Online Fast Adaptive Low-Rank Similarity Learning for Cross-Modal Retrieval

Yiling Wu, Shuhui Wang, and Qingming Huang, *Fellow, IEEE*

***Abstract*—The semantic similarity among cross-modal data objects, e.g., similarities between images and texts, are recognized as the bottleneck of cross-modal retrieval. However, existing batch-style correlation learning methods suffer from prohibitive time complexity and extra memory consumption in handling large-scale high dimensional cross-modal data. In this paper, we propose a Cross-Modal Online Low-Rank Similarity function learning (CMOLRS) method, which learns a low-rank bilinear similarity measurement for cross-modal retrieval. We model the cross-modal relations by relative similarities on the training data triplets and formulate the relative relations as convex hinge loss. By adapting the margin in hinge loss with pair-wise distances in feature space and label space, CMOLRS effectively captures the multi-level semantic correlation and adapts to the content divergence among cross-modal data. Imposed with a low-rank constraint, the similarity function is trained by online learning in the manifold of low-rank matrices. The low-rank constraint not only endows the model learning process with faster speed and better scalability, but also improves the model generality. We further propose fast-CMOLRS combining multiple triplets for each query instead of standard process using single triplet at each model update step, which further reduces the times of gradient updates and retractions. Extensive experiments are conducted on four public datasets, and comparisons with state-of-the-art methods show the effectiveness and efficiency of our approach.**

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

I. Introduction

Multimedia data with massive volumes and heterogeneous modalities are pervasive on the Web. As an appealing task, cross-modal retrieval [1]–[7] is to return a list of documents in another modalities given a query in one modality, e.g., retrieving images with a text query or retrieving texts with an image query. By retrieving cross-modal documents, users can find images that best illustrate the topic of a textual query, or textual descriptions that best explain the content of a visual query. The list of returned documents is to be ordered according to the similarities between database documents and the query. Learning a good similarity function that well reflects the semantic relevance between data objects of different modalities is the bottleneck problem in cross-modal retrieval. However, data in diverse modalities are presented in heterogeneous feature spaces, thus they usually have diversified statistical properties. The modality heterogeneity leads to great challenge in directly measuring the semantic relevance among massive cross-modal data objects. In this paper, we address several critical issues towards effective and efficient cross-modal retrieval.

First, a standard solution for measuring the semantic relation of multi-modal data is to project data into a shared latent space so that they can be compared [8]–[10]. For example, a common subspace is learned by CCA [11] which maximizes the correlation between the two projected modalities. A shared latent semantic model can be established by linear regression or other classification models from the feature space to the label space [12]–[14]. However, the relation among data objects, expressed by distance of data objects in the shared latent space with a predefined number of dimensions, is not learned directly towards ranking-based retrieval task. Different with approaches [8], [11], [15], [16], we consider similarity-based correlation learning strategy which measures the cross-modal similarity by a simple and flexible bilinear function. The similarity measurement can be optimized directly towards the goal of cross-modal retrieval where the data objects can be ranked by their similarity scores to the queries.

Specifically, the similarity function is learned by preserving relative similarity represented by a set of triplets. Each triplet consists of a query, a positive example and a negative example, where the positive example is more similar to the query than the negative example. Instead of absolute similarity [17], [18] that encodes only similar/dissimilar relation, the training loss on each triplet with relative similarity [19], [20] expresses more general multi-level semantic relation on multi-label data by comparing the similarities of different data pairs. To further deal with the complex distribution among cross-modal data, we propose an adaptive hinge loss which adjusts the margin of each triplet by considering the distance of data objects in both label space and feature space. By minimizing the loss of relative similarity ranking with adaptive margin, the learned similarity is more semantically consistent by encoding the local affinity structure in

Manuscript received January 17, 2018; revised March 3, 2019; accepted September 9, 2019. Date of publication September 20, 2019; date of current version April 23, 2020. This work was supported in part by the National Natural Science Foundation of China under Grants 61672497, 61931008, 61620106009, U1636214, and 61836002, in part by National Basic Research Program of China under 973 Program 2015CB351802, and in part by Key Research Program of Frontier Sciences of CAS under Grant QYZDJ-SSW-SYS013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Guo-Jun Qi. *(Corresponding author: Shuhui Wang.)*

Y. Wu and S. Wang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yiling.wu@vipl.ict.ac.cn; wangshuhui@ict.ac.cn).

Q. Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China, with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: qmhuang@ucas.ac.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2942494

both feature space and label space. Consequently, it can better adapt to content and semantic diversity of cross-modal data.

Second, most existing models are learned in batch styles [2], [8], [21], requiring training data storage beforehand and involving high computation complexity. Online learning is an alternative solution [22], [23] which can avoid the high computational burden in batch learning strategy. Grangier *et al.* [19] proposed an online similarity learning method for text-to-image retrieval by projecting visual representation to the tag space directly with a full-rank matrix of size $d^t \times d^v$ learned without any rank constraint, where d^t and d^v are the dimensions of textual and visual representations. It may lead to ill-posed solution on high dimensional cross-modal data, and the model training is very time-consuming when d^t and d^v are large. On the other hand, the low-rank assumption on the projected space is helpful in reducing the over-fitting risk, and has been widely applied in metric and similarity learning [24]–[27]. In cross-modal correlation learning aspect, Kang *et al.* [21] and Zhang *et al.* [28] impose nuclear-norm penalization on the bilinear similarity function or the mapping matrices. However, they still require computationally prohibitive SVD decomposition [24], [25] at each iteration.

We consider online low-rank similarity learning to gain more efficiency and flexibility in processing high dimensional cross-modal data. The bilinear similarity function can be seen as taking inner product of the projected representations, and learning a low dimensional subspace is equivalent to learning a low-rank similarity function. The low-rank constraint provides a natural regularization on the projected space to handle the high dimensionality. Practically, the online low-rank similarity learning problem is to perform online learning on the k-dimensional matrix manifold. Inspired by LORETA [29], we develop an efficient online similarity optimization algorithm which consists of a gradient step, followed by a second-order retraction back to the k-dimensional matrix manifold. With a rank k matrix, the complexity of the similarity calculation can be reduced to $O((d^t + d^v)k)$ ($k \ll d^t$, d^v). Since no SVD is needed at every iteration, the efficiency of the model learning process can be significantly enhanced.

To further speed up training, we propose a fast online learning algorithm. For all the training triplets in image-to-text or text-to-image directions, the triplet batch of the same query is combined to calculate the gradient and update the similarity function. By appropriately setting the number of sampled triplets for each query, a better trade-off between bias and variance can be reached on the gradient calculation using a small triplet batch. With the reduced number of gradient update and retraction steps, the training time is also reduced. The convergence behavior and stability can also be improved compared to using single triplet.

We propose Cross-Modal Online Low-Rank Similarity learning (CMOLRS) for cross-modal retrieval, which is a systematic extension of our previous work [30]. Key contribution is summarized as follows:

- We propose an adaptive hinge loss on cross-modal relative similarities. The bilinear similarity function is simple and flexible to encode the cross-modal semantic relation. The adaptive margin defined on both the label and feature space better deals with content divergence.
- We propose an online low-rank similarity learning framework to handle massive high dimensional cross-modal data. With a rank k matrix, the calculation of the similarity score reduces to $O((d^t + d^v)k)$ operations.
- We develop an optimization algorithm for online similarity learning in the low-rank matrix manifold. We further propose fast-CMOLRS, which combines triplets with the same query to reduce the times of gradient updates and retractions. Therefore, more efficiency and robustness is gained in processing large scale datasets.
- Experimental comparisons with the state-of-the-art show the effectiveness and efficiency of our approach.

Section II discusses related work. Section III presents the proposed cross-modal online low-rank similarity function learning method. Section IV provides extensive experimental evaluation. Section V concludes this study.

II. Related Work

A. Cross-Modal Retrieval

A number of tasks have emerged to exploit the relation and semantics among multi-modal data. Cross-modal retrieval is one of these interesting tasks which returns documents in different modality from the query. Partial least squares (PLS) [15], [31] and Canonical correlation analysis (CCA) [11] are two classical methods. PLS creates orthogonal score vectors by maximizing the covariance between two modalities. CCA learns a latent space by maximizing the correlating relationships between two modalities, which can be extended by nonlinear mappings using kernel trick into Kernel CCA [32]. A huge body of works are based on CCA [11], KCCA [32] and their variants. For example, Sharma *et al.* [8] combine popular supervised and unsupervised feature extraction techniques with CCA and KCCA to achieve closeness between multi-view samples of the same class. ml-CCA [33] extends CCA by using high level multi-label semantics to establish correspondences between different modalities. Tran *et al.* [34] put forward a new representation method that aggregates information on their aligned subspaces which are provided by the KCCA projections of both modalities.

