ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-399

A Novel Cross-Modal Topic Correlation Model for Cross-Media Retrieval

Yong Cheng, Fei Huang, Cheng Jin, Yuejie Zhang[1] and **Tao Zhang**[2]

Abstract. A novel cross-modal topic correlation model CMTCM is developed in this paper to facilitate more effective cross-modal analysis and cross-media retrieval for large-scale multimodal document collections. It can be modeled as a cross-modal topic correlation model which explores the inter-related correlation distribution over the deep representations of multimodal documents. It integrates the deep multimodal document representation, relational topic correlation modeling, and cross-modal topic correlation learning, which aims to characterize the correlations between the heterogeneous topic distributions of inter-related visual images and semantic texts, and measure their association degree more precisely. Very positive results were obtained in our experiments using a large quantity of public data.

1 INTRODUCTION

With the explosive growth of multimodal documents on the Web, how to seamlessly handle the complex structures of multimodal documents to achieve more effective cross-media retrieval has become an important research focus [1]. Usually, a multimodal document is exhibited in a form with different modalities (i.e., both visual and semantic), such as a web image with user defined annotation tags/narrative text descriptions, or a news article with paired visual images and textual illustrations. However, due to the semantic gap, there may be significant differences and independence among visual images and semantic texts for multimodal documents, which leads to the huge difficulty and high uncertainty in making full use of the corresponding relationships between the visual features (in images) and semantic features (in descriptions) [2]. Thus integrating multimodal information sources involved in multimodal documents to enable multimodal topic correlation has been the critical component for supporting cross-media retrieval.

Although multimodal topic correlation has been extensively studied for cross-media retrieval since recent years [3] [4], it still remains the necessity of optimal solutions and three inter-related issues should be addressed simultaneously: 1) valid construction and discovery of valuable document element to characterize visual images and textual descriptions for multimodal document representation; 2) reasonable topic correlation modeling to identify better correlations between visual images and textual descriptions of multimodal documents; and 3) cross-modal topic correlation learning to optimize the objective measurement for inter-related image-description correlations. To address the first issue, it is very important to explore the optimal document element that can achieve more precise and comprehensive visual and semantic feature expression for multimodal documents. To address the second issue, it is critical to establish a robust probabilistic topic model to maximize the likelihood of the observed multimodal documents in terms of the involved latent topics. To address the third issue, it is significant to map the attributes of different modalities into a common embedding space to efficiently maximize their statistical dependency and correlation.

Based on the above observations, a novel Cross-Modal Topic Correlation Model (CMTCM) is developed in this paper to facilitate more effective cross-media retrieval for large-scale multimodal document collections. Our scheme significantly differs from other earlier work as follows. a) The document element of "*deep word*" is created for encoding both the visual features in visual images (i.e., deep visual word) and the semantic features in textual descriptions (i.e., deep textual word) to obtain better deep multimodal document representation. Compared to the traditional visual word and textual word, the deep visual word that is closer to the visual image semantics can alleviate the problem of semantic gap to a great degree, and the deep textual word that integrates various relationship information among textual words can be more representative for expressing the specific semantics of textual descriptions. b) A relational topic correlation modeling scheme is designed to achieve more precise characterization of the inter-related multimodal correlations between visual images and textual descriptions, in which the topic generation and multimodal correlation learning are fused together to break the limitation of topic consistency in the traditional topic modeling. Different topic sets can be generated for different modality information, and at the same time the heterologous topic information from other modalities can be integrated in the topic generation process for one modality. c) An efficient learning mechanism for cross-modal topic correlation is established to achieve the objective decision-making for multimodal correlation, in which the deep topic features for different modalities are particularly mapped into a common space for mining their inter-related topic correlation. Compared to the traditional topic correlation learning strategies, the cross-modal topic correlation learning considers the heterologous property of topic for different modalities, and utilizes the specific mapping function to learn the topic correlation form different modalities. d) A new cross-modal topic correlation model is built by integrating the above deep multimodal document representation, relational topic correlation modeling, and cross-modal topic correlation learning, which can not

[1] School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China, email: {13110240027; 15210240036; jc; yjzhang}@fudan.edu.cn

[2] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China, email: taozhang@mail.shufe.edu.cn

only enable cross-media retrieval users to present on the multimodal query panel whatever they imagine in their mind, but also obtain the most relevant multimodal documents to the original query intention.

How to integrate multimodal information sources in topic correlation measurement for multimodal documents is an open issue, because it is hard to provide a common base for the correlations among multimodal documents because of the semantic gap. The main contribution of our work is that we effectively apply deep representation, relational modeling and cross-modal learning to enable cross-modal topic correlation model, which has provided a more reasonable base for us to integrate visual correlation with semantic correlation by determining an optimal correlated projection space. Such a cross-modal topic correlation model can be treated as an inter-related correlation distribution over deep representations of multimodal documents, in which the most important is to create more effective multimodal image-description topic correlation and measure what degree they are correlated. Our experiments on a large number of public data have obtained very positive results.

2 RELATED WORK

Topic correlation modeling is not a novel task, but has been the subject of extensive research in areas such as cross-media retrieval for large-scale multimodal documents. Earlier research placed the main emphasis on directly exploiting low-level visual features and simple semantic features to explore the image-description topic correlation [5] [6]. However, because of considering only limited shallow-level visual and textual implication in multimodal documents, such methods often demonstrate a poor performance. Recently, closer attention has been given to the methods that rely on cross-modal correlation mining with both deep-level visual and semantic features, that is, finding the high-level multimodal correlation to associate together visually and semantically correlated images and descriptions [7] [8]. Thus, there has been increasing research interests in leveraging the deep implication from multiple information sources and learning the cross-modal topic correlation for multimodal documents to satisfy more rigid requirements on the precision and efficiency for cross-media retrieval.

In recent years, there is some related research work for modeling the topic correlation between visual contents and semantic descriptions in multimodal documents. Blei et al. (2003) built a set of increasingly sophisticated models for a database of annotated images, culminating in correspondence latent Dirichlet allocation (Corr-LDA), a model that found conditional relationships between latent variable representations of sets of image regions and sets of words [9]. Wang et al. (2009) developed a probabilistic model that simultaneously learned the salient patterns among images that were predictive of their class labels and annotation terms, in which the supervised topic modeling (sLDA) was extended to classification problems and a probabilistic model of image annotation was embedded into the resulting supervised topic model [10]. Putthividhya et al. (2010) presented the topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA), a novel statistical topic model for the task of image and video annotation, which lay a latent variable regression approach to capture correlations between image or video features and annotation texts [11]. Rasiwasia et al. (2010) studied the problem of joint modeling for the text and image components of multimedia documents, in which the text component was represented as a sample from a hidden topic model learned with latent Dirichlet allocation, and images were represented as bags of visual (SIFT) features [12]. Nguyen et al. (2013) proposed a novel method for image annotation based on combining feature-word distributions which mapped from the visual space to the word space, and word-topic distributions which formed a structure to capture label relationships for annotation [13]. Niu et al. (2014) addressed the problem of recognizing images with weakly annotated text tags, in which the text tags were first encoded as the relations among the images, and then a semi-supervised relational topic model (ss-RTM) was proposed to explicitly model the image contents and their relations [14]. Wang et al. (2014) proposed a supervised multimodal mutual topic reinforce modeling (M3R) approach, which sought to build a joint cross-modal probabilistic graphical model for discovering mutually consistent semantic topics via the appropriate interactions between model factors (e.g., categories, latent topics and observed multi-modal data) [3]. Zheng et al. (2014) considered the application of DocNADE to deal with multimodal data in computer vision, and proposed a supervised variant of DocNADE (SupDocNADE), which can be used to model the joint distribution over an image's visual words, annotation words and class label [15]. Tian et al. (2015) presented a novel model that utilized the rich surrounding texts of images to perform image annotation, in which the words that described the salient objects in images were extracted by integrating the text analysis, and a new probabilistic topic model was built to jointly model image features, extracted words and surrounding text [16]. Wu et al. (2015) proposed a cross-modal learning to the rank approach called CML^{2}R to discover the latent joint representation of multimodal data, and they assumed that the correlations between the multimodal data were captured in terms of topics, and used a list-wise ranking manner to learn the discriminative ranking function [17]. Chen et al. (2015) addressed the image-text correspondence modeling gap by introducing Visual-Emotional LDA (VELDA), a novel topic model that captured image-text correlations through multiple evidence sources (namely, visual and emotional, yielding the method's namesake) for cross-modality image retrieval [4].

