A Probabilistic Topic-Connection Model for Automatic Image Annotation

Xin Chen[1], Xiaohua Hu[1], Zhongna Zhou[2], Caimei Lu[1], Gail Rosen[3], Tingting He[4], E.K. Park[5]

[1]College of Information Science and Technology, Drexel University, Philadelphia, PA, USA, [2]Dept. of ECE at University of Missouri in Columbia, MO, USA, [3]Dept. of ECE at Drexel University in Philadelphia, PA, USA, [4]Dept. of Computer Science at Central China Normal University in Wuhan, China, [5]CSI-CUNY in Staten Island, NY, USA

bruce.chen@drexel.edu, thu@ischool.drexel.edu, zz3kb@mizzou.edu, cl389@drexel.edu, gailr@ece.drexel.edu, tthe@mail.ccnu.edu.cn, ek.ek.park@gmail.com

ABSTRACT

The explosive increase of image data on Internet has made it an important, yet very challenging task to index and automatically annotate image data. To achieve that end, sophisticated algorithms and models have been proposed to study the correlation between image content and corresponding text description. Despite the success of previous works, however, researchers are still facing two major difficulties that may undermine their effort of providing reliable and accurate annotations for images. The first difficulty is lacking of comprehensive benchmark image dataset with high quality text descriptions. The second difficulty is lacking of effective way to represent the image content and make it associate with the text descriptions. In our paper, we aim to deal with both problems. To deal with the first problem, we utilize Wikipedia as external knowledge source and enrich the ontology structure of ImageNet database with comprehensive and highly-reliable text descriptions from Wikipedia articles. To address the second problem, we develop a Probabilistic Topic-Connection (PTC) model to represent the connection between latent semantic topic in text description and latent patterns from image feature space. We compare the performance of our model with the currently popular Correspondence LDA (Corr-LDA) model under the same automatic image annotation scenario using cross-validation. Experimental results demonstrate that our model is able to well represent the connection between latent semantic topics and latent patterns in image feature space, thus facilitates knowledge organization and understanding of both image and text descriptions.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database applications – Data mining; Image databases; I.2.6 **[ARTIFICIAL INTELLIGENCE]**: Learning

General Terms

Algorithms, Experimentation, Theory.

Keywords

Probabilistic models, topic learning, Gibbs sampling, image feature extraction, automatic image annotation.

CIKM'10, October 26–29, 2010, Toronto, Ontario, Canada.

1. INTRODUCTION

The prevalence of digital imaging device, such as digital camera and digital video camera, has brought an increasingly large amount of stored multimedia data, especially digital images. With nearly a million new images being added in one day, the Flickr.com now hosts over 3 billion shots of user-uploaded images. In comparison, Facebook.com, another famous online image sharing platform, has already hit 4.1 billion images on its site. Manually annotating such a huge amount of image dataset is time-consuming, laborious and prohibitively expensive. Therefore, it is very important to achieve some extend of automation in image annotation, which is currently a very difficult, yet long-term cost-efficient way to face the challenge of enormous explosion of digital images. Breakthroughs in automatic image annotation will help to organize the massive amount of digital images, promote developing and studying of image storage and retrieval systems, and serve for many other applications such as product searching, online studying, online image-sharing, etc.

Automatic image annotation is closely related to computer vision, image processing and content-based image retrieval [1, 2].During the last decade, we have seen great advance in developing automatic image annotation systems, related works involve considering image annotation as a clustering/categorization problem [3, 4], considering image annotation as an image searching problem [5], and considering image annotation as statistical modeling problem [6-9]. Despite the success of previous works, however, researchers are still facing two major difficulties that could undermine their effort of providing reliable and accurate annotation for images. The first difficulty is lacking of comprehensive benchmark image dataset with high quality text descriptions. The second difficulty is lacking of effective ways to represent the image content and associate it with the text descriptions.

High quality text descriptions of images play a critical role as training and benchmarking data in developing and evaluating an automatic image annotation system. Labeled image datasets such as Caltech 101/256 Categories [10, 11], PASCAL [12], LabelMe [13] have been popular with the computer vision community as benchmark training datasets; these datasets provide high quality image captions, yet lack of diversity, as only a limited number of object categories are covered. Images and captions from online data source like *Flickr.com* and *Facebook.com* have high diversity, however, due to their social networking purpose, image

captions are characterized by free-form text and user-sensitive descriptions, thus too noisy to be directly used as benchmark data.

The recently established ImageNet dataset [14] provides large scale ontology of image that is built upon the WordNet Structure [15]. Organized by an ontology structure, images in the ImageNet dataset are grouped into sets of cognitive synonyms (synsets), each expressing a distinct semantic concept. Due to its completeness and accuracy, the ImageNet dataset may potentially serves as benchmarking data for image annotation algorithms. One problem with ImageNet dataset is that it still lacks of comprehensive text descriptions for image data. Therefore, in our research, we utilize Wikipedia as external knowledge source and enrich the ontology structure of ImageNet database with comprehensive and highly-reliable text descriptions from Wikipedia articles.

In automatic image annotation, how to bridge over the "semantic gap" [1] between user and image features is another major challenge. Specifically, the first step of this problem is to identify a set of image features that well preserve the semantic consistency of image content.

In recent years, affine invariant local image detectors and descriptors [23-25] have exhibited very good performance in image categorization and semantic image retrieval across several well-known databases such as the *Caltech 101*, the *TRECVID* and the *Visual Object Classes (VOC)* datasets [3,4, 10-12]. The major goal of developing a local descriptor is to make it invariant under image variations (such as affine transform and illumination change), while maintaining high repeatability and discriminative power. Usually, it starts with a detection step which detects key points as local appearances of an image, then follows with a description step, in which local image patches containing these key-points are quantized into feature vectors in an affine invariant way. Comprehensive study conducted by Mikolajczyk and Schmid indicates that the SIFT descriptor [23] outperforms other affine invariant local descriptors due to its high robustness to image variations [26]. State-of-the-art image-representation methods further cluster and quantify local SIFT descriptors into visual code-words [24], and then apply text-like indexing schemes on 'bag-of-visual words' representation of images [9, 20-22].

One problem with SIFT descriptors is that since it is derived from key-points, it may become less discriminative due to the quantification and increase of image number [20]. With this consideration, it is suggested that researcher may combine both point features (such as SIFT descriptor) and region features (which are derived from local homogeneous parts in objects) to improve the understanding of image content [27, 28]. In our approach, we extract both SIFT features and Maximally Stable Extremal Region (MSER) [25], a widely used image region feature, to provide an efficient and robust representation of local image appearance.

After representing image content as combination of SIFT and MSER features, the second step of the problem is to uncover latent semantic topics from the co-occurrence patterns of image content and corresponding text descriptions. In the data mining and information retrieval community, there has been a long time focus on using probabilistic models to study the correlation between image and text descriptions. Specifically, the Correspondence LDA (CorrLDA) model [6], which is initially proposed by Blei et al. for automatic image annotation, provides a natural way to learn latent topics from text word and other entities (such as image features). As represented in Fig. 1, this model enforces great degree of correspondence between word and entity topics. It first generates latent topic for each text word, resulting in a document-level mixture of word topics (Fig. 1b). This document-level topic mixture then replicates itself as the composition of entity topics, which is used to supervise the generation of associated entities, resulting in a direct connection between word and entity topics. Most recent extensions of CorrLDA model, including sophisticated correspond topic models that extend to different kinds of entities (such as protein entities [8], visual words, and ontology-based biomedical concepts [9]), still follow the same generative process as the prototype CorrLDA model.

