Neurocomputing 168 (2015) 747–760

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Cluster-sensitive Structured Correlation Analysis for Web cross-modal retrieval

Shuhui Wang [a,*], Fuzhen Zhuang [a], Shuqiang Jiang [a], Qingming Huang [a,b], Qi Tian [c]

[a] Key Lab of Intellectual Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Department of Computer Science, University of Texas at San Antonio, TX 78249, USA

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form
23 March 2015
Accepted 12 May 2015
Communicated by Jinhui Tang
Available online 29 May 2015

Keywords:
Correlation learning
Cluster-sensitive
Structured correlation model
Correspondence missing

ABSTRACT

Modern cross-modal retrieving technology is required to find semantically relevant content from heterogeneous modalities. As previous studies construct unified dense correlation models on small scale cross-modal data, they are not capable of processing large scale Web data, because (a) the content of Web cross media is divergent; (b) the topic sensitive structure information in the high dimensional space is neglected; and (c) data should be organized as strictly corresponding pairs, which is not satisfied in real world scenarios. To address these challenges, we propose a cluster-sensitive cross-modal correlation learning framework. First, a set of cluster-sensitive correlation sub-models are learned instead of a unified correlation model, which better fits the content divergence in different modalities. We impose structured sparsity regularization on the projection vectors to learn a set of interpretable structured sparse correlation sub-models. Second, to compensate for the correspondence missing, we take full advantage of both intra-modal affinity and inter-modal co-occurrence. The projected coordinates of adjacent data within a modality tend to be similar, and the inconsistency of cluster-sensitive projection is minimized. The learned correlation model adapts to the content divergence and thus achieves better model generality and bias–variance trade-off. Extensive experiments on two large scale cross-modal data demonstrate the effectiveness of our approach.

1. Introduction

Millions of Web users produce diverse online content of multiple modalities everyday, e.g., textual documents and visual images. Instead of single media, knowledge is delivered by different modalities with rich context and structure information, which is known as *cross media* [1–3]. In this new information carrier, Web topics and events are described by semantically related documents from different modalities, providing complementary explanations from different aspects. For instance, the concept "tiger" can be described by a tiger head in an image, and textual description of the life of a tiger. On one side, Web users need to retrieve content of heterogeneous modalities. On the other side, a user-centric retrieving system should support more flexible query input and more versatile data retrieving. Therefore, it has become a very interesting yet challenging problem to develop effective cross-modal retrieving models for *cross media*.

As a well-established paradigm for modeling the cross-modal correlation, the low dimensional subspaces maximizing the correlation between two modalities can be learned by using canonical correlation analysis (CCA) [4] and partial least square (PLS) [5]. However, as much effort devoted to improving the correlation models [6–10], they are not capable of learning the correlation among cross-modal data from the Web. In general, the main technical challenges for developing robust correlation models can be analyzed from several aspects.

First, the topic and content distribution for cross media is complex and divergent. The research challenge is two folds:

- *Intra-modal divergence*: Given a topic or concept, the related documents are divergent within one modality. For example, the concept "*Apple*" may be related to content from multiple domains such as *food*, *plant*, *art*, *industry* and *hi-tech*, see Fig. 1. The intra-modal divergence poses difficulties in representing the wide range of content genres with a unified subspace.
- *Inter-modal divergence*: The physical structures are drastically different among features from different modalities. They are also drastically different among multiple features from one modality, as shown in the bottom part of Fig. 1. Therefore, it is hard to find the subspaces to directly calculate similarities among data from different modalities.

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.05.049
0925-2312/

In previous study on correlation learning [6,8,11–13], the unified subspace learning is the well-studied paradigm assuming the prior of projection function parameters that are *Gaussian* or *Laplacian*. They are not flexible in dealing with the content divergence in Web data. The intermediate shared latent topic spaces are learned with various probabilistic graphical models [14–16] to tackle the content divergence problem, while they suffer from the high computational cost of parameter inference. As another possible solution, localized approach [17] tends to achieve superfine correlation model with much more model parameters, but it is too sensitive to the ubiquitously existing noise.

The content divergence can also be observed on the high dimensional structured representation. For example, the vocabularies and writing styles (i.e., word frequencies and orders) of different textual documents are diversified, and the images can be represented by complementary visual features, such as color, texture, shape and Bag-of-Visual-Words. Intuitively, the importance of different feature dimensions should be topic dependent. For instance, the words (*black*, *white*) in textual representation have close relationship with color histogram in visual representation on "Apple products" related documents. However, the words (*chunk*, *leaf*) will be more related with the visual texture features on images describing "apple tree". Unfortunately, such topic specific relation cannot be well captured by global correlation models, even with complicated structured input and output regularization [11,18,19].

Another critical issue for correlation learning on Web data is correspondence missing. Specifically, there may be no explicit corresponding cross-modal documents. For example, on Wikipedia, there are many pages with only textual content but no images, and there are certain amount of textual paragraphs without a corresponding image description. However, the potential complementary cross-modal descriptions of these textual paragraphs may be found in other Web data corpus, e.g., social media photos. The correspondence missing is similar to the setting of semi-supervised learning, where a certain level of label information is assumed to be missing.

The correspondence information is usually supposed to be fully provided in existing study [6,10,20,21]. The one-to-one alignment of multiple modalities is enforced in the correlation learning objectives. This assumption is overly strict which makes the correlation models too sensitive to the noise and vacancy in the correspondence information. Introducing both intra-modal similarity and side information provides a good remedy for correspondence missing [16,22,23], but the potential power has not been fully released by existing global subspace learning strategies, and may only result in an over-smooth correlation models instead.

We address the challenges of content divergence and correspondence missing, and propose a new correlation learning approach

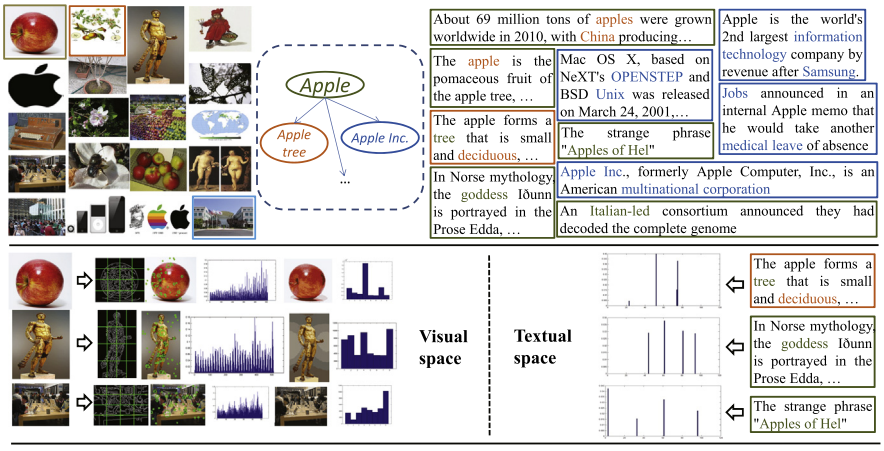

Fig. 1. The challenge of content divergence. In the upper part, given a keyword "*Apple*", there are multiple related topics such as "*Apple tree*" and "*Apple Inc.*", marked with different colored bounding boxes. In the bottom part, features with significant structure difference extracted from both modalities are shown.

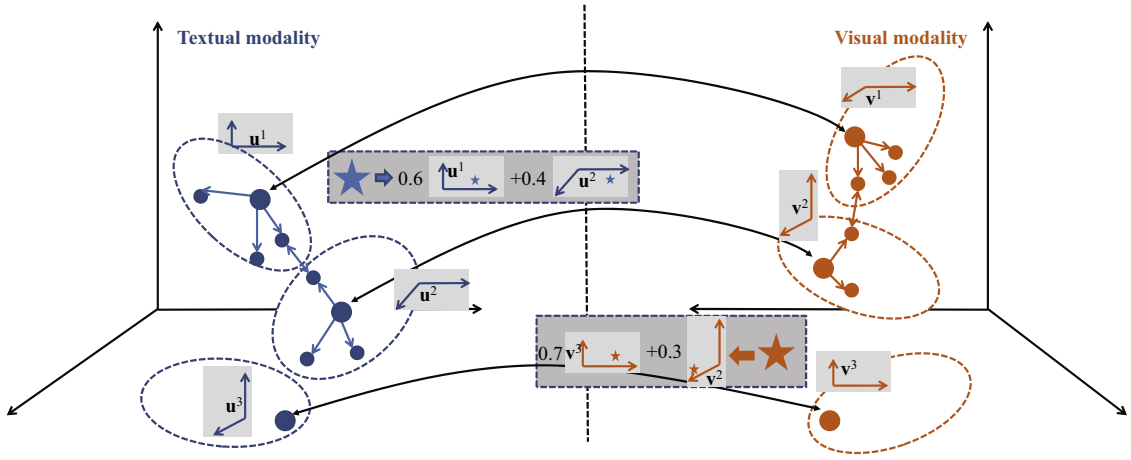

Fig. 2. The proposed correlation learning framework. The dashed ellipses denote data clusters in both modalities. The black double arrow lines represent the cross-modal correspondence information, which is propagated through the adjacent data, as shown by the arrow lines. The learned cluster-sensitive subspaces in gray color are illustrated at every ellipse (cluster). The new documents (stars) are projected into the new representation based on the aggregated projection.

based on an ensemble of multiple cross-modal sub-models, as shown in Fig. 2. First, we model the cross-modal correlation at the sub-topic level,[1] where the sub-topic structure is reflected by the cluster distribution on each modality. The cluster-sensitive transformation is determined by both the transformation function on each cluster (i.e., sub-topic) and the membership between the documents and the clusters. Compared to existing approaches, our method achieves a smaller model bias than global projection learning [4], and a smaller model variance than localized projection learning [17]. Therefore, such a bias–variance trade-off leads to a smaller expected model error. To further deal with high dimensional multi-modal representation on diversified content genres, we apply sparsity [6,13] and structured sparsity constraints [11,12] on each sub-model. Consequently, we obtain a set of *interpretable* cross-modal subspaces where each dimension is the combination of part of the feature dimensions.

Second, to compensate for the correspondence missing, we take full advantage of both intra-modal affinity and inter-modal co-occurrence which has been shown to be equally important by [8,22,23]. By encoding the intra-modal relation, the correspondence information can be appropriately propagated to the neighboring data to make the transformed representation more semantically consistent. Moreover, our method achieves better model generality by penalizing the unsmooth projection brought by multiple sub-models using the intra-modal affinity. Therefore, the correspondence missing can be firmly alleviated, and the model robustness can be enhanced.

