MULTI-VIEW FUSION THROUGH CROSS-MODAL RETRIEVAL

Limeng Cui[1] *Zhensong Chen*[2] *Jiawei Zhang*[3] *Lifang He*[4] *Yong Shi*[2] *Philip S. Yu*[5]

[1] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China
[2] School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China
[3] IFM Lab, Department of Computer Science, Florida State University, FL, USA
[4] Weill Cornell Medicine, Cornell University, NY, USA
[5] Department of Computer Science, University of Illinois at Chicago, IL, USA

ABSTRACT

Cross-modal retrieval, which takes text queries to retrieve relevant images or vice versa, has drawn much attention in recent years. This topic exhibits dual-heterogeneity: heterogeneity of different modalities and heterogeneous features obtained from multiple views. To address this issue, we propose an effective multi-view fusion method for cross-modal retrieval based on tensor modeling (CMTM) for cross-modal retrieval from the full-order feature interactions within the multimodal data. In order to facilitate integration of heterogeneous features from multiple views, we adopt the tensor structure to model the full-order interactions among the multi-view features effectively. Besides, a tensor factorization is applied to derive model parameters. Extensive experiments demonstrate the effectiveness of CMTM on cross-modal retrieval.

Index Terms— Cross modal retrieval, tensor modeling, multi-view learning

1. INTRODUCTION

In a typical cross-modal retrieval task, each type of data is treated as a single view, by using either deep model or shallow model. However, as multimedia data are often characterized by multiple types of descriptors, each of which describes certain aspects of object features. For example, handcrafted features and deep-learnt features characterize the different aspects of image data [1, 2, 3]. Similarly, explicit and latent features play different roles for text data characterization. Simply concatenating the multi-view features may result in that dense view dominate the feature space and potentially override the effect of sparse view. In this paper, we focus on the multi-view cross-modal retrieval problem, which benefits from fusing multiple views, and provide a more comprehensive information for understanding information entities.

The key challenge to this research is how to best reduce the heterogeneity among the different views. Since data are usually available in multiple views from a variety of feature subsets, each view has different statistical properties—for exmaple, the topic model vector of text is inherently dense, while representation of TF-IDF is naturally sparse. This makes it not applicable to apply algorithms designed for single-view data. Thus it is important to model the interactions/correlations between different views, wherein complementary information is contained.

Currently, many cross-modal methods have been proposed. Hashing methods can compress high-dimensional data into compact binary codes with similar binary codes for similar objects [4, 5]. Subspace learning is to find a common latent space in which the different modal features can be matched to each other [6, 7]. As different views can provide complementary information, methods based on multi-view learning have been proposed successively [8, 9]. But these approaches fail to fully explore the interactions between features across multiple views. In addition, label of paired data such as class and tag, which is related to each other through shared multimodal features, can be very helpful. Thus, a unified method which considers both information across different modalities and views needs to be investigated.

In order to overcome the issues mentioned above, we propose a novel cross-modal retrieval method based on tensor modeling, called CMTM, which considers the abundant interactions between features from different modalities and views. CMTM incorporates complementary features to characterize data and to take advantage of the shared information across different modalities. Specifically, we model the full-order interactions (dyadic, triadic, tetradic, and higher) among multiple labels and multiple views as a tensor structure, by taking the outer product of their respective feature spaces. A factorization is applied to prevent overfitting and deal with sparse data effectively. Then, we apply the alternating block coordinate descent method to optimize the objective function. We evaluate the performance of CMTM on four datasets.