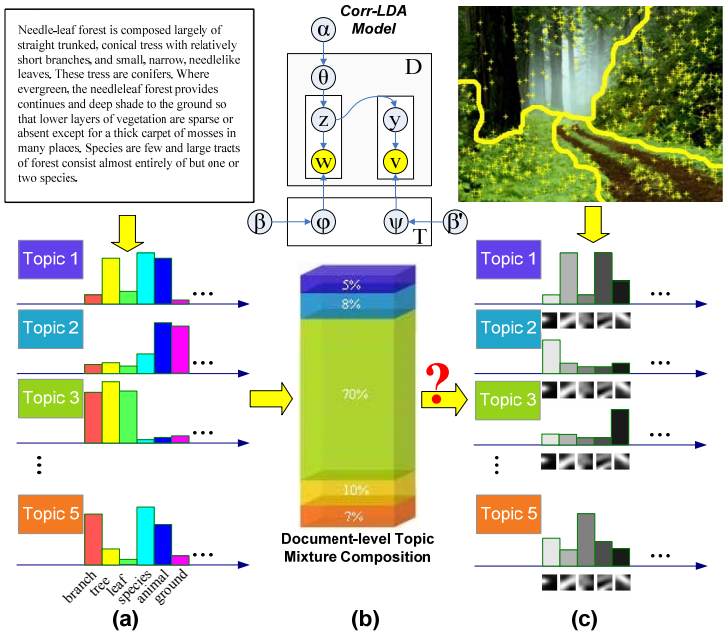


Fig. 1 Graphical illustration of generating latent topics for both image and text using CorrLDA model

Despite its great success in many data mining applications, the CorrLDA model may encounter some problems when dealing with both images and text descriptions. Since the text words and extracted image features have totally different characteristics, it is very possible that a word topic is connected to multiple entity topics each stands for a specific image feature pattern, instead of connected to only one entity topic.

To better explain this problem, let's consider a simple image-text modeling problem in Fig. 1, for simplicity, we name the entity topic of image feature as "visual topics". Assuming that we have a vocabulary of 6 words (branch, tree, leaf, species, animal and ground) and a total of 5 word topics each have a unique distribution of generating words (Fig. 1a). Take word topic 3 for example, it has high probability generating 'branch', 'tree' and 'leaf' while low probability generating 'species', 'animal' and 'ground', so it may be related to the concept of forest. As a comparison, topic 5, which has high probability of generating 'branch', 'species' and 'animal', may represents concept of branch splitting during animal species evolution. Now supposing that we have an image about needle-leaf forest and a piece of text

description that explain the needle-leaf forest, and that we choose to represent image content by visual code-words that are derived from SIFT descriptors. As we can see in Fig. 1, in the sense of single-word features, this example is almost 'uniform topic', which is mainly composed of topic 3 (Fig. 1b). However, in the sense of image feature representation, this example is not really a 'uniform topic' case. Although the image purely depicts the scene of needle-leaf forest, however, it still have multiple visual topics corresponding to different image regions such as trunks, leaves, path, grass, etc (Fig. 1c). For example, the visual topic "trunks" may favor some visual code-words that occur more frequently in trunks (e.g. vertical lines); similarly, visual topic "leaves" may in turn privilege other visual code-words such as blob-like structures. Since each region takes up similar portion of area in the image, there is no evidence that any of these visual topics be dominant in the entire image. Therefore, the shifting from word topic to visual topics is not as transparent as assumed in CorrLDA model.

In our research, we argue that each word topic is related to multiple visual topics, with different connection strength respectively. Based on this assumption, we propose a probabilistic Topic-Connection (PTC) model for automatic image annotation. The model is estimated via collapsed Gibbs sampling algorithm, while the parameter selection is achieved by studying the likelihood and perplexity. We compare the performance of our model with the Corr-LDA model under the same automatic image annotation scenario using cross-validation.

The remainder of this paper is organized as follows. In Section 2, we describe the procedure of indexing image and text description. In Section 3, we present the generative process of CorrLDA model and our probabilistic Topic-Connection model. Section 4 provides the collapse Gibbs sampling algorithms for inference and learning proposed probabilistic models. Section 5 reports the experimental results of the proposed method and compares our approach to the CorrLDA model. We conclude the paper in Section 6.

2. INDEXING OF IMAGES AND TEXT DESCRIPTIONS

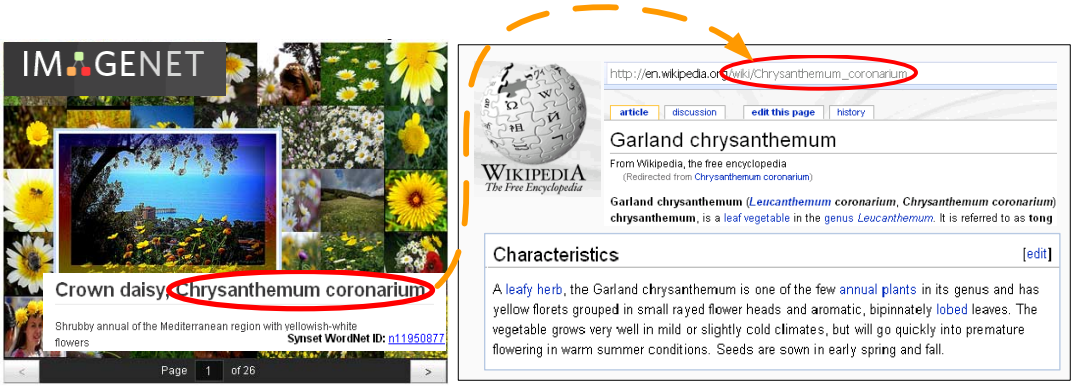
2.1 Text Description Enrichment and Indexing

ImageNet dataset [14] provides large scale ontology of image that is built upon the WordNet Structure. According to its latest release, ImageNet hosts a total of 15589 synsets of WordNet, with an average of 50-500 images under each synset. In our approach, we enrich image data from ImageNet dataset with high quality text descriptions from Wikipedia articles to provide benchmarking data set for automatic image annotation.

Wikipedia is one of the most comprehensive and well-formed electronic knowledge repositories on the web with millions of articles contributed collaboratively by volunteers. Because of its reliability, accuracy and neutral point of view, Wikipedia has been exploited as external knowledge source in many application of text mining [16-18]. Although Wikipedia is different from standard WordNet ontology, which is backed up by structured thesaurus, however, each article in Wikipedia only describes one single concept under a hierarchical categorization system. Therefore, the title of each article (which is a succinct phrase) still resembles an ontology term. This feature makes it possible to map a WordNet synset to a Wikipedia article (Fig. 2), which in turn provides text descriptions for images under this synset. In our research, we found matched Wikipedia articles for nearly 75 percent of the synsets we studied.

