Multi-modal Mutual Topic Reinforce Modeling for Cross-media Retrieval

Yanfei Wang, Fei Wu, Jun Song, Xi Li, and Yueting Zhuang
College of Computer Science, Zhejiang University
Hangzhou, Zhejiang, China
{yanfeiwang07, wufei, songjun54cm, xilizju, yzhuang}@zju.edu.cn

ABSTRACT

As an important and challenging problem in the multimedia area, multi-modal data understanding aims to explore the intrinsic semantic information across different modalities in a collaborative manner. To address this problem, a possible solution is to effectively and adaptively capture the common cross-modal semantic information by modeling the inherent correlations between the latent topics from different modalities. Motivated by this task, we propose a supervised multi-modal mutual topic reinforce modeling (M^3R) approach, which seeks to build a joint cross-modal probabilistic graphical model for discovering the mutually consistent semantic topics via appropriate interactions between model factors (e.g., categories, latent topics and observed multi-modal data). In principle, M^3R is capable of simultaneously accomplishing the following two learning tasks: 1) modality-specific (e.g., image-specific or text-specific) latent topic learning; and 2) cross-modal mutual topic consistency learning. By investigating the cross-modal topic-related distribution information, M^3R encourages to disentangle the semantically consistent cross-modal topics (containing some common semantic information across different modalities). In other words, the semantically co-occurring cross-modal topics are reinforced by M^3R through adaptively passing the mutually reinforced messages to each other in the model-learning process. To further enhance the discriminative power of the learned latent topic representations, M^3R incorporates the auxiliary information (i.e., categories or labels) into the process of Bayesian modeling, which boosts the modeling capability of capturing the inter-class discriminative information. Experimental results over two benchmark datasets demonstrate the effectiveness of the proposed M^3R in cross-modal retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

ACM Multimedia 2014, Nov. 3-7, 2014, Orlando, Florida, USA

http://dx.doi.org/10.1145/2647868.2654901.

Keywords

Multi-modal Analysis; Mutual Topic; Topic Reinforcement; Cross-media Retrieval

1. INTRODUCTION

Nowadays, many real-world applications involve multi-modal documents, where information inherently consists of data with different modalities, such as a web image with loosely related narrative text descriptions, or a news article with paired text and images. Since multi-modal documents exploit the symbiosis of multiple-modality data to deliver high-level semantics, it is desirable to disentangle the underlying latent correlations between data objects with different modalities in multi-modal documents and support similarity search across different modalities. However, the so-called *heterogeneity-gap* has been widely understood as a fundamental barrier of multi-modal document modeling. Therefore, effectively understanding the multi-modal data as well as their underlying semantic properties plays a crucial role in cross-media analysis.

To achieve the goal of multi-modal data understanding, a number of approaches have been proposed for multi-modal document modeling in the recent literature. According to the model-constructing mechanisms, these approaches can be typically categorized into two classes: statistical dependency modeling and probabilistic graphical modeling. Specifically, the first class of approaches mainly focuses on maximizing the statistical dependency (e.g., measured by the mutual information) of different modalities in the common latent space [12, 24, 23, 7]. In contrast, the second class of approaches is derived from the joint modeling of data objects with different modalities in a probabilistic manner[2, 22, 26]. These approaches tend to maximize the likelihood of observed multi-modal data in terms of their latent topics. They are usually based on some assumptions about how multi-modal data is correlated such as all modalities share same topic proportions, or have one-to-one topic correspondences, or have commonly sharing topics.

In order to learn discriminative correlated latent representations, a group of approaches integrate the side information (e.g., class labels) into the process of multi-modal document modeling, to enhance representation performance for within-class multi-modal data. For instance, canonical correlation analysis (CCA) and support vector machine (SVM)are combined in [8] to multi-view classification. A generic multi-view latent space Markov network was proposed in [5] to jointly maximize the likelihood of multi-view data and their supervising labels. By introducing the class label information, the

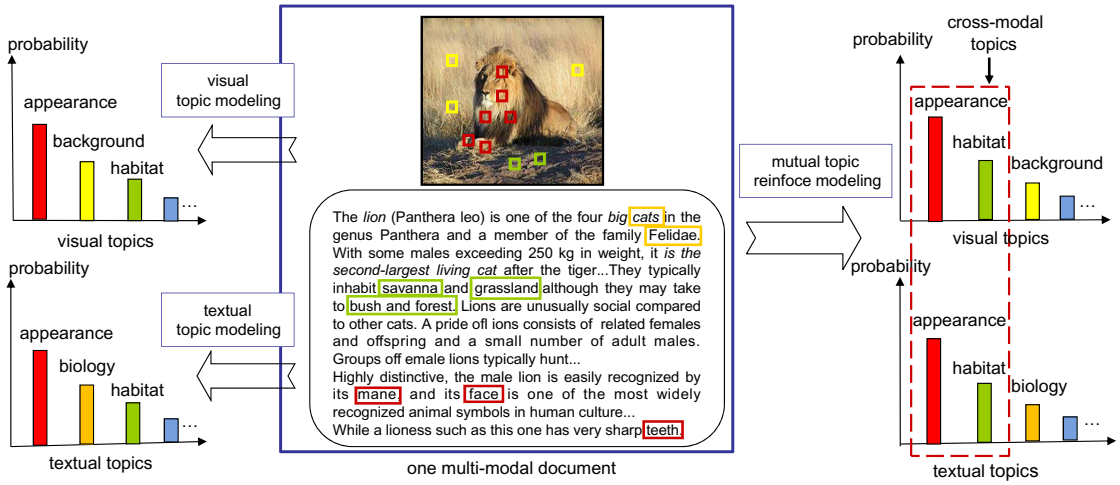

Figure 1: The intuitive illustration of multi-modal mutual topic reinforce modeling. Given one multi-modal document, its textual units or the visual units respectively describe individual text-specific topics (i.e., appearance,biology and habitat) or image-specific topics (i.e., appearance, background and habitat) with different probabilities. The proposed M^{3}R attempts to assign a high priority to the sharing cross-modal topics (i.e., appearance and habitat). The areas in the image and the words in the text highlighted by the same color share a same latent topic. (Figure best viewed in color)

generalized multi-view analysis (GMA) [24] extended original unsupervised two-view CCA to supervised multi-view counterpart. In [31], coupled dictionary learning (DL) was introduced to discover the correlations of multi-modal data. It is worth noting that various forms of side-information can potentially offer "free" supervision in many scenarios for multi-modal data modeling such as category belonging, user tagging, ratings and so on.