In a few literature, the retrieval problem is treated by classification methods, and the label space is directly used as the common space. Wang *et al.* [12] directly treat label space as the common space and perform linear regression to the common space. When performing linear regression, ℓ_{21}-norm and trace norm penalties are imposed for feature selection and enforcing the low-rank property on the projected representation, respectively. In LGCFL [13], ε-dragging is performed on the label space to drive the regression targets of different classes moving along opposite directions, and group sparsity constraints are imposed in the regression to learn the most discriminant groups. Deng *et al.* [14] propose a discriminative dictionary learning method augmented with common label alignment which first learns sparse codes and then regresses them to the label space.

Learning-to-rank techniques are also employed by [19], [35]–[37]. SSI [35] is a pairwise learning-to-rank algorithm that learns a bilinear model with stochastic gradient descent. Grangier [19] propose PAMIR to learn a linear mapping from visual space to text space trained by PA learning algorithm with pairwise ranking examples. Yao *et al.* [37] propose RCCA which jointly explores subspace learning and pairwise learning-to-rank technique. It initially finds a common subspace by CCA, then simultaneously learns a bilinear similarity function and adjusts the subspace to preserve the preference relations. In addition, a bi-directional listwise ranking loss is optimized in Bi-CMSRM [36]. Its latent embedding is discriminatively learned by structural risk minimization. Yang *et al.* [38] propose LRGA in which a local linear regression model is used to predict the ranking values of neighboring points of each data point. Then, a unified objective function globally enforces the alignment of local models to give each data point an optimal ranking value.

Deep neural networks are used for learning the non-linear mappings for two correlated modalities [9], [18], [39]–[41]. Masci [18] propose MMNN using a two-branch neural network with a loss that allows unified treatment of intra- and inter-modality similarity learning. Yan *et al.* [39] propose to learn the joint embeddings by complex nonlinear transformations of deep canonical correlation analysis (DCCA) [42] between two modalities. Wang *et al.* [40] propose a method named DSPE for learning joint embeddings of images and texts by using a large margin objective combining cross-view ranking constraints and intra-view neighborhood structure preservation constraints. BRNN [43] embeds fragments of images and fragments of sentences by deep neural networks into a common space. DVSH [44] is a hybrid deep architecture that constitutes a visual semantic fusion network for learning isomorphic hash codes, and two modality-specific hashing networks for learning hash functions. CCL [45] exploits both coarse-grained instances and fine-grained patches to build a multi-grained fusion hierarchical network. Peng *et al.* [46] propose modality-specific cross-modal similarity measurement (MCSM) by constructing independent semantic space for each modality, which directly generates modality-specific cross-modal similarity without explicit common representation.

In recent years, a few online hashing methods have been proposed. Xie *et al.* propose OCMH [23] in which hash codes are represented by the permanent SLC and dynamic transfer matrix. Thus, updating of hash codes is transformed to the efficient updating of SLC and transfer matrix. Another work proposed by Xie *et al.* is called DMVH [22], which can adaptively augment hash codes according to dynamic changes of image. Specifically, DMVH can increase the code length and update weight of each view in the online learning process.

As insufficient training data is a common problem in real world, Huang *et al.* [47] propose a deep cross-media knowledge transfer approach which transfers knowledge from a large-scale cross-media dataset to promote the model training on small-scale cross-media dataset.

B. Low-Rank Matrix Learning

Matrix low-rankness [24]–[26], [29] is a widely adopted regularization in multimedia analytics. However, the rank minimization problem is NP-hard, and is therefore ineffective for most practical applications. One solution is to replace $rank(X)$ by its convex envelope. It turns out that the convex envelope of $rank(X)$ on the set $\{X \in \mathbb{R}^{m\times n} : \|X\|_2 \leq 1\}$ is the nuclear-norm $\|X\|_*$ which is defined as the sum of the singular values of X. Nuclear-norm is widely used in matrix completion [24], [25]. Algorithms such as singular value thresholding (SVT) algorithm [24] and fixed-point continuation (FPC) algorithm [25] were thereafter proposed to solve nuclear-norm minimization problem.

The LogDet divergence [26], [27] has been used to learn low-rank matrix, which enjoys a range-space preserving property and positive semi-definiteness. However, traditional metric learning algorithms only focus on single-modal data, and the LogDet divergence is defined over the positive semi-definite matrices, which suffers difficulties in addressing the cross-modal similarity learning with different dimensions. To learn low-rank matrix efficiently, Shalit *et al.* [29] propose LORETA algorithm to perform online learning in the manifold of low-rank matrices. LORETA preserves rank of the input matrix by a second-order retraction back to the manifold after the gradient update.

A few cross-modal retrieval methods involve low-rank constraint. Kang *et al.* [21] impose nuclear-norm penalization on bilinear similarity function. Zhang *et al.* [28] adopt a nuclear-norm regularization to learn low-rank mapping matrices. However, both methods need SVD decomposition at each iteration. We learn the low-rank cross-modal similarity measurement in an online fashion which does not need the time consuming SVD decomposition at each online update step.

III. Approach

A. The Framework

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

Note that we construct two models respectively for image-to-text and text-to-image retrieval. Due to the intrinsic characteristics of low-rank constraint and online learning, it would be better to optimize the similarity towards a unique goal in our research context, e.g., optimizing image-to-text retrieval or text-to-image retrieval performance, so the similarity function can be more sufficiently learned online towards each retrieval direction. In the consequent sections, we describe our approach in text-to-image direction, while the whole process can also be applied to image-to-text direction.

The framework is shown in Fig. 1. Our method is composed of three main steps. In the first step, we sample one or several triplets from the training database by a textual (for text-to-image) or visual (for image-to-text) query. Then in Step 2, we calculate the adaptive margin on each of the triplets by considering distances in both the feature space and the label space. In Step 3,

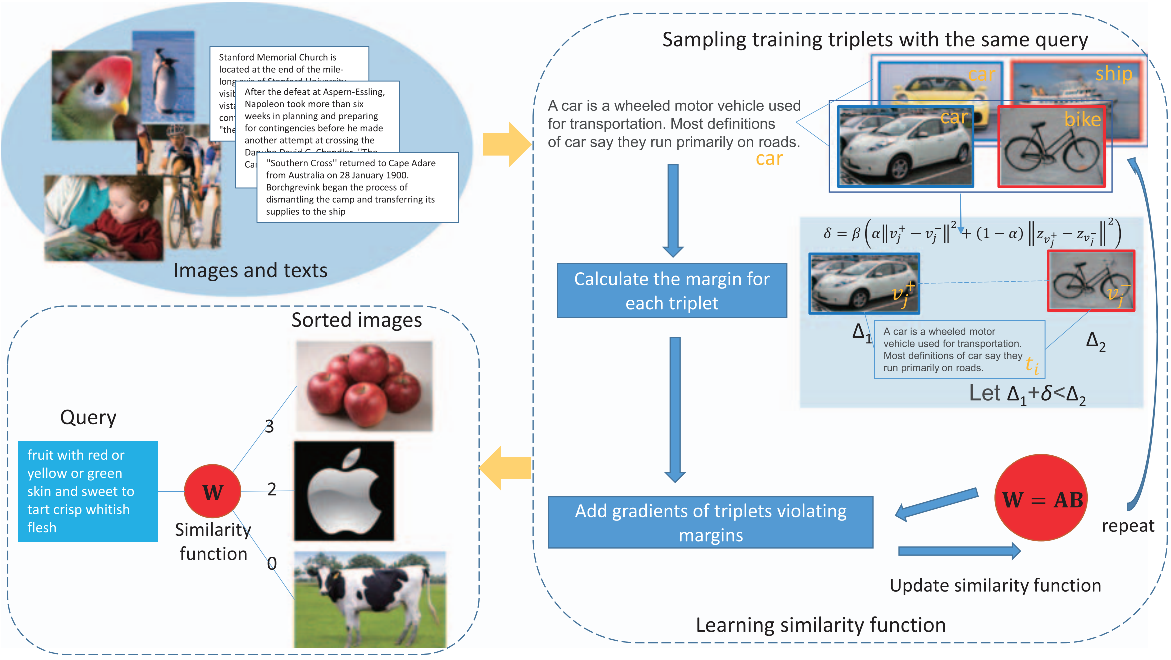

Stanford Memorial Church is located at the end of the mile-long
After the defeat at Aspern-Essling, Napoleon took more than six weeks in planning and preparing for contingencies before he made another attempt at crossing the
"Southern Cross" returned to Cape Adare from Australia on 28 January 1900. Borchgrevink began the process of dismantling the camp and transferring its supplies to the ship
Images and texts
Sampling training triplets with the same query
A car is a wheeled motor vehicle used for transportation. Most definitions of car say they run primarily on roads.
car
car
ship
car
bike
Calculate the margin for each triplet
$\delta = \beta\left(\alpha\|v_j^+ - v_j^-\|^2 + (1-\alpha)\left\|z_{v_j^+} - z_{v_j^-}\right\|^2\right)$
v_j^+
v_j^-
Δ_1
Δ_2
A car is a wheeled motor vehicle used for transportation. Most definitions of car say they run primarily on roads.
t_i
Let $\Delta_1+\delta<\Delta_2$
Sorted images
Query
fruit with red or yellow or green skin and sweet to tart crisp whitish flesh
W
Similarity function
3
2
0
Add gradients of triplets violating margins
W = AB
repeat
Update similarity function
Learning similarity function

Fig. 1. Framework of the proposed method.

we calculate the gradient using the sampled training triples to perform online update on the similarity matrix $\mathbf{W}$ using our approach. This process continues until the stoping criteria is reached or no triplet is generated.