2. PROPOSED METHOD

Assuming the problem is associated with training data $\mathcal{D} = \{((\mathbf{x_I}_1, \mathbf{x_T}_1), y_1), \cdots, ((\mathbf{x_I}_N, \mathbf{x_T}_N), y_N)\}$, a collection of images and corresponding text, where y_i is the label and N

is the number of samples. Each image $\mathbf{x_I}_i$ has representations in V_1 different views $\mathbf{x_I}_i = (\mathbf{x_I}_i^{(1)}, \cdots, \mathbf{x_I}_i^{(V_1)})$, where $\mathbf{x_I}_i^{(v_1)} \in \mathbb{R}^{d_{v_1}}$ is the image feature vector for the v_1-th view, and d_{v_1} is the dimensionality of the v_1-th view. Similarly, each text $\mathbf{x_T}_i$ is represented as $\mathbf{x_T}_i = (\mathbf{x_T}_i^{(1)}, \cdots, \mathbf{x_T}_i^{(V_2)})$, where $\mathbf{x_T}_i^{(v_2)} \in \mathbb{R}^{d_{v_2}}$ is the image feature vector for the v_2-th view with dimensionality d_{v_2}. The cross-modal retrieval problem aims at building a function $f : \mathcal{X}_I \rightarrow \mathcal{X}_T$ (similar for $f : \mathcal{X}_T \rightarrow \mathcal{X}_I$) using the image-text pairs $\{(\mathbf{x_I}_i, \mathbf{x_T}_i)\} \in \mathcal{X}_I \times \mathcal{X}_T$ as well as leveraging the complementary among different views.

To solve the cross-modal retrieval problem, it is straightforward to concatenate features from different views. However, transforming the multi-view data into a single-view data would fail to leverage the correlations between different views, which can provide complementary information. Although some kernel-based methods can utilize the high order interactions, they fail to explore the explicit correlations between features across multiple views. In the following, we introduce a framework for cross-modal retrieval, which intrinsically models the interactions in multimodal data among multiple views and different labels as a tensor structure.

2.1. Proposed Method

The data from each modal are available in multiple views from a variety of sources or feature subsets. Thus, we consider the instances from each modal are multi-view data. That is, $\mathbf{x}_I = (\mathbf{x}_I^{(1)}, \cdots, \mathbf{x}_I^{(V_1)}) = \{\mathbf{x}_I^{(v_1)}\}$, $\mathbf{x}_T = (\mathbf{x}_T^{(1)}, \cdots, \mathbf{x}_T^{(V_2)}) = \{\mathbf{x}_T^{(v_2)}\}$, where V_1 is the number of image views and V_2 is the number of text views.

Without losing generality, for a single view input vector $\mathbf{x}$ from label p, the linear model is given by $f_p(\mathbf{x}) = \mathbf{x}^T\mathbf{w}_p = \langle \mathbf{w}_p, \mathbf{x} \rangle$, where $\mathbf{w}_p$ is the label specific weight vector. We can extend this linear model to the fusion problem of multi-view data $\{\mathbf{x}^{(v)}\}_{v=1}^V$. Here, we consider fusing all interactions up to the full-order between V views.

Let $\mathbf{x}^{(v)} \in \mathbb{R}^{d_v}$ denote the the input multi-view data, where d_v is the dimensionality of the v-th view and V is the number of view. Similarly, denote $f_p(\{\mathbf{x}^{(v)}\}_{v=1}^V) = \langle \mathcal{W}, \mathcal{Z}_p \rangle$, where $\mathcal{Z}_p = \mathbf{z}^{(1)} \circ \mathbf{z}^{(2)} \circ \cdots \circ \mathbf{z}^{(V)} \circ \mathbf{e}_p$ is the full-order tensor. The *multi-view data fusion* function can be represented as

$$f_p(\{\mathbf{x}^{(v)}\}_{v=1}^V) = \sum_{s=1}^{P} \sum_{i_1=0}^{d_1} \cdots \sum_{i_V=0}^{d_V} w_{i_1,\cdots,i_V,s}(e_{p,s}\Pi_{v=1}^V \mathbf{z}_{i_v}^{(v)}) \quad (1)$$

where $\mathbf{z}^{(v)} = [1; \mathbf{x}^{(v)}]$, $\mathbf{e}_p = [0, \cdots, 0, 1, 0, \cdots, 0]^T \in \mathbb{R}^P$ (where P denotes the number of label) with only the p-th element is 1 to indicate the label, and $\mathcal{W} = \{w_{i_1,\cdots,i_V,s}\}$ is the weight tensor to be learned. It is worth noting that $w_{i_1,\cdots,i_V,s}$ with some indexes satisfying $i_v = 0$ encodes lower-order interactions among views whose indexes satisfy $i_v > 0$.

