ONLINE LOW-RANK SIMILARITY FUNCTION LEARNING WITH ADAPTIVE RELATIVE MARGIN FOR CROSS-MODAL RETRIEVAL

Yiling Wu[1,2]*, Shuhui Wang*[1]*, Weigang Zhang*[2,3]*, Qingming Huang*[1,2]

[1] Key Lab of Intell. Info. Process, Institute of Computing Technology, CAS, China
[2] School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
[3] School of Computer Science and Technology, Harbin Institute of Technology at Weihai, China
yiling.wu@vipl.ict.ac.cn wangshuhui@ict.ac.cn wgzhang@hit.edu.cn qmhuang@ucas.ac.cn

ABSTRACT

This paper presents a Cross-Modal Online Low-Rank Similarity function learning method (CMOLRS) for cross-modal retrieval, which learns a low-rank bilinear similarity measure on data from different modalities. CMOLRS models the cross-modal relations by relative similarities on a set of training data triplets and formulates the relative relations as convex hinge loss functions. By adapting the margin of hinge loss using information from feature space and label space for each triplet, CMOLRS effectively captures the multi-level semantic correlation among cross-modal data. The similarity function is learned by online learning in the manifold of low-rank matrices, thus good scalability is gained when processing large scale datasets. Extensive experiments are conducted on three public datasets. Comparisons with the state-of-the-art methods show the effectiveness and efficiency of our approach.

Index Terms— Cross-modality, similarity function learning, online learning, low-rank matrix

1. INTRODUCTION

With the advance of internet and multimedia technology, multimedia data with massive volumes and high dimensions are pervasive on the Web. An appealing task called cross-modal retrieval [1, 2, 3, 4, 5] has received considerable attention. Given a query in one modality, it returns a list of documents in another modality, *e.g.* retrieving images with a text query. The returned list of documents are to be ordered according to the similarities between database documents and the query document. Therefore, a good cross-modal similarity measure that well reflects the semantic relevance between different modalities plays an important role in cross-modal retrieval.

This work was supported in part by National Natural Science Foundation of China: 61672497, 61332016, 61620106009, 61650202, U1636214, 61572488, in part by National Basic Research Program of China (973 Program): 2015CB351802 and in part by Key Research Program of Frontier Sciences of CAS: QYZDJ-SSW-SYS013.

Generally, the high dimensionality of the multi-modal data can cause the so-called "curse-of-dimensionality" problem. A standard solution is to project data into a shared low dimensional latent space where data from different modalities can be compared. CCA [6] and PLS [7] are the most famous approaches to learn such low dimensional common space. However, the metrics used in the shared space are not learned towards information retrieval objectives. Therefore, we consider directly learning cross-modal similarity function in the paper. The cross-modal similarity function is learned from two aspects as follows.

First, the explosive growth of data necessitates the scalability of cross-modal approach on massive data. Most existing cross-modal retrieval models are learned in batch styles [1, 2, 4], requiring training data storage beforehand and involving high computational complexity. Online learning is a possible solution to improve the scalability. Grangier *et al.* [8] adapts the online Passive-Aggressive algorithm [9] to learn a similarity function for text-to-image retrieval by projecting the visual representation to the tag space. That procedure learns a full rank matrix of size $d^t \times d^v$ to support only text-to-image retrieval, where d^t and d^v are the dimensions of the textual and visual representations. However, it is not guaranteed in [8] that a low dimensional projected space is learned. Besides, naive methods of online low-rank matrix learning like repeated singular value decomposition or optimizing a factored representation of the low-rank matrix are either computationally expensive or numerically unstable. Therefore, existing approaches are not capable of handling high dimensional cross-modal data in real applications.

Second, existing approaches mainly use absolute similarity [10, 3] to express the cross-modal relation, *e.g.* a is similar to b but not similar to c. However, absolute similarity can only encode bi-level relation between data objects, *i.e.* are they similar or dissimilar. Relative similarity [8, 5] expresses multi-level relation by comparing the similarities of different data pairs, *e.g.* a is more similar to b than to c. In many applications, similarity is easy to be expressed in terms of relative similarity. For instance, in multi-label settings, da-

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

ta sharing more labels can be seen as more similar. Therefore, we learn the similarity function by preserving relative similarity. Note that for relative similarity, a triplet (a, b, c) gives a relative similarity relation, but the difference between pairwise similarities may vary on different triplets. To model such difference, instead of using a fixed margin used by existing approaches, we define an altered margin by considering the relevances of documents in both label space and feature space. Consequently, the adapted margin better deals with the content divergence compared to the fixed margin.