However, the aforementioned approaches to multi-modal modeling are generally incapable of explicitly and effectively modeling the intrinsic interactions between modalities and usually suffer from the modeling weakness in the following two aspects: 1) distinguishing the relative importance information on the latent topics for multi-modal data understanding; and 2) discovering the cross-modal topic consistency information. To alleviate the limitations, we propose a multi-modal mutual topic reinforce modeling (M^{3}R) approach that can adaptively encode the cross-modal topic consistency information by multi-modal reinforcement modeling. Here, "*cross-modal topic*" means the topics simultaneously remarked by multi-modal data (i.e., images and texts) within the same multi-modal documents. Using such cross-modal topic consistency information, M^{3}R is able to adaptively learn a set of more semantically meaningful latent topics via the interactions between multi-modal topics. As a result, the mutually consistent cross-modal topics (reflecting the same semantic information) are encouraged with a relatively high priority, while the remaining modality-specific topics are discouraged but still preserved. Therefore, the process of discovering the mutually consistent cross-modal topics is associated with the concept of mutual topic reinforcement.

Figure 1 gives out an intuitive illustration of M^{3}R. Given one multi-modal document about a "lion" with an image and corresponding text, the textual units (e.g., words or sentences) or the visual units (e.g., patches or regions) respectively describe individual modality-specific textual topics (i.e., appearance,biology and habitat) or visual topics (i.e., appearance, background and habitat) with different probabilities. However, some of textual units and visual units all remark several cross-modal topics (i.e., appearance and habitat). In principle, M^{3}R tends to assign a high priority to cross-modal topics by mutual topic reinforcement and at the same time preserves other modality-specific topics.

To obtain discriminative multi-modal data representations, we further incorporate the class label information into the process of probabilistic graphical modeling, and then formulate M^{3}R as a generative probabilistic model driven by a supervised learning manner. Therefore, the main contributions of our work are two-fold. First, we introduce the concept of multi-modal mutual topic reinforcement into multi-modal data understanding. Second, we establish a hybrid generative-discriminative probabilistic graphical model that adaptively learns a set of semantically meaningful latent topics via cross-modal mutual topic reinforcement in a supervised learning manner.

2. RELATED WORKS

As aforementioned, the modeling of multi-modal documents can be typically categorized into two classes: statistical dependency modeling and probabilistic graphical modeling. Representative works of the first class are Canonical Correlation Analysis (CCA) [12] and its extensions [24, 23] which project multi-modal data into a common (or shared) subspace so that the correlations between multi-modal data is preserved or maximized. For examples, after the maximally correlated subspace of text and image features is obtained by CCA, logistic regression is employed to cross-media retrieval in [23]. As a supervised kernelizable extension of CCA, Generalized Multiview Analysis [24] is conducted to map data in different modality spaces to a single (non)linear subspace. Although the aforementioned methods are able to

effectively discover the desired latent representations, they generally lack intuitive explanations since the obtained representations are projections of the multi-modal data in a common space without apparently interpretable meanings.

Representative works of probabilistic graphical modeling include multi-modal latent Dirichlet allocation (mmLDA)[2], Correspondence LDA (Corr-LDA) [2], Topic-regression Multi-modal LDA (tr-mmLDA) [22] and Factorized Multi-Modal Topic Model [26]. These approaches introduce shared latent variables which either indicate the topic proportions as in mmLDA or the indexes of topics as in CorrLDA [2]. Therefore, they either assume that all modalities share same topic proportions, or have one-to-one topic correspondences, or have commonly shared topics. Nevertheless, those predefined assumptions inherently restrain a more flexible application of cross-media retrieval in the setting involved uncontrolled multi-modal data. Thus, other topic model based methods such as Multi-modal Document Random Field (MDRF) [14] are proposed to deal with more realistic scenarios. All the above models learn the latent representations of the multi-modal data in an unsupervised manner though they are able to offer intuitive probabilistic interpretations. The most similar work with us is the so-called nonparametric Bayesian supervised multi-modal topic modal [16], which present a nonparametric Bayesian approach to learning upstream supervised topic models for analyzing multi-modal data. However, our proposed method is a downstream supervised topic model and encourages the learning of cross-modal topics through topic distribution interactions.

Other methods based on dictionary learning for multi-view/multi-modal retrieval such as factorized/coupled dictionary learning were proposed in [13][31]. There are also methods for cross-modal similarity metric learning based on learning to rank such as [18] and [28], where the authors proposed uni-directional and bi-directional cross-modal ranking methods via structural SVM. Hashing based methods for multi-modal similarity search were initiated by Bronstein et al. in CMSSH [4]. After that, CVH [15], MLBE [30] and sparse multi-modal hashing [29] were proposed respectively. These methods bear some resemblance with CCA and its extensions that they directly utilize intra-modality and inter-modality similarities to map multi-modal data into a comparable subspace and thus lack the interpretability for the latent representations.

Motivated by the recent remarkable advance of deep learning, several deep architectures have been conducted to learn the joint multi-modal representation. A multi-modal Deep Boltzmann Machine for learning a generative model for data with multiple modalities is proposed in [25]. Methods like stacked autoencoders [21] and deep CCA [1] have similar incentive. However, these methods all build their models on an unsupervised manner. In [10], a deep visual-semantic embedding model was presented to identify visual objects using both labeled image data as well as semantic information gleaned from unannotated text. However, this supervised deep learning method is for image labeling and can not be extended to multi-modal analysis in a straightforward manner.

3. THE MODEL OF M^{3}R

In this section, we illustrate the detailed information of our model. Notations and formulations are first presented followed by the generative process and model inference. Finally, we derive the prediction algorithm using our model for multi-modal retrieval.

3.1 Problem Formulation

With some training data of multi-modal documents and their corresponding labels, we aim to learn the latent representations in terms of mutually reinforced cross-modal topics as well as maintaining discriminative information of multi-modal documents. Terminologies from text modeling such as "words", "documents", and "vocabulary" are generalized in modeling data of other modalities and are used throughout the paper.

Suppose that we have a labeled training set of D multi-modal documents with M modalities from C categories: $\Omega = \{\boldsymbol{x}_d = (\mathbf{x}_{d1}, \cdots, \mathbf{x}_{dm}, \cdots, \mathbf{x}_{dM}, c_d)\}_{d=1}^{D}$, where $\mathbf{x}_{dm}$, which has N_{dm} words (e.g., textual words for texts or visual words for images) $\{x_{dmn}\}$, represents the unimodal document of m-th modality inside d-th multi-modal document while $c_d \in \{1, \cdots, C\}$ is the category of d-th multi-modal document. We assume that each word x_{dmn} from $\mathbf{x}_{dm}$ takes a discrete value from its modality-specific vocabulary $\{V_m\}_{m=1}^{M}$. There is no constraint that all the multi-modal document must have all M modalities, we just assume full correspondence for presentation convenience. This will be apparent in the section of generative process.