B. *Adaptive Relative Similarity*

The goal of this paper is to learn a similarity function for cross-modal data. When retrieving images by a text query, we rank the images in database according to the similarity scores with the text query using the learned similarity function. The similarity is formulated as a simple bilinear function:

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

where $\mathbf{W} \in \mathbb{R}^{d^t \times d^v}$. Different from traditional single-modal problem, $\mathbf{W}$ does not need to be square. This bilinear function can be seen as a linear function on joint feature of $t_i v_j^\top$, so the multiplicative interactions between the two modalities can be measured. Moreover, the negative correlations via negative values in $\mathbf{W}$ can be learned from the data [35]. It can also be seen as taking the inner product of the two modality-specific mappings. However, explicitly factorizing the matrix $\mathbf{W}$ leads to a non-convex problem that is not stable for online learning, so we take $\mathbf{W}$ as a whole during the learning process.

For retrieval task, the relative order of retrieved documents is more important than their absolute similarity to the query. We apply relative similarity to learn the similarity function, which optimizes pairwise ranking in order to achieve a better consistency in the final ranking. The relative relations in the training set are represented by a set of triplets $T = \{(t_i, v_j^+, v_j^-)\}$, where (t_i, v_j^+, v_j^-) encodes that t_i and v_j^+ are more relevant than t_i and v_j^-. Here t_i is called a query, v_j^+ is called a positive example and v_j^- is called a negative example. For triplet (t_i, v_j^+, v_j^-), the similarity score between text t_i and image v_j^+ is required to be larger than the score between text t_i and image v_j^- by a predefined margin δ,

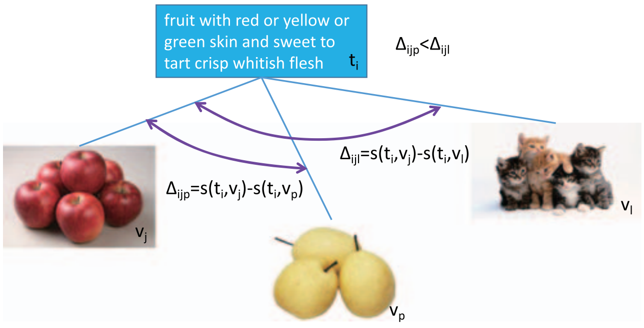

fruit with red or yellow or green skin and sweet to tart crisp whitish flesh t_i
$\Delta_{ijp}<\Delta_{ijl}$
Δ_{ijl}=s(t_i,v_j)-s(t_i,v_l)
Δ_{ijp}=s(t_i,v_j)-s(t_i,v_p)
v_j
v_l
v_p

Fig. 2. An example of differences in relative similarities. Δ_{ijp} and Δ_{ijl} denote the differences (margins) of similarity scores.

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

However, for different triplets, e.g., (t_i, v_j^+, v_j^-) and $(t_{i'}, v_{j'}^+, v_{j'}^-)$, the values of $s(t_i, v_j^+) - s(t_i, v_j^-)$ and $s(t_{i'}, v_{j'}^+) - s(t_{i'}, v_{j'}^-)$ may be different. For example, as shown in Fig. 2, given a text t_i describing "apples", images v_j depicting "apples", v_p depicting "pears" and v_l depicting "cats", the value of $s(t_i, v_j) - s(t_i, v_p)$ should be smaller 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. To model the relative similarity difference, we define margin to be a function of t_i, v_j^+ and v_j^-, and Eq. (2) becomes:

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

Obviously, the similarity scores of similar images to a given text will be similar. For the example in Fig. 2, since v_j and v_p are similar, $s(t_i, v_j)$ and $s(t_i, v_p)$ should not differ too much. Thus similar images should have smaller margin δ. When measuring the similarity of images which is a intra-modal problem, both semantic relevance and visual similarity should be considered. Therefore, we measure the distance between images in both the label space and the feature space by using a simple Euclidean distance, and define $\delta(t_i, v_j^+, v_j^-)$ as:

$$\delta(t_i, v_j^+, v_j^-) = \beta(\alpha\|v_j^+ - v_j^-\|^2 + (1-\alpha)\|z_{v_j^+} - z_{v_j^-}\|^2), \tag{4}$$

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

We reformulate the constraint in Eq. (2) as a standard hinge loss. The resulting loss function is given by:

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

On the whole, our goal is to minimize a global loss $\mathcal{L}$ that accumulates hinge losses in Eq. (5) over all possible triplets D_{train} on the training set:

$$\mathcal{L} = \sum_{(t_i, v_j^+, v_j^-) \in D_{train}} l(t_i, v_j^+, v_j^-). \tag{6}$$

With the adaptive margin, our model treats every triplet differently. This is somehow like hard negative mining which selects samples with larger losses and neglects samples than smaller losses. Our model selects triplets and ignores triplets according to the margins and relative similarity scores. For triplet with similar positive and negative examples, we give looser constraint, while, for triplet with dissimilar positive and negative examples, we give tighter constraint. Compared to the fixed margin setting, we select more triplets with dissimilar positive and negative examples and ignore triplets with similar positive and negative examples. Therefore, triplets with dissimilar positive and negative examples are paid with more attention, while small margin is given to triplet with similar positive and negative examples. compared to setting a fixed large margin, many triplets with similar positive and negative examples are omitted from the online updating process, which leads to acceleration on the training process.

Although our similarity function only explicitly relates intermodal data, the inter-modal relation and intra-modal relation have mutual impact and complementary nature to each other. For example, the intra-modal similarity can be used as the local affinity (manifold) for propagating inter-modal relation [17], [48]. On the other hand, if two images are both similar to a text, they should be related to each other as well [36]. In fact, we model the intra-modal relation by adaptive margin which considers both geometric structure and semantic similarity between two selected samples in a triplet in the same modality. Therefore, the proposed adaptive-margin-based empirical hinge loss provides strong guidance on the similarity learning procedure, and a more semantically consistent cross-modal similarity can be well established.

Algorithm 1: Online Low-Rank Matrix Learning With Rank-1 Gradient (LORETA-1)

```
Input: $\mathbf{A}_{i-1}, \mathbf{B}_{i-1}, \mathbf{A}_{i-1}^{\dagger}, \mathbf{B}_{i-1}^{\dagger}, p, q$
Output: $\mathbf{A}_i, \mathbf{B}_i, \mathbf{A}_i^{\dagger}, \mathbf{B}_i^{\dagger}$
  $a_1 = \mathbf{A}_{i-1}^{\dagger} \cdot p$, $b_1 = \mathbf{B}_{i-1}^{\dagger} \cdot q$
  $a_2 = \mathbf{A} \cdot a_1$
  $s = b_1^{\top} \cdot a_1$
  $a_3 = a_2(-\frac{1}{2} + \frac{3}{8}s + p(1 - \frac{1}{2}s))$
  $\mathbf{A}_i = \mathbf{A}_{i-1} + a_3 \cdot b_1^{\top}$
  $b_2 = (q^{\top}\mathbf{B}_{i-1}) \cdot \mathbf{B}_{i-1}^{\dagger}$
  $b_3 = b_2(-\frac{1}{2} + \frac{3}{8}s + q^{\top}(1 - \frac{1}{2}s))$
  $\mathbf{B}_i^{\top} = \mathbf{B}_{i-1}^{\top} + a_1 \cdot b_3$
  $\mathbf{A}_i^{\dagger}$=rank-1-pseudoinverse-update$(\mathbf{A}_{i-1}, \mathbf{A}_{i-1}^{\dagger}, a_3, b_1)$
  $\mathbf{B}_i^{\dagger}$=rank-1-pseudoinverse-update$(\mathbf{B}_{i-1}, \mathbf{B}_{i-1}^{\dagger}, b_3, a_1)$
```

C. Online Learning in the Low-Rank Manifold

Based on the adaptive margin based hinge loss, we learn a low-rank matrix $\mathbf{W}$. Adding a rank constraint, we have the following minimization problem:

$$\min_{\mathbf{W}} \sum_{(t_i, v_j^+, v_j^-) \in D_{train}} l(t_i, v_j^+, v_j^-), \quad \text{s.t. } \operatorname{rank}(\mathbf{W}) = k. \tag{7}$$

In order to scale to large datasets and facilitate model update on-demand, we consider online learning in which a triplet is provided at each iteration. However, the low-rank constraint makes the optimization difficult. Two naive approaches, i.e., 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 [29] to learn a low-rank matrix $\mathbf{W}$ by online learning on the manifold of low-rank matrices. The set of $d^t \times d^v$ matrices of rank k where $k \leq d^t, d^v$ is an $(d^t + d^v)k - k^2$ dimensional manifold embedded in $\mathbb{R}^{d^t \times d^v}$. LORETA [29] performs stochastic gradient descent on 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 $\mathcal{M}$ in representation space x. The mathematically ideal retraction is called exponential mapping which approximates the exponential mapping by second order. From Taylor expansion perspective, it provides more accurate approximating in learning nonlinearities. The best known example of a second-order retraction is the projection operation, while computing the projection is too costly on the manifold of low-rank matrices. In our algorithm, we use an alternative second-order retraction for computationally efficiency.