In this paper, we propose Cross-Modal Online Low-Rank Similarity Function learning (CMOLRS), a simple and flexible bilinear similarity function in the feature spaces by encoding the semantic similarity relation of cross-modal data with relative similarity and adaptive margin. The bilinear similarity function can be seen as taking inner product of the projected data, and learning a low dimensional subspace is equivalent to learning a low-rank similarity function. Enforcing low-rank constraint provides a natural regularization on the model to handle the high dimensionality. With a rank k matrix, the calculation of the similarity score reduces to $O((d^t + d^v)k)$ operations. To further reduce the training complexity, we enforce low-rank property on the online learning procedure of the similarity function. We adopt Loreta [11] which consists of a gradient step, followed by a second-order retraction back to the manifolds to perform online learning of low-rank matrices. Therefore, good scalability is gained when processing large scale datasets. Extensive experimental comparisons with the state-of-the-art methods show the effectiveness and efficiency of our approach.

2. RELATED WORK

With the increasing of multi-modal data, a number of tasks have been proposed to exploit multi-modal data [12, 13]. Cross-modal retrieval returns documents in different modality from the query. Most cross-modal retrieval techniques are based on latent space learning. CCA [6] aims at learning a latent space by maximizing the correlating relationships between two modality data. Sharma *et al.* [2] combine popular supervised and unsupervised feature extraction techniques with CCA to achieve closeness between multi-view samples of the same class. PLS [7] creates orthogonal score vectors by maximizing the covariance between different sets of variables. Hua *et al.* [14] build a semantic hierarchy to measure multi-level semantic relevance, and then use the relevance to guide the learning of aggregated local distances. There are also a few techniques directly use label space as the common space. In LGCFL [15], ε-dragging is performed on the label space to force the regression targets of different classes moving along opposite directions, and group sparse constrains are imposed in the regression process to learn the most discriminant groups. Deng *et al.* [16] propose a discriminative dictionary learning method that is augmented with common label alignment.

Another line of works extend linear mappings to nonlinear mappings by using deep learning methods. MMNN [3] is a coupled siamese neural network architecture that allows unified treatment of intra- and inter-modality similarity learning. Yan *et al.* [17] propose to learn the joint embeddings by deep canonical correlation analysis (DCCA)[18] which learns complex nonlinear transformations of two views of data. BRNN [19] embeds fragments of images and fragments of sentences by deep networks into a common space.

3. APPROACH

3.1. Problem formulation

Without loss of generality, we use image and text modalities for illustration and derive the algorithm in the direction of text-to-image. Let $\mathcal{T} = \{t_i, z_i^t\}_{i=1}^{N^t}$ denote the set of texts and their associated labels, where $t_i \in \mathbb{R}^{d^t}$ indicates a text, and $z_i^t \in \mathbb{R}^c$ is its corresponding label vector. Similarly, we have $\mathcal{V} = \{v_i, z_i^v\}_{i=1}^{N^v}$ as the set of images, where $v_i \in \mathbb{R}^{d^v}$ is an image and $z_i^v \in \mathbb{R}^c$ is its corresponding label vector.

The goal of this paper is to learn a similarity function between cross-modal data. The similarity function is formulated as a simple bilinear function:

$$s(t_i, v_j) = t_i^T W v_j, \tag{1}$$

where $W \in \mathbb{R}^{d^t \times d^v}$. This bilinear function can be seen as a linear function on the joint feature of $t_i v_j^T$, so the multiplicative interactions between the two modalities can be measured. Also the negative correlations via negative values in the weight matrix W can be learned from the data [20]. Although our similarity function only relates inter-modal data, the relations between different modalities also impact the relations in one modality. For example, if two images are similar to a text, they should be related to each other as well [21].

To learn the similarity function, relative similarity is used to guide the learning procedure. Another popular relation is absolute similarity. However, in many situations, relative similarity is easier to express than absolute one. For instance, in a multi-label setting, it is more intuitive to define instances sharing more labels to be more similar. What's more, relative similarity is a more general relationship that covers absolute one. Two instances that are absolutely similar can be seen as being more similar than other non-similar instances.