Unfortunately, all these existing approaches have not yet provided good solutions for the following crucial issues, which are tightly coupled with each other. **(1) Discovering Deep Information for Multimodal Document Representation** -- Most existing methods focus on exploiting the regular feature description of visual and semantic exhibition in multimodal documents, and do not consider the deep feature information in different modalities of the same multimodal document. This will result in a serious information loss problem for global visual semantics or inherent semantic associations, and forms the insufficient feature descriptions for multimodal documents. With the deep exploration of visual and semantic appearances for multimodal documents, it appears such a discovery mechanism can mitigate the lack of deep visual and semantic feature information. According to our best knowledge, no existing research has made full use of such both deep visual and semantic information to achieve more accurate multimodal document representation for topic correlation modeling. **(2) Relational Topic Correlation Modeling in Deep Level** -- Most existing work concerns finding the best topic correlation for multimodal documents underlying such an assumption that the latent topic sets for visual image and textual description of each multimodal document should be consistent, that is, the heterologous topic information is not considered for different modality information in the same multimodal document. However, for a multimodal document, the inter-related information in different modalities may not be completely incoordinate. With such an over-

strong assumption, the obvious noises will be introduced into the topic correlation measure between different modality information, and meanwhile the deep inclusion in multimodal content cannot be fully utilized. It's a very significant way to fuse multimodal topic feature information in the deep level, set a novel topic generation pattern, and form an optimal relational topic correlation modeling scheme under more reasonable assumptions. **(3) Cross-Modal Correlation Learning with Deep Topic Features** -- Most existing approaches concentrate on directly matching the topic distributions in different modalities to capture the inter-related correlation between visual image and textual description of the same multimodal document. It's due to a simple intuition that the more similar the topic distributions of different modalities are, the higher correlation they have. However, such a straightforward correlation learning strategy may lead to the imprecise correlation evaluation without the in-depth consideration of the deep topic features and the topic heterogeneity in different modalities. Cross-modal correlation learning could provide helpful hints on mining multimodal information for topic correlation modeling. Establishing multimodal associations between deep visual and semantic topic features may shed light on the in-depth understanding for multimodal documents. Thus, the explicit learning of cross-modal correlations between deep visual and semantic topic features becomes very important. From the viewpoint of multimodal document exploitation, it's a significant way to combine both deep visual and semantic topic abstractions for images and descriptions in a joint space and establish an effective cross-modal joint learning mechanism.

To tackle the above obstacles, we have developed a novel framework by integrating the deep multimodal document representation (i.e., mining the valuable multimodal feature information in the deep level), the relational topic correlation modeling (i.e., bridging the semantic gap between inter-related visual contents and semantic descriptions), and the cross-modal correlation learning (i.e., fusing the optimization mapping strategy to obtain more accurate multimodal topic correlation). In our study, we realize that a multimodal document usually appears with multiple correlated visual and semantic words and spans multimodal associations in both deep visual and semantic word levels. Our cross-modal topic correlation model aims at exploring the deep multimodal correlations involved in images and their descriptions to improve the reasoning ability for topic correlation. It's a new attempt on exploiting such deep feature representation, modeling and learning optimization strategies on cross-modal topic correlation model to facilitate cross-media retrieval.

3 DEEP MULTIMODAL DOCUMENT REPRESENTATION

The multimodal information is the significant expression and exhibition for multimodal document content, that is, the visual image and textual description in each multimodal document. To acquire the cross-modal topic correlation between the visual image and textual description in each multimodal document, the optimal basic element for multimodal document representation should be detected and represented more precisely. Thus the deep multimodal document representation is implemented to exploit multiple document elements in the deep level (i.e., deep visual word and deep textual word) and explore the multimodal associations between deep visual property elements and deep semantic expression elements, as shown in Figure 1.

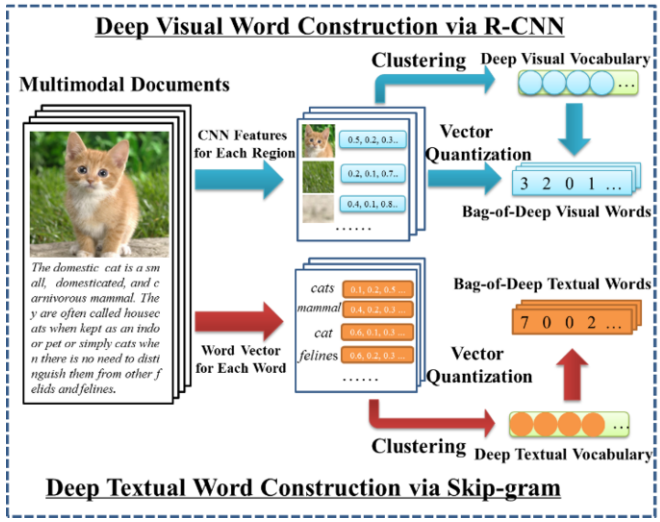


Figure 1. An instantiation for Deep Multimodal Document Representation.

3.1 Deep Visual Word Construction via R-CNN

The Region-based Convolution Neural Network (R-CNN) is a method combining region proposals with CNNs [18], which means extracting all the CNN features for all the regions in the image and is widely used in the field of computer vision [19] [20] [21]. R-CNN first uses selective search methods to generate possible object locations for each image in terms of image regions, and then extracts the feature vector from each region proposal based on CNN. For this purpose, each image region is converted to a fixed pixel size of 227×227, and all the features are computed through a network with five convolutional layers and two fully connected layers. The advantage of R-CNN is that the visual features extracted via CNN are closer to the image semantics, which can alleviate the problem of semantic gap to a certain degree. Furthermore, the regions contain the important spatial information in the image, and the visual words constructed by R-CNN can better represent the deep image contents. Thus we leverage R-CNN to construct the deep visual words for representing the deep visual semantic properties in images. Firstly, each image is represented in the form of bag-of-regions based on R-CNN, and each region can be viewed as a visual word. Since each region is represented as a feature vector, we use the Vector Quantization (VQ) method [22] to project the higher dimensional features into a sparse presentation. We construct the deep visual word vocabulary by clustering all the region features into a fixed-number of classes, and then project all the regions in the same image into the deep visual word vocabulary. Finally, each image can be represented in the form of bag-of-deep-visual-words. Compared to the traditional visual word descriptors like SIFT, the main advantage of deep visual word is that it can compute visual features with a hierarchical and multi-stage process, which is more informative and effective for visual recognition than using just low-level and superficial visual features.