However, the dimensionality of parameter tensor $\mathcal{W}$ is normally very high, which needs to be reduced for the sake of less computational cost and avoiding overfitting. Hence, based on the CP factorization [10], the $\mathcal{W}$ can be factorized into R factors: $\mathcal{W} = [\![\mathbf{\Theta}^{(1)}, \cdots, \mathbf{\Theta}^{(V)}, \mathbf{\Phi}]\!]$, where the factor matrix $\mathbf{\Theta}^{(v)} \in \mathbb{R}^{(d_v+1)\times R}$ is the shared structure matrix for the v-th view and the p-th row $\boldsymbol{\phi}_p$ of $\mathbf{\Phi}$ is the specific weight vector for the data from label p. Based on the above factorization representation, we rewrite the *multi-view data fusion function* as

$$\begin{aligned} & f_p(\{\mathbf{x}^{(v)}\}_{v=1}^V) \\ & = \sum_{s=1}^{P} \sum_{i_1=0}^{d_1} \cdots \sum_{i_V=0}^{d_V} (\sum_{r=1}^{R} \phi_{s,r} \Pi_{v=1}^V \boldsymbol{\theta}_{i_v,r}^{(v)})(e_{p,s}\Pi_{v=1}^V \mathbf{z}_{i_v}^{(v)}) \\ & = \sum_{r=1}^{R} \langle \boldsymbol{\theta}_r^{(1)} \circ \cdots \circ \boldsymbol{\theta}_r^{(V)} \circ \boldsymbol{\phi}_{:,r}, \mathbf{z}^{(1)} \circ \cdots \circ \mathbf{z}^{(V)} \circ \mathbf{e}_p \rangle \end{aligned} \quad (2)$$

Since $e_{p,s} = 1$ only when $p = s$, we have

$$f_p(\{\mathbf{x}^{(v)}\}_{v=1}^V) = \boldsymbol{\phi}_p \Pi_{v=1}^V * (\mathbf{z}^{(v)T}\mathbf{\Theta}^{(v)})^T \quad (3)$$

where $*$ is the Hadamard product. It should be noted that the first row of $\mathbf{\Theta}^{(v)}$ is always associated with $z_0^{(0)} = 1$ and represents the bias factors of the v-th view. Through the bias factors, the lower-order interactions are explored in the predictive function.

Considering that multi-view features have their own distinctive contributions, we add term $\mathbf{x}^T\mathbf{u}_p$ into the predictive function in Eq. (3), where $\mathbf{x}$ is the concatenated feature vector from multiple views and $\mathbf{u}_p$ is the label-specific weight:

$$f_p(\{\mathbf{x}^{(v)}\}_{v=1}^V) = \mathbf{x}^T\mathbf{u}_p + \boldsymbol{\phi}_p \Pi_{v=1}^V * (\mathbf{z}^{(v)T}\mathbf{\Theta}^{(v)})^T \quad (4)$$

2.2. Multi-view Cross-Modal Retrieval

Above as we have discussed the linear model for a single modal with multi-view data, then we will extend it to cross-modal retrieval learning model based on Eq. (4). The objective function for a given multi-view image-text pair $(\{\mathbf{I}_p^{(v_1)}\}, \{\mathbf{T}_p^{(v_2)}\})$ from label p is shown by

$$F_p(\{\mathbf{I}_p^{(v_1)}\}, \{\mathbf{T}_p^{(v_2)}\}) = \alpha_p f_p(\{\mathbf{x}_{p,I}^{(v_1)}\}) + (1-\alpha_p) f_p(\{\mathbf{x}_{p,T}^{(v_2)}\}) \quad (5)$$

where α_p is the inter-modal label-specific weight that used to trade off the influence between two modals.