In learning unambiguous semantic topics from text descriptions, polysemies and synonyms are the major barrier. In [9], the author suggests using both ontology-based biomedical concepts and single-word features to overcome the polysemy and synonym problems in biomedical literatures. In public domain, where ontology-based concept is not available and domain knowledge is rare, we propose to use multiword phrases in conjunction with unigram features. The multiword phrases usually have unchanged meanings, thus reduce the ambiguity in unigram 'bag-of-word' document model.

Therefore, the indexing of text descriptions involves two parts, i.e. the term indexing and the phrase indexing. The term indexing is simply achieved by calculating the term frequency of each word after lemmatizing and stop-word removal. In our approach, the Van Rijsbergen's stop-word list [30] is used to remove non-content-bearing terms. We utilize the statistical extraction tool Xtract [19] to identify frequent multiword phrases from the text description. The estimated precision is about 80%. Xtract uses four parameters, strength (k0), spread (U0), peak z-score (k1), and percentage frequency (T), to control the quantity and quality of the extracted phrases. In general, the bigger those parameters the higher quality but less quantity phrases will be produced. In our experiment, we set those four parameters to (1, 1, 4, 0.75) after extensive tuning and testing.

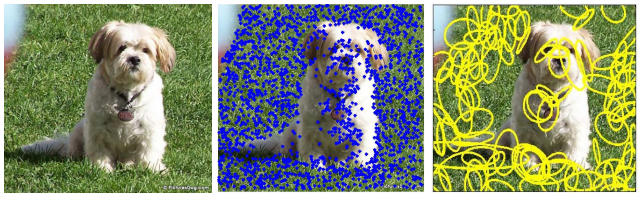


(a) ImageNet dataset: images under synset "Chrysamthemum coronarium" (b) Wikipedia article matched to the synset in (a)

Fig. 2 Graphical illustration of mapping a WordNet synset to a Wikipedia article to obtain text descriptions for images

2.2 Image Feature Extraction and Indexing

In our approach, we choose to represent the image content by both point-level SIFT features (Fig. 3b) and region-level Maximally Stable Extremal Region (MSER) features (Fig. 3c).



(a) Original image (b) SIFT features (c) MSER features

Fig. 3 Comparison of SIFT feature and MSER features

Following the framework in [9], we extract and index SIFT features as follows. Firstly, we employ the Difference-of-Gaussian (DoG) salient point detector [23] to detect salient points from images. The detection is achieved by locating scale-space extreme points in the difference-of-Gaussian images and the main orientations of salient points are determined by image gradient. Then, image patches containing the salient points are rotated to a canonical orientation and divided into 4×4 cells. In each cell, the gradient magnitudes at 8 different orientations are calculated. Consequently, each salient point is described by a 128-dimensional SIFT descriptor. In this way, each image in the training dataset is represented as a set of SIFT descriptors.

After that, K-mean clustering is performed to quantize all the extracted SIFT descriptors and produce a finite dictionary of appearance patterns called "code-book of visual words", with each cluster center as a unique "visual word". In our approach, SIFT descriptors extracted from training images are grouped into 2000 clusters, thus the vocabulary size of visual words is 2000. Finally, the indexing of SIFT features is accomplished by computing the term frequency of visual words with respect to each image document.

The Maximally Stable Extremal Region(MSER) is a widely used image feature to represent regions. Unlike the SIFT descriptors, which is derived from key-points, the detected MSER regions are local homogeneous parts in objects (Fig. 3c). Although the MSER detector output relatively smaller number of MSER features than SIFT descriptors, their distinctness is higher. Specifically, MSER detection begins with segmenting a set of image regions whose inner intensity value is less than certain thresholds while all intensities around the region boundary is greater than the same threshold. After that, a maximally stable extremal region is obtained when the area of the segments changes the least with respect to the threshold [25]. Extensive study reveal that the set of MSER regions is closed under continuous geometric transforms, thus providing an efficient affine invariant region detector for local image appearance [29]. We also extend MSER detector to multiple scales by constructing Gaussian pyramid and applying MSER detection separately in each resolution level.

After MSER detection, each detected elliptical region is normalized to circular patch of constant radius. In order to further improve its scale and affine invariant capability, we rectify each patch to canonical orientation following the coordinate transform in [29]. We set the size of final normalized patches as 21 pixels by 21 pixels.

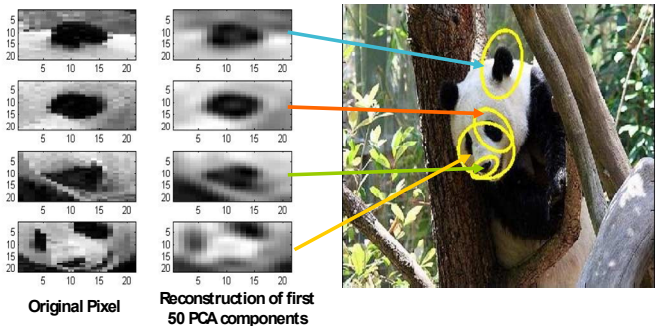


Fig. 4 PCA component number settings for the quantification of MSER normalized patches

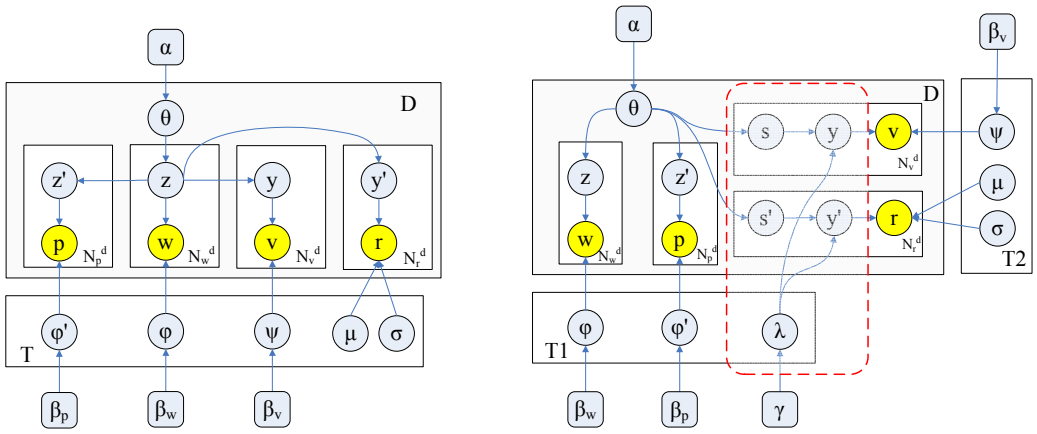
An effective descriptor for image patch representation should be compact while remaining highly distinctive. Thus we chose to perform principle component analysis (PCA) on MSER normalized patches to provide a robust and compact representation of image regions. More specifically, we construct covariance matrix for a total of 140,000 MSER normalized patches extracted from training dataset, each of which is a 21×21 dimensional vector of image intensity. Then, we perform eigen-decomposition on the covariance matrix to obtain the eigenvectors. We then obtain the projection matrix which is composed of first k principal components, i.e. eigenvectors corresponding to k largest eigen-values. In this way, we build the eigenspace for all the MSER normalized patches in our training dataset. To represent a new MSER patch, we simply multiply its 21×21 dimensional vector with the projection matrix to obtain its k dimensional projection. After extensive testing, we set k=50 (which means that all the MESR features are represented as 50-dimensional vectors). Reconstruction result in Fig. 4 shows that, when using the first 50 principle components, we are able to achieve significant dimension reduction of the MSER features without losing important details.