Specifically, at every iteration the algorithm suffering a loss, the gradient of the hinge loss in Eq. (5) is $\frac{\partial l(t_i, v_j^+, v_j^-)}{\partial \mathbf{W}} = t_i(v_j^+ - v_j^-)^{\top}$ which is a Rank-1 gradient. The online low-rank matrix update procedure with a Rank-1 gradient is summarized in Algorithm 1. For the computational cost of Algorithm 1, the bottleneck lies on the computation of the pseudo-inverse of the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$. Following a procedure developed by [49], we keep the pseudo-inverses $\mathbf{A}_{i-1}^{\dagger}$ and $\mathbf{B}_{i-1}^{\dagger}$ from the previous

Algorithm 2: Cross-Modal Online Low-Rank Similarity Learning (CMOLRS)

```
Input: $\mathbf{A}_0 \in \mathbb{R}_*^{d^t \times k}$, $\mathbf{B}_0 \in \mathbb{R}_*^{d^v \times k}$, s.t. $\mathbf{W}_0 = \mathbf{A}_0\mathbf{B}_0^\top$, $\mathbf{A}_0^\dagger$,
    $\mathbf{B}_0^\dagger$ are the pseudo-inverses of $\mathbf{A}_0$ and $\mathbf{B}_0$, step size $\eta \geq 0$
Output: $\mathbf{W}_N = \mathbf{A}_N\mathbf{B}_N^\top$
  for $i = 1, \dots, N$ do
    Randomly sample a triplet $(t_i, v_j^+, v_j^-)$
    if $s(t_i, v_j^+) < s(t_i, v_j^-) + \delta(t_i, v_j^+, v_j^-)$ then
      $p = -\eta t_i, q = v_j^+ - v_j^-$
    $\mathbf{A}_i, \mathbf{B}_i, \mathbf{A}_i^\dagger, \mathbf{B}_i^\dagger =$
    LORETA-1$(\mathbf{A}_{i-1}, \mathbf{B}_{i-1}, \mathbf{A}_{i-1}^\dagger, \mathbf{B}_{i-1}^\dagger, p, q)$
    end if
  end for
```

Algorithm 3: Fast Cross-Modal Online Low-Rank Similarity Learning (fast-CMOLRS)

```
Input: $\mathbf{A}_0 \in \mathbb{R}_*^{d^t \times k}$, $\mathbf{B}_0 \in \mathbb{R}_*^{d^v \times k}$, s.t. $\mathbf{W}_0 = \mathbf{A}_0\mathbf{B}_0^\top$, $\mathbf{A}_0^\dagger$,
    $\mathbf{B}_0^\dagger$ are the pseudo-inverses of $\mathbf{A}_0$ and $\mathbf{B}_0$, step size $\eta \geq 0$
Output: $\mathbf{W}_N = \mathbf{A}_N\mathbf{B}_N^\top$
  for $i = 1, \dots, N$ do
    Randomly sample a text $t_i$
    $q = 0$
    for $j = 1, \dots, J$ do
      Randomly sample a triplet $(t_i, v_j^+, v_j^-)$
      if $s(t_i, v_j^+) < s(t_i, v_j^-) + \delta(t_i, v_j^+, v_j^-)$ then
        $q = q + (v_j^+ - v_j^-)$
      end if
    end for
    if $q \neq 0$ then
      $p = -\eta t_i$
      $\mathbf{A}_i, \mathbf{B}_i, \mathbf{A}_i^\dagger, \mathbf{B}_i^\dagger =$
      LORETA-1$(\mathbf{A}_{i-1}, \mathbf{B}_{i-1}, \mathbf{A}_{i-1}^\dagger, \mathbf{B}_{i-1}^\dagger, p, q)$
    end if
  end for
```

round, and perform a Rank-1 update on them. The overall time and space complexities of Algorithm 1 are both $O((d^t + d^v)k)$ per gradient step, where k is the rank of $\mathbf{W}$, d^t and d^v are the dimensions of the two modalities respectively.

D. Sampling Strategy

One of the most important issues for online learning is how to generate the training samples for each model update step. Theoretically, in our research context, the actual scale of triplets is $O(M^3)$, where M is the total number of training images and texts. In practice, the number of triplets is prohibitively large and cannot be stored in memory. To facilitate large scale applications and support model update on-demand, we learn the similarity function by sampling triplets and feeding them into each step of the model update procedure as shown in Algorithm 1. The asymptotic property on using the gradient of the sampled triplets for model update can be guaranteed by the nature of addictive-loss-based models [50] and the characteristics of stochastic optimization [51].

We first randomly sample a text t_i as the query. Then we sample two images v_j^+ and v_j^- sharing different number of labels with the query. The image v_j^+ sharing more labels with the query text is taken as the positive example and the image v_j^- sharing less or even no labels is taken as the negative example. Triplets with negative examples sharing less labels to the query encode the multi-level semantic similarity relation, i.e., they are somehow semantically similar, while those sharing no labels with the query express explicit semantic similar and dissimilar relations, i.e., they are semantically dissimilar. The former case provides fine-grained semantic relation description for refining the similarity learning, while the latter case provides explicit semantic relation description for avoiding ambiguity in the similarity learning. Therefore, the two types of negative examples are equally important, and we empirically keep the ratio of triplets with negative example sharing no label and less labels with the query as 50%/50% on typical multi-label datasets.

Let $\mathbb{R}_*^{d \times k}$ denotes the set of $d \times k$ matrices of rank k. Based on the single triplet sampling strategy, the model learning procedure of standard CMOLRS is summarized in Algorithm 2. The time complexity of calculating a score with the rank k matrix $\mathbf{W}$ is $O((d^t + d^v)k)$. The time complexity of deciding whether a triplet violates the margin is $O((d^t + d^v)k)$. Along with the complexity analysis of Algorithm 1 in the above subsection, the time complexity of Algorithm 2 is $O((d^t + d^v)kN)$, where N is the number of triplets.

However, the model will be updated when a violated triplet is sampled and Algorithm 1 is performed once. To speed up the training procedure, at every iteration, we sample J triplets for a text query t_i. For a given text, the sum of gradient of the hinge losses of any number of triplets in Eq. (5) is still a rank-1 gradient. So we can update the model using J triplets by just a single call of Algorithm 1. Thus, with an appropriately setting of number of triplets, we reduce the number of gradient and retraction step, so as to reduce the training time. We call this method fast-CMOLRS. Algorithm 3 summarizes the fast-CMOLRS. We will show in Subsection IV-G that fast-CMOLRS results in reduced computational time without loss on retrieval performance.

IV. Experiments

We compare our methods with different baseline methods on WIKI, PASCAL VOC 2007, MIRFLICKR and NUS-WIDE. Codes are available at https://github.com/yiling2018/cmolrs.

A. Experimental Settings

We compared our methods with seven state-of-the-art methods. **CCA** [11] aims at learning a latent space by maximizing the correlation between two data modalities; **PLS** [15] creates orthogonal score vectors by maximizing the covariance between two data modalities; **GMLDA** [8] combines LDA with CCA to extend CCA with label information. **3view-CCA** [16] extends CCA by introducing labels as a third view. **ml-CCA** [33] extends CCA by using multi-label information to establish correspondences between two data modalities. **LRBS** [21] learns low-rank similarity measure by exploiting absolute

TABLE I
MAP@R ON WIKI

Task / Method	$R = 50$			$R = all$		
	image-to-text	text-to-image	average	image-to-text	text-to-image	average
CCA	0.377	0.376	0.377	0.274	0.254	0.264
PLS	0.454	0.513	0.484	0.355	0.327	0.341
GMLDA	0.416	0.518	0.467	0.407	0.353	0.380
3view-CCA	0.451	0.553	0.502	0.434	0.379	0.407
ml-CCA	0.441	0.558	0.500	0.436	0.381	0.409
LRBS	0.400	0.542	0.471	0.387	0.357	0.372
PL-ranking	0.465	0.532	0.499	0.400	0.365	0.383
CMOLRS	**0.476**	**0.594**	**0.535**	**0.454**	**0.414**	**0.434**
fast-CMOLRS	0.472	**0.594**	0.533	0.451	0.411	0.431

similarity with accelerated proximal gradient algorithm. **PL-ranking** [28] learns low-rank projections by optimizing both pairwise and listwise objectives.