In summary, we construct a set of transformation sub-models where the number of projection sub-models for each data modality is equal to the number of clusters. Data from each modality can be projected by the weighted combination of sub-models. The new presentation maximizes the correlation of different modalities and measures the intra-modal relation in a more appropriate manner. The advantage over existing correlation model is that the learned transformation is topic sensitive, leading to better adaptability to topic divergence. Our approach is more robust to noise compared to localized correlation model. It can also be seen as a generalization of unified correlation models [4] and localized models [17]. The trade-off between model bias and variance can be well controlled by adjusting the number of clusters. The key technical contributions can be summarized as follows:

- We propose a new correlation subspace learning method for cross modality retrieval. It better fits the content divergence by learning a set of cluster-sensitive correlation sub-models. By applying structured sparsity regularization, the learned projection is more interpretable compared to dense correlation models. Our method achieves better trade-off between model bias and variance.
- By encoding the intra-modal information with correlation sub-models, our model is more robust to the correspondence missing than traditional approaches that only leverage the content co-occurrence.
- Extensive experiments on two large scale cross-modal datasets demonstrate the advantages of our approach. With moderate model training complexity, our method achieves at least 20% higher performance in Mean Average Precision than state-of-the-art approaches.

The rest of the paper is organized as follows. In Section 2 we briefly review related works. In Sections 3 and 4 we introduce our approach and implementation details. We provide description on experiments in Section 5, and conclude this paper in Section 6.

[1] Each sub-topic represents a data subset with similar visual and textual representation.

2. Related work

CCA [4] is the first study on how to seek optimal basic vectors for two sets of variables to model the multi-modal correlation. It is used in various problems, such as cross language analysis [24] and Socio-Economic Transition [9]. PLS [5] aims to find a linear regression model by projecting the predicted variables and the observable variables to a new space, which is equivalent with CCA in many situations [25]. Such models are further extended to a regularized correlation learning framework [26]. Multiple subspace learning (e.g., GPCA [27]) deals with the intra-modal divergence by estimating multiple candidate subspaces using polynomial function fitting. It invokes explosion of polynomial combination and computation burden on large scale high dimensional data. Sparse correlation analysis [7,6,13] and structured sparse correlation models [12,11] are proposed to learn sparse correlation subspace. CCA can be used as a complimentary preprocessing for other learning tasks. Based on the subspaces learned by CCA, Rasiwasia et al. [10] propose to learn cross-modal topic classifiers to measure the semantic divergence of Web data, and Wu et al. [1] construct a semantic distance measurement model. Gong and Lazebnik [28] develop a binary codes learning approach which leverages the label information with CCA.

The main challenge for cross media analysis with multiple modalities is how to understand the relations among different modalities and how to leverage the multi-modal information [1,2,16]. Many studies use words to label the visual objects in the images [29–31]. Since they simply consider this problem as object detection or recognition task, the rich content in the text has been ignored. Latent Dirichlet Allocation (LDA) [31] is extended to multi-modal learning [15,32] which learns the latent topics to model the uncertainty of correlations. Jia et al. [16] propose a Markov random field method over LDA topic models. It does not require one-to-one correspondence as in [15], therefore it is more flexible in processing Web data. Xiao and Stibor [33] link the images and sounds via words and tag information by Corr-LDA [15].

In the research of cross modality retrieval, Bronstein et al. [20] propose a boosting based hash code learning where the "weak coders" and their weights are jointly learned for calculating the cross-modal weighted Hamming distance. Masci et al. [22] extend [20] by taking both intra-modal side information and inter-modal correlation into consideration based on multi-layered neuro-network encoder. Zhen and Yeung [23] develop a latent binary embedding method which learns the latent topics and the binary weights to model the observed intra-modal and inter-modal similarities. Rafailidis et al. [34] develop a unified intra-modal similarity and inter-modal similarity construction construction framework for large scale multi-modal retrieval.

Correlation learning can also be cast into manifold alignment [3,8,35] which leverages local adjacency and global geometric information. Mao et al. [36] develop a parallel field alignment approach for cross media retrieval. Zhai et al. [17] propose a localized multi-view semi-supervised metric learning method, which learns one localized correlation model for each labeled and unlabeled training sample. However, the computational cost for [17] is prohibitive when processing large data. To achieve better generality for real world data, cluster specific subspace learning, as proposed in this paper, can be a good trade-off between localized subspace learning and global subspace learning.

3. Approach

3.1. Overview

We are given two sets of data from two modalities $\mathbf{X} \in \mathbb{R}^{N_x \times D_x}$ and $\mathbf{Y} \in \mathbb{R}^{N_y \times D_y}$, respectively, where N_x or N_y represents the number

of training samples for each modality, and D_x or D_y represents the number of dimensions. Typically, we assume that both $\mathbf{X}$ and $\mathbf{Y}$ are zero-centered and norm-bounded. Without loss of generality, we denote each sample row vector as $\mathbf{x}_i$ or $\mathbf{y}_j$, and $\mathbf{X}_S$ or $\mathbf{Y}_S$ as the sub-matrices where the row entry indices are included in set S. Consistently, we denote the row indices subset with correspondence in $\mathbf{X}$ and $\mathbf{Y}$ as $\mathbf{X}_L$ and $\mathbf{Y}_L$, and the row indices subset without correspondence in $\mathbf{X}$ and $\mathbf{Y}$ as $\mathbf{X}_{U_x}$ and $\mathbf{Y}_{U_y}$, respectively. Note that in our study the sizes of U_x and U_y do not have to be equal. In this section, we only consider the first pair of canonical cluster-sensitive vectors. For learning multiple vectors, i.e., multi-dimension subspace learning, we will introduce a method in Section 4.

The learning stage includes the following steps (as illustrated in Fig. 2):

Step 1: Conduct dimension reduction and clustering by graph K-Means [37] independently of each modality. Then we obtain P cluster and Q clusters for $\mathbf{X}$ and $\mathbf{Y}$, respectively. Each cluster corresponds to a data subset with strong semantic correlation. Data of each modality is, respectively, softly assigned with the cluster membership by calculating the probability belonging to a cluster. As a result, the probabilistic membership matrices $\mathbf{W} \in \mathbb{R}^{N_x \times P}$ and $\mathbf{R} \in \mathbb{R}^{N_y \times Q}$ are obtained for both modalities.

Step 2: Calculate the graph Laplacians Θ_x and Θ_y using the intra-modal similarity and dissimilarity information.

Step 3: Learn H dimensional cluster-sensitive subspaces $\{(\mathbf{U}_h, \mathbf{V}_h),\ h=1,\ldots,H\}$ based on our proposed model in Section 3.4.

Given the unknown data from either modality, they are projected into the new representation based on their probabilistic cluster memberships and the corresponded transformed coordinates on the related subspaces. We introduce the correlation learning model in the consequent sections.

3.2. Structured Correlation Analysis

The aim of correlation analysis [4,6,7,11,12] is to find projection basis vectors that maximize the correlation between data from two modalities. The generalized formulation can be written as

$$\begin{cases} \max\limits_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{X}_L^T \mathbf{Y}_L \mathbf{v} \\ \text{s.t. } P_x(\mathbf{u}) \le c_1, \quad P_y(\mathbf{v}) \le c_2, \quad \|\mathbf{u}\|^2 \le 1, \quad \|\mathbf{v}\|^2 \le 1 \end{cases} \tag{1}$$

where $\mathbf{u}$ and $\mathbf{v}$ denote the learned vectors to maximize the correlation of the two modalities with the constraints. In fact, the difference of several correlation analysis models is determined by different types of P_x and P_y. For example, when

$$P_x(\mathbf{u}) = \mathbf{u}^T \mathbf{X}_L^T \mathbf{X}_L \mathbf{u}, \quad P_y(\mathbf{v}) = \mathbf{v}^T \mathbf{Y}_L^T \mathbf{Y}_L \mathbf{v}, \quad c_1 = c_2 = 1, \tag{2}$$

we refer to Eq. (2) as CCA [4]. The model solution can be done by generalized eigen-space decomposition [25], and the learned $\mathbf{u}$ and $\mathbf{v}$ are dense because the model prior defined by Eq. (2) is *Gaussian*.

With the situation that the cross-modal data dimension $D_x \gg N_x$ and $D_y \gg N_y$, sparse correlation analysis (SCA) [6] can be applied. The constraints can be formulated as

$$P_x(\mathbf{u}) = \|\mathbf{u}\|_1, \quad P_y(\mathbf{v}) = \|\mathbf{v}\|_1 \tag{3}$$

where c_1 and c_2 control the sparsity of the learned projection functions. To ensure the feasibility of the model solution, the parameters c_1 and c_2 should satisfy $0 < c_1, c_2 \le 1$. The learned projection vectors are sparse so that the partial covariances between two modalities can be identified. Furthermore, if there are M feature groups existing in $\mathbf{Y}$ modality, one can impose structured sparse regularization [38] on $\mathbf{v}$, e.g., $P_y(\mathbf{v}) = \rho \sum_m \varpi_m \|\mathbf{v}(g^m)\|_2 + (1-\rho)\|\mathbf{v}\|_1, 0 \le \rho \le 1$, where g_m denotes the m-th group in the feature representation, and ϖ_m denotes the weight of m-th feature group.[2] Then the formulation of structured sparse correlation analysis (StSCA) can be written as the following form:

$$\begin{cases} \min\limits_{\mathbf{u},\mathbf{v}} -\mathbf{u}^T \mathbf{X}_L^T \mathbf{Y}_L \mathbf{v} + \frac{\tau}{2} \mathbf{v}^T \mathbf{v} + \theta \left(\rho \sum\limits_m \varpi_m \|\mathbf{v}(g^m)\|_2 + (1-\rho)\|\mathbf{v}\|_1 \right) \\ \text{s.t. } \|\mathbf{u}\|_2^2 \le 1, \quad \|\mathbf{v}\|_2^2 \le 1, \quad \|\mathbf{u}\|_1 \le c_1 \end{cases} \tag{4}$$

3.3. Modeling intra-modal relation

Generally, the intra-modal relation can be described by density, affinity and category information [39]. One feasible way is to construct the graph Laplacian $\mathbf{L}$ on the local affinity matrix as Semi-Supervised Kernel Correlation Analysis (SSKCA) [8]. With graph Laplacian, the model propagates the correlation along the data manifold and finds highly correlated directions that are also located on high variance directions along the data manifold. When there is category or label information on the data, one can encode such information into the graph Laplacian as [39]

$$\Theta = \mathbf{L} + (\mathbf{1} - \Delta) \otimes \mathbf{S} \tag{5}$$

where $\otimes$ denotes the Hadamard (element-wise) product and $\mathbf{S}$ denotes the local affinity matrix. The vector $\mathbf{1}$ denotes matrix with all 1's and its size equals to $\mathbf{L}$. The elements in indicator matrix Δ denote if there is a similar(1) or dissimilar(-1) edge between the i-th and the j-th samples. It enforces that the projection from different categories have opposite signs. According to the analysis in [39], Θ is positive semi-definite.