The relative relations in the training set are represented by a set of triplets $T = \{(t_i, v_j, v_k)\}$, where each triplet (t_i, v_j, v_k) encodes that t_i and v_j are more relevant than t_i and v_k. For triplet (t_i, v_j, v_k), we want the similarity function score between text t_i and image v_j to be larger than the score between text t_i and image v_k by some enforced margin δ,

$$s(t_i, v_j) > s(t_i, v_k) + \delta. \tag{2}$$

However, for different triplets, *e.g.* (t_i, v_j, v_k) and (t_x, v_y, v_z), the differences $s(t_i, v_j) - s(t_i, v_k)$ and $s(t_x, v_y) -$

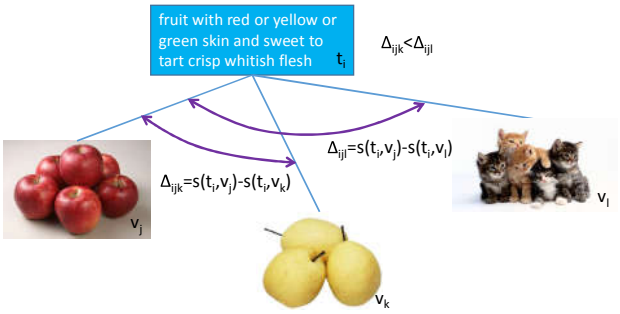


Fig. 1. An example of different differences of similarity functions. Δ_{ijk} denotes the difference of similarity functions

$s(t_x, v_z)$ may be different. For example, given a text t_i describing apples, and images v_j depicting apples, v_k depicting pears, v_l depicting cats, the difference $s(t_i, v_j) - s(t_i, v_k)$ should be lower than $s(t_i, v_j) - s(t_i, v_l)$, since apples and pears belong to a class fruit, while apples and cats do not. We show this in Figure 1. To model the difference of similarity functions, we define margin to be a function of v_j and v_k. Substituting this, Eq.(2) becomes:

$$s(t_i, v_j) > s(t_i, v_k) + \delta(v_j, v_k). \tag{3}$$

Obviously, the function scores of similar images to a given text will be similar. For the example in Figure 1, since v_j and v_k are similar, $s(t_i, v_j)$ and $s(t_i, v_k)$ should not differ too much. Thus similar images should have small margin. When measuring the similarity of images, semantic relevance and visual similarity should all be considered. Therefore, we compare images from label space and feature space, and adopt simple Euclidean distance to perform the comparison. We define $\delta(v_i, v_j)$ as:

$$\delta(v_i, v_j) = \beta(\alpha\|v_i - v_j\|^2 + (1-\alpha)\|z_i^v - z_j^v\|^2), \tag{4}$$

where α is a trade-off parameter determining the relative importance of label space and feature space, and β adjusts suitable margin for different datasets.

We reformulated the constraint in Eq.(4) as hinge loss in the standard way. The resulting loss function is given by:

$$l(t_i, v_j, v_k) = \max(0, s(t_i, v_k) - s(t_i, v_j) + \delta(v_j, v_k)). \tag{5}$$

Our goal is to minimize a global loss L that accumulates hinge losses in Eq.(5) over all possible triplets in the training set:

$$L = \sum_{(t_i, v_j, v_k) \in D_{train}} l(t_i, v_j, v_k). \tag{6}$$

In addition, high dimensional data are usually embedded in a low dimensional subspace. The bilinear similarity function can be seen as taking the inner product of the projected data, and learning a low dimensional subspace is equivalent to learning a low-rank similarity function. Also notice that calculating a score with the full rank bilinear similarity function in Eq.(1) takes $O(d^t d^v)$ operations. When the dimension of features is high, *e.g.* CNN features, the computational complexity is high. With a rank k matrix W, calculating a score reduces to $O((d^t + d^v)k)$ operations. Thus, we consider learning a low-rank matrix W. Adding a rank constraint, we have a minimization problem in the following form:

$$\min_W \sum_{(t_i, v_j, v_k) \in D_{train}} l(t_i, v_j, v_k), \quad s.t.\ \text{rank}(W) = k. \tag{7}$$