For GMLDA and PL-ranking which are designed for multi-class case, the category of every training pair is selected randomly from its multiple labels on PASCAL, MIR and NUS. For LRBS which considers only the similar/dissimilar relations, we consider samples sharing labels as similar, otherwise dissimilar on multi-label datasets. The parameters of each compared method are determined by validation to guarantee the best-case performance. For our two methods, we set parameter α to 0.5, parameter β to 20 and rank k to 32 on all the four datasets. For the step size η, we set it to 10 on all datasets. On the multi-label datasets PASCAL, MIR and NUS, we sample half of the triplets with negative examples sharing no label with the query. The number of triplets for CMOLRS and fast-CMOLRS are kept the same. For fast-CMOLRS, we sample 4 triplets for each query, i.e., $J = 4$ in Algorithm 3.

We evaluate our methods on two cross-modal retrieval tasks, i.e., retrieving images by text queries (text-to-image) and retrieving texts by image queries (image-to-text). To perform these tasks, we first estimate the similarity of each text-image pair in the testing set, and then for each query, we rank the documents based on their similarities. To evaluate the semantic consistency of our cross-modal similarity function, Mean Average Precision (MAP), a widely used metric in retrieval task, is adopted as the evaluation metric. Given a query and a set of R retrieved documents, Average Precision (AP) is defined as AP$= \frac{\sum_{r=1}^{R} P(r)\delta(r)}{\sum_{r=1}^{R} \delta(r)}$, where $P(r)$ denotes the precision of the top r retrieved documents, and $\delta(r) = 1$ if the r-th retrieved document is relevant with the query and $\delta(r) = 0$ otherwise. We average the AP values over all queries in the query set to obtain the MAP measure. MAP@R measures MAP score at top R retrieved samples, and we set R to 50 for the top 50 retrieved samples and R to all for all the retrieved samples. Besides, we present precision-scope curve to evaluate the performance of different methods. The scope is specified by the number (K) of top-ranked documents presented to the users. The precision-scope curve clearly shows the change of precision with different K.

B. Results on WIKI Dataset

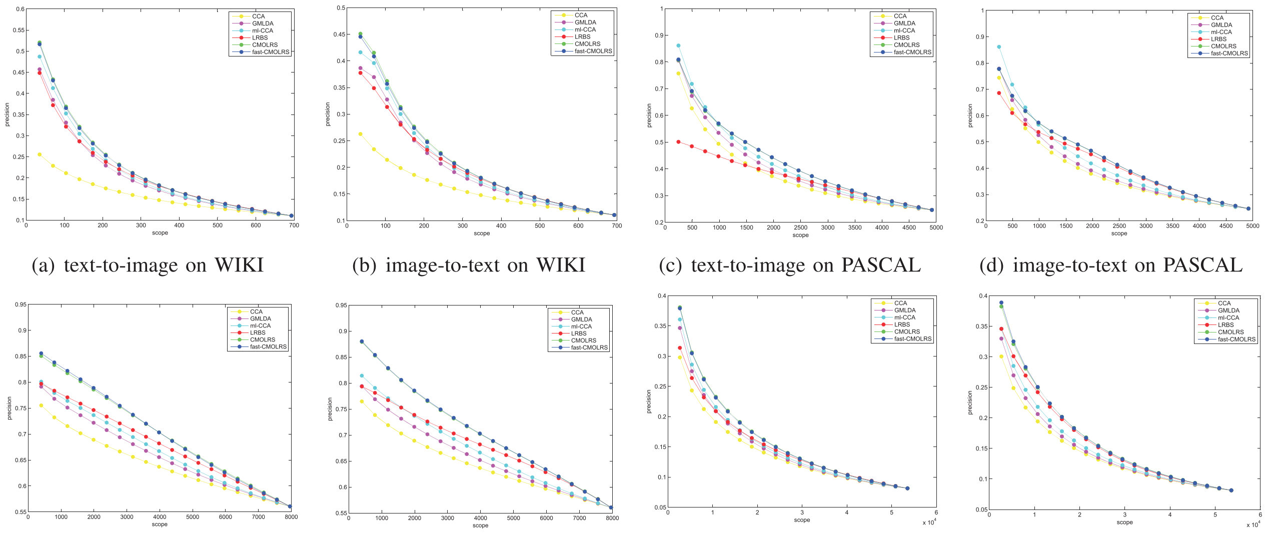
WIKI dataset [2], generated from Wikipedia's featured articles collection, contains 2,866 text-image pairs. Each pair is labeled with exactly one of 10 semantic classes. 2,173 pairs are taken as training set and 693 pairs as testing set as in [2]. For the text representations, we use the 5,000-dim bag-of-words (BoW) representations with TF-IDF weighting provided by [36]. For image representations, we extract 4,096-dim CNN features in 'fc7' with CaffeNet [52] pre-trained on ImageNet.

Table I shows the MAP scores on WIKI. The precision-scope curves are in Fig. 3(a) and 3(b). In Table I, fast-CMOLRS performs equally or slightly worse than CMOLRS measured by MAP@50 and MAP@all. From Fig. 3(a) and 3(b) we find that the performance curves of CMOLRS and fast-CMOLRS are almost the same. Notice that WIKI is a multi-class dataset, the positive sample and negative sample have different class labels and the label distance in Eq. (4) is fixed. The feature distance in visual space or textual space plays the main role in determining the adaptive margin. In this situation, on small scale multi-class dataset, the training triplets contain limited ranking patterns for model training, and combining multiple triplets for gradient calculation in fast-CMOLRS would not bring more information in per step model update compared to using only one triplet in CMOLRS. However, the experimental observation is totally different on large scale multi-label datasets.

C. Results on PASCAL VOC Dataset

PASCAL VOC 2007 [53], collected from Flickr, contains 9,963 images. Although the original user tags were not kept, Hwang *et al.* [54] collected tags using Amazon Mechanical Turk. We use the 399-dim tag frequency features provided by them [54] for text representations. The 4,096-dim CNN image features from 'fc7' layer of CaffeNet are used for image representations. Ground-truth annotation of the images which have 20 classes are used for label representations. The 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 II shows the MAP scores of different approaches on PASCAL dataset. The corresponding precision-scope curves are plotted in Fig. 3(c) and 3(d). From Table II, we can see that CMOLRS and fast-CMOLRS outperform other methods with remarkable performance gains. Curves in Fig. 3(c) and 3(d) show that our methods achieve better precision at most of the scope positions. The observations indicate that on multi-label dataset, our methods can encode the multi-label information using relative similarity and adaptive margin, thus they obtain better model capacities towards more semantically consistent ranking. In comparison, GMLDA and LGCFL can just model multi-class information, and LRBS can only



(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 MIRFLICKR (f) image-to-text on MIRFLICKR (g) text-to-image on NUSWIDE (h) image-to-text on NUSWIDE

Fig. 3. The precision-scope curves of different methods on all benchmark datasets.

TABLE II
MAP@R ON PASCAL VOC 2007

Task / Method	$R = 50$			$R = all$		
	image-to-text	text-to-image	average	image-to-text	text-to-image	average
CCA	0.830	0.887	0.859	0.665	0.655	0.660
PLS	0.819	0.745	0.782	0.643	0.538	0.591
GMLDA	0.829	0.914	0.872	0.694	0.695	0.695
3view-CCA	0.853	0.925	0.889	0.694	0.698	0.696
ml-CCA	0.854	0.924	0.889	0.712	0.718	0.715
LRBS	0.790	0.544	0.667	0.636	0.484	0.560
PL-ranking	0.610	0.813	0.712	0.549	0.600	0.575
CMOLRS	0.867	**0.933**	0.900	**0.741**	0.743	0.742
fast-CMOLRS	**0.868**	**0.933**	**0.901**	**0.741**	**0.745**	**0.743**

TABLE III
MAP@R ON MIRFLICKR

Task / Method	$R = 50$			$R = all$		
	image-to-text	text-to-image	average	image-to-text	text-to-image	average
CCA	0.803	0.798	0.801	0.669	0.671	0.670
PLS	0.845	0.778	0.812	0.677	0.668	0.673
GMLDA	0.826	0.835	0.831	0.690	0.698	0.694
3view-CCA	0.853	0.855	0.854	0.697	0.703	0.700
ml-CCA	0.841	0.840	0.841	0.708	0.713	0.711
LRBS	0.780	0.826	0.803	0.721	0.735	0.728
PL-ranking	0.809	0.838	0.824	0.663	0.671	0.667
CMOLRS	0.915	0.877	0.896	0.760	0.765	0.763
fast-CMORLS	**0.918**	**0.884**	**0.901**	**0.761**	**0.766**	**0.764**

model a two-level similar/dissimilar relation, thus they suffer from limited power in exploiting the complicated cross-modal relation.

