-text dataset. For cross-modal retrieval problem, specific projections to common space are learned on training dataset. Then, we project the data from different modalities into the common space. For the test set, we take the data from one modality as the query set, and the data from another modality as the database set.

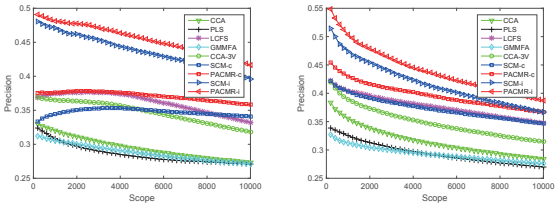
3.1. Datasets

The Pascal VOC dataset [12] consists of 5,011/4,952 (training/ testing) image-tag pairs, which can be categorized into 20 different classes. We select images with only one object as the way in [20], resulting in 2,808 training and



(a) Image query (b) Text query

Figure 1. precision-scope curve on the VOC dataset



(a) Image query (b) Text query

Figure 2. precision-scope curve on the NUS dataset

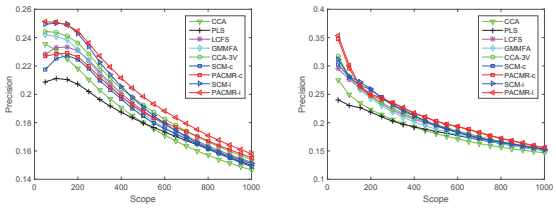
2,841 testing data. The image features are 512-dimensional Gist features [16], and the text features are 399-dimensional word frequency features with both features provided in [12] for fair comparison.

The NUS-WIDE dataset [4] is currently the largest dataset used for cross modal retrieval where each image is associated with kinds of user tags. To guarantee that each class has abundant training samples like [28], we select image-tag pairs that belong to one of the 21 largest classes with each pair exclusively belonging to one of the 21 classes, which results in 72,219 image-text pairs. The images are represented with a 500-dimensional SIFT feature vectors [15], and the textual tags are represented with 1000-dimensional tag occurrence feature vectors. We take 50% of the data as the training set and the remaining as the testing set.

The Wiki image-text dataset [18], generated from Wikipedia's "featured article", consists of 2,866 image-text pairs. In each pair, the text is an article describing people, places or some events and the image is closely related to the content of the article. Each pair is labelled with one of 10 semantic classes. We split it into a training set of 1,300 pairs (130 pairs per class) and a testing set of 1,566 pairs. The representation of the text with 10 dimensions is derived from a LDA model [2]. The images are represented by the 128 dimensional SIFT descriptor histograms [15].

3.2. Experimental Setting

We compare the proposed PACMR approach with PLS [19], CCA [10], SCM [18], CCA-3V [9], LFCS [24], GMMFA [20] in terms of common cross-modal retrieval tasks: (1) Image query vs. Text database, (2) Text query vs. Image database. Among these methods, PLS and CCA are two classical methods which use pairwise information to learn a common latent subspace across multi-



(a) Image query (b) Text query

Figure 3. precision-scope curve on the Wiki dataset

modal data. Besides, SCM, LFCS, GMMFA and CCA-3V are four supervised methods which further exploit the label information. Regarding metric selection[2], Cosine distance is adopted to measure the similarity between text-image pairs of these methods when inner product distance is applied with SCM and our proposed method.

To evaluate the performance of the proposed method, the mean average precision (MAP) [18] is used to evaluate the overall performance of the tested algorithms. To compute MAP, we first evaluate the average precision (AP) of a set of R retrieved documents by $AP = \frac{1}{T}\sum_{i=1}^{R} P(r)\delta(r)$ where T is the number of relevant documents in the retrieved set, $P(r)$ denotes the precision of the top r retrieved documents, and $\delta = 1$ if the rth retrieved document is relevant (where relevant means belonging to the class of the query) and $\delta = 0$ otherwise. The larger the MAP, the better the performance. Besides the MAP, we also use precision-scope curve [18] to evaluate the effectiveness of different methods explicitly. For NUS-WIDE and PASCAL VOC, PCA is used to reduce the dimensions of the original features. For PACMR, the length of PAR are fixed as 200, 500 and 500 for three datasets respectively.

3.3. Results

Table 1 shows the MAP scores achieved by two unsupervised methods, four supervised methods and the proposed method (PACMR) on three benchmark datasets while SCM and PACMR adopt inner product distance additionally. It can be observed that the proposed method significantly outperforms its several counterparts for both forms of cross-modal retrieval tasks through traditional cosine distance metric. Under inner product metric, the average MAP values of both SCM and PACMR are improved by 10.4%, 11.2% and 10.9% respectively, and PACMR-i is consistently better than SCM-i and other algorithms.

Further analyses of the results are presented in Figures 1, 2 and 3, which show the corresponding precision-scope curves of all approaches reported. Our method again significantly outperforms other algorithms for both forms of cross-modal retrieval. In Figure 1, both PACMR-i and PACMR-c are superior to other algorithms in both Figure 1(a) and Figure 1(b). For the NUS-WIDE dataset in Figure 2, PACMR-i obtains the best performance with huge advantages while PACMR-c gains better performance than other algorithms except SCM-i for both the text query and the image query. Concerning the Wiki dataset shown in Figure 3, PACMR-i is also the best performing method in both retrieval tasks. Besides, the performance of PACMR-c for image query is still better than SCM-c even though inferior to CCA-3V and GMMFA. However, the curve for the text query of PACMR-c is absolutely superior to other algorithms such as CCA-3V and GMMFA. It is worth noting that PACMR-i can always obtain the best performance among these algorithms while PACMR-c is slightly inferior to other algorithms just on the Wiki dataset, this may be because the length of features is too short for favorable retrieval performance.