Let $\boldsymbol{\pi}_{p,I} = \Pi_{v_1=1}^{V_1} * (\alpha_p \mathbf{z}_{p,I}^{(v_1)T}\mathbf{\Theta}_I^{(v_1)})^T$ and $\boldsymbol{\pi}_{p,T} = \Pi_{v_2=1}^{V_2} * ((1-\alpha_p)\mathbf{z}_{p,T}^{(v_2)T}\mathbf{\Theta}_T^{(v_2)})^T$ for convenience. Then, according to Eq. (5), we get the final predictive function for cross-modal retrieval as follows:

$$F_p(\{\mathbf{I}_p^{(v_1)}\}, \{\mathbf{T}_p^{(v_2)}\}) = \mathbf{x}_p^T\mathbf{u}_p + \boldsymbol{\phi}_p \boldsymbol{\pi}_p \quad (6)$$

where $\mathbf{x}_p = [\alpha_p \mathbf{x}_{p,I}; (1-\alpha_p)\mathbf{x}_{p,T}]$, $\mathbf{u}_p = [\mathbf{u}_{p,I}; \mathbf{u}_{p,T}]$, $\boldsymbol{\phi}_p = [\boldsymbol{\phi}_{p,I}, \boldsymbol{\phi}_{p,T}]$ and $\boldsymbol{\pi}_p = [\boldsymbol{\pi}_{p,I}; \boldsymbol{\pi}_{p,T}]$.

We name this model as **C**ross-**M**odal using **T**ensor **M**odeling (CMTM). Clearly, the parameters of the interactions among different labels and multiple views are jointly factorized. Since the dependencies exist when the interactions share the same labels or features, the joint factorization benefits parameter estimation under sparsity. Therefore, the model parameters can be effectively learned without direct observations of such interactions especially in highly sparse data. Further, there is no need to construct the input tensor physically since the weight tensor $\mathcal{W}$ is factorized. Moreover, the model complexity is $O(R(V_1+V_2+d_U+P)+\sum_{p=1}^{P} N_p^f)$, where $N_p^f = N_p^I + N_p^T$, and N_p^I and N_p^T are the number of image features and text features in the p-th label respectively. It is linear in the number of parameters, which can help save memory and also speed up the learning procedure.

Then, we can write the optimization model as follows:

$$\begin{aligned}\min \mathcal{R} = \ & \sum_{p=1}^{P} \mathcal{L}_p(F_p(\{\mathbf{I}_p^{(v_1)}\}, \{\mathbf{T}_p^{(v_2)}\}), \mathbf{y}_p) \\ & + \lambda\Omega_\lambda(\mathbf{\Phi}, \{\mathbf{\Theta}_I^{(v_1)}\}, \{\mathbf{\Theta}_T^{(v_2)}\}) + \gamma\Omega_\gamma(\mathbf{U})\end{aligned} \tag{7}$$

where $\mathcal{L}_p$ is the prescribed loss function, $\{\mathbf{\Theta}_I^{(v)}\}$, $\{\mathbf{\Theta}_T^{(v)}\}$, $\mathbf{U}$, $\mathbf{\Phi}$ can be obtained by solving the problem, λ and γ are parameters to be tuned. The regularization terms Ω_λ and Ω_γ can be Forbenius norm, $\ell_{2.1}$ norm, or others.

2.3. Optimization Procedure

The optimization problem stated in Eq. (7) is hard to be directly solved due to its non-convexity with all the parameters. Therefore, we apply the alternating block coordinate descent approach [11] to solve our model.

STEP 1: With the $\mathbf{U}$, $\mathbf{\Phi}$, $\boldsymbol{\alpha}$, and $\mathbf{\Theta}_T^{(v_2)}$ fixed, the minimization over $\mathbf{\Theta}_I^{(v_1)}$ is given by

$$\frac{\partial \mathcal{R}}{\partial \mathbf{\Theta}_I^{(v_1)}} = \sum_{p=1}^{P} \frac{\partial \mathcal{L}_p}{\partial F_p} \frac{\partial F_p}{\partial \mathbf{\Theta}_I^{(v_1)}} + \lambda \frac{\partial \Omega_\lambda(\mathbf{\Theta}_I^{(v_1)})}{\partial \mathbf{\Theta}_I^{(v_1)}} \tag{8}$$

where $\frac{\partial \mathcal{L}_p}{\partial F_p} = \frac{1}{N_p}[\frac{\partial \ell_{p,1}}{F_{p,1}}, \cdots, \frac{\partial \ell_{p,N_p}}{F_{p,N_p}}]^T \in \mathbb{R}^{N_p}$.