3. PROBABILISTIC TOPIC MODELS FOR IMAGE AND TEXT DESCRIPTION

In this section, we introduce the proposed Probabilistic Topic-Connection (PTC) model that addresses the problem in Corr-LDA model which we mentioned in Section 1.

We begin this section with the introduction of extended CorrLDA model, which is the state-of-the-art in modeling image and associated text description [8, 9]. By presenting the generative process of CorrLDA model, we make explicit its problem of topic replicating from text to image. Then, in a series of steps, we show how we address the problem in CorrLDA model by introducing new latent semantic variables and relations to the generative process. In Figure 5, we provide graphical representations of both extended CorrLDA model and our model, in which we highlight the innovation part of proposed model by red dash line.

Following the convention in depicting graphical representation of topic models, we use round nodes to represent random variables, in which the light blue nodes stand for latent random variables, while the yellow nodes denote observed ones during the model training. The rounded boxes are used to represent fixed hyper-parameters of the model, while the edges illustrate the conditional dependency underlying the generative process.



(a) Correspondence LDA (CorrLDA) model **(b) Probabilistic Topic-Connection (PTC) model**

Fig. 5 Graphical representation of the extended CorrLDA model and the Probabilistic Topic-Connection (PTC) model

For clarity, we name each image-text pair as one *document*. Some notations to be used in the two topic models are listed as follows: D is the number of documents, T is the anticipated number of latent topics, N_w^d is the total number of text words in document d, N_p^d denotes the total number of extracted multiple-word phrases in document d, while N_v^d and N_r^d represents the total number of extracted visual words and MSER regions in document d, respectively. In the model, parameters α, β_w, β_p *and* β_v are fixed hyper-parameters for the Dirichlet distributions. In our approach, we assume symmetric priors, with α, β_w, β_p *and* β_v being scalar parameters. Detailed explanations of notations used in Figure 5 and following discussions are summarized in Table 1.

3.1 Extended Correspondence LDA Model

The CorrLDA model is a generative model for correlated multiple type entity data. Similar to other generative models, the CorrLDA model assume that the observed data is generated by some parameterized random variable known as "latent topics". Specifically, a "word topic" (denotes by 'z' in Fig. 5a) is used to derive the generation of the text words from a topic-specific word distribution (e.g. for a word topic that is related to the concept of forest, the corresponding word distribution will have high probability generating words like 'branch', 'tree' and 'leaf'). In a text document, the word topics (which usually relate to some semantic concepts) play an intermediate role between basic elements (words) and high level semantic meanings.

The "visual topic" (denotes by 'y' in Fig. 5a) is a visual counterpart of the word topic, each had a unique distribution over image features. Specifically, each visual topic is formalized as cluster of features that represent similar image appearance or fit the same distribution in the image feature space. For example, the visual topic "trunks" may favor some image patterns that occur more frequently in trunks (e.g. vertical lines), while visual topic "leaves" may privilege some blob-like image patterns.

After identifying latent topics and assigning topic labels to each entity in a document, each document may in turn be represented by a document-level mixture of latent topics. The document-level topic mixture is defined as a probability distribution of latent topics with respect to each document, specifying which word topics are most likely to be generated from observed text description, or which kind of visual topics are most likely to be generated from observed image features.

When the instances from one entity type (say, the text words) serves as description of other entity types (such as tags, image features), the CorrLDA model directly use the latent topics of the former entities to generate the later entities. Consequently, both types of entities will share the same document-level topic composition, resulting in strong correspondence between them.

The generative process for the extend CorrLDA model (Fig.5a) is:

1. For the d^{th} (d=1…D) document, sample $\theta_d \sim Dir(\alpha)$
2. For the k^{th} (k=1…T) topic, sample $\varphi_k \sim Dir(\beta_w)$, $\varphi'_k \sim Dir(\beta_p)$ and $\psi_k \sim Dir(\beta_v)$.
3. For each of the N_w^d words w_i in document d:
   a) Sample a topic $z_i \sim Mult(\theta_d)$
   b) Sample $w_i \mid z_i = k \sim Mult(\varphi_k)$
4. For each of the N_p^d phrases p_i in document d:
   a) Sample a topic $z'_i \sim Uniform(z_{w_1}, ..., z_{w_{N_w^d}})$
   b) Sample $p_i \mid z'_i = k \sim Mult(\varphi'_k)$
5. For each of the N_v^d visual words v_i in document d:
   a) Sample a topic $y_i \sim Uniform(z_{w_1}, ..., z_{w_{N_w^d}})$
   b) Sample $v_i \mid y_i = k \sim Mult(\psi_k)$
6. For each of the N_r^d MSER region feature r_i in document d:
   a) Sample a topic $y'_i \sim Uniform(z_{w_1}, ..., z_{w_{N_w^d}})$
   b) For the n^{th} dimension of the MSER feature r_i^n
      i. Sample $r_i^n \mid y'_i = k \sim N(\mu_{k,n}, \sigma_{k,n}^2)$

In the step 1 of the generative process, a T-dimensional topic-prior vector $\boldsymbol{\theta_d}$ is sampled for each document d, with the k^{th} dimension of the vector represents the prior probability of the k^{th} topic in d. For each document d, the generative process of the N_w^d words is achieved by sampling topics from the document-topic multinomial distribution (with prior $\boldsymbol{\theta_d}$) and sampling words from the topic-word multinomial distribution (with prior $\boldsymbol{\varphi_k}$). After that, instead of being sampled from their own topics, all the other

entities (phrases, visual words, etc) are sampled from the same topic as words.

3.2 Probabilistic Topic-Connection Model

A closer look into the generative process of extended CorrLDA model reveals that, the document-level topic composition is only decided by the single-word feature, even though other entities such as the multiple word phrases also serve as a part of description. What's more, in the extended CorrLDA model, the image features are generated from the word topics. However, as we discussed in Section 1, each word topic may be related to multiple visual topics, enforcing word topics to image features may ignore such a relation and make topic modeling results inconsistent with the underlying image patterns.