As aforementioned, the modality correlations in multi-modal documents are adaptively learned by the interactions of latent topic distributions while discriminations are acquired by integration of label information into the probabilistic graphic model. Figure 2 illustrates our model as a graphic model. The shaded nodes indicate observations, while the others represent latent variables. The dashed edges indicate that the topic proportion of each unimodal document is determined not only by the prior, but also by the topic proportions of other unimodal documents with the same multi-modal documents as it. Note that M^{3}R systematically models the observed multi-modal data, the class labels and the interactions between cross-modal latent topics.

Following the notations in Figure 2, our model follows the tradition of latent Dirichlet allocation (LDA) that the topic proportions $\boldsymbol{\pi}_{d1}, \cdots, \boldsymbol{\pi}_{dM}$ are generated from Dirichlet distribution with hyper-parameter α while the topic z_{dmn} of the word in a document with certain modality is drawn from a multinomial distribution. Each word x_{dmn} is drawn from the corresponding topic-word multinomial distribution ϕ_{mk} while ϕ_{mk} is drawn from a Dirichlet distribution with prior $\beta_{1:M}$. In addition to that, we introduce the correlation (interaction) among topic proportions inside a multi-modal document as well as the supervision (label) information.

3.2 Correlation and Supervision Modeling

Given the d-th multi-modal document $\boldsymbol{x}_d$, there must be some correlations between different modality data $\{\boldsymbol{x}_{dm}\}_{m=1}^{M}$ to synthetically express the whole semantic delivered by $\boldsymbol{x}_d$. For example, given one multi-modal document with an image and corresponding text, the image and corresponding text are complementary to each other for describing the same semantics embedded in the multi-modal document. Here we assume that if two data objects in one multi-modal document are similar or correlated, the same should be with their topic proportions. Therefore, we learn the correlations and further reinforce the cross-modal topics of different modalities inside the multi-modal document through topic

proportion similarities. If $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$ are topic proportions of two data objects in one multi-modal document respectively, the similarity between topic proportions $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$ can be calculated by the potential function as follows [14]:

$$\Psi(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) = \exp(-\lambda f(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)) \tag{1}$$

where λ is a positive scaling factor for the potential function, and $f(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j)$ is the symmetric KL-divergence which can be defined as follows:

$$\begin{aligned} f(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) &= \frac{1}{2}(D_{KL}(\boldsymbol{\pi}_i||\boldsymbol{\pi}_j) + D_{KL}(\boldsymbol{\pi}_j||\boldsymbol{\pi}_i)) \\ &= \frac{1}{2}\sum_{k=1}^{K}(\pi_{ik}\log\frac{\pi_{ik}}{\pi_{jk}} + \pi_{jk}\log\frac{\pi_{jk}}{\pi_{ik}}) \end{aligned} \tag{2}$$

where K is the dimension of topic proportions $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_j$, i.e., the number of topics.

To model the discriminative (label) information, we referred to downstream supervised models such as supervised latent Dirichlet allocation (LDA) and its variants [3] [27], which have gained much success in many applications such as document or image classification in uni-modal scenarios. Inspired by these works, the label information of the multi-modal documents is modeled through the softmax function. We generate the labels using the softmax regression based on the empirical topic frequencies of the multi-modal documents which are the concatenations of the empirical topic frequencies for unimodal documents within the multi-modal documents.

Thus, referred to Figure 2, the parameters of our model include the hyper-parameter α for topic proportions, a set of M hyper-parameter $\beta_{1:M}$ for topic-words distributions in different modalities, a parameter λ for the correlation (interaction) term, and a set of C class coefficients $\boldsymbol{\eta}_{1:C}$. Each coefficient $\boldsymbol{\eta_c}$ is a $M \times K$-vector of real values.

3.3 The Generative Process

For generating a unimodal document within a multi-modal document, we start with choosing the category label of the multi-modal document, then we generate the topic proportion of this unimodal document, next the topic of each word of this unimodal document is assigned, finally we generate the word according to the topic. This process is repeated until we draw the whole multi-modal document. Denoting Dirichlet and Multinomial distributions as "Dir" and "Multi", the whole generative process can be described as follows:

1. For each topic k in each modality m, sample the V_m dimensional topic-word distribution $\boldsymbol{\phi}_{mk} \sim \text{Dir}(\boldsymbol{\phi}|\beta_m)$
2. For each multi-modal document $\boldsymbol{x}_d$:

   a) Draw the M topic proportions $\boldsymbol{\pi}_{d1}, \cdots, \boldsymbol{\pi}_{dM}$ of $\boldsymbol{x}_d$ from the distribution

$$p(\boldsymbol{\pi}_{d1}, \cdots, \boldsymbol{\pi}_{dM}) = \frac{1}{Z}\exp(-\lambda \sum_{m_i=1}^{M}\sum_{m_j \neq m_i} f(\boldsymbol{\pi}_{dm_i}, \boldsymbol{\pi}_{dm_j})) \times \prod_{m=1}^{M}\text{Dir}(\boldsymbol{\pi}_{dm}|\alpha)$$

   where Z is a normalization constant. If some modalities are missing in a multi-modal document, we just adjust M to the accurate number of modalities in the multi-modal document. For example, in the retrieval stage, $M = 1$ and the first term (similarity term) is disregarded.

   b) For each word x_{dmn} in $\mathbf{x}_{dm}$ of the multi-modal document $\boldsymbol{x}_d$:

   Draw a topic $z_{dmn} \sim \text{Multi}(z|\boldsymbol{\pi}_{dm})$;

   Draw a word $x_{dmn} \sim \text{Multi}(x|\boldsymbol{\phi}_{mz_{dmn}})$.

   c) Draw class label $c_d|\boldsymbol{z}_{(1\ldots M)1:N_{dm}} \sim$ softmax $(\bar{\boldsymbol{z}}_d, \boldsymbol{\eta})$, where $\bar{\boldsymbol{z}}_d = [\bar{z}_{d1} \cdots \bar{z}_{dm} \cdots \bar{z}_{dM}]$ and $\bar{z}_{dm} = \frac{1}{N_{dm}}\sum_{n=1}^{N_{dm}} z_{dmn}$ denotes the empirical topic frequencies. The softmax function provides the following distribution,

$$p(c|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}) = \exp(\eta_c^T \bar{\boldsymbol{z}}_d)/\sum_{l=1}^{C}\exp(\eta_l^T \bar{\boldsymbol{z}}_{\boldsymbol{d}})$$

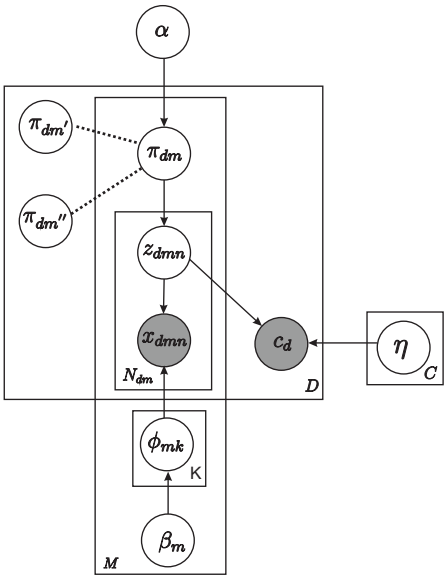

Figure 2: The graphical structure of our model. Given one multi-modal document $\boldsymbol{x}_d$, the dashed edges denote the topic proportion of each uni-modal document (π_{dm} here) is determined not only by the prior, but also by the topic proportions of other uni-modal documents (such as $\pi_{dm'}$ and $\pi_{dm''}$) in the same multi-modal document.