Consequently, the model with structured sparsity is formulated as

$$\begin{cases} \min\limits_{\mathbf{u},\mathbf{v}} \frac{1}{2} \|\mathbf{X}_L \mathbf{u} - \mathbf{Y}_L \mathbf{v}\|^2 + \frac{\lambda_1}{2} \mathbf{u}^T \mathbf{X}^T \Theta_x \mathbf{X} \mathbf{u} + \frac{\lambda_2}{2} \mathbf{v}^T \mathbf{Y}^T \Theta_y \mathbf{Y} \mathbf{v} \\ \quad + \theta_1 \|\mathbf{u}\|_1 + \theta_2 \left(\rho \sum\limits_m \varpi_m \|\mathbf{v}(g^m)\|_2 + (1-\rho)\|\mathbf{v}\|_1 \right) \\ \text{s.t. } \|\mathbf{u}\|_2^2 \le 1, \quad \|\mathbf{v}\|_2^2 \le 1 \end{cases} \tag{6}$$

where Θ_x and Θ_y denote the graph Laplacian of $\mathbf{X}$ and $\mathbf{Y}$, respectively. We denote the model in Eq. (6) as semi-supervised structured sparse correlation analysis (SSStSCA).

3.4. Cluster-sensitive correlation learning

Suppose we have divided $\mathbf{X}$ into P clusters and $\mathbf{Y}$ into Q clusters. Each $\mathbf{x}_i$ and $\mathbf{y}_j$ from the dataset have a probabilistic cluster membership vector:

$$\begin{aligned} &\mathbf{x}_i : [w_i^1, \ldots, w_i^p, \ldots, w_i^P], \quad \sum_{p=1}^{P} w_i^p = 1, \quad w_i^p \ge 0 \\ &\mathbf{y}_j : [r_j^1, \ldots, r_j^q, \ldots, r_j^Q], \quad \sum_{q=1}^{Q} r_j^q = 1, \quad r_j^q \ge 0 \end{aligned} \tag{7}$$

We refer to the whole cluster membership for $\mathbf{X}$ and $\mathbf{Y}$ as $\mathbf{W} \in \mathbb{R}^{N_x \times P}$ and $\mathbf{R} \in \mathbb{R}^{N_y \times Q}$, respectively, where the i-th row corresponds to the cluster membership vector for the i-th data. We need to learn the following cluster-sensitive projection vectors for both $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{U} = [\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^P], \quad \mathbf{V} = [\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^Q] \tag{8}$$

[2] $\rho = 0$ indicates lasso, and $\rho = 1$ indicates group lasso. Typically, $\rho = 0.5$. $\varpi_m = \sqrt{|g^m|}$.

The projected coordinates for all the data, including data with or without correspondence, are defined by

$$\begin{cases} \mathbf{f}: f(i)=\mathbf{x}_i\left(\sum_{p=1}^{P} w_i^p \mathbf{u}^p\right), & i=1,\ldots N_x \\ \mathbf{g}: g(j)=\mathbf{y}_j\left(\sum_{q=1}^{Q} r_j^q \mathbf{v}^q\right), & j=1,\ldots N_y \end{cases} \tag{9}$$

To learn $\mathbf{U}$ and $\mathbf{V}$, we rewrite the objective function as follows:

$$\begin{cases} \min_{\mathbf{U},\mathbf{V}} \frac{1}{2}\|\mathbf{f}_L-\mathbf{g}_L\|^2+\frac{\lambda_1}{2}\mathbf{f}^T\Theta_x\mathbf{f}+\frac{\lambda_2}{2}\mathbf{g}^T\Theta_y\mathbf{g}+\theta_1\sum_{p=1}^{P}\|\mathbf{u}^p\|_1 \\ \quad +\theta_2\sum_{q=1}^{Q}\left(\rho\left(\sum_m \varpi_m\|\mathbf{v}^q(g^m)\|_2\right)+(1-\rho)\|\mathbf{v}^q\|_1\right) \\ \text{s.t. } \|\mathbf{u}^p\|_2^2\le 1, \quad \|\mathbf{v}^q\|_2^2\le 1,\ p=1,\ldots,P, q=1,\ldots,Q \end{cases} \tag{10}$$

where $\mathbf{f}_L$ and $\mathbf{g}_L$ denote the projected coordinates for the labeled pairs. λ_1 and λ_2 represent the weights of the intra-modal smoothness constraints. θ_1 and θ_2 denote the importance of the regularization of the correlation model vectors. Similar with the discussion in [6,11], the proposed model is bi-convex. To effectively solve Eq. (10), we design an alternating optimization method which alternatively updates and optimizes each $\mathbf{u}^p$ or $\mathbf{v}^q$ at each time, which will be detailed in the next section.

Besides, all the other correlation analysis methods, including CCA [4], SCA [6] and StSCA (Eq. (4)), can be developed into cluster-sensitive version. Specifically, CCA, SCA and StSCA do not need to consider the manifold assumption on cross modal data, therefore the cluster-sensitive extension is even simpler.

We would like to provide some justifications on the rationality of our cluster-sensitive correlation learning model. Previous studies show that the structured sparsity regularization is useful only when the numbers of cross-modal data dimensions (D_x or D_y) are far larger than the number of cross-modal data items (N_x or N_y). Specifically, according to empirical study, a typical problem type is $D_x \simeq 3N_x$ or $D_y \simeq 3N_y$. Based on this, we can see that the global correlation analysis approach ($P=1$ and $Q=1$) is not suitable because both the numbers of data dimension and data items are very large for many Web applications. With the extension of cluster-sensitive correlation learning (see Eq. (10)), the equivalent number of data items for each sub-problem is significantly reduced at least 10 times (under a typical setting that $P, Q = 10$–200). Therefore, the feature selection with structured sparsity regularization will come to work, and the cluster-sensitive correlation model better fits the structure in the features and encodes the intra-modal density information in a more fine-grained manner, leading to better trade-off between model bias and variance.

In the visual modality, we can represent each image with multiple features that describe different properties of the visual content. Therefore, we use the structured sparsity penalty on the projection functions of visual modality, which performs feature selection on both group level and dimension level. In the textual modality, we represent each textual document with the standard bag-of-word representation. Therefore, we use the sparse regularization on the projection functions of textual modality. Note that it is possible to incorporate more kinds of textual representation, e.g., the n-gram language model. However, according to our empirical experiment, we find that using more kinds of textual features is helpful in nothing but only increasing the computational burden. The reason can be three-folds. First, a single TF-IDF feature is sufficient to encode the textual content information. Second, the dimensions of TF-IDF on our datasets are already extremely high. Incorporating more dimensions makes the correlation problem much harder to solve. Third, to better perform feature selection on the projections of textual modality, the semantic relation among different words should be used. However, this is beyond the scope of this paper.

4. Model solution

4.1. Sub-problem optimization

Due to the complex cluster-sensitive projection functions, the objective function in (10) cannot be directly minimized with any existing off-the-shelf convex optimization toolbox because a set of projection vector pairs other than one pair should be learned in our cluster-sensitive model. We borrow the idea from [6] and develop a special purpose bilateral optimization for our model, which alternatively optimizes $\mathbf{u}^p$ and $\mathbf{v}^q$. First, given an initialized and l_2-norm normalized $\mathbf{U}_0$ and $\mathbf{V}_0$, we optimize $\mathbf{U}$ with $\mathbf{V}$ fixed. To optimize $\mathbf{U}_0$, we randomly select index p and optimize the corresponding $\mathbf{u}^p$ with other $\mathbf{u}^{p'}: p'\neq p, p=1,\ldots,P$ fixed. Then we optimize $\mathbf{V}$ with previously optimized $\mathbf{U}$ fixed in a similar way. We randomly select index q and optimize the corresponding $\mathbf{v}^q$ with other $\mathbf{v}^{q'}: q'\neq q, q=1,\ldots,Q$ fixed. This bilateral optimization process continues until the model converges to a local optimal solution.

When we optimize $\mathbf{u}^p$, we solve the following sub-problem:

$$\begin{cases} \min_{\mathbf{u}^p} \frac{1}{2}\sum_{i\in L}(w_i^p\mathbf{x}_i\mathbf{u}^p+\mathbf{x}_i\mathbf{b}^i-g(i))^2+\frac{\lambda_1}{2}\left((\mathbf{u}^p)^T\mathbf{B}_1\mathbf{u}^p+2\mathbf{A}_1\mathbf{u}^p\right)+\theta_1\|\mathbf{u}^p\|_1 \\ \text{s.t. } \|\mathbf{u}^p\|_2^2\le 1. \end{cases} \tag{11}$$

where

$$\mathbf{B}_1=\mathbf{X}^T\left(\Theta_x\otimes\left(\mathbf{w}^p(\mathbf{w}^p)^T\right)\right)\mathbf{X}, \quad \mathbf{A}_1=\mathbf{f}^T\Theta_x\left(\mathbf{X}\otimes\left(\mathbf{w}^p\mathbf{1}^{D_x}\right)\right),$$
$$\mathbf{b}^i=\sum_{p'=1,p'\neq p}^{P} w_i^{p'}\mathbf{u}^{p'} \tag{12}$$

where $\mathbf{w}^p$ denotes the p-th column of $\mathbf{W}$, and $\mathbf{1}^{D_x}$ denotes the row vector with D_x ones.

When we optimize $\mathbf{v}^q$, we solve the following sub-problem:

$$\begin{cases} \min_{\mathbf{v}^q} \frac{1}{2}\sum_{j\in L}(r_j^q\mathbf{y}_j\mathbf{v}^q+\mathbf{y}_j\mathbf{c}^j-f(j))^2+\frac{\lambda_2}{2}((\mathbf{v}^q)^T\mathbf{B}_2\mathbf{v}^q+2\mathbf{A}_2\mathbf{v}^q) \\ \quad +\theta_2\left(\rho\sum_m \varpi_m\|\mathbf{v}^q(g^m)\|_2+(1-\rho)\|\mathbf{v}^q\|_1\right) \\ \text{s.t. } \|\mathbf{v}^q\|_2^2\le 1. \end{cases} \tag{13}$$

where

$$\mathbf{B}_2=\mathbf{Y}^T\left(\Theta_y\otimes\left(\mathbf{r}^q(\mathbf{r}^q)^T\right)\right)\mathbf{Y}, \quad \mathbf{A}_2=\mathbf{g}^T\Theta_y\left(\mathbf{Y}\otimes\left(\mathbf{r}^q\mathbf{1}^{D_y}\right)\right),$$
$$\mathbf{c}^j=\sum_{q'=1,q'\neq q}^{Q} r_j^{q'}\mathbf{v}^{q'} \tag{14}$$

where $\mathbf{r}^q$ denotes the q-th column of $\mathbf{R}$, and $\mathbf{1}^{D_y}$ denotes the row vector with D_y ones.