3.2. Online learning in the low rank manifold

In this subsection, we consider optimizing Eq.(7). In order to scale to large datasets, we concern online learning in which a triplet is provided at each iteration. However, the low-rank constraint makes the optimization difficult. Two naive approaches, repeated singular value decomposition of the matrix and optimizing a factored representation of the low-rank matrix are either computationally expensive or numerically unstable. To learn efficiently, we apply Loreta [11] here.The set of $n \times m$ matrices of rank k where $k \leq m, n$ is an $(n+m)k - k^2$ dimensional manifold embedded in $\mathbb{R}^{n \times m}$. Loreta [11] performs stochastic gradient descent in the manifold of low-rank matrices. It consists of a gradient step, followed by a second-order retraction back to the manifold. A retraction R_x is a mapping from the tangent space $T_x\mathcal{M}$ to the manifold $\mathcal{M}$ at x. The mathematically ideal retraction is called the exponential mapping, and second-order retractions approximate the exponential mapping to second order. The best known example of the second-order retraction is the projection operation, while computing the projection is too costly for the manifold of low-rank matrices. Loreta uses a specific second-order retraction that can be computed efficiently.

At every iteration the algorithm suffers a loss, the gradient of the hinge loss in Eq.(5) is $\frac{\partial l(t_i, v_j, v_k)}{\partial W} = t_i(v_j - v_k)^T$ which is a rank-1 gradient, so we apply Loreta-1 which is a specific form of Loreta for rank-1 gradient. Algorithm 1 summarizes the proposed algorithm for Cross-Modal Online Low-Rank Similarity function learning (CMOLRS). Let $\mathbb{R}_*^{n \times k}$ denotes the set of $n \times k$ matrices of rank k.

For the computational cost of Algorithm 1, the bottleneck lies on the computation of the pseudo-inverse of the matrices A_i and B_i. Following a procedure developed by [22], we can keep the pseudo-inverses $A_{i-1}^\dagger$ and $B_{i-1}^\dagger$ from the previous round, and perform a rank-one update to them. The overall time and space complexity of Loreta-1 are both $O((d^t + d^v)k)$ per gradient step, where k is the rank of W, d^t and d^v are the dimensions of the two modalities respectively.

4. EXPERIMENTS

In this section, we compare different methods on three public datasets, Wiki, Pascal VOC 2007 and NUS-WIDE.

4.1. Experimental settings

We compared our proposed method CMOLRS with five cross-modal methods, namely CCA [6], PLS [7], GMLDA

Algorithm 1 Cross-Modal Online Low-Rank Similarity Function Learning

Input: $A_0 \in \mathbb{R}_*^{d^t \times k}$, $B_0 \in \mathbb{R}_*^{d^v \times k}$, s.t. $W_0 = A_0 B_0^T$, $A_0^\dagger$, $B_0^\dagger$ are the pseudo-inverses of A_0 and B_0, step size $\eta \geq 0$
Output: $W_N = A_N B_N^T$
for $i = 1, \ldots, N$ **do**
  Randomly sample a triplet (t_i, v_j, v_k)
  if $s(t_i, v_j) < s(t_i, v_k) + \delta(v_j, v_k)$. **then**
    $p = -\eta t_i$, $q = v_j - v_k$
    $a_1 = A_{i-1}^\dagger \cdot p$, $b_1 = B_{i-1}^\dagger \cdot q$
    $a_2 = A \cdot a_1$
    $s = b_1^T \cdot a_1$
    $a_3 = a_2(-\frac{1}{2} + \frac{3}{8}s + p(1 - \frac{1}{2}s))$
    $A_i = A_{i-1} + a_3 \cdot b_1^T$
    $b_2 = (q^T B_{i-1}) \cdot B_{i-1}^\dagger$
    $b_3 = b_2(-\frac{1}{2} + \frac{3}{8}s + q^T(1 - \frac{1}{2}s))$
    $B_i^T = B_{i-1}^T + a_1 \cdot b_3$
    $A_i^\dagger$=rank-1-pseudoinverse-update$(A_{i-1}, A_{i-1}^\dagger, a_3, b_1)$
    $B_i^\dagger$=rank-1-pseudoinverse-update$(B_{i-1}, B_{i-1}^\dagger, b_3, a_1)$
  end if
end for