D. Results on MIRFLICKR Dataset

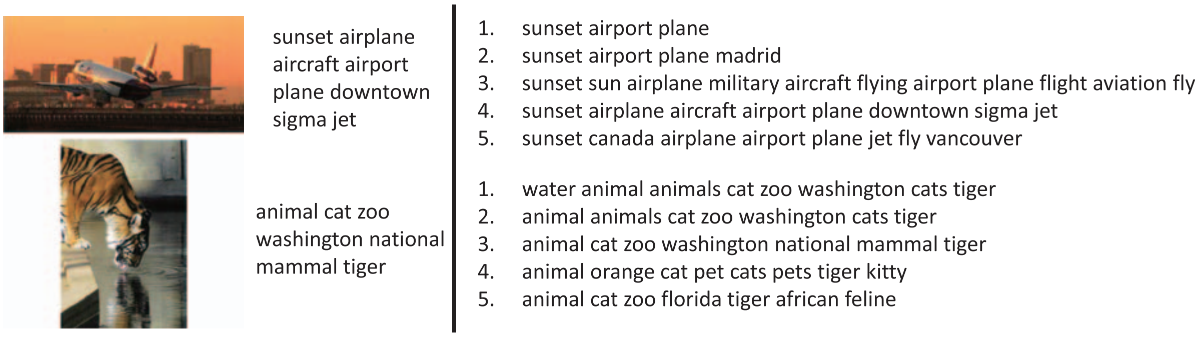
MIRFLICKR dataset [55] contains 25,000 images along with the user assigned tags. Each image-text pair is assigned with multiple labels from a total of 38 classes. We use the publicly available 2,000-dim tag frequency text features [56], and extract 4,096-dim CNN image features of ‘fc7’ from CaffeNet [52]. Following the training-testing split [55] and removing images without tags, we have 12,144 pairs for training and 7,958 pairs for testing.

Table III shows the MAP scores of different approaches on MIRFLICKR dataset. LRBS performs better than PL-ranking when setting R to all, but performs worse than PL-ranking when setting R to 50. 3view-CCA and GMLDA outperform CCA since the former two utilize label information while the latter performs without using label information. As shown in [33], ml-CCA outperforms 3view-CCA and GMLDA, since ml-CCA makes use of multi-label information.

It is apparent in Table III that our methods outperform others on both the two R values. Fast-CMOLRS performs better than

TABLE IV
MAP@R ON NUS-WIDE

Task / Method	$R = 50$ image-to-text	$R = 50$ text-to-image	$R = 50$ average	$R = all$ image-to-text	$R = all$ text-to-image	$R = all$ average
CCA	0.471	0.555	0.513	0.299	0.267	0.283
PLS	0.468	0.581	0.525	0.330	0.279	0.305
GMLDA	0.506	0.697	0.602	0.372	0.323	0.348
3view-CCA	0.536	0.714	0.625	0.375	0.321	0.348
ml-CCA	0.528	0.715	0.622	0.388	0.328	0.358
LRBS	0.492	0.507	0.500	0.330	0.255	0.293
CMOLRS	0.568	0.719	0.644	**0.415**	0.340	0.378
fast-CMOLRS	**0.574**	**0.733**	**0.654**	0.414	**0.348**	**0.381**



sunset airplane aircraft airport plane downtown sigma jet

1. sunset airport plane
2. sunset airport plane madrid
3. sunset sun airplane military aircraft flying airport plane flight aviation fly
4. sunset airplane aircraft airport plane downtown sigma jet
5. sunset canada airplane airport plane jet fly vancouver

animal cat zoo washington national mammal tiger

1. water animal animals cat zoo washington cats tiger
2. animal animals cat zoo washington cats tiger
3. animal cat zoo washington national mammal tiger
4. animal orange cat pet cats pets tiger kitty
5. animal cat zoo florida tiger african feline

Fig. 4. Examples of top 5 results of image-to-text retrieval on NUS-WIDE dataset. The first two columns are the image queries and the paired texts of image queries. The third column is the retrieved ranked results.

CMOLRS on average. Since fast-CMOLRS combines multiple triples for gradient calculation at each model update step, the similarity function tends to be learned more sufficiently and the performance more stable than CMOLRS. The behavior difference between fast-CMOLRS and CMOLRS can be explained by the difference between mini-batch stochastic gradient descent [57] using multiple training data in each model update step and canonical stochastic gradient descent [51] using a single training data in each step. The precision-scope curves are plotted in Fig. 3(e) and 3(f). We see that CMOLRS and fast-CMOLRS outperform other methods with steeper slopes on the precision-scope curves. Specifically, the precisions of our methods are much higher than others when the scope is small. Note that the size of MIRFLICKR is much larger than PASCAL and WIKI, our methods can better capture the diversified relations among cross-modal data objects, so more promising performances are gained on processing larger scale data.

E. Results on NUS-WIDE Dataset

NUS-WIDE [58], also crawled from Flickr, contains 269,648 images associated with multiple tags. Each image is labeled with 81 semantic concepts. We take the 1,000-dim tag frequency features as text features and the 4,096-dim output of ‘fc7’ layer from CaffeNet as image features. We use the original train-test split provided in the dataset for training and testing. By selecting the images with at least one tag, we obtain 79,659 images for training and 53,550 images for testing [4].

Table IV shows the MAP scores of different approaches on NUS-WIDE. PL-ranking is not compared here due to its prohibitively high time complexity using a single PC. From the table, we can see that our methods outperform the compared methods on $R = 50$ and $R = all$. The corresponding precision-scope curves are plotted in Fig. 3(g) and 3(h). Again, the promising results of our methods confirm their effectiveness on large dataset which can be attributed to its simple form in similarity function and its ability in taking full advantage of relative similarity with adaptive margin. The large number of relative similarity training triplets provide rich cross-modal relations for our online similarity learning strategy. In comparison, LRBS does not perform well on this dataset since it only models similar and non-similar relations which are not suitable for dataset containing a large number of labels. Also, LRBS is trained in batch style and requires $O(M^2)$ memory complexity to store the relation matrix, while our methods require no extra memory for data storage.

We illustrate retrieved examples of fast-CMOLRS in Fig. 5 and Fig. 4. Our method can ensure that the top ranked textual examples are semantically consistent to the query, as the words in each textual example describe at least one object in the image query. Our method can also identify those images expressing consistent semantics to the queries but with drastically different visual appearances.

F. Properties of the Factorization

Our low-rank matrix update method actually learns a factorization $\mathbf{A}$ and $\mathbf{B}$ of the similarity matrix $\mathbf{W}$ to ensure the low-rankness. We can project images and texts into a latent space with the learned factorization matrices $\mathbf{A}$ and $\mathbf{B}$. We conduct experiments to analyze the properties of the projected image and text representations of Algorithm 3. Since

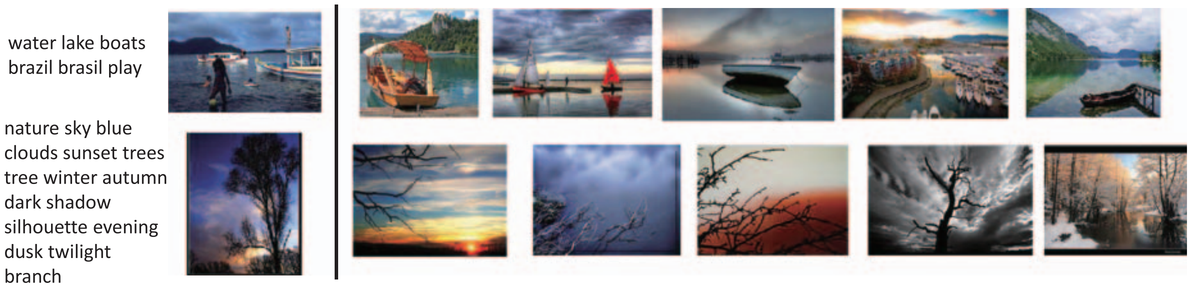


Fig. 5. Examples of top 5 results of text-to-image retrieval on NUS-WIDE dataset. The first two columns are the text queries and the paired images of text queries. The third column is the retrieved ranked results.