Both the correspondence inconsistency and Laplacian regularizer in Eqs. (11) and (13) are convex and Lipschitz continuous. Therefore, many existing optimization methods can be used. Since the regularization on $\mathbf{v}^q$ is complicated, we use the smooth proximal gradient method [19] which is suitable for optimizing general structured sparse learning problems. When solving Eqs. (11) and (13) with [19], we need to calculate the gradient of the correspondence inconsistency and the penalties using graph Laplacian accordingly. With the sub-problem optimization, the whole bilateral optimization process is illustrated in Algorithm 1. Based on the analysis in [6,11,13,17], a local-optimal solution is guaranteed to be achieved.

Algorithm 1. The bilateral optimization method.

Input:
Data: X, **Y**, $\mathbf{U}_0$, $\mathbf{V}_0$, **W**, **R. Parameters:** T, λ_1, λ_2, θ_1, θ_2.
Output: U,**V**
for $t=1$ **to** T **do**

for $p'=1$ **to** P
Select p from 1:P without replacement
Solve the sub-problem with respect to $\mathbf{u}^p$ (Eq. (11))
Project and ensure $\|\mathbf{u}^p\|_2^2 \le 1$
end for
for $q'=1$ **to** Q **do**
Select q from 1:Q without replacement
Solve the sub-problem with respect to $\mathbf{v}^q$ (Eq. (13))
Projectand ensure $\|\mathbf{v}^q\|_2^2 \le 1$
end for
end for

4.2. Multi-dimensional subspace learning

The methods discussed in Section 3 learn one pair of dimensional projection functions. In practise, we usually need to learn multi-dimensional subspaces from the original features in order to capture more complicated correlation among cross-modal data. Unlike CCA which can be solved by a generalized eigen-space decomposition, only one projection vector can be obtained for each learning phase of our proposed model. A practical solution for learning multi-dimensional subspace is to learn the projection vector based on the current data first, then subtract the corresponding component of projection from it, which is known as "deflation process" in the literature [13].

However, the deflation process in our paper is slightly different from those in [6] or [13], as we learn a set of cluster-sensitive projection vectors. Therefore, based on the symmetric deflation process in [13], we propose a new deflation process as described in Algorithm 2. The key differences between our proposed deflation process and existing work [13] are the way of calculating the coordinates of the subtracted components and the way of removing the components from each sub-model, rather than from a unified model in [13]. For other cluster-sensitive extension, the proposed deflation process can also be applied.

Algorithm 2. The proposed multi-dimension subspace learning.

$\mathbf{X}_0=\mathbf{X}$, $\mathbf{Y}_0=\mathbf{Y}$
for $h=1$ **to** H **do**
Learn $\mathbf{U}_h$ and $\mathbf{V}_h$ using Algorithm 1
for $i=1$ **to** N_x **do**

$$\xi^h(i)=\mathbf{x}_i^{h-1}\left(\sum_{p=1}^{P} w_i^p\mathbf{u}_h^p\right)/\|\sum_{p=1}^{P} w_i^p\mathbf{u}_h^p\|_2^2$$

end for

$$\kappa^h=\left(\mathbf{X}^{h-1}\right)^T\xi^h/\|\xi^h\|_2^2,\ \mathbf{X}^h=\mathbf{X}^{h-1}-\xi^h(\kappa^h)^T$$

for $j=1$ **to** N_y **do**

$$\omega^h(j)=\mathbf{y}_j^{h-1}\left(\sum_{q=1}^{Q} r_j^q\mathbf{v}_h^q\right)/\|\sum_{q=1}^{Q} r_j^q\mathbf{v}_h^q\|_2^2$$

end for

$$\tau^h=\left(\mathbf{Y}^{h-1}\right)^T\omega^h/\|\omega^h\|_2^2,\ \mathbf{Y}^h=\mathbf{Y}^{h-1}-\omega^h(\tau^h)^T$$

end for
return$(\mathbf{U}_h, \mathbf{V}_h)$, $h=1,\ldots,H$

4.3. Time complexity

The time cost of the model training mainly depends on (1) the complexity C of each sub-problem optimization on visual modality, and the number of sub-problems P and Q; (2) the number of iterations T to traverse all the sub-problems; (3) the number of dimensions H. Furthermore, the complexity of each sub-problem optimization is mainly determined by the proximal gradient computation and the maximal iterations T_n for sub-model updating. According to [19], the time complexity for each proximal gradient calculation is $O(D_y^2+M)$, where $M \ll D_y$ denotes the number of feature groups. For the sub-problems in textual modality, the model is directly solved with the so-called soft-thresholding operation [7] with $O(D_x^2+D_x)$ complexity. The overall time complexity of our model is $O(D_y^2T_nQHT+D_x^2PHT)$. The practical training time in our experiment is only about 1/5 of the estimated time, since the results of the last iteration can be used as the warm start of the next iteration. Consequently, the optimization of each sub-problem will terminate with less iterations than T_n given a predefined stoping criterion.

4.4. Graph Laplacian and clustering

For textual modality, we construct sparse graph Laplacian Θ_x on the returned nearest neighbors using the cosine similarity. For visual modality, as the images are represented by multiple features, we conduct k-NN search and construct the sparse graph Laplacian Θ_y on the average similarity. We use graph K-Means [37] for dimension reduction and clustering on both visual and textual modalities. First, it learns the low dimensional representation, which is the eigenspaces of the smallest eigen-values (except 0) of the graph Laplacian. Then, K-Means clustering is applied on the low dimensional representation, and a set of cluster centers is obtained. For each sample $\mathbf{x}_i$, we select the top 5 nearest centers, and calculate the Nadaraya–Watson kernel regression parameters using the Gaussian kernel as the probabilistic cluster membership as

$$w_i^p=\frac{K_h(\overline{\mathbf{x}}_i,\mathbf{d}_p)}{\sum_{p'\in\langle i\rangle}K_h(\overline{\mathbf{x}}_i,\mathbf{d}_{p'})}\quad \forall p\in\langle i\rangle \tag{15}$$

where $\overline{\mathbf{x}}_i$ denotes the representation after the dimension reduction, and $\mathbf{d}_p$ denotes the center of p-th cluster. The notation $\langle i\rangle$ is the set saving the indexes of 5 nearest center of $\overline{\mathbf{x}}_i$. Similar calculation can be done to get $\mathbf{R}$ for $\mathbf{Y}$.

For query data outside the training database, we use the out-of-sample extension [40] to obtain the reduced low dimensional representation and their corresponding probabilistic cluster membership. The multi-dimensional projection can be performed by the same formulation as in Eq. (9) based on the learned $(\mathbf{U}_h, \mathbf{V}_h)$, $h=1,\ldots,H$.

5. Experiments

Datasets: We conduct experiments on two datasets: the ImageClef 2010 dataset and the dataset collected by [41] from Wikipedia. The ImageClef data consists of 223 065 image and text document pairs after noise cleansing, where images and texts with identical document IDs imply that they serve as the complementary descriptions of each other, i.e., the correspondence information. We randomly select 50K images with their corresponding text as the multi-modal test dataset and the rest as the training database. Among the training data, we randomly keep 45% as the training pairs with correspondence, remove the correspondence information of 45% as the training data without correspondence, and the rest 10% as the validation pairs for parameter tuning. Consequent experiments are all conducted based on this experimental setting.

The Wikipedia dataset [41] is a collection of 6382 Wikipedia webpages. They are categorized into 11 topic categories by the wikipedia website. Such category information can be encoded with the graph Laplacian as described in Section 3.2. We split each page into several paragraphs, and each image in the page is linked to the paragraph where it was originally placed. The whole dataset

contains 74 961 paragraphs and 35 149 images. Each visual or textual document is attached with a 11-dim category label vector and textual-visual correspondence (one-to-many or possibly no correspondence). There are only 23 490 text documents having the correspondence information with the images, indicating that the correspondence missing is prominent. We select 5K images and 5K paragraphs with the true correspondence information as the test dataset for image-to-text and text-to-image retrieval, respectively, and others are served as the training data.

Features: We represent the textual paragraphs by TF-IDF after a stop word removal and calculate 9 types of visual feature, such as color histogram, wavelet feature, PHOG, GIST and Dense SIFT with sparse coding and 3 level spatial max-pooling (codebook size: 500). The overall visual feature dimension is about 20K. There are 29 feature groups in the visual feature. Among them, there are 21 groups from Dense SIFT with sparse coding and spatial pooling feature with each feature group corresponds to the histogram feature of different parts of spatial block. We extract TF-IDF features on the text documents of both datasets with a stop word removal process, and the final textual feature dimensions for ImageClef and Wikipedia are 50K and 73K, respectively. We conduct Latent Dirichlet allocation (LDA) on each textual document. The number of the latent topics is 800 for ImageClef and 1200 for Wikipedia.

Comparison: We perform experiment comparison with the following approaches:

(1) CCA [4] and its cluster-sensitive version (CSCCA);
(2) sparse correlation analysis[6] (SCA) and its cluster-sensitive version (CSSCA);
(3) structured sparse correlation analysis (StSCA) and its cluster-sensitive version (CSStSCA);
(4) semi-supervised kernel correlation analysis [8] (SSKCA) with RBF kernel for each visual feature and average kernel for representing the visual similarity;
(5) multi-layer neuro-network (MMNN) [22];
(6) multi-modal latent binary embedding (MLBE) [23];
(7) cross-view hashing (CVH) [21] where the learned representation is not quantized into binary codes;
(8) generalized multi-view LDA (GMLDA) [42];
(9) our proposed method.

For (1)–(3) and (5)–(8), we can only use the labeled training pairs for their model training. For (4) and (9), we use both the labeled training data and unlabeled training data for model training. For the parameter setting of both the compared methods and ours, we conduct a tuning process on the training datasets by 5-folds cross validation. Unless specially discussed in the rest of the paper, all the parameters are set according to the (near-) optimal performance in the tuning process.