With this consideration, in our new model, we allow each word topic to connect to multiple visual topics, with different prior probabilities. This model also allow for different number of word topics and visual topics. The generative process for the Probabilistic Topic-Connection (PTC) Model (Fig.5b) is:

1. For the d^{th} (d=1…D) document, sample $\theta_d \sim Dir(\alpha)$
2. For the k^{th} (k=1…T_1) text topic, sample $\varphi_k \sim Dir(\beta_w)$, $\varphi'_k \sim Dir(\beta_p)$ and $\lambda_k \sim Dir(\gamma)$.
3. For the j^{th} (j=1…T_2) visual topic, sample $\psi_j \sim Dir(\beta_v)$.
4. For each of the N_w^d words w_i in document d:
   a) Sample a text topic $z_i \sim Mult(\theta_d)$
   b) Sample $w_i \mid z_i = k \sim Mult(\varphi_k)$
5. For each of the N_p^d phrases p_i in document d:
   a) Sample a text topic $z'_i \sim Mult(\theta_d)$
   b) Sample $p_i \mid z'_i = k \sim Mult(\varphi'_k)$
6. For each of the N_v^d visual words v_i in document d:
   a) Sample an indicator $s_i \sim Mult(\theta_d)$
   b) Sample a visual topic $y_i \mid s_i = k \sim Mult(\lambda_k)$
   c) Sample $v_i \mid y_i = j \sim Mult(\psi_j)$
7. For each of the N_r^d MSER region feature r_i in document d:
   a) Sample an indicator $s'_i \sim Mult(\theta_d)$
   b) Sample a visual topic $y'_i \mid s'_i = k \sim Mult(\lambda_k)$
   c) For the n^{th} dimension of the MSER feature $r_i^{(n)}$
      i. Sample $r_i^{(n)} \mid y'_i = j \sim N(\mu_{j,n}, \sigma^2_{j,n})$

As represented in the generative process, new latent variables had been introduced to allow for more flexible sampling of word topics and visual topics. Specifically, latent variable $\boldsymbol{s}$ and $\boldsymbol{s'}$ play the role as word topic indicators, while latent variable $\boldsymbol{\lambda_k}$ serves as the prior probabilities of word topic k connecting to any visual topics. For a given image feature, the model firstly sample a word topic indicator, then sample the visual topic according to the priori distribution of corresponding word topic connecting to different visual topics.

3.3 An Explanation of Data Distributions

In both models, we deal with four types of entities: single-word, multiple word phrases, visual words and MSER region features. The topic-specific single-word distribution is modeled as a multinomial distribution over W different words, denoted by Multi($\boldsymbol{\varphi}$), in which $\boldsymbol{\varphi}$ is a W-dimensional prior vector. Similarly, we model the topic-specific multiple word phrases distribution as multinomial distribution. The visual words have similar statistical properties with text words, thus assumed to follow multinomial distributions, too. As mentioned in Section 2.2, each 50-dimensional MSER feature vector is obtained from the projection to the first 50 principle components of image patch eigenspace. Therefore, each dimension of MSER feature is real-valued and thus follows a Gaussian distribution, whose mean and variance are topic-specific. For all multinomial distributions, the Dirichlet distribution is used as prior, which yields posterior probability also in the form of Dirichlet distribution. For Gaussian distribution, the standard way is to draw its parameters from the conjugate prior, i.e. $\mu \sim N(\mu_0, r_0^2)$, $\sigma^2 \sim Inv-\chi^2(\nu_0, \sigma_0^2)$. In our approach, for the algebraic convenience in calculation, we place a non informative prior over the parameters μ and σ^2, in which the actual mean and variance are approximated by sample mean and variance.

Table 1. Notations in proposed topic model

d, w, p, v, r	Instances of variables: *d* for document, *w* for word, *p* for phrase, *v* for visual word, *r* for MSER region
D, W, P *V*	Total number of documents, vocabulary size of words, phrases, visual words
z, z', y, y'	Indicator for word topics (*z, z'*) and visual topics (*y, y'*)
T1,T2	The selected number of word topics and visual topics.
N_w^d, N_p^d, N_v^d, N_r^d	The number of word tokens, phrases, visual words and MSER regions contained in document d
$C_{kd}^{T_1D}$, $C_{kd,-i}^{T_1D}$	The number of times that word topic k has occurred in document d, with/without counting the current instance
C_{wk}^{WT} $C_{wk,-i}^{WT}$	The number of times that word *w* is assigned to word topic *k*, without counting the current instance.
C_{pk}^{PT} $C_{pk,-i}^{PT}$	The number of times that phrase *p* is generated from word topic *k*, with/without counting the current instance.
$C_{vj}^{VT_2}$ $C_{vj,-i}^{VT_2}$	Number of times that visual word *v* is generated from visual topic *j*, with/without counting the current instance.
$C_{jk}^{T_2T_1}$ $C_{jk,-i}^{T_2T_1}$	The number of times that word topic *k* connects to visual topic *j*, with/without counting the current instance.
$C_{rj,-i}^{RT_2}$	The number of times that MSER region r is generated from visual topic *j*, except current assignment;
θ	A $D \times T$ matrix that indicates the document-topic distribution.
α, β_v, β_v, β_v, γ	Hyper-parameters of Dirichlet distributions.
λ	A $T_1 \times T_2$ matrix that indicates the connection from word topic to visual topic
$\mu_{j,n}, \sigma^2_{j,n}$	Parameters of the nth Gaussian distribution with respect to visual topic *j*
$\bar{u}_{j,n}$	Sample mean of the nth Gaussian distribution with respect to visual topic *j*
$s^2_{j,n}$	Sample variance of the nth Gaussian distribution with respect to visual topic *j*

4. COLLAPSE GIBBS SAMPLING FOR PROPOSED TOPIC MODEL

In recent years, several methods have been developed for estimating the latent variable in topic model, such as the

variational expectation maximization, expectation propagation, and Collapse Gibbs sampling [31]. Compared to the other two methods, Gibbs sampling is less computationally intensive, and often yields relatively simple algorithms for approximate inference [31]. With this consideration, we perform the Collapse Gibbs Sampling procedure for model estimation. In the Gibbs Sampling process, a Markov chain is constructed and converges to the posterior distribution on latent topics. The transition between successive states in the Markov chain is modeled by repeatedly drawing a topic for each observed entity from the conditional probability. Due to the space limit, we only introduce our implementation of the Gibbs Sampling for proposed PTC model. For the extended CorrLDA model, our implementation is similar with that outlined in [8] and [9].

Given the generative process in Section 3.2, our objective is to compute the entity-topic posterior probability and sample topic for each entity from posterior probability. Thus, we derive the posterior sampling equations as follows, in which we follow the standard notations detailed in Table 1.

Sampling a word topic (z_i) for a given word (w_i)

$$\cdot\ p(z_i = k \mid w_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \mathbf{z}', \alpha, \beta_w) \propto \frac{C^{T_1 D}_{kd,-i} + \alpha}{\sum_{k'} C^{T_1 D}_{k'd,-i} + T_1\alpha} \frac{C^{WT}_{wk,-i} + \beta_w}{\sum_{w'} C^{WT}_{w'k,-i} + W\beta_w} \quad (1)$$

The above posterior probability is obtained by integrating out (collapsing) all the latent variables φ_k and θ_d separately.