3.2 Deep Textual Word Construction via Skip-gram

The Skip-gram model is an efficient method to learn the distributed representations of textual words in a vector space from large amounts of unstructured text data [23], which has achieved better performance in a wide range of natural language processing tasks [20] [24]. Its training objective aims at learning deep word vector

representations that are good at predicting the nearby textual words, which can capture more precise syntactic and semantic relationships among textual words and group similar textual words together. Compared to other learning methods for textual word vector, the advantage of Skip-gram is that the training process is extremely efficient for massive text data since it does not involve dense matrix multiplications. Thus we leverage the Skip-gram model to construct the deep textual word for better representing the deep textual semantic properties in textual descriptions.

Let D^T be the textual description part of the whole multimodal document corpus, W denotes all the raw textual words in D^T, and V is the textual word vocabulary. For each textual word w in W, I_w and O_w are the input and output vector representations for w, $Context(w)$ represents the nearby textual words of w, here the context window size is set as 5. We define the set of all the input and output vectors for each textual word as a long vector $\omega \in R^{2*|V|*dim}$ and dim is the dimension number of the input or output vector, thus the objective function of Skip-gram can be described as:

$$\begin{aligned} \mathrm{BSG}(\omega) &= \arg\max_{\omega} \frac{1}{|W|}\sum_{i=1}^{|W|}\sum_{j=1}^{Context(w_i)} \log P(w_j|w_i) \\ &= \arg\max_{\omega} \frac{1}{|W|}\sum_{i=1}^{|W|}\sum_{j=1}^{Context(\mathrm{w_i})} \frac{\exp(O_{w_j}\cdot I_{w_i})}{\sum_{k=1}^{|V|}\exp(O_{w_k}\cdot I_{w_i})} \end{aligned} \quad (1)$$

Since the computing cost is extremely high for the standard softmax formulation of Skip-gram, the Negative Sampling is utilized to compute $\log P(w_j|w_i)$ approximatively.

$$\log P(w_j|w_i) = \log\sigma\left(O_{w_j}\cdot I_{w_i}\right) + \sum_{k=1}^{m} E_{w_k\sim P(w)}\log\sigma\left(O_{w_j}\cdot I_{w_i}\right) \quad (2)$$

where $\sigma(\cdot)$ is the sigmoid function; and m is the number of negative samples, each sample is drawn from the noise distribution $P(w)$ based on the textual word frequency. With the learned textual word vector representations, we quantize all these textual word vectors by using the K-means clustering to obtain a discrete set of text terms, which form the new deep textual word vocabulary. Since the textual word vector considers the relationships between textual words, the clustering algorithm allocates the textual words with the high semantic similarity to one new textual word, and all these new textual words constitute the deep textual word vocabulary. Thus each description can be represented in the form of bag-of-deep textual words. Compared to the raw textual word, the main advantage of deep textual word is the consideration of the semantic relationships among raw textual words, which makes the deep textual words more representative to describe textual contents.

4 RELATIONAL TOPIC CORRELATION MODELING VIA CROSS-MODAL LEARNING

The general consideration for multimodal image-description topic correlation is that the topic distribution for the visual appearances of visual image is heterologous but related with the distribution for the semantic exhibitions of textual description. To achieve more precise topic correlation of multimodal documents, it's very useful to establish an effective topic correlation modeling pattern for evaluating the intrinsic image-description association degree among all the multimodal documents in the whole database. Thus a relational topic correlation model is built to fine measure the image-description association, in which the cross-modal learning mechanism is especially conducted over multimodal topic distributions to facilitate more refined evaluation for multimodal topic correlation, as shown in Figure 2.

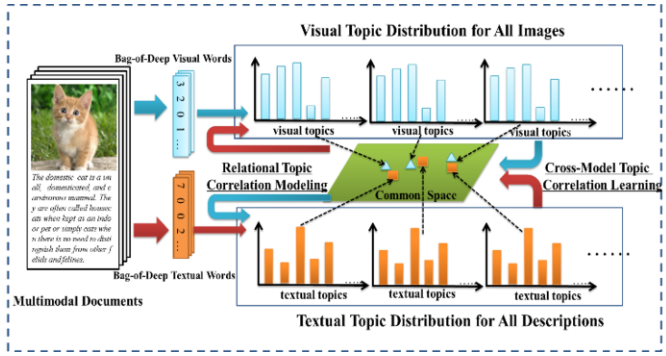


Figure 2. An instantiation for Topic Correlation Modeling.

4.1 Relational Topic Correlation Modeling

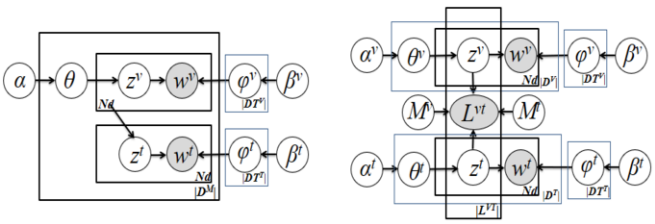
Our main purpose aims at building a joint probabilistic model to maximize the likelihood of the observed multimodal documents. We assume that each deep word is generated from one latent topic, and such a topic is derived from a multinomial distribution over all the topics. The major difference of our modeling is that the latent topic sets for different modalities are different, that is, the number and meaning of the topics from different modalities are different. Such an intuitive perception is because that in many cases the visual image and the textual description in the same multimodal document are semantically correlated but not consistent on latent topics.

We consider splitting the multimodal document collection D^M into three parts, that is, the visual image set D^V, the textual description set D^T, and the linkage set L^{VT} which indicates the multimodal image-description associations. D^V is composed of the deep visual word set DW^V and DV^V is the deep visual vocabulary, while D^T is composed of the deep textual word set DW^T and DV^T is the deep textual vocabulary. For $l^{vt}\in L^{VT}$, l^{vt}=1 means that the visual image $d^v \in D^V$ and the textual description $d^t \in D^T$ is relevant, otherwise irrelevant when l^{vt}=0. Given DT^V is the visual topic set, DT^T is the textual topic set, α and β are two hyper-parameters for the topic proportion and the topic-deep-word distribution, θ, θ^v and θ^t are the topic distributions for each multimodal document d, its visual image d^v and its textual description d^t, φ is the topic-deep-word distribution for each topic, z is the actual deep-word-related topics generated from θ, $Dir(\ )$ and $Mult(\ )$ denotes the Dirichlet and multinomial distribution, n denotes the n^{th} deep word, and N_d is the total number of deep words in the multimodal document d, the basic framework of our modeling is shown as follows.

1. For each visual topic $t^v \in DT^V$, sample the topic-deep-visual-word distribution over the deep visual vocabulary, i.e., $\varphi^v{}_{t^v}\sim Dir(\varphi^v|\beta^v)$.
2. For each textual topic $t^t \in DT^T$, sample the topic-deep-textual-word distribution over the deep textual vocabulary, i.e., $\varphi^t{}_{t^t}\sim Dir(\varphi^t|\beta^t)$.
3. For each visual image $d^v \in D^V$:
   (a) Sample the visual topic distribution $\theta^v{}_{d^v}\sim Dir(\theta^v|\alpha^v)$.
   (b) For each deep visual word $w^v{}_{d^v,n^v}$:
      i. Sample the visual topic assignment $z^v{}_{d^v,n^v}\sim Mult(\theta^v{}_{d^v})$.
      ii. Sample the deep visual word $w^v{}_{d^v,n^v}\sim Mult(\varphi^v{}_{z^v{}_{d^v,n^v}})$.
4. For each textual description $d^t \in D^T$:
   (a) Sample the textual topic distribution $\theta^t{}_{d^t}\sim Dir(\theta^t|\alpha^t)$.
   (b) For each deep textual word $w^t{}_{d^t,n^t}$:
      i. Sample the textual topic assignment $z^t{}_{d^t,n^t}\sim Mult(\theta^t{}_{d^t})$.
      ii. Sample the deep textual word $w^t{}_{d^t,n^t}\sim Mult(\varphi^t{}_{z^t{}_{d^t,n^t}})$.
5. For each linkage $l^{vt}\in L^{VT}$ for the relationship between the visual image d^v and the textual description d^t:

(a) Sample the linkage indicator l^{vt}~$TCor(l^{vt}|\overline{z_{d^v}}, \overline{z_{d^t}}, M^v, M^t)$, where $\overline{z_{d^v}}$ and $\overline{z_{d^t}}$ are the empirical topic frequencies for d^v and d^t, $\overline{z_{d^v}} = \frac{1}{Nd^v}\sum_{n^v=1}^{Nd^v} z^v{}_{d^v,n^v}$, $\overline{z_{d^t}} = \frac{1}{Nd^t}\sum_{n^t=1}^{Nd^t} z^t{}_{d^t,n^t}$, $M^v \in R^{|T^V|*dim}$ and $M^t \in R^{|T^T|*dim}$ are two mapping matrices to map the visual and textual topic distributions into one common space with the dimension of *dim*, and $TCor(l^{vt})$ denotes the topic correlation between d^v and d^t, $TCor(l^{vt}=1)$ is the topic correlation, while $TCor(l^{vt}=0)$ is the pairwise topic uncorrelation.

Compared to the classical Corr-LDA, our modeling does not treat the multimodal document as a single one, but deals with the visual image and the textual description separately, which makes the limitation for the topic sets of different modalities looser and allows different constituents and properties for such topic sets, then the linkage L^{VT} is utilized to link two empirical topic distributions of the visual image and the textual description, as shown in Figure 3:



(a) Corr-LDA-based Modeling (b) Our Modeling Mechanism

Figure 3. Comparison between the Corr-LDA modeling and ours.

Based on the above modeling assumption, a joint probabilistic topic model can be built to maximize the likelihood of the observed multimodal documents, which is defined as:

$$\begin{aligned}&\mathrm{P}(D^V, D^T, LT^{VT})\\&= \mathrm{P}(\varphi^V, \varphi^T, \theta^V, \theta^T, Z^V, Z^T, DW^V, DW^T, L^{VT}|\alpha^v, \alpha^t, \beta^v, \beta^t, M^v, M^t)\\&= \prod_{t^v\epsilon T^V} Dir(\varphi^v{}_{t^v}|\beta^v)\prod_{t^t\epsilon T^T} Dir(\varphi^t{}_{t^t}|\beta^t) *\\&\prod_{d^v\epsilon D^V} Dir(\theta^v{}_{d^v}|\alpha^v)\prod_{w^v{}_{d^v,n^v}} Mult(z^v{}_{d^v,n^v}|\theta^v{}_{d^v})Mult(w^v{}_{d^v,n^v}|z^v{}_{d^v,n^v}) *\\&\prod_{d^t\epsilon D^T} Dir(\theta^t{}_{d^t}|\alpha^t)\prod_{w^t{}_{d^t,n^t}} Mult(z^t{}_{d^t,n^t}|\theta^t{}_{d^t})Mult(w^t{}_{d^t,n^t}|z^t{}_{d^t,n^t}) *\\&\prod_{l^{vt}\epsilon L^{VT}} TCor(l^{vt}|\overline{z_{d^v}}, \overline{z_{d^t}}, M^v, M^t)\end{aligned} \quad (3)$$

where the first part means the generation of topic-deep-word distributions, the middle two parts indicate the generation of deep visual and textual words, and the last part represents the generation of image-description linkages.

Our relational topic correlation model considers the heterogeneity of the multimodal topics, and exploits the linkage probability function *TCor*( ) to associate the topic distributions of different modalities, which can break the constraint of the topic consistency in traditional multimodal topic models.

4.2 Cross-Modal Correlation Learning

Due to the topic heterogeneity for different modalities, directly learning the topic correlation over multimodal topic distributions becomes computationally intractable. Thus we develop a specific cross-modal learning mechanism by projecting multimodal topic distributions into a common space and making sure that the cross-modal correlation can be maximized.

As the important part in computing the multimodal topic correlation probability function $TCor(l^{vt})$, the mapping matrices M^v and M^t aim at mapping the heterogeneous topic distributions into one common space. For the visual topic distribution $\overline{z_{d^v}}$ and the textual topic distribution $\overline{z_{d^t}}$, f^v and f^t are two new feature vectors in the common space for $\overline{z_{d^v}}$ and $\overline{z_{d^t}}$. We can compute $TCor(l^{vt})$ with the correlation measurement between f^v and f^t based on two commonly-used vector correlation evaluation patterns, shown as follows:

$$TCor(l^{vt}|\overline{z_{d^v}}, \overline{z_{d^t}}, M^v, M^t) = \begin{cases} \begin{cases} \mathrm{sigmoid}(f^v \cdot f^t), & l^{vt}=1 \\ 1-\mathrm{sigmoid}(f^v \cdot f^t), & l^{vt}=0 \end{cases} \\ \begin{cases} 0.5+0.5*\mathrm{cosine}(f^v, f^t) & l^{vt}=1 \\ 0.5-0.5*\mathrm{cosine}(f^v, f^t) & l^{vt}=0 \end{cases} \end{cases}$$
$$f^v = \overline{z_{d^v}} * M^v, f^t = \overline{z_{d^t}} * M^t \quad (4)$$

where *Pattern* 1 utilizes the sigmoid function to map the dot product value into [0, 1], and *Pattern* 2 computes the topic correlation by normalizing the cosine similarity of two vectors.

Based on the generated multimodal topic distributions, the cross iterative learning is explored to further learn the cross-modal topic correlation more precisely. We consider using Maximum Likelihood Estimate (MLE) to optimize M^v and M^t by maximizing the log probability of Formula (4), and the objective function for cross-modal learning is defined as:

$$\begin{aligned}&\mathrm{F}(M^v, M^t)\\&= \begin{cases} \underset{(M^v,M^t)}{\mathrm{argmax}} \sum_{l^{vt}=1} \log\frac{1}{1+e^{-(f^v \cdot f^t)}} + \sum_{l^{vt}=0} \log\frac{e^{-(f^v \cdot f^t)}}{1+e^{-(f^v \cdot f^t)}} \\ \underset{M^v,M^t}{\mathrm{argmax}} \sum_{l^{vt}=1} \log\left(0.5+\frac{f^v \cdot f^t}{2*|f^v|*|f^t|}\right) + \sum_{l^{vt}=0} \log\left(0.5-\frac{f^v \cdot f^t}{2*|f^v|*|f^t|}\right) \end{cases}\end{aligned} \quad (5)$$

With the above objective function, M^v and M^t can be computed by using the gradient descent strategy. It's worth noting that in the actual training process the training numbers of positive image-description linkages ($l^{vt}=1$) and negative linkages ($l^{vt}=0$) are imbalance, the number of negative ones is far more than that of positive ones. To solve this problem, we randomly sample negative linkages under the constraint that the pairwise image and description are from different classes, and set the proportion for positive and negative linkages as 1:1. Such cross-modal learning can bridge the gap between the heterogeneous topic distributions via mapping the multimodal topic distributions into the learned common space, in which the topic correlation can be integrated to the whole modeling.

4.3. Related Model Inference

Since the exact inference of topic model is generally intractable, some approximate strategies are usually conducted in the model inference. Thus we adopt the collapsed Gibbs sampling to infer the model parameters due to its simplicity and effectiveness [25].