[2], Bi-CMSRM [21], LGCFL [15]. For GMLDA and LGCFL, the category of every training pair is selected randomly from its multiple labels. For methods involving random sampling, we perform 3 runs and take the average performance for comparison. All the experiments were performed in Matlab on an Intel Core i7 3.6GHz Windows 8 machine with 32GB RAM. The dimensionality of common space for all methods except LGCFL which fixes label space as its common space is set to 8, 32, 32 on Wiki, Pascal VOC 2007 and NUS-WIDE respectively. For our CMOLRS, we set parameter α to 0.5 on all datasets and set parameter β to 1, 1, 3 on Wiki, Pascal VOC 2007 and NUS-WIDE respectively. Triplets are sampled randomly, with 1e6 triplets on Wiki and Pascal VOC 2007, 1e7 triplets on NUS-WIDE respectively. Early stopping can be used to further improve the performance.

We evaluate our CMOLRS method on two cross-modal retrieval tasks, *i.e.* retrieving images by text queries and retrieving texts by image queries. To evaluate the semantic consistency of our cross-modal similarity function, Mean Average Precision(MAP) which is a widely used metric in the retrieval task is used as the evaluation metric. On NUS-WIDE, we only evaluate MAP of top 100 retrieval results. Besides, we present precision-scope curve to evaluate the performance of different methods.

4.2. Results on the Wiki dataset

Wiki dataset [1], generated from Wikipedia's featured articles collection, contains 2,866 text-image pairs. Each pair is labelled with exactly one of 10 semantic classes. 2,173 pairs are taken as training set and 693 pairs are taken as testing set as in [1]. We use the publicly available 10-dim LDA text features, and extract 4096-dim CNN image features in 'fc7' using Caffe[23] with the pre-trained architecture learned on ImageNet.

Table 1. MAP on the Wiki dataset.

method	txt2img	img2txt	average
CCA	0.344	0.371	0.358
PLS	0.380	0.406	0.393
GMLDA	0.356	0.386	0.371
Bi-CMSRM	0.356	0.400	0.378
LGCFL	**0.392**	0.428	0.410
CMOLRS	0.391	**0.435**	**0.413**

Table 2. MAP on the Pascal VOC 2007 dataset.

method	txt2img	img2txt	average
CCA	0.664	0.659	0.662
PLS	0.655	0.681	0.668
GMLDA	0.677	0.678	0.668
Bi-CMSRM	0.724	0.670	0.697
LGCFL	0.734	0.734	0.734
CMOLRS	**0.736**	**0.737**	**0.737**

Table 1 shows the MAP scores achieved by compared methods and our method on the Wiki dataset. We observe that our method outperforms its several counterparts. CMOLRS achieves comparable performance to LGCFL in text-to-image task. However, LGCFL is a batch learning algorithm, while CMOLRS is an online learning method. The corresponding precision-scope curves are plotted in Figures 2(a) and 2(b). Our method performs well on the Wiki dataset which is a small multi-class dataset.

4.3. Results on the Pascal VOC dataset

Pascal VOC 2007 [24], collected from the Flickr, contains 9,963 images. Although the original user tags were not kept, Hwang *et al.* [25] collected tags using Mechanical Turk. We use the 399-dim tag frequency features provided by them for text representations. 4,096-dim CNN image features from 'fc7' layer are used for image representations. Groundtruth annotation of the images which have 20 classes are used for label representations. Original train-test split provided in the dataset is used for training and testing. After removing images without tags, we get a training set with 5,000 images and a test set with 4,919 images.

Table 2 shows the MAP scores of different approaches on the Pascal dataset. The comparison shows that our method outperforms its counterparts. The corresponding precision-scope curves are plotted in Figures 2(c) and 2(d). From the figures, we see that CMOLRS compares favorably to other methods. This may be caused by multi-label information contained in this dataset. Our method can represent the multi-label information easily using relative similarity, while GMLDA and LGCFL can just model multi-class information.