Sampling a word topic (z'_i) for a multiple word phrase (p_i)

$$\cdot\ p(z'_i = k \mid p_i = p, \mathbf{z'}_{-i}, \mathbf{p}_{-i}, \mathbf{z}, \alpha, \beta_p) \propto \frac{C^{T_1 D}_{kd,-i} + \alpha}{\sum_{k'} C^{T_1 D}_{k'd,-i} + T_1\alpha} \frac{C^{PT}_{pk,-i} + \beta_p}{\sum_{p'} C^{PT}_{p'k,-i} + P\beta_p} \quad (2)$$

Sampling a visual topic (y_i) for a visual word feature (v_i)

$$\cdot\ p(y_i = j, s_i = k \mid v_i = v, \mathbf{y}_{-i}, \mathbf{v}_{-i}, \mathbf{y}', \mathbf{z}', \mathbf{z}, \gamma, \beta_v) \propto \frac{C^{T_1 D}_{kd}}{N^d_w} \frac{C^{T_2 T_1}_{jk,-i} + \gamma}{\sum_{j'} C^{T_2 T}_{j'k,-i} + T_2\gamma} \frac{C^{VT_2}_{vj,-i} + \beta_v}{\sum_{v'} C^{VT_2}_{v'j,-i} + V\beta_v} \quad (3)$$

Sampling a visual topic (y'_i) for a MSER region (r_i), in which $r_i = r = (r^{(1)}, \ldots, r^{(n)}, \ldots, r^{(50)})^T$.

$$\cdot\ p(y'_i = j, s'_i = k \mid r_i = r, \mathbf{y'}_{-i}, \mathbf{r}_{-i}, \mathbf{y}', \mathbf{z}', \mathbf{z}, \gamma) \propto \frac{C^{T_1 D}_{kd}}{N^d_w} \frac{C^{T_2 T_1}_{jk,-i} + \gamma}{\sum_{j'} C^{T_2 T}_{j'k,-i} + T_2\gamma} \prod_n t_{C^{RT_2}_{rj,-i}-1}\left(r^{(n)} \mid \bar{u}_{j,n}, s^2_{j,n} / C^{RT_2}_{rj,-i}\right) \quad (4)$$

In equation 4, the term in the form of $t_{n-1}\left(r^{(n)} \mid \bar{u}, s^2/n\right)$ is the student-t density with mean $\bar{u}$, variance s^2/n and n-1 degree of freedom. As we place a non-informative prior over the Gaussians, the mean and variance of each Gaussian are purely determined by their sufficient statistics (i.e. the sample mean $\bar{u}$ and sample variance s^2 , respectively). As a result, the student-t density function in equation 4 provides the confidence of drawing the value of $r^{(n)}$ from a topic-specific Gaussian distribution (please refer to [32] Ch. 3.2 for detailed derivation of this conclusion).

During the Gibbs Sampling processes based on above posterior distributions calculations, we may also update single latent variables in the following manner:

$$E[\theta_{kd} \mid \mathbf{z}, \mathbf{z}', \alpha] = \frac{C^{T_1 D}_{kd} + \alpha}{\sum_{k'} C^{T_1 D}_{k'd} + T_1\alpha}$$

$$E[\varphi_{wk} \mid \mathbf{z}, \mathbf{w}, \beta_w] = \frac{C^{WT}_{wk} + \beta_w}{\sum_{w'} C^{WT}_{w'k} + W\beta_w}$$

$$E[\varphi'_{pk} \mid \mathbf{z}', \mathbf{p}, \beta_p] = \frac{C^{PT}_{pk} + \beta_p}{\sum_{p'} C^{PT}_{p'k} + P\beta_p} \quad (5)$$

$$E[\psi_{vj} \mid \mathbf{y}, \mathbf{v}, \beta_v] = \frac{C^{VT_2}_{vj} + \beta_v}{\sum_{v'} C^{VT_2}_{v'j} + V\beta_v}$$

$$E[\lambda_{jk} \mid \mathbf{z}, \mathbf{z}', \mathbf{y}, \mathbf{y}', \gamma] = \frac{C^{T_2 T_1}_{jk} + \gamma}{\sum_{j'} C^{T_2 T}_{j'k} + T_2\gamma}$$

5. EXPERIMENTAL RESULTS

In this section, we apply the proposed PTC model to topic learning and compare the performance of our model with that of the extended Correspondence LDA (Corr-LDA) model under the same image annotation scenario using cross-validation. The performance of automatic image annotation is evaluated by perplexity and annotation accuracy.

5.1 Data Collection and Settings

The image dataset used in our study is downloaded from the ImageNet database (http://www.image-net.org/) under the granted access permission, following the term of access. The ImageNet is built on the hierarchical ontology structure provided by WordNet, in which each node involves a group of images that depict a particular concept named as a synonym set, or "synset". Specifically, we download a total of 508 synsets under the "flower" subtree, 1473 synsets under the mammal subtree and 1118 synsets under the tree subtree. Following the term mapping schema in Section 2.1, we map each synset to a Wikipedia article that describes the same concept. Then, we parse the structured content of Wikipedia articles and apply a rule-based method to identify the explanative sections. Unrelated sections such as "External_links" and "References" are removed from the articles. After that, to ensure the quality of text description, we filter out articles with insufficient words (<200 words). The qualified articles then serve as text description for corresponding ImageNet synsets. In total, we obtain comprehensive text descriptions for 1452 synsets (330, 562 and 560 synsets for subtrees "flower", "mammal" and "tree", respectively).

For each of the 1452 synsets, we randomly select 5 images from the corresponding image group and adjust them to normalized size (640×480 pixels). After that, we replicate the text descriptions to each of the 5 images, resulting in 5 image-text pairs. As introduced in Section 2, we make index for single-words and multiple word phrases in the text descriptions, and extract visual-word features as well as MESR region features from images. In total, we indexed 5,699,505 word tokens which belong

to 35,744 different words, 624,205 multiple word phrases from a total number of 13078 unique phrases, 7,945,075 visual words (an average of 1095 visual words per image) from a vocabulary size of 2000, and a total of 924,924 MSER region features (an average of 127 MSER regions per image). The original dataset is divided into 5 subsets with equal size. Of the 5 subsets, one subset (20%) is retained as the validation data for testing the model, and the remaining 4 subsets (80%) are used as training data. For image annotation evaluation, the cross-validation process repeats 5 times, with each of the 5 subsets used once as the validation data. After that, we take the average results for evaluation.

5.2 Topic Learning and Representation

The estimation of the proposed probabilistic topic model is achieved by performing Gibbs Sampling over training dataset until convergence (generally, the model takes less than 100 iterations to converge). Once the topic model is estimated from the training dataset, we will be able to evaluate it by log-likelihood and visualize the uncovered latent topics.

5.2.1 Likelihood Comparison

Log-likelihood is one of the standard criteria for generative model evaluation. It provides a quantitative measurement of how well a topic model fits the training data. The score of log-likelihood (which is a negative number) is the higher the better. In practice, the log-likelihood of elements given latent topics can be calculated by integrating out all the latent variables.