The Gibbs sampling aims at inferring the latent topic for each deep word in each multimodal document. We first compute the marginal probability distribution of the observed deep words, topic assignments and linkages by integrating the other latent variables, shown as follows:

$$\begin{aligned}&\mathrm{P}(Z^V, Z^T, DW^V, DW^T, L^{VT})\\&= \int\cdot\cdot\int\left(P(\varphi^V, \varphi^T, \theta^V, \theta^T, Z^V, Z^T, DW^V, DW^T, L^{VT})\right) d\varphi^V d\varphi^T d\theta^V d\theta^T\\&=\\&\prod_{d^v\epsilon D^V}\frac{\Gamma(|DT^V|\alpha^v)}{\Gamma(\alpha^v)^{|DT^V|}}\frac{\prod_{t^v\epsilon DT^V}\Gamma(m^v{}_{d^v,t^v}+\alpha^v)}{\Gamma(\sum_{t^v\epsilon DT^V} m^v{}_{d^v,t^v}+|DT^V|\alpha^v)}\prod_{d^t\epsilon D^T}\frac{\Gamma(|DT^T|\alpha^t)}{\Gamma(\alpha^t)^{|DT^T|}}\frac{\prod_{t^t\epsilon DT^T}\Gamma(m^t{}_{d^t,t^t}+\alpha^t)}{\Gamma(\sum_{t^t\epsilon DT^T} m^t{}_{d^t,t^t}+|DT^T|\alpha^t)} *\\&\prod_{t^v\epsilon DT^V}\frac{\Gamma(|DV^V|\beta^v)}{\Gamma(\beta^v)^{|DV^V|}}\frac{\prod_{w^v\epsilon DW^V}\Gamma(n^v{}_{t^v,w^v}+\beta^v)}{\Gamma(\sum_{w^v\epsilon DW^V} n^v{}_{t^v,w^v}+|V^V|\beta^v)}\prod_{t^t\epsilon DT^T}\frac{\Gamma(|DV^T|\beta^t)}{\Gamma(\beta^t)^{|DV^T|}}\frac{\prod_{w^t\epsilon DW^T}\Gamma(n^t{}_{t^t,w^t}+\beta^t)}{\Gamma(\sum_{w^t\epsilon DW^T} n^t{}_{t^t,w^t}+|DV^T|\beta^t)} *\\&\prod_{l^{vt}\epsilon L^{VT}} TCor(l^{vt}|\overline{z_{d^v}}, \overline{z_{d^t}}, M^v, M^t)\end{aligned} \quad (6)$$

where $m_{d,t}$ is the number of the topic t that occurs in the related document d, and $n_{t,w}$ is the number of the deep word w assigned to t. Based on this probability distribution, we can further deduce the single-variable probability distribution of the topic assignment z for the Gibbs sampling. The sampling rules for z^v and z^t are defined as:

$$\begin{aligned}&\mathrm{P}\left(z^v{}_{d^v,n^v} = t^v \middle| Z^{-d^v,n^v}, DW^V, DW^T, L^{VT}\right) \propto \mathrm{P}(Z^T, Z^V, DW^V, DW^T, L^{VT})\\&\propto \frac{\hat{m}^v{}_{d^v,t^v}+\alpha^v}{\sum_{t^v\epsilon DT^V}\hat{m}^v{}_{d^v,t^v}+|DT^V|\alpha^v}\frac{\hat{n}^v{}_{t^v,w^v{}_{d^v,n^v}}+\beta^v}{\sum_{w^v\epsilon DV^V}\hat{n}^v{}_{t^v,w^v}+|DV^V|\beta^v}\prod_{\substack{l^{vt}\epsilon L^{VT}\\ d^v \text{ in } l^{vt}}} TCor(l^{vt}|\overline{z_{d^v}}, \overline{z_{d^t}}, M^v, M^t)\\&\mathrm{P}\left(z^t{}_{d^t,n^t} = t^t \middle| Z^{-d^t,n^t}, DW^V, DW^T, L^{VT}\right) \propto \mathrm{P}(Z^T, Z^V, DW^V, DW^T, L^{VT})\\&\propto \frac{\hat{m}^t{}_{d^t,t^t}+\alpha^t}{\sum_{t^t\epsilon DT^T}\hat{m}^t{}_{d^t,t^t}+|DT^T|\alpha^t}\frac{\hat{n}^t{}_{t^t,w^t{}_{d^t,n^t}}+\beta^t}{\sum_{w^t\epsilon DV^T}\hat{n}^t{}_{t^t,w^t}+|DV^T|\beta^t}\prod_{\substack{l^{vt}\epsilon L^{VT}\\ d^t \text{ in } l^{vt}}} TCor(l^{vt}|\overline{z_{d^v}}, \overline{z_{d^t}}, M^v, M^t)\end{aligned} \quad (7)$$

where $\hat{m}_{d,t}$ denotes the occurrence number of the topic t in the document d excluding the current deep word; similarly $\hat{n}_{t,w}$ denotes the occurrence number of the deep word w assigned to the topic t excluding the current deep word. As described in Formula (5), the mapping matrices M^t and M^v can be updated in each sampling iteration by using the gradient descent method to get the optimized values. With the statistics of the topic assignment z acquired in the sampling process, the other latent variables like φ^V, φ^T, θ^V, θ^T can be computed as:

$$\theta^{v}{}_{d^v,t^v} = \frac{m^{v}{}_{d^v,t^v}+\alpha^v}{\sum_{t^v\in DT^V} m^{v}{}_{d^v,t^v}+|DT^V|\alpha^v}, \theta^{t}{}_{d^t,t^t} = \frac{m^{t}{}_{d^t,t^t}+\alpha^t}{\sum_{t^t\in DT^T} m^{t}{}_{d^t,t^t}+|DT^T|\alpha^t}$$
$$\varphi^{v}{}_{t^v,w^v} = \frac{n^{v}{}_{t^v,w^v}+\beta^v}{\sum_{w^v\in DV^V} n^{v}{}_{t^v,w^v}+|DV^V|\beta^v}, \varphi^{t}{}_{t^t,w^t} = \frac{n^{t}{}_{t^t,w^t}+\beta^t}{\sum_{w^t\in DV^T} n^{t}{}_{t^t,w^t}+|DV^T|\beta^t} \quad (8)$$

5 EXPERIMENT AND ANALYSIS

5.1 Dataset and Evaluation Metrics

Our dataset is established based on two benchmark datasets of *Nus-Wide* (*Nus*) [26] and *Wiki_10cats* (*Wiki*) [12]. The *Nus-Wide* dataset is collected from the *Flickr* website, which contains 269,648 images with 1,000-dimensional tags and 81-dimensional concepts. Each image in *Nus-Wide* is annotated with several user-defined tags. As the work in [3], we only select those image-text pairs that belong to the 10 largest categories. As a result, we get 20,000 image-annotation pairs for training, 1,000 pairs for verification and 4,000 pairs as testing queries. As for *Wiki_10cats*, all the image-text pairs are collected from the *Wikipedia*'s "*featured articles*", which is a continually updated collection of articles selected by *Wikipedia*'s editors. These articles are accompanied by one or more pictures from Wikimedia Commons. In our work, 1,866 *Wikipedia* multimodal documents from the 10 most populated categories are selected for our experiment, with 2,173 pairs for training, 200 pairs for verification and 693 pairs as testing queries. In addition, each image/annotation or description in both datasets is represented as a 500-dimensional bag of deep visual/textual words by a specific grid search method.