Let $\boldsymbol{\pi}_I^{(-v_1)} = \Pi_{v_1'=1, v_1' \neq v_1}^{V_1} * (\mathbf{z}_I^{(v_1')T} \mathbf{\Theta}_I^{(v_1')})^T \in \mathbb{R}^R$. Then, we have $\mathbf{\Pi}_{p,I}^{(-v_1)} = [\boldsymbol{\pi}_{I,1}^{(-v_1)}, \cdots, \boldsymbol{\pi}_{I,N_p}^{(-v_1)}]^T$. Therefore,

$$\frac{\partial \mathcal{L}_p}{\partial F_p} \frac{\partial F_p}{\partial \mathbf{\Theta}_I^{(v_1)}} = \alpha_p \mathbf{Z}_{p,I}^{(v_1)}((\frac{\partial \mathcal{L}_p}{\partial F_p} \boldsymbol{\phi}_{\boldsymbol{p},\boldsymbol{I}}) * \mathbf{\Pi}_{p,I}^{(-v_1)}) \tag{9}$$

Similarly, with the $\mathbf{U}$, $\mathbf{\Phi}$, $\boldsymbol{\alpha}$, and $\mathbf{\Theta}_I^{(v_1)}$ fixed, we can minimize $\mathbf{\Theta}_T^{(v_2)}$ through

$$\frac{\partial \mathcal{L}_p}{\partial F_p} \frac{\partial F_p}{\partial \mathbf{\Theta}_T^{(v_2)}} = (1-\alpha_p) \mathbf{Z}_{p,T}^{(v_2)}((\frac{\partial \mathcal{L}_p}{\partial F_p} \boldsymbol{\phi}_{\boldsymbol{p},\boldsymbol{T}}) * \mathbf{\Pi}_{p,T}^{(-v_2)}) \tag{10}$$

where $\mathbf{\Pi}_{p,T}^{(-v_2)} = [\boldsymbol{\pi}_{T,1}^{(-v_2)}, \cdots, \boldsymbol{\pi}_{T,N_p}^{(-v_2)}]^T$.

Table 1. Mean Average Precision (MAP) and Precision@100 (P@100) for task $I \rightarrow T$ on four datasets

Method	Wiki		NUS-WIDE		MIRFlickr		Pascal VOC	
	MAP	P@100	MAP	P@100	MAP	P@100	MAP	P@100
JRL	**0.3387**	0.2558	0.5499	0.5018	0.5842	0.6041	0.2270	0.2886
SMFH	0.2653	0.2168	0.5996	0.4673	0.6123	0.6113	0.3063	0.3033
CMFH	0.2208	0.2492	0.4491	0.4764	0.5643	0.6257	0.3135	0.3287
LSSH	0.1497	0.2078	0.4129	0.4358	0.5610	0.6324	0.4477	0.4225
SCM_orth	0.1331	0.1349	0.6903	0.6010	0.5789	0.6583	0.4040	0.4058
SCM_seq	0.2459	0.2276	**0.7107**	0.6834	0.6234	0.6045	0.4868	0.4517
SePH	0.2891	0.2633	0.5687	0.5398	0.6783	0.7071	0.4780	0.4277
DCMH	0.3064	0.2875	0.6824	0.7012	0.6768	0.7513	0.4896	0.4813
CMTM	0.2927	**0.3184**	0.7093	**0.7448**	**0.6921**	**0.7727**	**0.5194**	**0.5072**

Table 2. Mean Average Precision (MAP) and Precision@100 (P@100) for task $T \rightarrow I$ on four datasets