Given the parameters α, $\boldsymbol{\eta}$, $\boldsymbol{\beta}$ and λ, following the generative process above, we can now write the joint probability of a corpus with D multi-modal documents as follows:

$$\begin{aligned} &p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\phi}) \\ =&\frac{1}{Z}\prod_{d=1}^{D} p(c_d|\bar{z}, \eta)\exp(-\lambda \sum_{m_i=1}^{M}\sum_{m_j \neq m_i} f(\boldsymbol{\pi}_{dm_i}, \boldsymbol{\pi}_{dm_j})) \\ &\prod_{m=1}^{M}\Big(\prod_{k=1}^{K}\text{Dir}(\boldsymbol{\phi}_{mk}|\beta_m)\Big)\text{Dir}(\boldsymbol{\pi}_{dm}|\alpha) \\ &\Big(\prod_{n=1}^{N_{dm}}\text{Multi}(z_{dmn}|\pi_{dm})\text{Multi}(x|\boldsymbol{\phi}_{mz_{dmn}})\Big) \end{aligned} \tag{3}$$

3.4 Inference

In this section, we come to the learning strategy of our model. The hidden variables of the model are the topic-words distribution parameters $\boldsymbol{\phi}$, topic proportions $\boldsymbol{\pi}$, and the topic assignments $\boldsymbol{z}$ for the whole corpus. As with other topic models, the exact inference of the model is in general intractable. Some commonly used approximate methods are usually conducted in parameter inference as substitutes, such as variational inference[2], expectation propagation[20], or Gibbs sampling[11]. In this paper, we utilize the collapsed Gibbs sampling method for its simplicity and effectiveness.

With the joint probability of Eq.(3) for a corpus, we then conduct collapsed Gibbs sampling for inference. Gibbs sampling samples the topic assignment for one word based on its conditional probability with the observations and the topic assignments for the other words given, while the latent variables $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ are integrated out. We only perform Gibbs sampling on the $\boldsymbol{z}$ in this case. The probability of observations and sampling variable conditioned on the hyper-parameters can be calculated by integrating out the latent variables:

$$\begin{aligned}
&\int p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\phi}) d\boldsymbol{\pi} d\boldsymbol{\phi} \\
=&\frac{1}{Z} \prod_{d=1}^{D} \exp(\eta_{c_d}^T \bar{\boldsymbol{z}}_d) / \sum_{l=1}^{C} \exp(\eta_l^T \bar{\boldsymbol{z}}_d) \times \\
&\prod_{m=1}^{M} \frac{\Gamma(\sum_{k=1}^{K} \alpha)}{\prod_{k=1}^{K} \Gamma(\alpha)} \frac{\prod_{k=1}^{K} \Gamma(n_{dmk} + \alpha)}{\Gamma(\sum_{k=1}^{K} (n_{dmk} + \alpha))} \\
&\prod_{k=1}^{K} \frac{\Gamma(\sum_{v=1}^{V} \beta_m)}{\prod_{v=1}^{V} \Gamma(\beta_m)} \frac{\prod_{v=1}^{V} \Gamma(n_{mkv} + \beta_m)}{\Gamma(\sum_{v=1}^{V} (n_{mkv} + \beta_m))} \\
&\int \exp(-\lambda \sum_{m_i=1}^{M} \sum_{m_j \neq m_i} f(\boldsymbol{\pi}_{dm_i}, \boldsymbol{\pi}_{dm_j})) d\boldsymbol{\pi}
\end{aligned} \tag{4}$$

where n_{mkv} is the occurrence number of word v in topic k for modality m while n_{dmk} is the occurrence number of words assigned to topic k of $\mathbf{x}_{dm}$ in multi-modal document $\boldsymbol{x}_d$, respectively.

Noticing that the topic proportions for different modality data inside the multi-modal document are coupled, which makes the integration difficult. Inspired by [2], we introduce an empirical topic proportion instead of the original one to relax the coupled topic proportions of our model. We use the empirical model for the rest of our paper. We define the empirical topic proportion distribution given the topic assignments as follows:

$$\hat{\pi}_{dmk} = \frac{n_{dmk} + \alpha}{\sum_{k=1}^{K} (n_{dmk} + \alpha)} \tag{5}$$

where n_{dmk} is the occurrence number of topic k of $\mathbf{x}_{dm}$ in multi-modal document $\boldsymbol{x}_d$.

At this point, the markov chain updates of the topic assignment for one word based on the observations and the topic assignments for the other words can be derived as follows:

$$\begin{aligned}
&p(z_{dmn} = k | \mathcal{D}, \mathbf{z}_{-x_{dmn}}, \alpha, \boldsymbol{\beta}, \boldsymbol{\eta}, \lambda) \propto \\
&\frac{n_{dmk} + \alpha}{\sum_{k=1}^{K} (n_{dmk} + \alpha)} \times \frac{n_{mkx} + \beta_m}{\sum_{x=1}^{V} (n_{mkx} + \beta_m)} \\
&\times \frac{\exp(\eta_{c_d}^T \bar{\boldsymbol{z}}_{d,z=k}) / \sum_{l=1}^{C} \exp(\eta_l^T \bar{\boldsymbol{z}}_{d,z=k})}{\exp(\eta_{c_d}^T \bar{\boldsymbol{z}}_{d,-z}) / \sum_{l=1}^{C} \exp(\eta_l^T \bar{\boldsymbol{z}}_{d,-z})} \\
&\prod_{m_j \neq m} \exp\Big(\lambda f(\hat{\boldsymbol{\pi}}_{dm,-z}, \hat{\boldsymbol{\pi}}_{dm_j}) - \lambda f(\hat{\boldsymbol{\pi}}_{dm,z=k}, \hat{\boldsymbol{\pi}}_{dm_j})\Big)
\end{aligned} \tag{6}$$

where n_{mkx} is the occurrence number of word x_{dmn} in topic k for modality m while n_{dmk} is the occurrence number of words assigned to topic k of $\mathbf{x}_{dm}$ in multi-modal document $\boldsymbol{x}_d$ respectively, both exclude the current word. $\hat{\boldsymbol{\pi}}_{dm,-z}$ is the empirical topic distribution for $\mathbf{x}_{dm}$ in multi-modal document $\boldsymbol{x}_d$ excluding the current word, and $\hat{\boldsymbol{\pi}}_{dm,z=k}$ is the empirical topic distribution for modality m of multi-modal document d when the topic for the current word is k.