To evaluate the effectiveness of our algorithm, the ground truth image-description correlation is considered to measure the official criteria of Precision-Recall (*P-R*) curves and Mean Average Precision (*MAP*) for cross-media retrieval. Cross-media retrieval allows different retrieval manners with the original queries in different modalities, that is, image query-to-text retrieval (return all the relevant texts for the given image query) or text query-to-image retrieval (return all the relevant images for the given text query). To measure the average performance of different retrieval manner, *AMAP* (the average *MAP* for both retrieval manners) is also used as an evaluation criterion in our experiment. The ranking score for such different retrieval manners can be defined as:

$$RankingScore(image\ query - to - text) = RankingScore(d^t|d^v) = \frac{TCor(l^{vt}=1|\theta^{v}{}_{d^v},\theta^{t}{}_{d^t},M^v,M^t)}{\sum_{d\in D^T} TCor(l^{vt}=1|\theta^{v}{}_{d^v},\theta^{t}{}_{d,M^v,M^t})}$$
$$RankingScore(text\ query - to - image) = RankingScore(d^v|d^t) = \frac{TCor(l^{vt}=1|\theta^{v}{}_{d^v},\theta^{t}{}_{d^t},M^v,M^t)}{\sum_{d\in D^V} TCor(l^{vt}=1|\theta^{v}{}_{d^v},\theta^{t}{}_{d,M^v,M^t})} \quad (9)$$

where M^v and M^t are obtained through Formula (5) in the training process; and θ_d is the topic distribution for the test document d, which can be calculated by aggregating all the word-topic distribution for each word in d based on the topic-word distribution φ obtained through Formula (8) in the training process.

5.2 Experiment on Cross-Modal Topic Correlation Model

Our cross-modal topic correlation model is created by integrating Deep Multimodal Document Representation (DMDR), Relational Topic Correlation Modeling (RTCM), and Cross-Modal Correlation Learning (CMCL). To show the effect of each part, we focus on investigating the whole performance of our cross-modal topic correlation model for cross-media retrieval. We compare the performance rising speeds for different scheme settings of DMDR, RTCM and CMCL, which implies the effectiveness difference between the general topic correlation modeling without DMDR, RTCM or CMCL and our proposed modeling with the integration of these three components. The related experimental results for cross-modal topic correlation model with different scheme settings are shown in Figure 4~6.

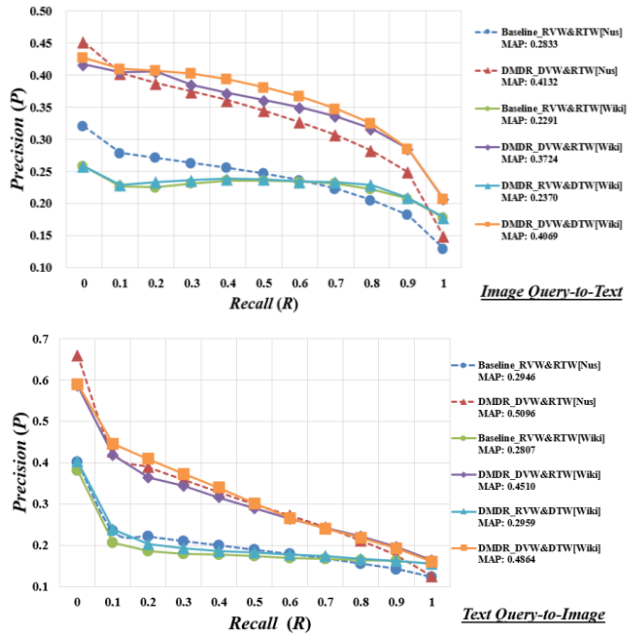


Figure 4. The experimental results on our model with different scheme settings of DMDR, in which we take the representation with raw visual and textual words as the baseline (*Baseline_RVW&RTW*) and make a comparison to DMDR with deep visual and raw textual words (*DMDR_DVW&RTW*), DMDR with raw visual and deep textual words (*DMDR_RVW&DTW*) and DMDR with deep visual and textual words (*DMDR_DVW&DTW*).

It can be seen from Figure 4 that for the cross-modal topic correlation model with DMDR on *Nus-Wide* and *Wiki_10cats*, we can obtain the best cross-media retrieval performance (*MAP*=0.5096) in the evaluation pattern of fusing *DMDR_DVW&DTW* with RTCM and CMCL. In comparison with the baseline pattern using the raw visual and textual word information for multimodal document representation, the performance could be greatly promoted by successively adding the deep visual and textual word representation, which confirms the obvious advantage of our deep multimodal document representation for cross-modal topic correlation model. Through comparing the baseline with two patterns using single deep visual or textual word information, our model can still gain the significant advantage for the performance on both *Nus-Wide* and *Wiki_10cats*. Meanwhile, we can find the performance with single deep visual word information appears better, which shows the beneficial effect of deep visual features on cross-modal topic correlation model. Comparing the results on *Nus-Wide* and

Wiki_10cats, the results on *Nus-Wide* appear less performant on the whole *P-R* curves, while better on the *MAP* values that are measured with the performance statistics for the top-50 ranking results. Overall, the performance difference for *Nus-Wide* and *Wiki_10cats* is not obvious, which reflects the performance advantage to some degree. Due to the differences between these two datasets, we do not compare two patterns of *DMDR_RVW&DTW* and *DMDR_DVW&DTW* on *Nus-Wide*. *Nus-Wide* is a typical social annotated image dataset with discrete tags in each textual annotation, and these discrete annotation tags are independent and meaningful for describing images, so we just use the raw text words in our experiment. While *Wiki_10cats* is a *Wikipedia* featured article dataset with successive narrations in each textual description, there are a lot of redundant information in the raw documents, so the deep textual words are applied in *Wiki_10cats*. As shown in Figure 4, the same conclusions as above can be drawn from the *P-R* curves and *MAP* values for both *Image Query-to-Text* and *Text Query-to-Image* retrieval on *Nus-Wide* and *Wiki_10cats*, which show the consistence of our model on the performance indicators for different cross-media retrieval manners. These results are consistent with what we expect given deep affluent feature descriptions for multimodal documents.

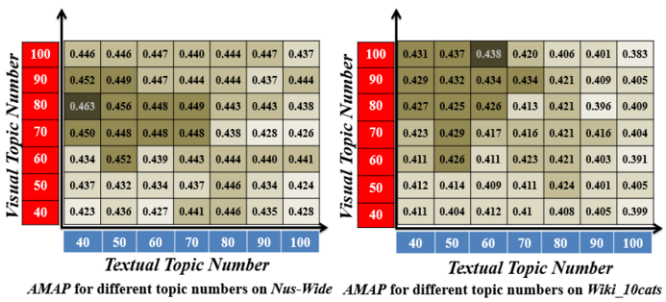


Figure 5. The experimental results on our model with different scheme settings of RTCM, in which we introduce the density graph to show the *AMAP* values for the average performance of cross-media retrieval with different numbers of visual and textual topics.

It can be viewed from Figure 5 that the best performance can be achieved on both *Nus-Wide* and *Wiki_10cats* when the numbers for visual and textual topics are under different settings. This confirms the advantage of our modeling mechanism with the basis assumption that the topics in different modalities are heterologous. Meanwhile, we can find that the best performance can be obtained when the number of visual topics is more than the number of textual topics, which indicates the deep feature information involved in the visual image is richer than that in the textual description on both *Nus-Wide* and *Wiki_10cats*. In addition, we also observe that the best performance on *Nus-Wide* can be acquired when the visual and textual topic numbers are set as 40 and 80 respectively, while on *Wiki_10cats* the best performance can be implemented when the visual and textual topic numbers are set as 60 and 100 respectively. It's obvious that the topic numbers utilized on *Nus-Wide* is smaller than those on *Wiki_10cats*, which is also due to the structural differences between these two datasets as mentioned above, and the contents in *Wiki_10cats* are more complicated and diverse.