Platform: All the experimental evaluations are conducted on a server with Intel (R) Xeon (R) Processor E7-4870 (30M Cache, 2.40 GHz, 6.40 GT/s Intel (R) QPI, 10 Cores), 128 GB main memory and 10,000 RPM server-level hard disks. The programs are implemented with Matlab and C++.

Table 1
The performance of cross-modal retrieval (in MAP) on ImageClef.

Method	img-to-txt	txt-to-img	Method	img-to-txt	txt-to-img
CCA	0.0517	0.0642	CSCCA	0.1283	0.1738
SCA	0.0524	0.0668	CSSCA	0.1443	0.1769
StSCA	0.0578	0.0653	CSStSCA	0.1459	0.1836
SSKCA	0.1830	0.2068	–	–	–
MMNN	0.1743	0.1952	MLBE	0.1689	0.1967
CVH	0.1896	0.2098	GMLDA	0.1843	0.2065
Ours (sng)	0.0783	0.0948	**Ours**	**0.2545**	**0.3164**

5.1. Evaluation metric

For cross-modal retrieval (image-to-text and text-to-image), we treat test data from one modality as the query, and their ground-truth correspondent data from other modality as the retrieving target. After performing cross-modal projection, we calculate the distances between queries and all the training database. Based on the ranking results, we evaluate the performance by Mean Average Precision (MAP).

5.2. Performance on cross-modal retrieval

We record the performance of all the compared models for image-to-text retrieval and text-to-image retrieval with the optimal parameter setting. The parameters of our method are set as $\{\theta_1=0.1,\theta_2=0.08,\rho=0.5,\lambda=0.4,P=Q=200,H=100\}$ for ImageClef, and $\{\theta_1=0.15,\ \theta_2=0.1,\ \rho=0.5,\ \lambda=0.8,\ P=Q=200,\ H=120\}$ for Wikipedia. For the cluster-sensitive versions of other approaches, we set $P=Q=200$ and use the optimal setting of other parameters for fair comparison. The experimental results on the two datasets are reported in Table 1 and 2, respectively. Since there is no category information on ImageClef, we only evaluate our approach with the original graph Laplacian (i.e., only the similarity information) in Table 1. Besides, we also evaluate our method when $P=Q=1$ on both datasets, and denote such unified model with sng.

In Table 2, we evaluate two versions of our method using different graph Laplacians because Wikipedia dataset provides category information for cross-modal data. Specifically, we denote our method using ordinary graph Laplacian with ord, and denote our method using dissimilarity in graph Laplacian with dis. The results imply that our cluster-sensitive model significantly outperforms other correlation learning approaches on real world data. Specifically, when $P=Q=1$, our method only performs at the same level with CCA, SCA and StSCA, and underperforms SSKCA which learns the nonlinear correlation from the data. When we increase the number of clusters and encode the category information, the performance of cross-modal retrieval is significantly enhanced, as observed in Table 2. Moreover, our method outperforms other cluster-sensitive extensions of benchmark approaches, since we incorporate the intra-modal affinity information into the model to penalize the unsmooth projection of neighborhood data.

We also compare our method with MMNN [22], MLBE [23], CVH [21] and GMLDA [42] in both Tables 1 and 2. We see that the four state-of-the-art approaches perform better than baseline approaches based on single correlation maximization paradigm, i.e., CCA (CSCCA), SCA (CSSCA) and StSCA (CSStSCA). But some of them (e.g., MLBE and MMNN) under-perform the SSKCA approach as SSKCA take full advantage of both intra-modal and inter-modal relations with the well-established kernel learning framework. CVH consistently performs the best among the four state-of-the-art approaches. The result may be explained by the usage of both intra-modal similarity and inter-modal correspondence on labeled data in CVH. However, all the compared approaches underperform our cluster-sensitive correlation learning approach, because our model takes advantage of cross-modal data without the correspondence information while the state-of-the-art approaches cannot. Such “unlabeled” data can be served as the sample set to enhance the estimation accuracy of local density information in each modality. Therefore, the correspondence information can be propagated via the neighborhoods more effectively.

5.3. Cluster sensitivity

In our model, the parameters λ_1 and λ_2 determine the relative importance of intra-modal graph Laplacian regularization. We set $\lambda=\lambda_1=\lambda_2$ as we treat different modalities as equally important for cross-modal learning. Moreover, the penalty of intra-modal smoothness interacts with the number of sub-models. Therefore, we conduct experiment to study how the intra-modal information influences the model capacity. To this end, we record the validation performance on different λ and different number of clusters on both visual and textual modalities. For ImageClef data, the penalties of structure sparsity regularization and number of dimensions are set as $\theta_1=0.1$, $\theta_2=0.08$, $\rho=0.5$ and $H=100$. For Wikipedia data, $\theta_1=0.15$, $\theta_2=0.1$, $\rho=0.5$ and $H=120$.

As Fig. 3 shows, the performances of image-to-text and text-to-image on the validation sets of both datasets are enhanced when the number of clusters (P and Q, we set $P=Q$ in this paper) is increased, which means that the cluster-sensitive correlation model better deals with the topic divergence in real world data, with the cost of more computation burden. For different settings of cluster numbers, the best performance with respect to λ is usually achieved when $\lambda=0.2$–0.8 on both datasets. For the rest of the experiments, we set $\lambda=0.4$ for ImageClef and $\lambda=0.8$ for Wikipedia, and the number of clusters (sub-models) is set to be 200 for both datasets. Note that when the amount of training data for different modalities are very imbalanced, setting different cluster numbers for different modalities may better adapt to the intra-modal data distribution.

5.4. The influence of structure sparsity

In this section, we conduct experiment on how the structure sparsity regularization encourages the correlation sub-models to

Table 2
The performance of cross-modal retrieval (in MAP) on Wikipedia.

Method	img-to-txt	txt-to-img	Method	img-to-txt	txt-to-img
CCA	0.0429	0.057	CSCCA	0.098	0.1032
SCA	0.0435	0.065	CSSCA	0.1265	0.1438
StSCA	0.0454	0.079	CSStSCA	0.1304	0.1496
SSKCA (ord)	0.1231	0.1698	SSKCA (dis)	0.1484	0.1746
MMNN	0.1401	0.1712	MLBE	0.1334	0.1687
CVH	0.1396	0.1782	GMLDA	0.1405	0.1776
Ours(sng, ord)	0.0633	0.0849	ours(sng, dis)	0.0646	0.0901
Ours (ord)	**0.1564**	**0.2678**	**Ours** (dis)	**0.1715**	**0.2840**

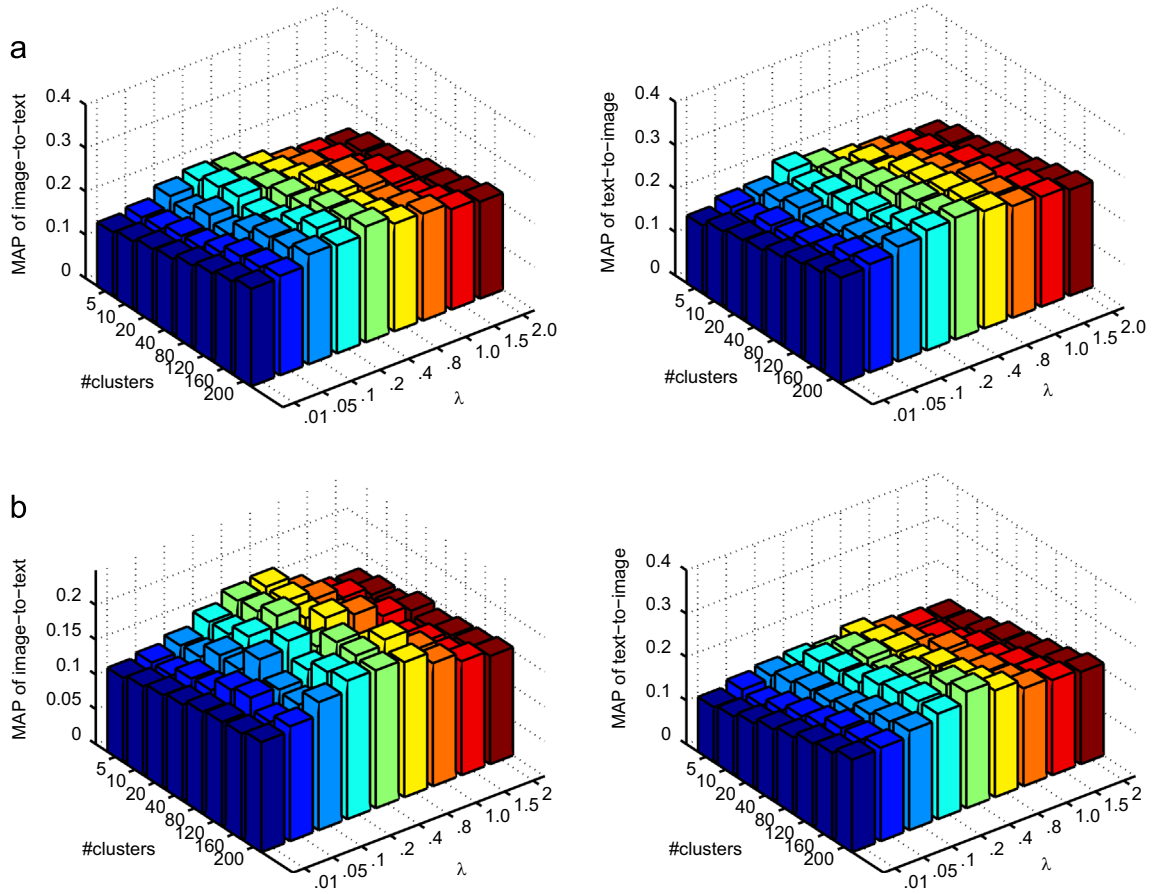

Fig. 3. Sensitivity on different λ and number of clusters on (a) ImageClef and (b) Wikipedia.