TABLE V
THE MAP@all SCORES OF INTRA-MODAL RETRIEVAL ON PASCAL

	text-to-text	image-to-image	average
CCA	0.590	0.738	0.664
PLS	0.597	0.608	0.603
GMLDA	0.603	0.797	0.701
3view-CCA	0.616	0.781	0.698
ml-CCA	0.629	0.800	0.714
fast-CMOLRS-i2t	0.631	0.693	0.662
fast-CMOLRS-t2i	0.622	0.793	0.708

our models are learned for image-to-text and text-to-image retrievals respectively, there are two types of factorization results using models trained in different directions. We mark the fast-CMOLRS model trained with image-to-text direction triplets as fast-CMOLRS-i2t, and text-to-image direction triplets as fast-CMOLRS-t2i, respectively.

Based on the learned projections, our model can perform intra-modal retrieval. The performance provides experimental evidence on how models handle intra-modal content divergence. We use the learned mapping matrices for image-to-image and text-to-text retrieval on PASCAL. The results are shown in Table V. We do not compare LRBS because there is no explicit factorization for the learned model.

The results in Table V validate that our methods are able to capture intra-modal relations even though we trained our model with inter-modal relative similarity constraints. In fact, if two images are similar to a text, they should be related to each other as well [36]. The proposed adaptive margin also provides a complementary way in modeling the intra-modal relation. For example, the adaptive margin gives a reference value of distance for v_j^+ (t_j^+) and v_j^- (t_j^-). When the relative similarity pairs in a triple far deviate the reference value, the learned similarity function will be penalized and updated to reduce the deviation. Note that the reference distance value between v_j^+ (t_j^+) and v_j^- (t_j^+) is calculated using both original features and labels, the proposed method can be seen as a topology and semantic preserving mechanism which propagates the affinity and semantic information in the original data to the learned low dimensional representations, which has been proved to be effective in representation learning [59] and metric learning [60] perspectives.

We visualize the learned representation with t-SNE [61] on PASCAL in Fig. 6 for better understanding of the factorized similarity function. Using the model trained in text-to-image direction, we visualized the learned representation of "bike" and "car" categories in a 2-dimensional space. The dots represent the projected representations of texts and images that are projected using $\mathbf{A}$ and $\mathbf{B}$, respectively. Dots in blue color represent data with "bike" label only, dots in green color represent data with "car" label only, and dots in red color represent data with both "bike" and "car" labels. From Fig. 6, we can see that the projected data points of different categories are well separated, while the modality heterogeneity seems to be removed as the projected images and texts are almost evenly distributed in the same region. The visualization results further indicate that our models are able to accurately associate images and texts with the same semantic meaning and enhance the intra-topic/category coherence simultaneously.

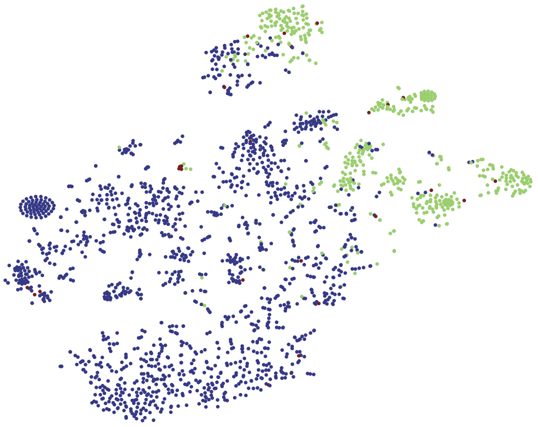
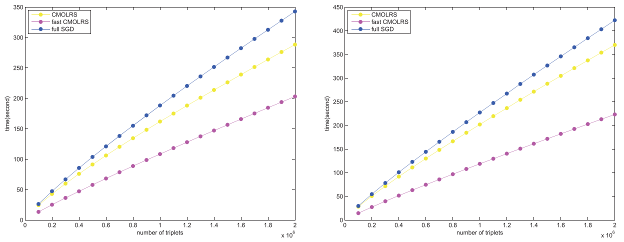


Fig. 6. 2-Dim visualization of the learned data representation of classes "bike" and "car" on PASCAL dataset (best viewed in color).

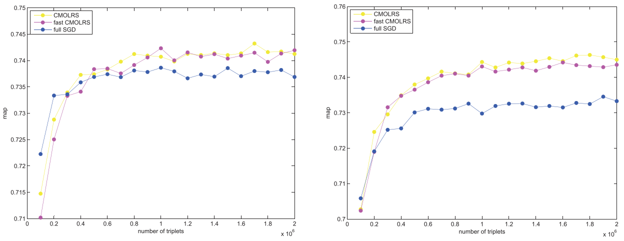
G. Comparison of Full Model and Low-Rank Models

We compare the time consumed by full-SGD, CMOLRS and fast-CMOLRS with respect to the number of sampled triplets. The experiments are conducted on PASCAL. Our hardware configuration includes an Intel Core i7 3.6 GHz CPU and a 32 GB RAM. Here full-SGD refers to learning Eq. (6) by stochastic gradient descent (SGD). Fig. 7 shows the time consumed on different numbers of triplets, and Fig. 8 shows the performance of



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

Fig. 7. Comparison of full-SGD, CMOLRS and fast-CMOLRS with respect of consumed time on PASCAL dataset.



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

Fig. 8. Comparison of full-SGD, CMOLRS and fast-CMOLRS with respect of MAP on PASCAL dataset.

cross-modal ranking in terms of MAP@all with respect to the number of triplets.

The training time of each method is linear to the number of sampled triplets. However, it is apparent from the figures that CMOLRS and fast-CMOLRS are faster than full-SGD. The enhancement in efficiency may be attributed to the low-rank constraint imposed on both of them. In the learning procedure, we need to calculate the similarity scores for each triplet to decide whether the triplet violates the margin. For full-SGD, calculating a similarity score with the full rank bilinear similarity function in Eq. (1) takes $O(d^t d^v)$ operations. However, calculating a score with a rank k matrix $\mathbf{W}$ in CMOLRS and fast-CMOLRS only takes $O((d^t + d^v)k)$ operations, where $k \ll d^t, d^v$. Fast-CMOLRS learns faster than CMOLRS, as it reduces the number of calling Algorithm 1 by combining a batch of triplets as explained in Section III-D.

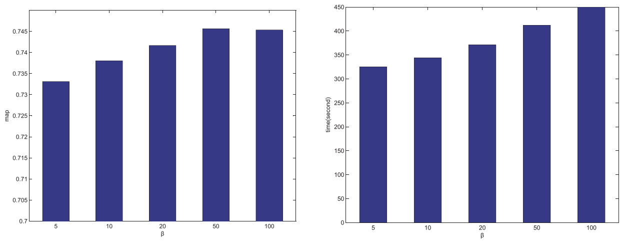
In Fig. 8, the performance of full-SGD is inferior to CMOLRS and fast-CMOLRS. The performance of CMOLRS and fast-CMOLRS are comparable in most cases. Notice that the low-rank constraint is not imposed by full-SGD, the results in Fig. 8 further show that low-rank constraint not only speeds up training, but also improves model generality.

H. Analysis on Adaptive Margin

We analyze the effect of adaptive margin on fast-CMOLRS. First, we sample 2×10^6 triplets on PASCAL dataset. When running fast-CMOLRS with those triplets in text-to-image direction, the MAP score is 0.745, the minimum of margin is 10, the maximum of margin is 114.78, and the time consumed is 273.42 seconds. Then, we conduct experiments by running fast-CMOLRS on PASCAL with fixed margin of 10 and 114.78 to show the impact of adaptive margin. The results are shown in Table VI. When fixing margin to 10, the MAP score is 0.727 with running time of 252.73 seconds. When fixing margin to 114.78, the MAP score is 0.746 with running time of 290.91 seconds. We can see from the experimental results that small margin leads to smaller training time but lower MAP score. On the contrary, large margin results in larger training time but higher performance. However, compared with adaptive margin model, the performance gain with a fixed larger margin is small, and the time consumed by the latter is much longer than the former. The experimental results ensure the advantage of adaptive margin, which selects important triplets to update the model and neglects unimportant triplets.

TABLE VI
MAP@all WITH DIFFERENT MARGINS ON PASCAL

margin	MAP@all	time consumed(seconds)
adaptive	0.745	273.42
10	0.727	252.73
114.78	0.746	290.91



(a) MAP@all (b) Time consumed by training 2×10^6 triplets

Fig. 9. MAP of fast-CMOLRS with different β on PASCAL.