It can be found from Figure 6 that for the cross-modal topic correlation model with CMCL on *Nus-Wide* and *Wiki_10cats*, we can obtain the best performance (*MAP*=0.5096) in the evaluation pattern of fusing *CMCL_Sigmoid* with DMDR and RTCM. In comparison with the baseline Corr-LDA model [9], which has the tight restriction that the topic distributions for different modalities

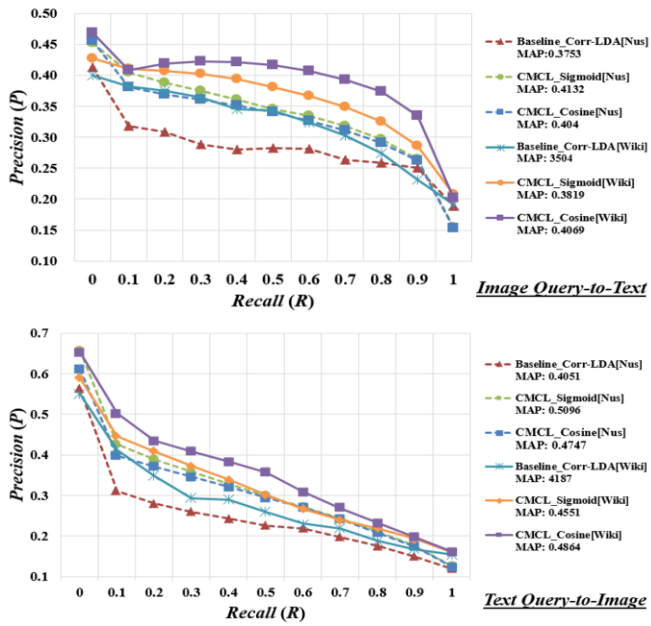


Figure 6. The experimental results on our model with different scheme settings of CMCL, in which we introduce the *Corr-LDA* (without learning but restricting the same topic distributions in different modalities) as the baseline (*Baseline_Corr-LDA*) and make a comparison with the models using two kinds of learning pattern, that is, CMCL with the sigmoid function to compute the topic correlation probability *TCor*(*CMCL_Sigmoid*), and CMCL with the cosine function to compute *TCor*(*CMCL_Cosine*).

must be same, the whole performance could be greatly promoted by integrating the learning with *CMCL_Sigmoid* or *CMCL_Cosine* in the topic correlation modeling, which confirms the obvious advantage of our cross-modal correlation learning scheme. Comparing two learning patterns of *CMCL_Sigmoid* and *CMCL_Cosine*, the *CMCL_Sigmoid*-based learning achieves the better performance than the *CMCL_Cosine*-based one on *Nus-Wide*, while on *Wiki_10cats* the *CMCL_Cosine*-based learning performs better due to the different dataset structure from *Nus-Wide*. The same conclusions as above can also be drawn from the *P-R* curves and *MAP* values for both *Image Query-to-Text* and *Text Query-to-Image*.

From all the above observations, it's worth noting that our cross-modal topic correlation model is available and presents more impactful ability for discovering the meaningful deep multimodal features and correlations. Our framework can not only significantly improve the cross-modal topic correlation measurement, but also greatly enhance the cross-media retrieval in different manners. The same conclusions from two different datasets of *Nus-Wide* and *Wiki_10cats* show the consistence of our model on different data sources. An instantiation of some cross-media retrieval results with our cross-modal topic correlation model is shown in Figure 7.

5.3 Comparison with Existing Approaches

Compared to the common topic correlation methods in recent years, our approach is a new exploration for taking full advantage of deep information in cross-modal topic correlation measurement. To give full exhibition to the superiority of our topic correlation model, we have also performed a comparison between our method and the other existing classical approaches in recent years. Three approaches developed by Blei *et al*. (2003) [9], Pereira *et al*. (2014) [27] and Wang *et al*. (2014) [3] respectively are analogous with ours to some

Figure 7. An instantiation of some retrieval results with our model.

extent, and then we accomplished them on the same dataset. The experimental results are presented in Table 1, which reflect the difference of power among these four approaches.

Table 1. The comparison results between our and the other approaches.

Dataset	Approach	Evaluation Pattern	MAP		AMAP
			Text Query-to-Image	Image Query-to-Text	
Nus-Wide	***Corr-LDA*** **(Blei et al., 2003)** **(*Blei*)**	*Original*	0.2513	0.2444	0.2479
		Original+DMDR	**0.4051**	**0.3753**	**0.3902**
	LDA-KCCA **(Pereira et al., 2014)** **(*Pereira*)**	*Original*	0.3021	0.2726	0.2874
		Original+DMDR	**0.4214**	**0.3829**	**0.4022**
	M³R **(Wang et al., 2014)** **(*Wang*)**	*Original*	0.2631	0.2714	0.2673
		Original+DMDR	**0.4611**	**0.4092**	**0.4352**
	Our Approach	***No_DMDR +RTCM+CMCL***	**0.2946**	**0.2833**	**0.2890**
		DMDR+RTCM+CMCL	**0.5096**	**0.4132**	**0.4614**
Wiki_10cats	***Corr-LDA*** **(Blei et al., 2003)** **(*Blei*)**	*Original*	0.2261	0.2157	0.2209
		Original+DMDR	**0.4187**	**0.3504**	**0.3846**
	LDA-KCCA **(Pereira et al., 2014)** **(*Pereira*)**	*Original*	0.2563	0.2268	0.2415
		Original+DMDR	**0.4154**	**0.3587**	**0.3871**
	M³R **(Wang et al., 2014)** **(*Wang*)**	*Original*	0.2387	0.2135	0.2261
		Original+DMDR	**0.4394**	**0.375**	**0.4072**
	Our Approach	***No_DMDR +RTCM+CMCL***	**0.2807**	**0.2291**	**0.2549**
		DMDR+RTCM+CMCL	**0.4864**	**0.4069**	**0.4467**

It can be found from Table 1 that for the topic correlation models on *Nus-Wide* and *Wiki_10cats* by Blei/Pereira/Wang *et al.*'s approaches, we can obtain the best *AMAP* values of 0.3902, 0.4022 and 0.4352 in the evaluation pattern of *Original+DMDR* respectively. The main reason is that when considering the deep multimodal feature attributes all these three approaches can explore more precise multimodal topic distribution information for topic correlation measurement and then the relatively better *AMAP* values can be obtained on both *Nus-Wide* and *Wiki_10cats* in comparison with their original performance exhibitions (i.e., *Blei*/*Pereira*/*Wang* (*Original*)). This obviously confirms the prominent role of our DMDR in the topic correlation modeling. Comparing the results of *Blei*/*Pereira*/*Wang* (*Original*) and our baseline model with *No_DMDR+RTCM+CMCL*, we can find the best *AMAP* value of 0.2890 appears in the results of our model, which is obviously higher than those *AMAP* values of 0.2479, 0.2874 and 0.2673 for *Blei*/*Pereira*/*Wang* (*Original*) respectively. This implies that our cross-modal topic correlation modeling mechanism with RTCM and CMCL is feasible for facilitating more effective topic correlation evaluation and optimization. Compared to the improved Blei/Pereira/Wang *et al.*'s approaches that integrate with DMDR, our model with DMDR, RTCM and CMCL can still present the better performance on both *Nus-Wide* and *Wiki_10cats*, and the best *AMAP* value of 0.4614 differs greatly from those of Blei/Pereira/Wang *et al.*'s. This indicates that our approach is really superior to Blei/Pereira/Wang *et al.*'s, and also further confirms that our cross-modal topic correlation model with DMDR, RTCM and CMCL is exactly a better way for determining cross-modal image-description topic correlation and can support cross-media retrieval with queries in different modalities more effectively. From the view of computational load performance, the retrieval efficiency of our model is high during the process of cross-media retrieval, and can meet the demand of real-time response.