In our study, we are interested in which topic model is more suitable to study the latent patterns of image features. Thus, instead of calculating word-likelihood, we choose to evaluate the log-likelihood of visual words for both models. In the proposed probabilistic topic-connection (PTC) model, The marginal likelihood of visual words v given all the visual topics y is p($\boldsymbol{v}|\boldsymbol{y}$), which can be calculated by integrating out latent variables $\boldsymbol{\psi}$:

$$p(\mathbf{v}\mid\mathbf{y}) = \prod_{j=1}^{T_2}\left[\int_{\psi_j} p(\mathbf{v}\mid y_j,\psi_j)\,p(\psi_j\mid y_j)\,d\psi_j\right] = \left[\frac{\Gamma(V\beta_v)}{\Gamma(\beta_v)^V}\right]^{T_2}\cdot\prod_{j=1}^{T_2}\frac{\prod_v(C_{vj}^{VT_2}+\beta_v)}{\Gamma(\sum_{v'}C_{v'j}^{VT_2}+V\beta_v)} \quad (6)$$

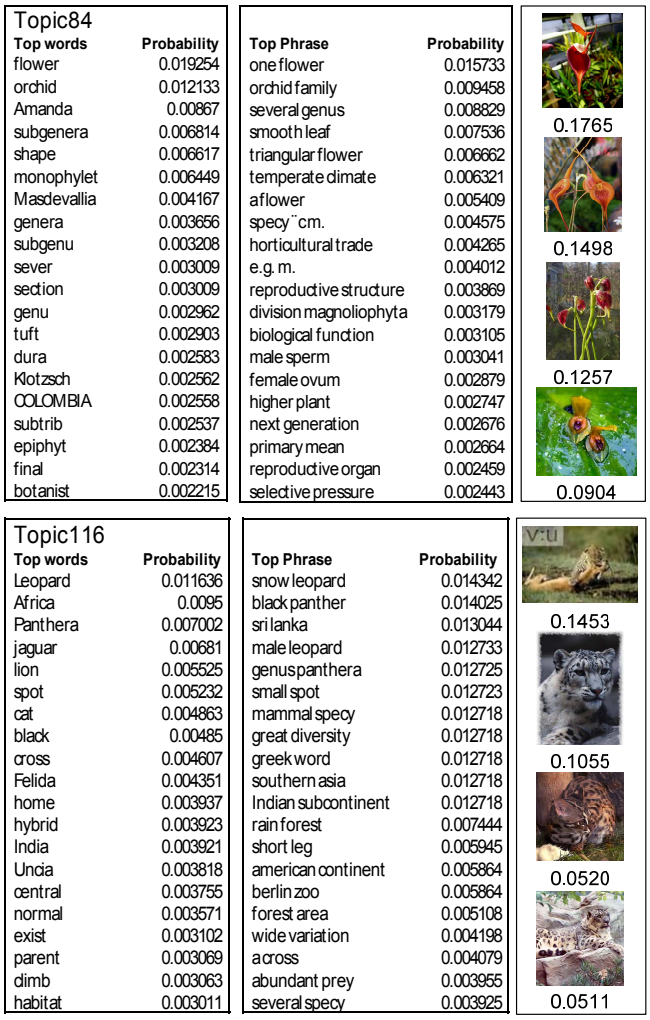
The final log-likelihood of visual words is obtained by taking the logarithm of eq. (6) and averaging the resulting summation by V.

For the extended Corr-LDA model, the log-likelihood can be calculated in a similar way, the only difference is, instead of using their own latent topics, the visual words in Corr-LDA model directly use latent topics generated from text words.

In Fig. 7a, we plot the log-likelihood for both models under different topic number (to make this comparison fair, the number of word topic and visual topics are made equal). It shows that our model has higher log-likelihood than Corr-LDA model, which means that our model fits training data better. It also shows that the log-likelihood of both models increase as the number of topic increase, which suggests that a relatively greater topic number may potentially fit the training data better. However, it should be noted that there is a trade-off between topic numbers and convergence time of model estimation, and the unbounded increase of topic number may results in an over-fitting problem.

5.2.2 Illustration of Uncovered Latent Topics

In order to better interpret the uncovered latent topics, we visualize the word topics by providing the top-ranked words, top-ranked phrases and most related images. As represented in Fig. 6, the words and phrases are sorted by their probability of being generated from a word topic, while images are sorted by the probability of containing that word topic (by counting the topic indicator variables of image features).



Topic84

Top words	Probability
flower	0.019254
orchid	0.012133
Amanda	0.00867
subgenera	0.006814
shape	0.006617
monophylet	0.006449
Masdevallia	0.004167
genera	0.003656
subgenu	0.003208
sever	0.003009
section	0.003009
genu	0.002962
tuft	0.002903
dura	0.002583
Klotzsch	0.002562
COLOMBIA	0.002558
subtrib	0.002537
epiphyt	0.002384
final	0.002314
botanist	0.002215

Top Phrase	Probability
one flower	0.015733
orchid family	0.009458
several genus	0.008829
smooth leaf	0.007536
triangular flower	0.006662
temperate climate	0.006321
a flower	0.005409
specy¨cm.	0.004575
horticultural trade	0.004265
e.g. m.	0.004012
reproductive structure	0.003869
division magnoliophyta	0.003179
biological function	0.003105
male sperm	0.003041
female ovum	0.002879
higher plant	0.002747
next generation	0.002676
primary mean	0.002664
reproductive organ	0.002459
selective pressure	0.002443

0.1765

0.1498

0.1257

0.0904

Topic116

Top words	Probability
Leopard	0.011636
Africa	0.0095
Panthera	0.007002
jaguar	0.00681
lion	0.005525
spot	0.005232
cat	0.004863
black	0.00485
cross	0.004607
Felida	0.004351
home	0.003937
hybrid	0.003923
India	0.003921
Uncia	0.003818
central	0.003755
normal	0.003571
exist	0.003102
parent	0.003069
climb	0.003063
habitat	0.003011

Top Phrase	Probability
snow leopard	0.014342
black panther	0.014025
sri lanka	0.013044
male leopard	0.012733
genus panthera	0.012725
small spot	0.012723
mammal specy	0.012718
great diversity	0.012718
greek word	0.012718
southern asia	0.012718
Indian subcontinent	0.012718
rain forest	0.007444
short leg	0.005945
american continent	0.005864
berlin zoo	0.005864
forest area	0.005108
wide variation	0.004198
across	0.004079
abundant prey	0.003955
several specy	0.003925

0.1453

0.1055

0.0520

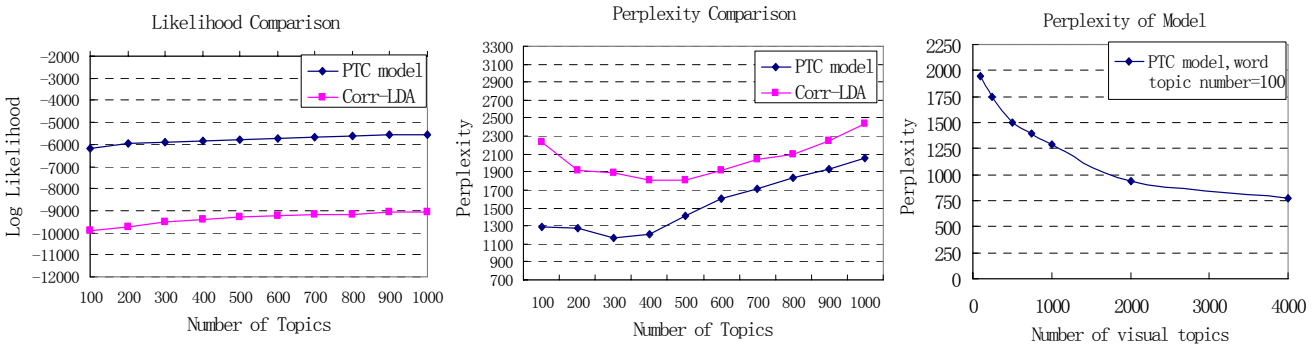
0.0511

Fig. 6 Illustration of uncovered latent topics by PTC model

In Fig. 6, we present two examples of uncovered latent word topics. The former one is a topic related to the concept of "orchid", while the later one is a topic related to the concept of "leopard". By providing a combination of words, multiple word phrases and images, it becomes much easier to interpret the domain knowledge captured by each topic. As we can see, the uncovered latent topics show high consistency to semantic concepts.