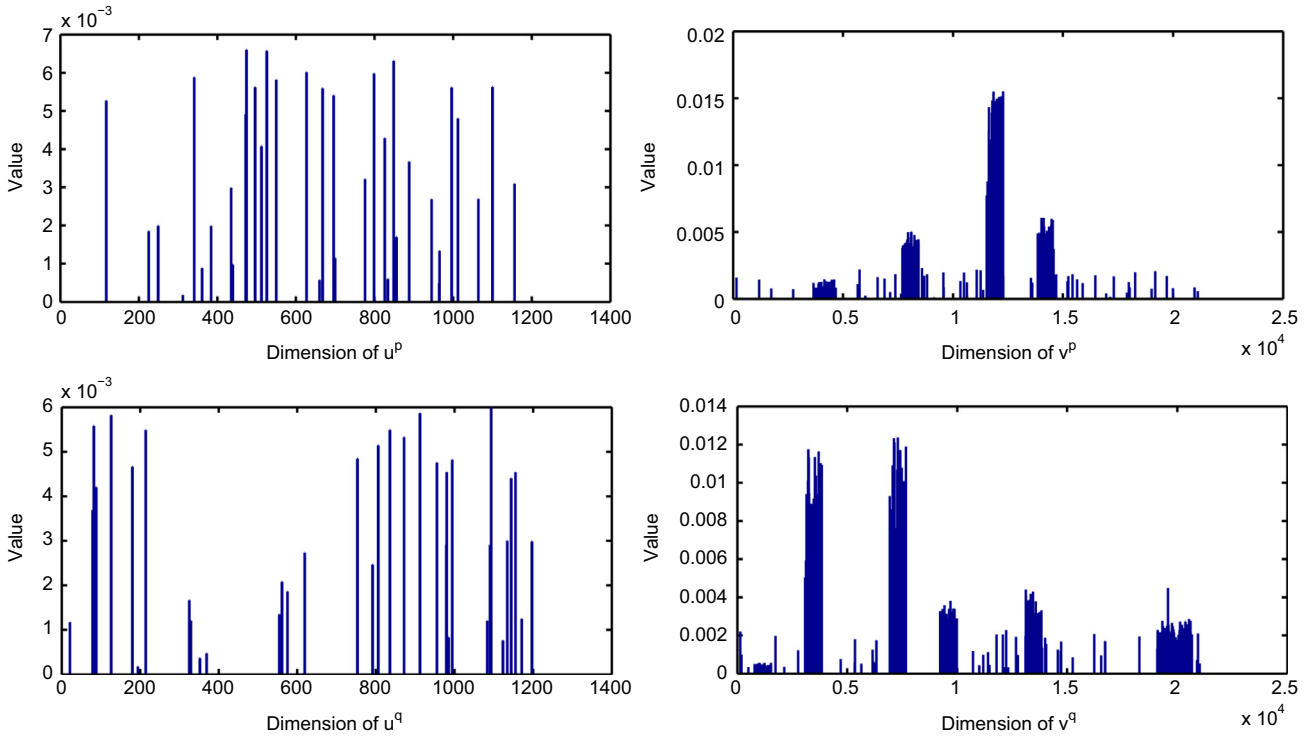

Fig. 4. Examples of the learned sub-models by our proposed method. $\mathbf{u}^p$ ($\mathbf{u}^{p'}$) and $\mathbf{v}^q$ ($\mathbf{v}^{q'}$) denote the sub-models of textual and visual modality, respectively.

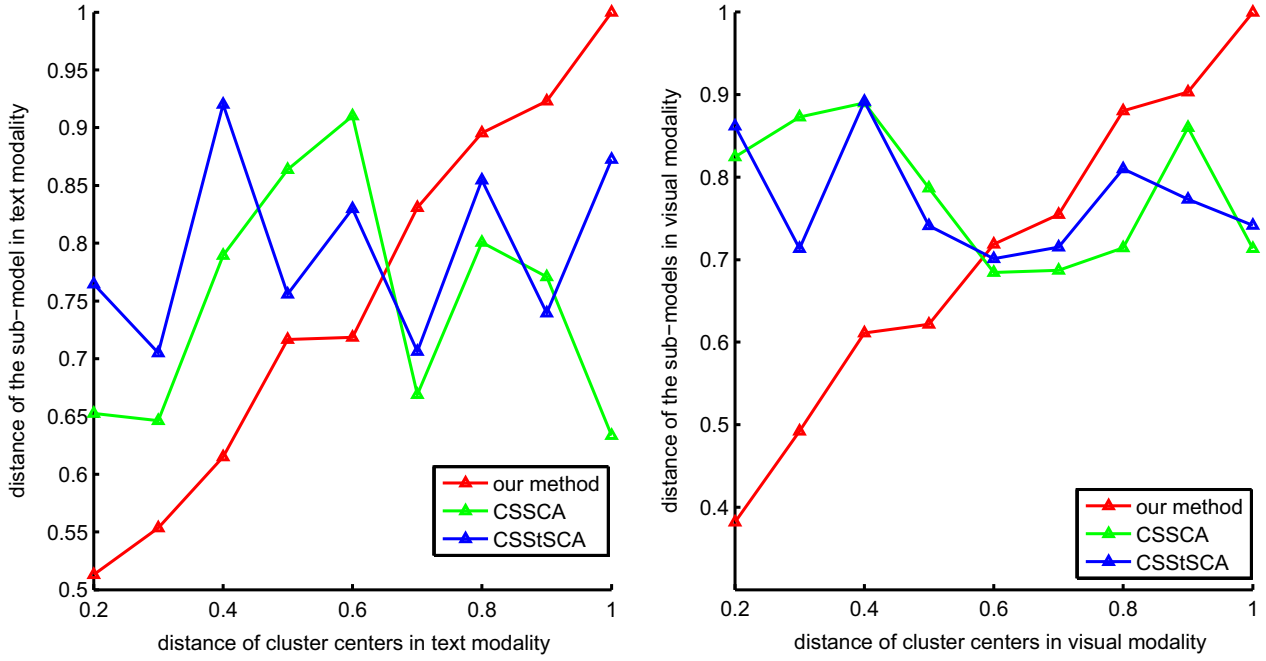

Fig. 5. The average l_2 norm distance of sub-models with respect to the distance of cluster centers.

select parts of correlated feature dimensions. By careful tuning using the validation data, we set the parameters as $\{\theta_1=0.1, \theta_2=0.08, \rho=0.5, \lambda=0.4, P=Q=200\}$ for ImageClef and $\{\theta_1=0.15, \theta_2=0.1, \rho=0.5, \lambda=0.8, P=Q=200\}$ for Wikipedia.

In Fig. 4, we plot the learned projection vector pairs $\{\mathbf{u}^p, \mathbf{v}^q\}$ and $\{\mathbf{u}^{p'}, \mathbf{v}^{q'}\}$ belonging to different sub-models on Wikipedia data. By imposing the sparse regularization, $\mathbf{u}^p$ ($\mathbf{u}^{p'}$) is sparse since only several dimensions are non-zero. The percentage of non-zero dimension is about 5% of the total dimension number. By imposing the structure sparse regularization, $\mathbf{v}^q$ ($\mathbf{v}^{q'}$) is structured sparse, as some groups of dimensions tend to be jointly non-zero, where the percentage of non-zero groups is about 5–20%. Besides, there are dimensions with zero values in the non-zero groups because of the sparsity constraint in Eq. (13), and some dimensions outside these groups (usually less than 100) are non-zero.

We see from Fig. 4 that the selected correlated dimensions for different sub-models are very different. The observation means that different feature dimensions play differently for different

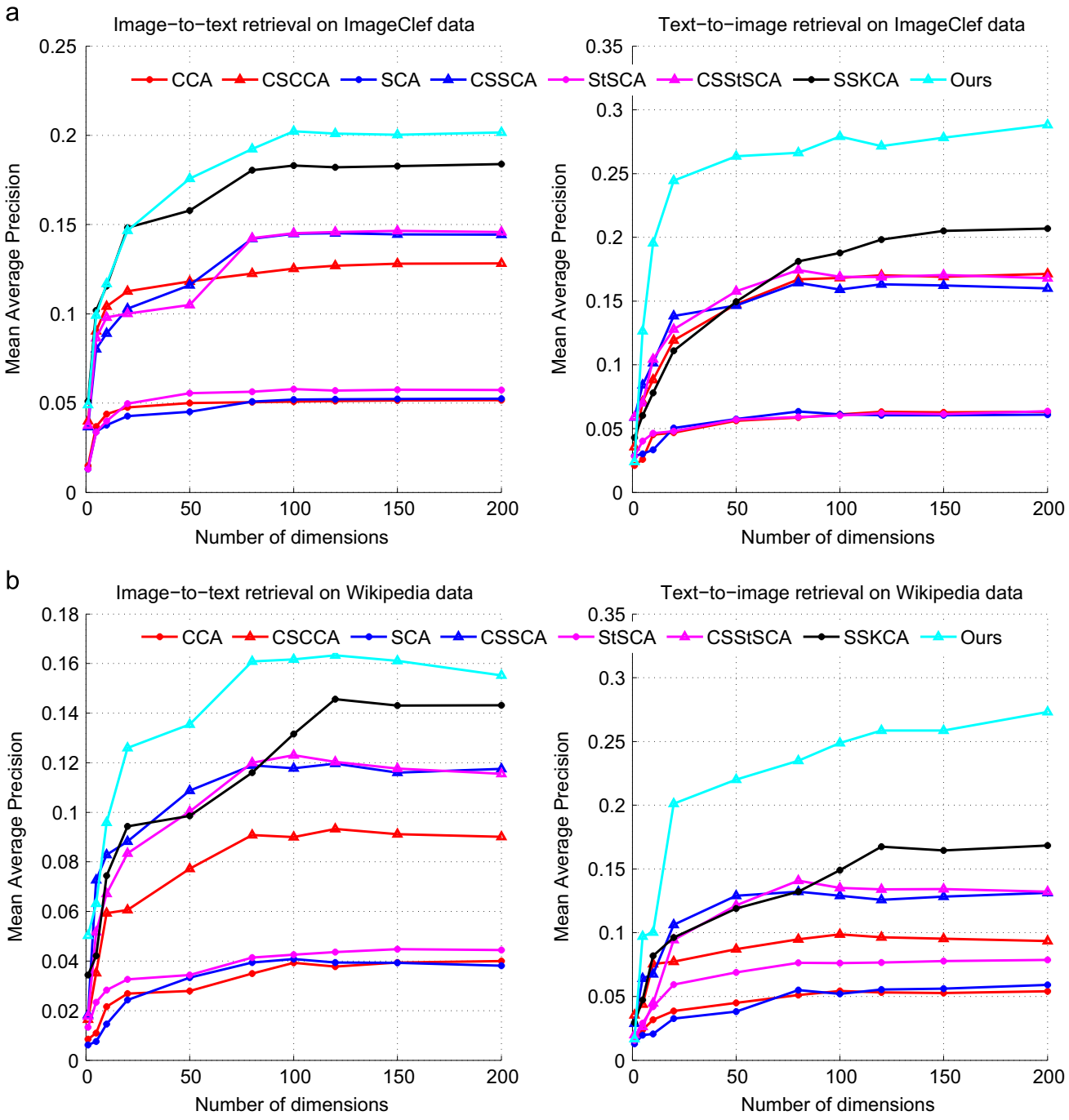

Fig. 6. The performance of cross-modal retrieval with different dimensions on both (a) ImageClef and (b) Wikipedia datasets.

subtopics (clusters) in the Web content. To further show how the learned sub-models are related with the topic divergence, we record the average l_2-norm of sub-models with respect to the distance of data centers on both text and visual modalities in Fig. 5. We see that the divergence of sub-models is consistently larger when the divergence of data centers increases, owing to the smoothness penalty in our model. When the divergence of data centers is small, the divergence of sub-models will not go to zero because the cluster membership of the data still differs. When it is large, the sub-models turn out to be totally different, as the average l_2-norm distance is close to 1. In contrast, the average l_2-norm distances of sub-models learned by the other two cluster-sensitive methods (CSSCA and CSStSCA) are not influenced by the cluster distances. Since they do not incorporate the intra-modal smoothness, the local affinity structure of cross-modal data after the projection would inevitably be disrupted.