I. Effect of α and β

We conducted experiments on PASCAL dataset to examine the effects of α and β in Eq. (4). The results in image-to-text direction is shown in Fig. 9. The performance of the other direction is similar. In our experiments, fast-CMOLRS is not sensitive to α. The training time increases as β increases, since a larger β leads to larger margins and stronger constraints imposed on the triplets. With a larger β, more triplets will have violated relative similarity relations (i.e., more triplets with non-zero losses), and thus the model will be updated more times. However, when β increases, MAP@all of CMOLRS first increases and then decreases. This indicates that a large β makes the learned similarity more accurate, while an overly large β will inevitably lead to over-fitting. Therefore, an appropriately setting of β will be beneficial to achieve a better performance on real-world application.

J. Effect of Step Size η

We analyze the settings of step size η in Algorithm 3. To investigating the impact of η, we conduct experiment on PASCAL dataset with different η values. The results in text-to-image direction are shown in Fig. 10, and we obtained similar results on image-to-text direction. We can see from the figure that as other gradient-based methods, the performance of fast-CMOLRS is

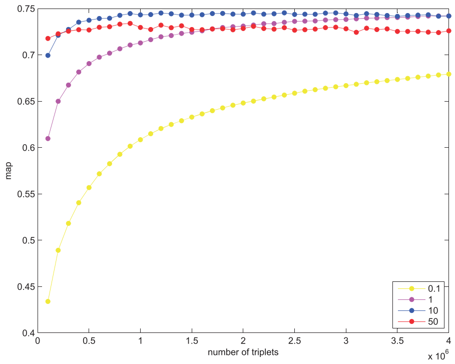


Fig. 10. MAP@all scores with different η on PASCAL.

TABLE VII
MAP@all SCORES WITH DIFFERENT RANKS ON PASCAL

	rank	8	16	32	64
text-to-image	CCA	0.616	0.665	0.659	0.643
	ml-CCA	0.647	0.712	0.701	0.685
	CMOLRS	**0.705**	**0.734**	**0.741**	**0.741**
image-to-text	CCA	0.598	0.654	0.645	0.627
	ml-CCA	0.645	0.718	0.700	0.684
	CMOLRS	**0.705**	**0.741**	**0.743**	**0.741**

highly related to step size. When η is small (i.e., $\eta = 0.1$), the MAP score increases slowly. When η is large (i.e., $\eta = 50$), the MAP score increases very fast, but can not converge to optimal value. When η is suitable (i.e., $\eta = 10$), the MAP increases very fast and stably. This shows that a suitable step size is important for fast-CMOLRS.

K. Effect of Rank k

We also conduct experiments to examine the effect of the rank of $\mathbf{W}$. We ran CCA, ml-CCA and CMOLRS with different rank settings on PASCAL dataset. Table VII shows their MAP scores under different rank settings. From the table, we can see that CMOLRS outperforms CCA and ml-CCA with different ranks. The performances of the compared approaches achieve the highest when setting rank to 16 or 32. This may be attributed to the intrinsic semantic dimensionality of the PASCAL dataset which contains 20 labels, i.e., the dimension of the label space is 20. All the compared approaches suffer a significant performance drop when setting the rank to 8. When setting rank to 16, 32 and 64, the performance of CMOLRS does not change very much. Compared to CCA and ml-CCA, CMOLRS performs more stably to the change of ranks.

V. CONCLUSION

We present CMOLRS, an online low-rank similarity learning approach for large scale high dimensional cross-modal data. The similarity is modeled by a set of relative similarity with adapted margin. We develop an algorithm for online learning on the manifold of low-rank matrices to learn the similarity function efficiently. To further speed up the online training, we propose fast-CMOLRS which updates multiple triplets at each step. Experiments clearly demonstrate the promising power of our method over state-of-the-art methods.

REFERENCES

[1] Y.-T. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 221–229, Feb. 2008.

[2] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia.*, 2010, pp. 251–260.

[3] M. Lazaridis, A. Axenopoulos, D. Rafailidis, and P. Daras, "Multimedia search and retrieval using multimodal annotation propagation and indexing techniques," *Signal Process.: Image Commun.*, vol. 28, no. 4, pp. 351–367, 2013.

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

[5] P. Daras, S. Manolopoulou, and A. Axenopoulos, "Search and retrieval of rich media objects supporting multiple multimodal queries," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 734–746, Jun. 2012.

[6] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Cross-modal retrieval using multiordered discriminative structured subspace learning," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1220–1233, Jun. 2017.

[7] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[8] S. Abhishek, K. Abhishek, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 2160–2167.

[9] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.

[10] X. Wu, Y. Qiao, X. Wang, and X. Tang, "Bridging music and image via cross-modal ranking analysis," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1305–1318, Jul. 2016.

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

[12] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2088–2095.

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

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

[15] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection*. Berlin, Germany: Springer, 2006, pp. 34–51.

[16] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[17] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 1198–1204.

[18] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, Apr. 2014.

[19] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, Aug. 2008.

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

[21] C. Kang *et al.*, "Cross-modal similarity learning: A low rank bilinear formulation," in *Proc. Int.Conf. Inf. Knowl. Manage.*, 2015, pp. 1251–1260.

[22] L. Xie, J. Shen, J. Han, L. Zhu, and L. Shao, "Dynamic multi-view hashing for online image retrieval," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3133–3139.

[23] L. Xie, J. Shen, and L. Zhu, "Online cross-modal hashing for web image retrieval," in *Proc. 30th AAAI Conf. Artif. Intell.*, pp. 294–300, 2016.

[24] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[25] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and Bregman iterative methods for matrix rank minimization," *Math. Program.*, vol. 128, no. 1, pp. 321–353, 2011.

[26] B. Kulis, M. A. Sustik, and I. S. Dhillon, "Low-rank kernel learning with Bregman matrix divergences," *J. Mach. Learn. Res.*, vol. 10, pp. 341–376, 2009.

[27] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 209–216.

[28] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Pl-ranking: A novel ranking method for cross-modal retrieval," in *Proc. ACM Multimedia*, 2016, pp. 1355–1364.

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

[30] Y. Wu, S. Wang, W. Zhang, and Q. Huang, "Online low-rank similarity function learning with adaptive relative margin for cross-modal retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2017, pp. 823–828.

[31] J. He, B. Ma, S. Wang, Y. Liu, and Q. Huang, "Cross-modal retrieval by real label partial least squares," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 227–231.

[32] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[33] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4094–4102.

[34] T. Quynh Nhi Tran, H. Le Borgne, and M. Crucianu, "Aggregating image and text quantized correlated components," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2046–2054.

[35] B. Bai *et al.*, "Learning to rank with (a lot of) word features," *Inf. Retrieval*, vol. 13, no. 3, pp. 291–314, 2010.

[36] F. Wu *et al.*, "Cross-media semantic representation via bi-directional learning to rank," in *Proc. ACM Multimedia*, 2013, pp. 877–886.

[37] T. Yao, T. Mei, and C.-W. Ngo, "Learning query and image similarities with ranking canonical correlation analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 28–36.

[38] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 175–184.

[39] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3441–3450.

[40] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5005–5013.

[41] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian, "Joint global and co-attentive representation learning for image-sentence retrieval," in *Proc. ACM Multimedia*, 2018, pp. 1398–1406.

[42] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[43] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[44] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1445–1454.

[45] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2018.

[46] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5585–5599, Nov. 2018.

[47] X. Huang and Y. Peng, "Deep cross-media knowledge transfer," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8837–8846.

[48] G. Song, S. Wang, Q. Huang, and Q. Tian, "Multimodal similarity gaussian process latent variable model," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4168–4181, Sep. 2017.

[49] C. D. Meyer, Jr., "Generalized inversion of modified matrices," *SIAM J. Appl. Math.*, vol. 24, no. 3, pp. 315–323, 1973.

[50] J. Friedman *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *Ann. Stat.*, vol. 28, no. 2, pp. 337–407, 2000.

[51] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.

[52] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[53] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[54] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval.," in *Proc. British Mach. Vis. Conf.*, 2010, vol. 1, no. 2, pp. 1–12.

[55] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.

[56] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[57] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Math. Program.*, vol. 127, no. 1, pp. 3–30, 2011.

[58] T.-S. Chua *et al.*, "Nus-wide: a real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, p. 48.

[59] T. Martinetz and K. Schulten, "Topology representing networks," *Neural Netw.*, vol. 7, no. 3, pp. 507–522, 1994.

[60] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[61] L. van der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Yiling Wu received the B.S. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2013, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2019. She is currently working with Huawei Ltd., Shenzhen, China. Her research interests include image-text retrieval and deep learning.

Shuhui Wang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include large-scale web data mining, visual semantic analysis, and machine learning.

Qingming Huang (F'18) received the B.S. degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Professor with the University of Chinese Academy of Sciences, Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 400 academic papers in international journals and conferences. His research interests include multimedia content analysis, image processing, computer vision, pattern recognition, and machine learning.