After getting $\boldsymbol{z}$, we can estimate $\boldsymbol{\pi}$ according to Eq.(5) and $\boldsymbol{\phi}$ as follows:

$$\hat{\phi}_{mkv} = \frac{n_{mkv} + \beta_m}{\sum_{x=1}^{V} (n_{mkv} + \beta_m)} \tag{7}$$

For determining the hyper-parameters, we choose to automatically update the hyper-parameters. We initialize α to be identical for all the dimensions and $\boldsymbol{\beta}$ to be identical for all the modalities, then we update them according to the training data as follows [19]:

$$\begin{aligned}
&\alpha \leftarrow \\
&\frac{\alpha \Big[\sum_{d=1}^{D} \sum_{m=1}^{M} \sum_{k=1}^{K} \Big(\Psi(n_{dmk} + \alpha) - \Psi(\alpha) \Big) \Big]}{K \Big[\sum_{d=1}^{D} \sum_{m=1}^{M} \Big(\Psi(\sum_{k=1}^{K} (n_{dmk} + \alpha)) - \Psi(\sum_{k=1}^{K} \alpha) \Big) \Big]}
\end{aligned} \tag{8}$$

$$\begin{aligned}
&\beta_m \leftarrow \\
&\frac{\beta_m \Big[\sum_{k=1}^{K} \sum_{v=1}^{V} \Big(\Psi(n_{mkv} + \beta_m) - \Psi(\beta_m) \Big) \Big]}{V_m \Big[\sum_{k=1}^{K} \Big(\Psi(\sum_{v=1}^{V} (n_{mkv} + \beta_m)) - \Psi(V_m \beta_m) \Big) \Big]}
\end{aligned} \tag{9}$$

where $\Psi(\cdot)$ is the digamma function $\Psi(x) = \frac{d}{dx} \ln \Gamma(x)$. For the updating of η, we resort to the general gradient descent method of softmax regression parameter update. We summarize the learning process in Algorithm 1.

3.5 Multi-modal Retrieval

After the training stage, we aim to verify whether the proposed M^{3}R can find the corresponding modality data given the query of certain modality and rank the corresponding modality data with the same category as the query forward than those with other categories, when it is applied to cross-media retrieval. Suppose we are given a query $\boldsymbol{x}$ which consists of N words, i.e. $\boldsymbol{x} = \{x_1, x_2, \cdots, x_N\}$ from p-th modality and aim to find its similar data from the q-th modality by cross-media retrieval. First we compute the topic proportion of all the q-th modality data, then we rank the q-th modality data using scores of the likelihood for the

Algorithm 1 The Learning Process of M^{3}R

Input: $\Omega = \{\boldsymbol{x}_d = (\mathbf{x}_{d1}, \cdots, \mathbf{x}_{dm}, \cdots, \mathbf{x}_{dM}, c_d)\}_{d=1}^{D}$
Randomly initialize topics $\boldsymbol{z}$, set values of document-topic occurrence number vector $\boldsymbol{ndmk} = \{\cdots, n_{dmk}, \cdots\}$ and topic-word occurrence number vector $\boldsymbol{nmkv} = \{\cdots, n_{mkv}, \cdots\}$ to 0
Initialize $\alpha, \beta_{1:M}$ and $\eta_{1:C}$
Calculate $\boldsymbol{nmkv}$ and $\boldsymbol{ndmk}$ according to Ω and $\boldsymbol{z}$
repeat
  for $d \leftarrow 1$ to D **do**
    for $m \leftarrow 1$ to M **do**
      for $n \leftarrow 1$ to N_{dm} **do**
        $k \leftarrow z_{dmn}$
        $n_{dmk} \leftarrow n_{dmn} - 1$
        $n_{mkv} \leftarrow n_{mkv} - 1$
        calculate $\bar{\boldsymbol{z}}_{d,-z}$ and $\hat{\boldsymbol{\pi}}_{dm,-z}$
        for $k \leftarrow 1$ to K **do**
          calculate $\bar{\boldsymbol{z}}_{d,z=k}$ and $\hat{\boldsymbol{\pi}}_{dm,z=k}$
          calculate $p(z_{dmn} = k)$ according to Eq.(6)
        end for
        sample a topic k of current word according to $\boldsymbol{p} = \{\cdots, p(k), \cdots\}$
        $n_{dmk} \leftarrow n_{dmn} + 1$
        $n_{mkv} \leftarrow n_{mkv} + 1$
        $z_{dmn} \leftarrow k$
      end for
    end for
  end for
  update $\alpha, \beta_{1:M}$ and $\eta_{1:C}$
until convergence or maximum iteration
calculate $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\phi}}$ according to Eq.(5) and Eq.(7).
return $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\phi}}$;
Output:
Latent representations for multi-modal documents $\hat{\boldsymbol{\pi}}$, the topic-words distributions $\hat{\boldsymbol{\phi}}$

query document which can be calculated as follows:

$$\begin{aligned} s_i = p(\boldsymbol{x}|\boldsymbol{\pi}_i) &= \prod_{n=1}^{N} p(x_n|\boldsymbol{\pi}_i) \\ &= \prod_{n=1}^{N}\sum_{k=1}^{K} p(x_n|z=k)p(z=k|\boldsymbol{\pi}_i), \end{aligned} \tag{10}$$

where $\boldsymbol{\pi}_i$ is the latent topic proportion for the i-th document of the q-th modality data, while $p(x_n|z=k)$ looks into the topic-word distribution $\hat{\boldsymbol{\phi}}_{pk}$ of p-th modality learned in the training stage. Note that the marginal probabilities $p(x_n|\boldsymbol{\pi}_i)$ can be pre-computed during learning time, so we use this method rather than the KL-divergence between the topic proportions of the two modalities to avoid time-consuming inference step for each query.

4. EXPERIMENTS

In this section, we evaluate the performance of our proposed method when applied to cross-media retrieval (specified to *image-query-texts* and *text-query-images*). We first introduce the data sets and evaluation criteria we adopted, and then elaborate parameter setting and tuning in our experiments. At last, we compare our method with other state-of-the-art algorithms and demonstrate the results.