Table 3. MAP@100 on the NUS-WIDE dataset.

method	CCA	PLS	LGCFL	CMOLRS
txt2img	0.531	0.579	0.597	**0.674**
img2txt	0.458	0.446	0.473	**0.533**
average	0.495	0.513	0.535	**0.604**

Table 4. Comparison of MAP scores of CMOLRS with different rank on the Pascal dataset.

rank	8	16	32	64
txt2img	0.695	0.734	0.739	0.737
img2txt	0.705	0.733	0.738	0.738

4.4. Results on the NUS-WIDE dataset

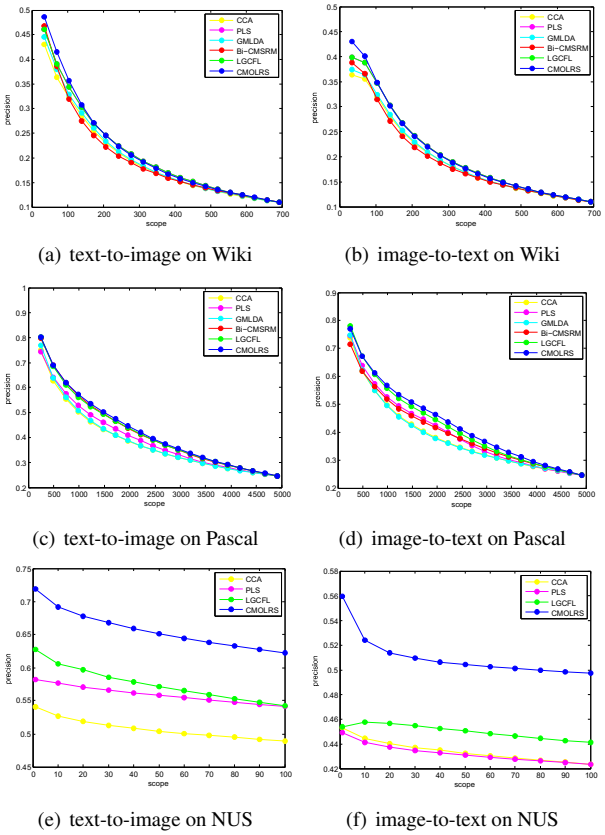
NUS-WIDE dataset [26], crawled from the Flickr website, contains 269,648 images associated with their tags. Each image is labeled with 81 underlying semantic concepts. We take the 1,000-dim tag frequency features as text features and take the 4,096-dim output of 'fc7' layer from CNN as image features. We use the original train-test split provided in the dataset for training and testing. By keeping the images that have at least one tag and one concept, we get 79,659 images for training and 53,550 images for testing.

Table 3 shows the MAP@100 scores of different approaches on the NUS-WIDE dataset. We don't compare GMLDA and Bi-CMSRM here, because of the prohibitively high space complexity of them. From the table, we can see that our method significantly outperforms the compared methods. The corresponding precision-scope curves are plotted in Figures 2(e) and 2(f). The promising results confirmed the effectiveness of CMOLRS on large dataset. The effectiveness of CMOLRS can be attributed to its simple form and its ability to take full advantage of relative similarity. The large number of relative similarity training triplets provide rich cross-modal relations for our online similarity learning strategy. Figure 4 shows examples of retrieved images using text queries. Figure 3 shows examples of retrieved texts using image queries.

4.5. Further Analysis

We also conducted experiments to examine the effect of the rank of W. We ran CMOLRS with different ranks on the Pascal dataset. Table 4 shows the MAP scores of CMOLRS with different ranks. When setting ranks to 16, 32 and 64, the performance varies little. When setting the rank to small value (*i.e.* 8), performances are inferior.

Since we initialize A_0, B_0 and sample triplets randomly, the performance is not fixed. We ran 5 times of CMOLRS on the Pascal dataset. The results are show in Figure 5. It can be seen that the performance varies little. Thus the performance of CMOLRS is not sensitive to initial values and training triplets.



(a) text-to-image on Wiki (b) image-to-text on Wiki

(c) text-to-image on Pascal (d) image-to-text on Pascal

(e) text-to-image on NUS (f) image-to-text on NUS

Fig. 2. Performance of different methods on all benchmark datasets based on precision-scope curve.

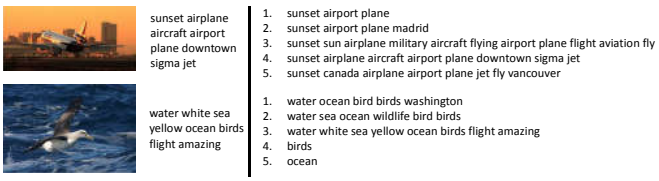


Fig. 3. Examples of top 5 results of image-to-text retrieval on the NUS-WIDE dataset. The first two columns are the image queries and the paired texts of image queries.