Method	Wiki		NUS-WIDE		MIRFlickr		Pascal VOC	
	MAP	P@100	MAP	P@100	MAP	P@100	MAP	P@100
JRL	0.2499	0.2536	0.5132	0.5210	0.6074	0.5708	0.2464	0.1942
SMFH	0.6136	0.5172	0.5574	0.4958	0.6213	0.5856	0.3086	0.3012
CMFH	0.5484	0.5213	0.4726	0.3787	0.5724	0.5409	0.3156	0.2953
LSSH	0.2719	0.2360	0.5231	0.5342	0.5791	0.5543	0.4962	0.4721
SCM_orth	0.1393	0.1272	0.6888	0.5263	0.5816	0.5622	0.4526	0.4866
SCM_seq	0.2410	0.2045	0.7406	0.6872	0.6345	0.6134	0.5455	0.5280
SePH	0.6421	0.5871	0.6873	0.6342	0.7271	0.6583	0.5826	0.5439
DCMH	0.6424	0.6082	0.7234	0.7519	0.7433	0.7284	0.6019	0.5316
CMTM	**0.6587**	**0.6215**	**0.7611**	**0.7767**	**0.7567**	**0.7553**	**0.6074**	**0.5464**

STEP 2: With all the $\mathbf{U}$, $\boldsymbol{\alpha}$, $\mathbf{\Theta}_I^{(v_1)}$, and $\mathbf{\Theta}_T^{(v_2)}$ fixed, we have

$$\frac{\partial \mathcal{R}}{\partial \mathbf{\Phi}} = [(\frac{\partial \mathcal{L}_1}{\partial F_1})^T \mathbf{\Pi}_1; \cdots; (\frac{\partial \mathcal{L}_P}{\partial F_P})^T \mathbf{\Pi}_P] + \lambda \frac{\partial \Omega_\lambda(\mathbf{\Phi})}{\partial \mathbf{\Phi}} \tag{11}$$

where $\mathbf{\Pi}_p = [\boldsymbol{\pi}_{p,1}, \cdots, \boldsymbol{\pi}_{p,N_p}]^T, \forall p \in [1:P]$.

STEP 3: With all the $\mathbf{\Phi}$, $\boldsymbol{\alpha}$, $\mathbf{\Theta}_I^{(v_1)}$, and $\mathbf{\Theta}_T^{(v_2)}$ fixed, the partial derivative of $\mathcal{R}$ w.r.t. $\mathbf{U}$ is given by

$$\frac{\partial \mathcal{R}}{\partial \mathbf{U}} = [\mathbf{X}_1 \frac{\partial \mathcal{L}_1}{\partial F_1}, \cdots, \mathbf{X}_P \frac{\partial \mathcal{L}_P}{\partial F_P}] + \gamma \frac{\partial \Omega_\gamma(\mathbf{U})}{\partial \mathbf{U}} \tag{12}$$

where $\mathbf{X}_p = [X_{p,I}; X_{p,T}]$ is the concatenated feature.

STEP 4: When we obtain the $\mathbf{U}$, $\mathbf{\Phi}$, $\mathbf{\Theta}_I^{(v_1)}$, and $\mathbf{\Theta}_T^{(v_2)}$, the partial derivative of $\mathcal{R}$ w.r.t. $\boldsymbol{\alpha}$ is given by

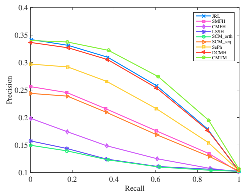
$$\frac{\partial \mathcal{R}}{\partial \boldsymbol{\alpha}} = [(\frac{\partial \mathcal{L}_1}{\partial F_1})^T \mathbf{\Delta}_1, \cdots, (\frac{\partial \mathcal{L}_P}{\partial F_P})^T \mathbf{\Delta}_P)] \tag{13}$$

where $\mathbf{\Delta}_p = \mathbf{F}_{p,I} - \mathbf{F}_{p,T}, \forall p \in [1:P]$ and $\mathbf{\Delta}_p \in \mathbb{R}^{N_p}$.