5.5. Number of dimensions

We evaluate how the number of dimensions influences the performance of cross-modal correlation learning. We set the parameters as $\{\theta_1 = 0.1, \theta_2 = 0.08, \rho = 0.5, \lambda = 0.4, P = Q = 200\}$ for ImageClef and $\{\theta_1 = 0.15, \theta_2 = 0.1, \rho = 0.5, \lambda = 0.8, P = Q = 200\}$ for Wikipedia. For other benchmark approaches, their parameters

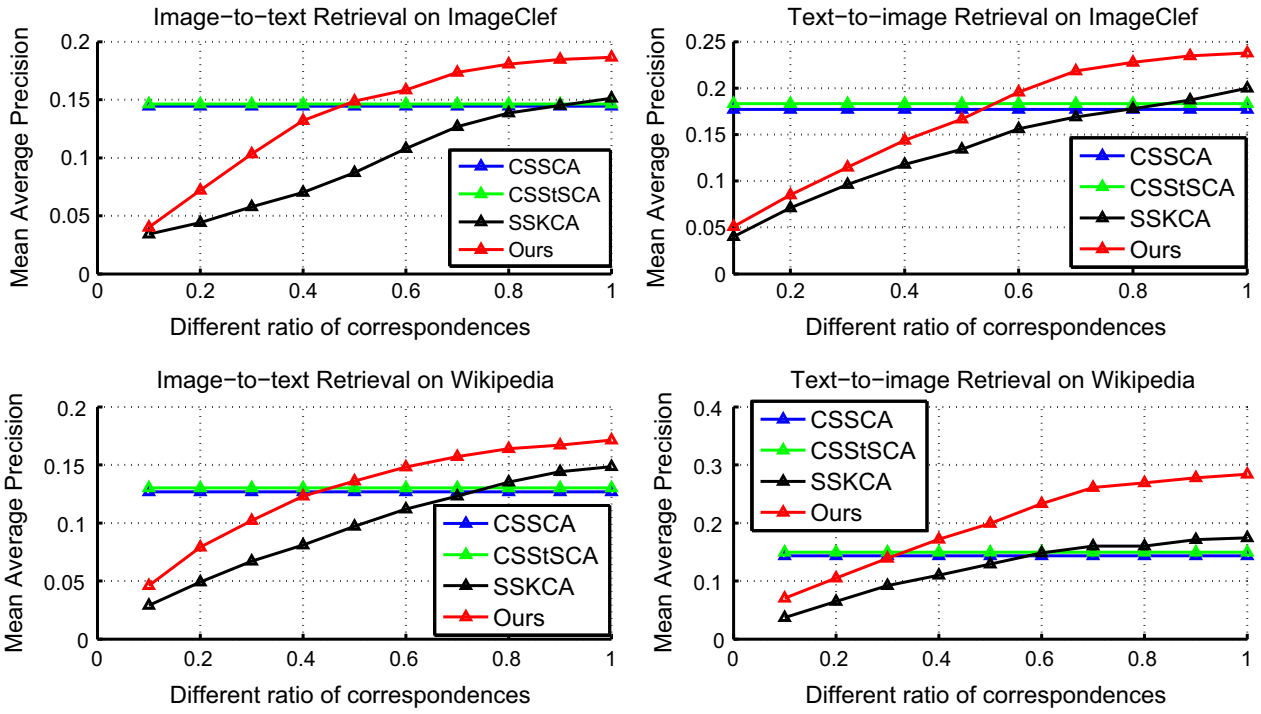

Fig. 7. The cross-modal retrieval performance with different ratios of correspondences on the training data on ImageClef and Wikipedia datasets.

Fig. 8. Some top ranked examples of text-to-image (left part) and image-to-text (right part) retrieval on Wikipedia. Items marked with red dots are the ground-truth corresponding target documents in the top ranked lists. Readers may zoom in this figure for more details. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

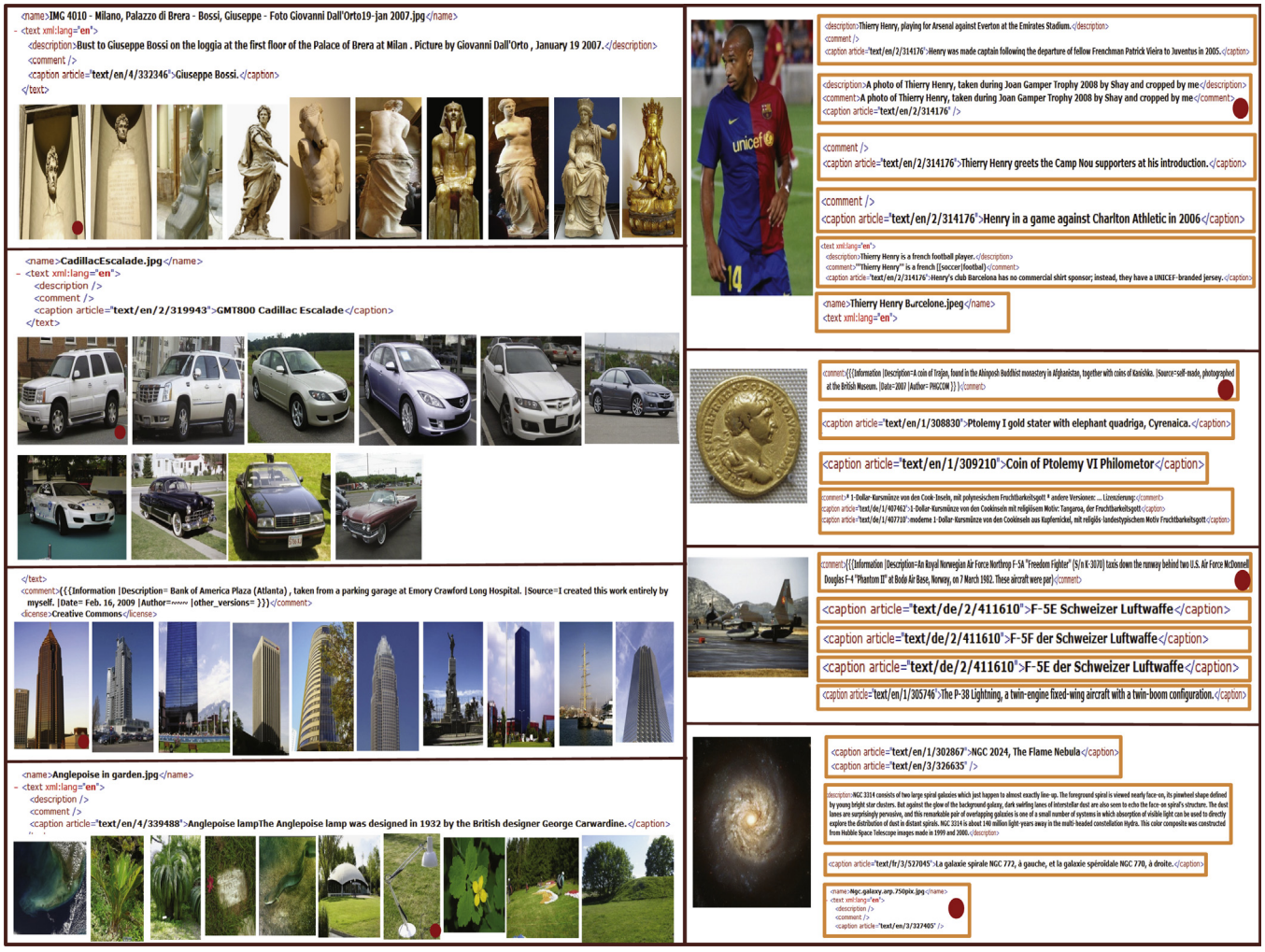

Fig. 9. Some top ranked examples of text-to-image (left part) and image-to-text (right part) retrieval on ImageClef. Items marked with red dots are the ground-truth corresponding cross-modal target documents in the top ranked lists. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

are optimally tuned. The performance of cross-modal retrieval with respect to the number of dimensions is recorded in Fig. 6. Notice that when the dimensions increase from several to dozens, the retrieving performances of all the methods are enhanced significantly. However, when the dimension number achieves hundreds, e.g, 100, the performance of our method on ImageClef data would change slightly or even drop a little. Similar observation can be obtained for all the other approaches. Although the performance is still enhanced when the number of dimensions achieves 200 on the bottom right sub-figure in Fig. 6, it should be pointed out that more computation burden is also introduced. Therefore, considering both effectiveness and efficiency, the number of dimensions for our method can be set as $H=100$ on ImageClef and $H=120$ on Wikipedia. This setting also means that we learn a set of low-rank cluster-sensitive correlation models for cross-modal learning.

5.6. Number of correspondences

In this section, we only use the training data partition with correspondence to evaluate the impact of the number of correspondences on different models. To this end, we compare our method with CSSCA, CSStSCA and SSKCA. For CSSCA and CSStSCA, we use all the correspondence information on the data partition to train the models. For SSKCA and our method, we remove different numbers of correspondence information from the data partition, and use the data to train the correlation model. We set the parameters as $\{\theta_1=0.1, \theta_2=0.08, \rho=0.5, \lambda=0.4, P=Q=200, H=100\}$ for ImageClef and $\{\theta_1=0.15, \theta_2=0.1, \rho=0.5, \lambda=0.8, P=Q=200, H=120\}$ for Wikipedia. The performances on validation data with different ratios of correspondences are recorded in Fig. 7. We see that when the number of correspondences increases, the performances of both SSKCA and our method are increased. They outperform CSSCA and CSStSCA early before the ratio is up close to 1. The reason can be explained by the use of the graph Laplacian, with which the projected coordinate can be sufficiently propagated to the neighborhood data within the modality. When the ratio is increased to 1 (i.e., no correspondence information is missing), the intra-modal smoothness still leads to performance enhancement, especially for our approach, because it penalizes the inconsistency of cluster-sensitive projection.