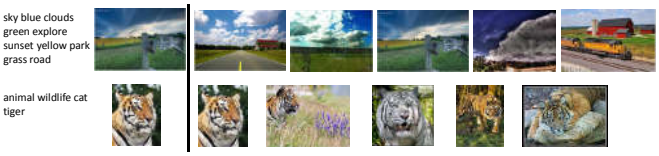
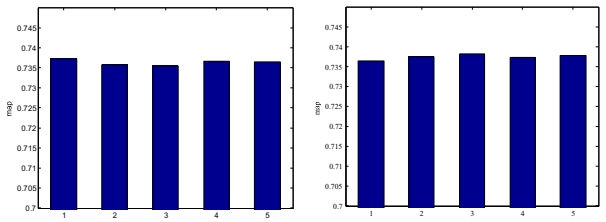


Fig. 4. Examples of top 5 results of text-to-image retrieval on the NUS-WIDE dataset. The first two columns are the text queries and the paired images of text queries.



(a) text-to-image (b) image-to-text

Fig. 5. Effects of initial values and training triplets on performance of CMOLRS in 5 runs on Pascal

5. CONCLUSION

We presented CMOLRS, an approach to learn a cross-modal similarity function for large scale high dimensional cross-modal data. The cross-modal similarity relation is modeled by a set of relative similarities with adapted margins. An algorithm for online learning in the manifold of low-rank matrices is applied to efficiently learn the similarity function. Extensive experiments clearly demonstrate the superiority of our proposed approach over the state-of-the-art methods.

6. REFERENCES

[1] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. RG Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM Multimedia*, 2010, pp. 251–260.

[2] S. Abhishek, K. Abhishek, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *CVPR*, 2012, pp. 2160–2167.

[3] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *TPAMI*, vol. 36, no. 4, pp. 824–830, 2014.

[4] Cuicui Kang, Shengcai Liao, Yonghao He, Jian Wang, Wenjia Niu, Shiming Xiang, and Chunhong Pan, "Cross-modal similarity learning: A low rank bilinear formulation," in *CIKM*. ACM, 2015, pp. 1251–1260.

[5] Z. Kuang and K.-Y. K. Wong, "Relatively-paired space analysis: Learning a latent common space from relatively-paired observations," *IJCV*, vol. 113, no. 3, pp. 176–192, 2015.

[6] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[7] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, latent structure and feature selection*, pp. 34–51. Springer, 2006.

[8] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *TPAMI*, vol. 30, no. 8, pp. 1371–1384, 2008.

[9] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *JMLR*, vol. 7, pp. 551–585, 2006.

[10] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval.," in *AAAI*, 2013.

[11] U. Shalit, D. Weinshall, and G. Chechik, "Online learning in the embedded manifold of low-rank matrices," *JMLR*, vol. 13, no. 1, pp. 429–458, 2012.

[12] Jun Yu, Dacheng Tao, Meng Wang, and Yong Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE transactions on cybernetics*, vol. 45, no. 4, pp. 767–779, 2015.

[13] Jun Yu, Xiaokang Yang, Fei Gao, and Dacheng Tao, "Deep multimodal distance metric learning using click constraints for image ranking," *IEEE transactions on cybernetics*, 2016.

[14] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *TMM*, vol. 18, no. 6, pp. 1201–1216, 2016.

[15] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *TMM*, vol. 17, no. 3, pp. 370–381, 2015.

[16] C. Deng, J. Tang, X.and Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *TMM*, vol. 18, no. 2, pp. 208–218, 2016.

[17] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *CVPR*, 2015, pp. 3441–3450.

[18] Galen Andrew, Raman Arora, Jeff A Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *ICML (3)*, 2013, pp. 1247–1255.

[19] A. Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[20] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger, "Learning to rank with (a lot of) word features," *Information retrieval*, vol. 13, no. 3, pp. 291–314, 2010.

[21] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," in *ACM Multimedia*, 2013, pp. 877–886.

[22] Carl D Meyer, Jr, "Generalized inversion of modified matrices," *SIAM Journal on Applied Mathematics*, vol. 24, no. 3, pp. 315–323, 1973.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimeida*, 2014, pp. 675–678.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[25] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval.," in *BMVC*, 2010, vol. 1, p. 5.

[26] 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 *CIVR*. ACM, 2009, p. 48.