4.1 Data Sets

One of our experimental data sets is the Wiki Text-Image data from Wikipedia feature articles [23]. Wiki Text-Image contains 2866 text-image pairs from ten different categories with each image associated with a text snippet describing the image. After SIFT features [17] are extracted, k-means clustering is conducted to obtain the representation of bag-of-visual-words (abbreviated as $BoVW$) [9] for each image. Each text is represented as a 5000-dimensional bag-of-textual-words (abbreviated as $BoTW$) vector by term frequency. In this dataset, there are on average 117.5 surrounding textual words for each image. We randomly choose 1/5 pairs of the dataset for test with the remaining pairs for training.

The other data set we used is the NUS-WIDE data set [6], which contains 133,208 images with 1000-dimensional tags and 81-dimensional concepts. Each image with its annotated tags in NUS-WIDE can be treated as a pair of image-text data while the concepts are regarded as the labels. We only select those pairs that belong to the 10 largest categories (concepts) with each pair has a unique category (concept). As a result, we get 26813 paired data samples and then we randomly choose 1/5 of them for test with the remaining pairs for training. We use the 500-dimension BoVW based on SIFT features for the representation of each image and 1000-dimensional tags for the representation of each text as the authors supplied. There are on average 7.7 surrounding textual words per image for this data set.

4.2 Evaluation Metrics

In the experiments, we aim to evaluate the quantitative performance of different methods in the following two aspects:(1) evaluation of the category relevance between query data and the retrieved results. A retrieved result is considered relevant if it belongs to the same category as the query data [24]. (2) investigation of cross-modal relevance for image-text pairs. The relevant retrieved result refers to the corresponding unique data object paired with the query [14]. The former aspect reveals the capability of learning discriminative cross-modal latent representations while the latter one can indicate the ability of learning correlated latent concepts. In this paper, we use three metrics regarding the two aspects as follows:

MAP : MAP is defined here to measure whether the retrieved data belong to the same category as the query (*relevant*) or does not belong to the same category (*irrelevant*). Given a query (one image or one text) and a set of its corresponding R retrieved results, the Average Precision is defined as

$$AP = \frac{1}{L}\sum_{r=1}^{R} prec(r)\delta(r), \tag{11}$$

where L is the number of relevant data in the retrieved set, $prec(r)$ represents the precision of the top r retrieved data (i.e., the ratio of the relevant results in the top r retrieved results). $\delta(r) = 1$ if the r-th retrieved data object is relevant to the query and $\delta(r) = 0$ otherwise. MAP is defined as the average AP of all the queries. Same as [30], we set $R = 50$ in the experiments.

Percentage: Since there is only one ground-truth match for each image/text regarding the above mentioned aspect (2), to evaluate the multi-modal performance we can resort to the position of the ground-truth textt/image in the

ranked list obtained. In general, one query image (or text) is considered correctly retrieved if its corresponding ground-truth text (or image) appears in the first t percent of the ranked list obtained by submitting the query according to [14]. Percentage is ratio of correctly retrieved query samples among all the query samples. t is set to equal to 0.2 in our experiments.

MRR: We also use Mean Reciprocal Rank (MRR) to evaluate the performances of different methods in our experiments regarding the position of the corresponding unique ground-truth paired with the query. The definition of Mean Reciprocal Rank (MRR) is as follows:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}, \tag{12}$$

where $|Q|$ is the number of query samples, $rank_i$ represents the position of the corresponding unique ground-truth paired with the ith query in the retrieved list.

4.3 Compared Methods

We make a comparison with three state-of-the-art supervised cross-modal methods as follows:

- **SLDA-KL**: sLDA [3] is first individually employed to obtain the latent representations (i.e.,topic proportions) of each image or text. When one image (text) is submitted, we get its nearest neighboring image (text), take the corresponding paired text (image), then obtain the ranked list of retrieved texts (images) via the symmetric KL-divergence between the paired text (image) and the retrieved texts (images) in term of their topic proportions.
- **Generalized Multiview Analysis (GMA)**[24]: GMA is a supervised method in cross-modal retrieval which utilizes both pair-wised and label information of multi-modal data. As stated by authors , GMA is a supervised kernelizable extension of CCA and maps data in different modality spaces to a single subspace.
- **Supervised coupled dictionary learning with gro -up structures for Multi-Modal retrieval (SliM^2)** [31]: SliM^2 is a supervised dictionary learning approach with group structures utilizing the class information to jointly learn discriminative multi-modal dictionaries as well as mapping functions between different modalities for multi-modal retrieval.

4.4 Parameter Tuning

As indicated before, the parameters of our model are the hyper-parameter α for topic proportions, a set of M hyper-parameter $\beta_{1:M}$ for topic-words distributions in different modalities, the parameter of potential function λ , and a set of C class coefficients $\eta_{1:C}$. The hyper-parameters α and $\beta_{1:M}$ can be learned directly from the training data using Eq.(8) and Eq.(9). The initial values of α and $\beta_{1:M}$ are set to the value commonly used in text modeling[11]. The class coefficients $\eta_{1:C}$ can also be learned in the training process but their initial values may influence the results. We set the initial values of $\eta_{1:C}$ same for all categories and all dimensions making it a variable base on a single value η. Then we perform grid-search for λ and η. The setting of λ and η is 500 and 0.6 on Wiki data set while 40 and 6 on NUS-WIDE data set, respectively.

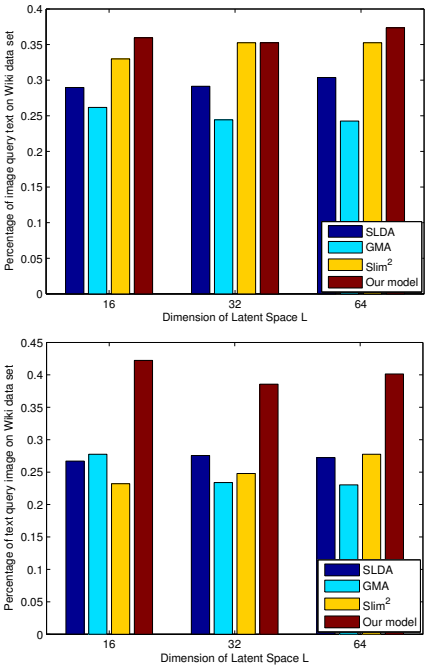

Figure 3: The Performance comparison of different methods according to Percentage metric with varied dimension of latent space on Wiki data set. Top is for image query text while bottom is for text query image.

4.5 Results

4.5.1 *Performance Comparison*

The performances of each algorithm are shown in Table 1 and Table 2 in terms of MAP, Percentage and MRR.