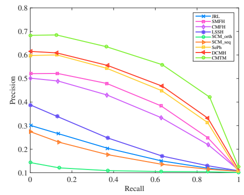
3. EXPERIMENTS

We conduct extensive experiments to evaluate the efficacy of the proposed model with several state-of-the-art cross-modal retrieval methods on four widely-used benchmark datasets.

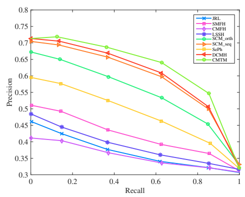
Wiki dataset is collected from Wikipedia consisting of 2,866 multimedia documents. Totally 10 categories are considered in this dataset and each image-text pair is labeled by one of them. Documents are considered to be similar if they belong to the same category. **NUS-WIDE** dataset [12] contains 81 concepts, which can be regarded as labels. Each



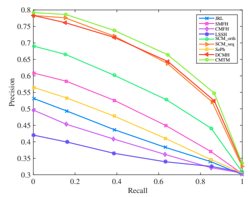
(a) Wiki (I→T)



(b) Wiki (T→I)



(c) NUS-WIDE (I→T)



(d) NUS-WIDE (T→I)

Fig. 1. Precision-recall curves of cross-modal retrieval on Wiki and NUS-WIDE.

image-text pair is annotated by at least one concept. Pairs are considered to be similar if they share at least one concept. We select 1000 most frequently used tags from 5,018 unique tags in this dataset. **MIRFlickr** [13] originally contains 25,000 instances collected from Flickr, each being an image with its associated textual tags. Each instance is manually annotated with some of 24 provided unique labels. For each instance, the image view is represented with a 150-D edge histogram and the text view as a 500-D feature vector derived from PCA on its binary tagging vector. **Pascal VOC** [14] consists of 5011/4952 (training/ testing) image-tag pairs, which are categorized into 20 different classes. Since some images are multi-labeled, researchers usually select images with only one object as the way in [15], resulting in 1865 training and 1905 testing data. The image features include histograms of bag-of-visual-words, GIST and color. The text features are 399-D tag features.

Evaluation Protocols: For Wiki dataset, each image is represented by a 128-D SIFT histogram and a 1000-D CNN feature generated by Alexnet [16], which is pretrained on ImageNet. Each text is represented by a 500-D bag-of-words feature and a 10-D topics vector generated by Latent Dirichlet Allocation (LDA) model [17]. For other datasets, each image is also represented by both shallow and deep feature.

For our method, the dimension of latent factors $R = 20$, the maximum number of iteration is set as 200. Grid search is applied to select optimal values for each regularization hyperparameter from $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ and $\gamma \in \{10^{-3}, 10^{-2}, \cdots, 10^{4}\}$. For the dataset we compose training set, validation set and testing set for each label. Since these four datasets have been divided into training set and testing set, we randomly select 20% of samples from testing set for each task as validation set. Validation sets are used for hyperparameter tuning for our method, and each of the validation and testing sets does not overlap with any other set so as to ensure the sanity of our experiments.

Compared Methods: We compare the performance of our approach CMTM with seven state-of-the-art cross-modal methods, including two subspace learning methods **JRL** [18] and **SMFH** [7], and five hashing methods **CMFH** [19], **LSSH** [5], **SCM** [20], **SePH** [21] and **DCMH** [22] .

Follow [20, 23], we evaluate the retrieval performance based on three metrics: Mean Average Precision (MAP), Precision@100 (P@100) and precision-recall curves of two cross-modal retrieval tasks: image query on text database (I $\rightarrow$ T) and text query on image database (T $\rightarrow$ I). The code length is 32 bits for hashing based methods.

Quantitative Results: Table 1 and Table 2 show the MAP and P@100 results of all the comparison method on Wiki, NUS-WIDE, MIRFlickr and Pascal VOC datasets. The best results are presented in bold figure and the second best results are marked by underline. We can observe that the proposed CMTM method substantially outperforms all state-of-the-art methods for two cross-modal tasks on almost all the benchmarks datasets, which demonstrates its effectiveness. An interesting observation is that our method performs better than deep method DCMH. We assume that our model can use both hand-crafted features and deep-learned features from multiple views and exploit the complex feature correlations effectively.