5.7. Cross-modal retrieval examples

We illustrate some examples of cross-modal retrieval on both datasets in Figs. 8 and 9. Given text queries, our method correctly rank the ground-truth corresponded images as top 1 result, which can be seen in the examples on both figures. Besides, our method also identifies images having similar visual content and semantics with the ground-truth documents. Given image queries, our method finds a set of text documents with similar keywords, as shown in the right part of Figs. 8 and 9.

6. Conclusions

We propose a cluster-sensitive structured correlation learning framework for cross-modal retrieval. Multiple cluster-sensitive correlation sub-models are learned instead of a unified correlation model, which better fits the content divergence in different modalities. By using structure sparsity regularization on the projection vectors, a set of interpretable structure sparse correlation sub-models are obtained. To deal with correspondence information missing, we take full advantage of both intra-modal affinity and inter-modal co-occurrence. The corresponding smoothness assumption imposed on each modality guarantees that the projected coordinates of adjacent data within a modality tend to be similar. Extensive experiments are conducted on large scale cross-modal data, and the results demonstrate the effectiveness of our approach. For future work, we will study how to extend the proposed method for dealing with more diversified modalities, such as video and audio.

Acknowledgment

This work was supported in part by National Basic Research Program of China (973 Program): (2012CB316400) and (2015 CB351802), 863 program of China: (2014AA015202), and National Natural Science Foundation of China (NSFC): (61025011), (61303160), (61332016), (61390511), (61322212), (61473273) and (61429201). This work was supported in part to Prof. Qi Tian by ARO Grant W911NF-12-1-0057 and Faculty Research Awards by NEC Laboratories of America.

References

[1] F. Wu, H. Zhang, Y. Zhuang, Learning semantic correlations for cross media retrieval, in: ICIP, 2006.
[2] Y. Zhuang, Y. Yang, F. Wu, Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval, Trans. Multimed. 10 (2) (2008) 221–229.
[3] Y. Zhuang, Y. Yang, F. Wu, Y. Pan, Manifold learning based cross-media retrieval: a solution to media object complementary nature, J. VLSI Signal Process. 46 (2–3) (2007) 153–164.
[4] H. Hotelling, Relations between two sets of variates, Biometrika 28 (34) (1936) 321–372.
[5] H. Wold, Partial least squares, in: Samuel Kotz, Norman L. Johnson (Eds.), Encyclopedia of Statistical Sciences, vol. 6, 1985, pp. 581–591.
[6] D. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.
[7] D.R. Hardoon, J. Shawe-Taylor, Sparse canonical correlation analysis, Mach. Learn. 83 (2011) 331–353.
[8] M.B. Blaschko, C.H. Lampert, A.Gretton, Semi-supervised Laplacian regularization of kernel canonical correlation analysis, in: ECML-PKDD, 2008.
[9] A. Tenenhaus, S. Gif-Sur-Yvette, M. Tenenhaus, H.P. Jouy-En-Josas, Regularized generalized canonical correlation analysis, Psychometrika 76 (2) (2011) 257–284.
[10] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM Multimedia, 2010.
[11] X. Chen, H. Liu, J.G. Carbonell, Structured sparse canonical correlation analysis, in: AISTATS, 2012.
[12] S. Virtanen, A. Klami, S. Kaski, Bayesian CCA via Group Sparsity, in: ICML, 2011.
[13] MLAKimAnh Lê Cao/ Debra Rossouw/ Christèle RobertGranié/ Philippe Besse, / Debra Rossouw, and / Philippe Besse. "A Sparse PLS for Variable Selection when Integrating Omics Data." Statistical Applications in Genetics & Molecular Biology 7.1(2008):1-32.
[14] N. Chen, J. Zhu, F. Sun, E.P. Xing, Large-margin predictive latent subspace learning for multi-view data analysis, Trans. Pattern Anal. Mach. Intell. 34 (12) (2012) 2365–2378.
[15] D. Blei, M. Jordan, Modeling annotated data, in: SIGIR, 2003.
[16] Y. Jia, M. Salzmann, T. Darrell, Learning cross-modality similarity for multinominal data, in: ICCV, 2011.
[17] D. Zhai, H. Chang, S. Shan, X. Chen, W. Gao, Multi-view metric learning with global consistency and local smoothness, ACM Trans. Intell. Syst. Technol. 3 (6) (2011) 53–75.
[18] D.K.H. Lim, B.McFee, G.Lanckriet, Robust structural metric learning, in: ICML, 2013.
[19] X. Chen, Q. Lin, S. Kim, J.G. Carbonell, E.P. Xing, Smoothing proximal gradient method for general structured sparse learning. ⟨arXiv:1202.3708v1⟩, 2012.
[20] M.M. Bronstein, A.M. Bronstein, F. Michel, N. Paragios, Data fusion through cross-modality metric learning using similarity-sensitive hashing, in: CVPR, 2010.
[21] S. Kumar, R. Udupa, Learning hash functions for cross-view similarity search, in: IJCAI, 2011.
[22] J. Masci, M.M. Bronstein, A.A. Bronstein, J. Schmidhuber, Multimodal similarity-preserving hashing ⟨arXiv:1207.1522⟩, 2012.
[23] Y. Zhen, D.-Y. Yeung, A probabilistic model for multimodal hash function learning, in: KDD, 2012.
[24] A. Vinokourov, J. Shawe-Taylor, N. Cristianini, Inferring a semantic representation of text via cross-language correlation analysis, in: NIPS, 2003.
[25] Shawe-Taylor, John, D. R. Hardoon, and S. Szedmak. "Canonical correlation analysis; An overview with application to learning methods." Neural Computation 16.12(2007):págs. 2639-2664.
[26] H.D. Vinod, Canonical ridge and econometrics of joint production, J. Econom. 4 (1976) 147–166.
[27] R. Vidal, Y. Ma, S. Sastry, Generalized principal component analysis (gpca), Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1945–1959.
[28] Y. Gong, S. Lazebnik, Iterative quantization: a procrustean approach to learning binary codes, in: CVPR, 2011.
[29] K. Barnard, P. Duygulu, D. Forsyth, N. deFreitas, D.M. Blei, M.I. Jordan, Matching words and pictures, J. Mach. Learn. Res. 3 (2003) 1107–1135.
[30] C. Wang, D. Blei, F. Li, Simultaneous image classification and annotation, in: CVPR, 2009.
[31] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[32] D. Putthividhy, H. Attias, S. Nagarajan, Topic regression multi-modal Latent Dirichlet Allocation for image annotation, in: CVPR, 2010.
[33] H. Xiao, T. Stibor, Toward artificial synesthesia: linking images and sounds via words, in: NIPS Workshop on Machine Learning for Next Generation Computer Vision Challenges, 2010.
[34] D. Rafailidis, S. Manolopoulou, P. Daras, A unified framework for multimodal retrieval, Pattern Recognit. 46 (2013) 3358–3370.
[35] Y. Yang, Y. Zhuang, F. Wu, Y. Pan, Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval, Trans. Multimed. 10 (3) (2008) 437–446.
[36] X. Mao, B. Lin, D. Cai, X. He, J. Pei, Parallel field alignment for cross media retrieval, in: ACM Multimedia, 2013, pp. 897–906.
[37] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, E.Y. Chang, Parallel spectral clustering in distributed systems, Trans. Pattern Anal. Mach. Intell. 3 (33) (2011) 568–586.
[38] J. Friedman, T. Hastie, R. Tibshirani, A note on the group lasso and a sparse group lasso ⟨arXiv:1001.0736v1⟩, 2010.
[39] A.B. Goldberg, X. Zhu, S. Wright, Dissimilarity in graph-based semi-supervised classification, in: AISTAT, 2007.
[40] Y. Bengio, J.-F. Paiement, P. Vincent, Out-of-sample Extensions for lle, isomap, mds, Eigenmaps, and Spectral Clustering, Technical Report 1238, 2003.
[41] W. Xiong, S. Wang, C. Zhang, Q. Huang, WIKI-CMR: a web cross modality database for studying and evaluation of cross modality retrieval methods, in: ICME, 2013.
[42] A. Sharmay, A. Kumar, Hal Daume III, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: CVPR, 2013.

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 Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include semantic image analysis, image and video retrieval and large-scale web multimedia data mining.

Fuzhen Zhuang received the B.S. degree in Computer Science from Chongqing University, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently an Associate Professor in the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include transfer learning, machine learning, data mining, and parallel classification algorithms.

Shuqiang Jiang received the M.S. degree from the College of Information Science and Engineering, Shandong University of Science and Technology, Shandong, China, in 2000, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He is currently a Full Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 90 papers on the related research topics.

Qingming Huang received the B.S. degree in Computer Science and Ph.D. degree in Computer Engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a Professor with the Graduate University of the Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. He has authored or coauthored nearly 200 academic papers in prestigious international journals and conferences. His research areas include multimedia video analysis, video adaptation, image processing, computer vision, and pattern recognition. Dr. Huang is a reviewer for IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Communications. He has served as program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, and PSIVT.

Qi Tian received the B.E. degree in Electronic Engineering from Tsinghua University, China, in 1992 and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois, Urbana-Champaign in 2002. He is currently a Full Professor in the Department of Computer Science at the University of Texas at San Antonio (UTSA). His research interests include multimedia information retrieval and computer vision. He has published over 150 refereed journal and conference papers. His research projects were funded by NSF, ARO, DHS, SALSI, CIAS, and UTSA and he also received faculty research awards from Google, NEC Laboratories of America, FXPAL, Akiira Media Systems, and HP Labs. He took a one-year faculty leave at Microsoft Research Asia (MSRA) during 2008–2009. He was the author of a Top 10% Best Paper Award in MMSP 2011, a Best Student Paper in ICASSP 2006, and a Best Paper Candidate in PCM 2007. He received 2010 ACM Service Award. He has been serving as Program Chairs, Organization Committee Members and TPCs for numerous IEEE and ACM Conferences including ACM Multimedia, SIGIR, ICCV, and ICME. He is the Guest Editors of IEEE Transactions on Multimedia, Journal of Computer Vision and Image Understanding, Pattern Recognition Letter, EURASIP Journal on Advances in Signal Processing, Journal of Visual Communication and Image Representation, and is in the Editorial Board of IEEE Transactions on Circuit and Systems for Video Technology (TCSVT), Journal of Multimedia (JMM) and Journal of Machine Visions and Applications (MVA). He is a Senior Member of IEEE and a Member of ACM.