5.4 Analysis and Discussion

Through the analysis for the topic correlated image-description linkages with failure, it can be found that the modeling quality is highly related to the following aspects. (1) The modeling effect is closely associated with the appropriate representation for multimodal document. It's easier to introduce superficial and noisy information for images and descriptions, which will seriously affect the whole retrieval performance. (2) For multimodal topic model modeling, it's helpful to integrate the topic generation and cross-modal topic correlation analysis into one model, which can adaptively generate the latent topics related to both visual contents and textual information. (3) In some multimodal documents, the textual description has very weak correlation to the visual image content, which leads to the huge semantic topic gap between images and descriptions. It's hard for such documents to successfully implement precise cross-modal topic correlation measurement. This may be the stubbornest problem. (4) The modeling effect is greatly influenced by the topic distribution and number, but such important information may be changed dynamically. It's better to establish the adaptive strategy for finding the optimal settings.

6 CONCLUSIONS AND FUTURE WORK

A new cross-modal topic correlation model is implemented to exploit multimodal correlations between visual images and textual descriptions to enable more effective cross-media retrieval for large-scale multimodal documents. The deep words are conducted to discover deep features for multimodal document representation. A relational topic correlation modeling scheme is designed to achieve more precise characterization of multimodal correlations. An efficient learning mechanism is established to achieve more objective decision-making for cross-modal image-description topic correlation. Our future work will focus on adding supervised information to our model and making our system available online, so that more Internet users can benefit from our research.

7 ACKNOWLEDGMENTS

This work is supported by National Natural Science Fund of China (61572140), Shanghai Municipal R&D Foundation (16511105402& 16511104704), Shanghai Philosophy Social Sciences Planning Project (2014BYY009), and Zhuoxue Program of Fudan University. Yuejie Zhang is the corresponding author.

REFERENCES

[1] R.Datta, D.Joshi, J.Li, and J.Z.Wang. 'Image Retrieval: Ideas, Influences, and Trends of the New Age', *ACM Computing Surveys* (*CSUR*) 40(2), Article 5, (2008).

[2] J.P.Fan, X.F.He, N.Zhou, J.Y.Peng, and R.Jain. 'Quantitative Characterization of Semantic Gaps for Learning Complexity Estimation and Inference Model Selection', *IEEE Transactions on Multimedia* 14(5):1414-1428, (2012).

[3] Y.F.Wang, F.Wu, J.Song, X.Li, and Y.T.Zhuang. 'Multi-modal Mutual Topic Reinforce Modeling for Cross-media Retrieval', In *Proceedings of MM 2014*, 307-316, (2014).

[4] T.Chen, H.M.SalahEldeen, X.N.He, M.Y.Kan, and D.Y.Lu. 'VELDA: Relating an Image Tweet's Text and Images', In *Proceedings of AAAI 2015,* (2015).

[5] K.Barnard, P.Duygulu, D.Forsyth, N.Freitas, D.M.Blei, and M.I.Jordan. 'Matching Words and Pictures', *Journal of Machine Learning Research*. 3:1107-1135, (2003).

[6] X.Wang, Y.Liu, D.Wang, and F.Wu. 'Cross-media Topic Mining on Wikipedia', In *Proceedings of MM 2013*, 689-692, (2013).

[7] A.Frome, G.S.Corrado, J.Shlens, S.Bengio, J.Dean, M.A.Ranzato, and T.Mikolov. 'DeViSE: A Deep Visual-Semantic Embedding Model', In *Proceedings of NIPS 2013*, (2013).

[8] F.X.Feng, X.J.Wang, and R.F.Li. 'Cross-modal Retrieval with Correspondence Autoencoder', In *Proceedings of MM 2014*, 7-16, (2014).

[9] D.M.Blei, and M.I.Jordan. 'Modeling Annotated Data', In *Proceedings of SIGIR 2003*, 127-134, (2003).

[10] C.Wang, D.M.Blei, and L.Fei-Fei. 'Simultaneous Image Classification and Annotation', In *Proceedings of CVPR 2009*, 1903-1910, (2009).

[11] D.Putthividhya, H.T.Attias, and S.S.Nagarajan. 'Topic Regression Multi-Modal Latent Dirichlet Allocation for Image Annotation', In *Proceedings of CVPR 2010*, 3408-3415, (2010).

[12] N.Rasiwasia, J.C.Pereira, E.Coviello, G.Doyle, G.R.G.Lanckriet, R.Levy, and N.Vasconcelos. 'A New Approach to Cross-Modal Multimedia Retrieval', In *Proceedings of MM 2010*, 251-260, (2010).

[13] C.T.Nguyen, N.Kaothanthong, T.Tokuyama, and X.H.Phan. 'A Feature-Word-Topic Model for Image Annotation and Retrieval', *ACM Transactions on the Web* 7(3), Article 12, (2013).

[14] Z.X.Niu, G.Hua, X.B.Gao, and Q.Tian. 'Semi-supervised Relational Topic Model for Weakly Annotated Image Recognition in Social Media', In *Proceedings of CVPR 2014,* 4233-4240*,* (2014).

[15] Y.Zheng, Y.J.Zhang, and H.Larochelle. Topic Modeling of Multimodal Data: An Autoregressive Approach. In *Proceedings of CVPR 2014*, 1370-1377*,* (2014).

[16] J.Tian, Y.Huang, Z.Guo, and X.Qi. 'A Multi-Modal Topic Model for Image Annotation Using Text Analysis', *In IEEE Signal Processing Letters* 22(7): 886-890, (2014).

[17] F.Wu, X.Jiang, X.Li, and S.Tang. Cross-Modal Learning to Rank via Latent Joint Representation, *In IEEE Transactions on Image Processing* 24(5): 1497-1509, (2015).

[18] R.Girshick, J.Donahue, T.Darrell, and J.Malik. 'Rich feature hierarchies for accurate object detection and semantic segmentation', In *Proceedings of CVPR 2014,* 580-587, (2014).

[19] B.Hariharan, P.Arbelaez, R.Girshick, and J.Malik. 'Simultaneous Detection and Segmentation', In *Proceedings of ECCV 2014,* 297-312, (2014).

[20] A.Karpathy, A.Joulin, and L.Fei-Fei. 'Deep Fragment Embeddings for Bidirectional Image Sentence Mapping', In *Proceedings of NIPS 2014*, (2014).

[21] N.Zhang, J.Donahue, R.Girshick, and T.Darrell. 'Part-Based R-CNNs for Fine-Grained Category Detection', In *Proceedings of ECCV 2014,* 834-849. (2014).

[22] J.Sivic, and A.Zisserman. 'Video Google: A Text Retrieval Approach to Object Matching in Videos', In *Proceedings of ICCV 2003*, 2:1470-1477, (2003).

[23] T.Mikolov, I.Sutskever, K.Chen, G.Corrado, and J.Dean. 'Distributed Representations of Words and Phrases and their Compositionality', In *Proceedings of NIPS 2013*, (2013).

[24] D.Y.Tang, F.R.Wei, B.Qin, M.Zhou, and T.Liu. Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. In *Proceedings of COLING 2014*, 172-182, (2014).

[25] T.L. Griffiths and M. Steyvers, 'Finding scientific topics', *In Proceedings of The National Academy of Sciences of USA*, pp. 5228-5235, (2004).

[26] 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 CIVR 2009*, (2009).

[27] J.C.Pereira, E.Coviello, G.Doyle, N.Rasiwasia, G.R.G.Lanckriet, R.Levy, and N.Vasconcelos. 'On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval', *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 36(3): 521-535, (2014).