The precision-recall curves for the two cross-modal tasks I $\rightarrow$ T and T $\rightarrow$ I are shown in Figure 1 respectively. As it is shown, the proposed CMTM method achieves the best performance at almost all recall levels for both T $\rightarrow$ I and I$\rightarrow$ T tasks on both dataset, except when recall is close to zero.

4. CONCLUSION

In this paper, we present a novel tensor modeling based method for cross-modal retrieval (CMTM), which can learn the shared structure across the different modalities and model the full-order interactions among different features obtained in multiple views. Our method builds upon multi-view features and models the correlations across them as a tensor structure. Moreover, the labels of paired data are embedded within the multimodal interactions. Experimental results prove the effectiveness of our method.

Source codes will be publicly available.

5. ACKNOWLEDGMENT

The work is supported by the National Natural Science Foundation of China under Grant No.: 61672313, 61503253, 91546201 and 71331005, the National Science Foundation under Grant No.: IIS-1526499 and CNS-1626432, and Natural Science Foundation of Guangdong Province under Grant No.: 2017A030313339.

6. REFERENCES

[1] Lu Jin, Shenghua Gao, Zechao Li, and Jinhui Tang, "Hand-crafted features or machine learnt features? together they improve rgb-d object recognition," in *2014 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2014, pp. 311–319.

[2] Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud, and Jean-Luc Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," in *Multimedia*. ACM, 2015, pp. 1263–1266.

[3] Jinhui Tang, Lu Jin, Zechao Li, and Shenghua Gao, "Rgb-d object recognition via incorporating latent data structure and prior knowledge," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1899–1908, 2015.

[4] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Multimedia*. ACM, 2013, pp. 143–152.

[5] Jile Zhou, Guiguang Ding, and Yuchen Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 415–424.

[6] Hong Liu, Rongrong Ji, Yongjian Wu, and Gang Hua, "Supervised matrix factorization for cross-modality hashing," pp. 1767–1773, 2016.

[7] Jun Tang, Ke Wang, and Ling Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3157–3166, 2016.

[8] Saehoon Kim, Yoonseop Kang, and Seungjin Choi, "Sequential spectral learning to hash with multiple representations," *ECCV*, pp. 538–551, 2012.

[9] Li Liu, Mengyang Yu, and Ling Shao, "Multiview alignment hashing for efficient image search," *IEEE Transactions on image processing*, vol. 24, no. 3, pp. 956–966, 2015.

[10] Tamara G Kolda and Brett W Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[11] Yangyang Xu and Wotao Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[12] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*. ACM, 2009, p. 48.

[13] Mark J Huiskes and Michael S Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 39–43.

[14] Sung Ju Hwang and Kristen Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 6, pp. 1145–1158, 2012.

[15] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs, "Generalized multiview analysis: A discriminative latent space," in *CVPR*, 2012, pp. 2160–2167.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Multimedia*. ACM, 2010, pp. 251–260.

[18] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2014.

[19] Guiguang Ding, Yuchen Guo, and Jile Zhou, "Collective matrix factorization hashing for multimodal data," in *CVPR*, 2014, pp. 2075–2082.

[20] Dongqing Zhang and Wu-Jun Li, "Large-scale supervised multimodal hashing with semantic correlation maximization.," in *AAAI*, 2014, vol. 1, p. 7.

[21] Zijia Lin, Guiguang Ding, Jungong Han, and Jianmin Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE transactions on cybernetics*, vol. 47, no. 12, pp. 4342–4355, 2017.

[22] Qing-Yuan Jiang and Wu-Jun Li, "Deep cross-modal hashing," in *CVPR*, 2017, pp. 3270–3278.

[23] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang, "Semantics-preserving hashing for cross-view retrieval," in *CVPR*, 2015, pp. 3864–3872.