Here, two kinds of cross-media retrieval tasks are evaluated: (1) submitting one image to retrieve texts (image-query-texts), (2) submitting one text to retrieve images (text-query-images). From Table 1 and Table 2, we can see our proposed method achieves the best average performance in almost all of metrics for two data sets. Compared to the second best methods, we gain relatively 7.8%, 7.8% and 0.84% average performance improvement in terms of MAP, Percentage and MRR respectively on wiki data set while gain relatively 7.8% and 33% average performance improvement in terms of Percentage and MRR respectively on NUS-WIDE data set.

For NUS-WIDE data set, GMA and SliM^2 performs better than M^{3}R in the average performance of MAP metric. The reason is probably that one image in NUS-WIDE is associated with about only seven words in average, which restricts the power of our proposed algorithm. Moreover, for cross-media retrieval, the Percentage and MRR metrics are considered as more accurate indicators of true performance than the MAP metric. Since the underlying motivation of cross-media retrieval is to learn the correlations of data objects with different modalities and perform similarity search across different modalities, the Percentage and MRR metrics which evaluate the performance according to the position of the corresponding unique ground-truth paired with the query is more close to the goal of cross-modal retrieval than

Table 1: The performance comparison on the Wiki data set. The results shown in boldface are best results.

Wiki	Image query Texts			Text query Images			Average Performance		
	MAP	Percentage	MRR	MAP	Percentage	MRR	MAP	Percentage	MRR
SLDA	0.2116	0.3037	0.0369	0.2146	0.2723	0.0241	0.2131	0.2880	0.0305
GMA	0.2074	0.2792	0.0153	0.2542	0.2827	0.0208	0.2308	0.2810	0.0181
SliM^2	**0.2548**	**0.4084**	**0.0454**	0.2021	0.3106	0.0261	0.2285	0.3595	0.0358
M^{3}R	0.2298	0.3735	0.0321	**0.2677**	**0.4014**	**0.0400**	**0.2488**	**0.3875**	**0.0361**

Table 2: The performance comparison on the NUS-WIDE data set. The results shown in boldface are best results.

Wiki	Image query Texts			Text query Images			Average Performance		
	MAP	Percentage	MRR	MAP	Percentage	MRR	MAP	Percentage	MRR
SLDA	0.1976	0.2396	0.0022	0.2078	0.2640	0.0035	0.2027	0.2518	0.0029
GMA	0.2202	0.3765	0.0043	**0.4199**	0.3752	0.0045	**0.3201**	0.3759	0.0044
SliM^2	**0.3154**	**0.4639**	0.0057	0.2924	0.3877	0.0045	0.3039	0.4258	0.0051
M^{3}R	0.2445	0.3896	**0.0065**	0.3044	**0.4853**	**0.0071**	0.2742	**0.4375**	**0.0068**

the MAP metric evaluating the performance according to the labels.

For text-query-images task,our proposed method also achieves the best performance in almost all of metrics for two data sets except for NUS data set measured by MAP, which can be explained by the aforementioned same reason.

For image-query-texts, SliM^2 is superior except for NUS-WIDE data set measured by MRR. SliM^2 performs cross-media retrieval by the minimization of reconstruction error through coupled multi-modal dictionary learning. The performance by SliM^2 is noticeably unbalanced. The reason is that SliM^2 is prone to be over-fitting over one modality data and at the same time be under-fitting over another modality. Mathematically speaking, the minimization of reconstruction error in image-query-text retrieval direction will unavoidably increase the reconstruction error in text-query-image direction. However, our approach achieves an attractive balance performance of cross-media retrieval due to its intrinsic power of learning a unified space (e.g., topic space) via multi-modal mutual topic reinforce modeling, in which the pair-correspondence of images and text documents ensure an equal contribution to the learned metric, which is essential in practical interest.

We also do performance comparison of different methods in best parameter setting according to Percentage metric with varied dimension of latent space on Wiki data set as illustrate on Figure 3. The results on top row is for image query text while results on bottom row is for text query image. From Figure 3, we can see our method outperform all the other methods and are relatively stable to the dimension of latent space.

4.5.2 Results Demonstration

As one of topic model based method, our model has the advantage to mine the interpretable latent topics. Since textual topics bear more explicit semantics and visual topics can be hardly illustrated, we only demonstrate the latent topics in text modality. We select some topics (indicated by their corresponding topical words) that apparently have some meanings related to the categories and illustrate them in Table 3 with each topic assigned to its most relevant category. The table indicates that our method is able to discover the latent topics. For example, the topic related to music consists of words like "Punk","video" and "bands" which all reveals the semantics of music from different aspects.

Figure 4 illustrates one example of image query text and one example of text query image over Wiki image-text data set. The retrieved results by different methods are compared.

The query image and the query text come from a paired document from the "geography" category. The underlined semantics are mainly about "Fanno creek", "park" and "trail". For the example of image-query-texts, we use the corresponding images together with keywords of the retrieved texts to demonstrate the results. Though all of retrieved texts (and their corresponding images) come from the "geography" category same as the query image but the third retrieved text of GMA is irrelevant to the query image in semantics. For the example of text-query-images, The retrieved images by SliM^2 and our method (M^{3}R) all come from "geography" category, while there are images coming from other categories rather than "geography" by the other two methods. Retrieved results irrelevant to the query in semantics are marked with red color. From the results, we can observe that our proposed method are more semantically correlated with the query both for images and texts.

5. CONCLUSION

M^{3}R is proposed in this paper for multi-modal data understanding. M^{3}R is a supervised multi-modal mutual topic reinforce modeling (M^{3}R) approach which can learn correlated yet discriminative latent representations for multi-modal data by introducing of topic interaction and label information. We have demonstrated the superior performance of M^{3}R in terms of several metrics on two data sets for cross-modal retrieval. M^{3}R gains interpretable latent representations for multi-modal retrieval and is effective for cross-modal retrieval in terms of MAP, Percentage and MRR.

6. ACKNOWLEDGEMENTS

This work is supported in part by National Basic Research Program of China (2012CB316400), NSFC (NO. 61103099, 61105074), 863 program (2012AA012505), the Fundamental Research Funds for the Central Universities and Chinese Knowledge Center of Engineering Science and Technology

Table 3: Exemplar topics from the Wiki dataset. We assign each learned topic to its most probable category. Topic words are sorted by their importance values in the descending order.