[2]In Table 1 and Figure 1, 2, 3, -c means cosine distance (default metric) while -i means inner product distance.

4. Conclusion

In this paper, we have proposed a simple yet effective PAR based cross-modal retrieval method to learn heterogeneous projections from different modalities. PAR plays a similar role as embedding, preserving origin manifold structure through anchor points with low computation costs. Compared with traditional subspace methods, semantic information space is directly treated as common space which can avoid the time-consuming optimization brought by subspace learning methods. Due to the boosted performance brought by inner product distance, one of our future work is to generate corresponding binary codes for cross-modal retrieval.

Acknowledgement

This work is funded by the Youth Innovation Promotion Association, Chinese Academy of Sciences (Grant No. 2015190), and the National Natural Science Foundation of China (Grant No. 61473289).

References

[1] M. Alexis and F. Jurie. Cmml: A new metric learning approach for cross modal matching. In *Asian Conference on Computer Vision*, pages 14–pages, 2012. 1

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. 1, 3

[3] B. Chen, L. Xing, J. Liang, N. Zheng, and J. C. Principe. Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion. *IEEE Signal Processing Letters*, 21(7):880–884, 2014. 3

[4] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, page 48, 2009. 3

[5] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval.

Dataset	Query	CCA	PLS	LCFS	GMMFA	CCA-3V	SCM-c	PACMR-c	SCM-i	PACMR-i
Pascal VOC	Image	0.2665	0.2662	0.3336	0.3272	0.3657	0.3906	0.4306	0.4115	**0.4580**
	Text	0.2184	0.2180	0.2485	0.2635	0.2964	0.2860	0.3474	0.2927	**0.3557**
	Average	0.2424	0.2421	0.2910	0.2953	0.3311	0.3383	0.3890	0.3521	**0.4068**
NUS-WIDE	Image	0.2930	0.2798	0.3830	0.3100	0.3604	0.3574	0.3849	0.4218	**0.4513**
	Text	0.2866	0.2717	0.3460	0.3015	0.3324	0.3531	0.3760	0.3795	**0.4046**
	Average	0.2898	0.2757	0.3645	0.3058	0.3464	0.3552	0.3805	0.4007	**0.4280**
Wiki	Image	0.2512	0.2438	0.2764	0.2743	0.2759	0.2753	0.2785	0.3199	**0.3229**
	Text	0.1986	0.1943	0.2135	0.2166	0.2238	0.2237	0.2342	0.2240	**0.2366**
	Average	0.2249	0.2190	0.2450	0.2454	0.2499	0.2495	0.2564	0.2719	**0.2797**

Table 1. MAP Comparison on three benchmark datasets, and the blue and underscored values are the best performances within cosine metric with the red and bold values being the best within inner product metric.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(3):521–535, 2014. 3

[6] R. Fan, K. Chang, C.-J. Hsieh, X. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008. 2, 3

[7] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007. 2

[8] J. Gareth, W. Daniela, H. Trevor, and T. Robert. *An introduction to statistical learning*. Springer, 2013. 2

[9] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233, 2014. 1, 3

[10] 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. 1, 3

[11] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008. 2

[12] S. J. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1145–1158, 2012. 3

[13] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *IEEE International Conference on Computer Vision*, 2011. 1

[14] W. Liu, J. He, and S.-F. Chang. Large graph construction for scalable semi-supervised learning. In *International Conference on Machine Learning*, pages 679–686, 2010. 2

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 3

[16] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. 3

[17] N. Quadrianto and C. H. Lampert. Learning multi-view neighborhood preserving projections. In *International Conference on Machine Learning*, pages 425–432, 2011. 1

[18] 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 International Conference on Multimedia*, pages 251–260, 2010. 1, 3, 4

[19] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*, pages 34–51. Springer, 2006. 1, 3

[20] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, 2012. 1, 3

[21] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2012. 1

[22] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013. 2

[23] D. Wang, Q. Yin, R. He, L. Wang, and T. Tan. Multi-view clustering via structured low-rank representation. In *ACM International Conference on Information and Knowledge Management*, 2015. 1

[24] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *IEEE International Conference on Computer Vision*, pages 2088–2095, 2013. 1, 3

[25] K. Wang, W. Wang, R. He, L. Wang, and T. Tan. Multimodal subspace learning with joint graph regularization for cross-modal retrieval. In *Asian Conference on Pattern Recognition*, pages 236–240, 2013. 1

[26] W. Wang, B.-C. Ooi, X. Yang, D. Zhang, and Y. Zhuang. Effective multi-modal retrieval based on stacked auto-encoders. *Proceedings of the VLDB Endowment*, 7(8):649–660, 2014. 1

[27] Q. Yin, S. Wu, R. He, and L. Wang. Multi-view clustering via pairwise sparse subspace representation. *Neurocomputing*, 156:12–21, 2015. 1

[28] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *International ACM SIGIR Conference on Research & Development on Information Retrieval*, pages 395–404, 2014. 3