Category	Topic Words
Literature	Literature Poets Capture Inspired Tradition God Admiral Volume Duke Piece
Media	Episode Davis Filming Cast Murray Hollywood Drama Actors Angeles Round
Music	Punk Video Studio Bands Award Albums Tracks Billboard POP Guitar
Sport	Football Cup Ball Marshall Conference Players Wales Professional Competition Africa
Warfare	Infantry Battalion Regiment Corps Units Artillery Marine Lieutenant Brigade Ridge

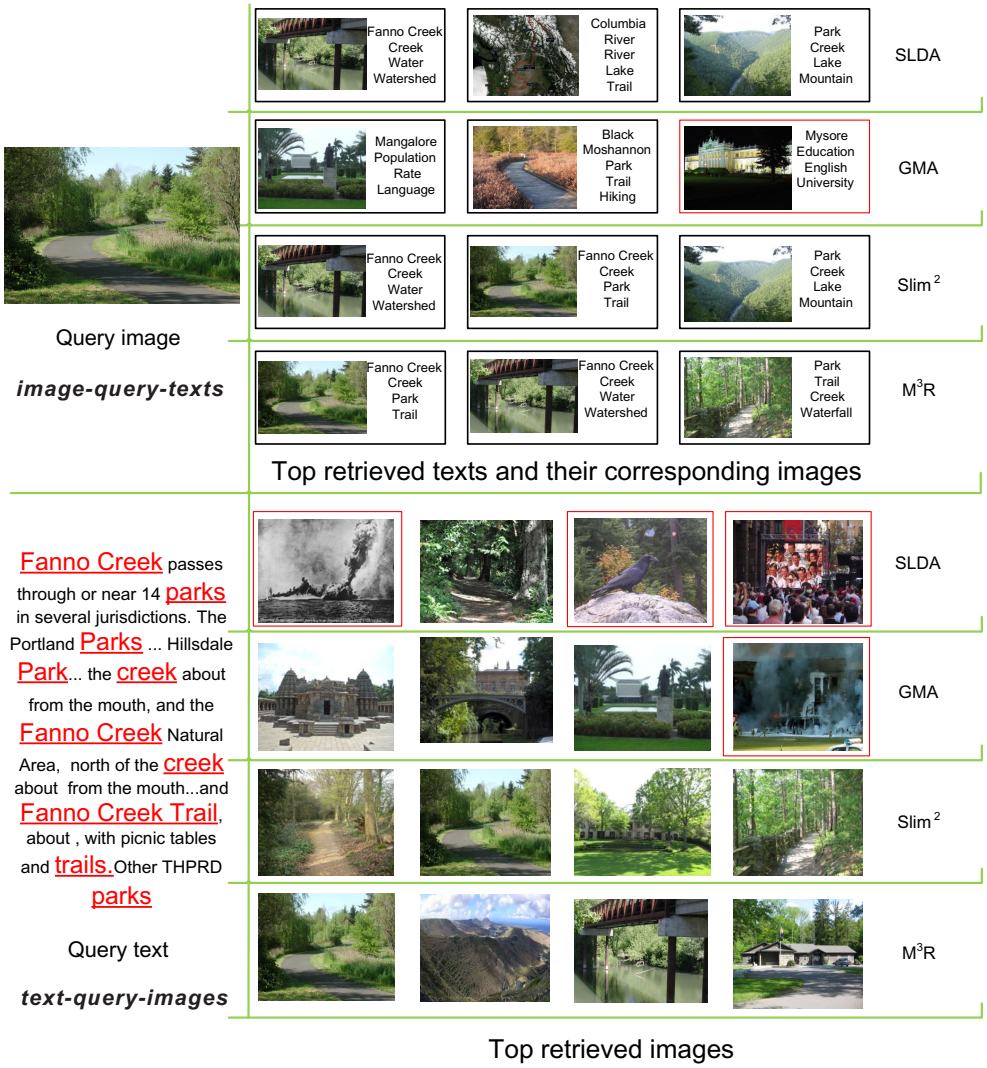

Figure 4: Two examples of image-query-texts and text-query-images over Wiki data set by four algorithms. For image-query-texts, we use the corresponding images of retrieved texts and key words of the texts to demonstrate the results. The query image and text are paired data from the "geography" category. Retrieved results irrelevant to the query in semantics are marked with red color.

(CKCEST) and Program for New Century Excellent Talents in University.

7. REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proceedings of International Conference on Machine Learning*, pages 1247–1255, 2013.

[2] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[3] D. M. Blei and J. D. McAuliffe. Supervised topic models. *arXiv preprint arXiv:1003.0783*, 2010.

[4] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality

metric learning using similarity-sensitive hashing. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3594–3601, 2010.

[5] N. Chen, J. Zhu, F. Sun, and E. P. Xing. Large-margin predictive latent subspace learning for multiview data analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(12):2365–2378, 2012.

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

[7] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *Transactions of Pattern Analysis and Machine Intelligence*, 36(3):521–535, March 2014.

[8] J. Farquhar, D. Hardoon, H. Meng, J. S. Shawe-taylor, and S. Szedmak. Two view learning: Svm-2k, theory and practice. In *Advances in neural information processing systems*, pages 355–362, 2005.

[9] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Computer Vision and Pattern Recognition Workshop, IEEE Conference on*, pages 178–178, 2004.

[10] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[11] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[12] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[13] Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. *Advances in Neural Information Processing Systems*, 23:982–990, 2010.

[14] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 2407–2414, 2011.

[15] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *Proceedings of international joint conference on Artificial Intelligence*, pages 1360–1365, 2011.

[16] R. Liao, J. Zhu, and Z. Qin. Nonparametric bayesian upstream supervised multi-modal topic models. In *Proceedings of the ACM international conference on Web search and data mining*, pages 493–502, 2014.

[17] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the IEEE international conference on Computer vision*, volume 2, pages 1150–1157, 1999.

[18] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang. A low rank structural large margin method for cross-modal ranking. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 433–442, 2013.

[19] T. Minka. Estimating a dirichlet distribution. *Technical report, MIT*, 2000.

[20] T. P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[21] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 689–696, 2011.

[22] D. Putthividhy, H. Attias, and S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 3408–3415, 2010.

[23] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, pages 251–260, 2010.

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

[25] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2231–2239, 2012.

[26] S. Virtanen, Y. Jia, A. Klami, and T. Darrell. Factorized multi-modal topic model. *arXiv preprint arXiv:1210.4920*, 2012.

[27] C. Wang, D. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 1903–1910, 2009.

[28] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang. Cross-media semantic representation via bi-directional learning to rank. In *Proceedings of ACM international conference on Multimedia*, pages 877–886, 2013.

[29] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang. Sparse multi modal hashing. *IEEE Trans. Multimedia*, 2013. doi: 10.1109/TMM.2013.2291214.

[30] Y. Zhen and D.-Y. Yeung. A probabilistic model for multimodal hash function learning. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 940–948, 2012.

[31] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *Proceedings of Conference on Artificial Intelligence*, pages 1070–1076, 2013.