5.3 Image Annotation and Evaluation

Since we are targeting modeling the connection between image and text, we are interested in its ability of predicting missing text descriptions from image features. With this consideration, we conduct an automatic image annotation experiment, in which we compare our model with the Corr-LDA model. During the experiment, the text description of testing data is considered as unknown (missing). The predictive ability of both models is evaluated by both perplexity and annotation accuracy.



(a) Likelihood comparison **b) Perplexity comparison (# of visual topics=1000)** **(c) Perplexity vs. visual topic numbers**

Fig. 7 The likelihood and perplexity comparison of the proposed PTC model and the extended Corr-LDA model

5.3.1 Perplexity Comparison

The perplexity is a standard criterion for topic models that evaluates how well the model predicts the new data. Specifically, the perplexity of a set of testing documents $d \in D_{test}$ is defined as the exponential of the negative normalized per-word predictive log-likelihood using parameters from the trained topic model. The score of perplexity is the lower the better.

With uncovered latent topics from training image-text pairs, the problem of estimating topic priors in testing images can be approximated by performing Gibbs sampling over observations of image features, while keeping all the topic-entity conditional probability fixed. It should be noted that we need to know the posterior probability of word topic indicators given visual topics: $p(\mathbf{s}|\mathbf{y})$ when estimating the new document-level word topic prior. In our study, this probability is approximated by counting the number of evidences across the training dataset.

Upon the convergence of the Gibbs sampling process over testing data, the word perplexity of testing image-text pairs is:

$$Perplexity = \exp\left[\frac{-\sum_{d_{test}} \log p(\mathbf{w}^{\mathbf{d}}, \mathbf{p}^{\mathbf{d}} \mid \mathbf{v}^{\mathbf{d}}, \mathbf{r}^{\mathbf{d}})}{\sum_{d_{test}} (N_w^d + N_p^d)}\right] \quad (7)$$

One advantage of our model is that it assigns different topic numbers to different types of data, which makes this model more suitable to deal with image and associated text. Fig.7b represents the perplexity comparison between our model and the Corr-LDA model as the increase of word topic number, in which the number of visual topics in our model is fixed to 1000. It shows that the perplexity of our model is consistently lower than Corr-LDA model, which suggests that our model is 'less surprised' by the testing data, thus demonstrates better performance. Also, it shows that the predictive ability of our model may benefit from greater visual topic number, as it tend to have lower perplexity as the visual topic number increases (Fig. 7c)

5.3.2 Annotation Accuracy Comparison

Upon the convergence of the Gibbs sampling process over testing data, the probability of annotating the 'missing' words and phrases of an image can be calculated via the production of document-level word topics prior probability and the topic-word/phrase conditional probability. Words and phrases with highest probability are then used as the annotation. After that, the image annotations are compared with the ground truth, in which the cross-validation process repeats 5 times, and the results are averaged to produce the final annotation accuracy. In our study, the annotation accuracy of our model and Corr-LDA model are compared under their best performance (i.e. 1000 visual topics and 300 word topics for PTC model, and 500 word topics for Corr-LDA model). The experiment result (Fig. 8) shows that our model consistently outperforms Corr-LDA in both word and phrase annotations.

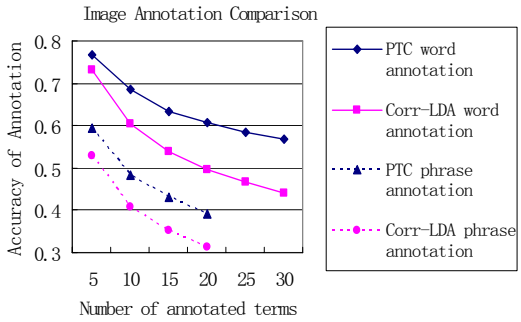


Fig. 8 Image annotation comparison

6. CONCLUSIONS

In this paper, a probabilistic topic-connection model is proposed to deal with the problem of modeling images and associated text description. Specifically, new latent variables have been introduced to allow for more flexible sampling of word topics and visual topics, in which one word topic may connect to multiple visual topics. The proposed model provides better representation of the connection between latent semantic topics and latent image patterns, thus achieves better performance in the task of automatic image annotation compared to the traditional Corr-LDA model.

7. ACKNOWLEDGEMENT

This research work is supported in part from the NSF Career grant IIS 0448023, NSF CCF 0905291, NSF IIP 0934197, NSFC 90920005 "Chinese Language Semantic Knowledge Acquisition and Semantic Computational Model Study," and the Program of Introducing Talents of Discipline to Universities B07042 (China).

8. REFERENCES

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content-based image rerieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] Lew, M. S., et al. Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl., 2006.

[3] Jia Li and James Z. Wang, ``Real-time Computerized Annotation of Pictures," Proceedings of the ACM Multimedia Conference, pp. 911-920, ACM, Santa Barbara, CA, October 2006.

[4] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories from Google's Image Search," Proc. Int'l Conf. Computer Vision, vol. II, pp. 1816-1823, Oct. 2005.

[5] Changhu Wang, Lei Zhang, Hong-Jiang Zhang. Learning to Reduce the Semantic Gap in Web Image Retrieval and Annotation, in Proc. of the 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR), Singapore, July

[6] David M. Blei, Michael I. Jordan: Modeling annotated data. SIGIR 2003: 127-134

[7] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, Nuno Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 3, pp. 394-410, Mar. 2007

[8] Amr Ahmed, Eric P. Xing, William W. Cohen, Robert F. Murphy, Structured Correspondence topic models for mining captioned figures in biomedical literature, Proceedings of the 15th ACM SIGKDD International conference on Knowledge discovery and data mining, June 28-July 01, 2009, Paris, France.

[9] X. Chen, C. Lu, Y. An, and P. Achananuparp. Probabilistic Models for Topic Learning from Images and Captions in Online Biomedical Literatures. In the Proceedings of 18th ACM Conference on Information and Knowledge Management (CIKM'09).

[10] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. PAMI, 28(4):594–611, April 2006.

[11] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, Caltech, 2007.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/ challenges/VOC/voc2008/workshop/.

[13] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. IJCV, 77(1- 3):157–173, May 2008.

[14]J. Deng, W. Dong, R. Socher, L. –J. Li and L. Fei-Fei, ImageNet: A Larget-Scale Hierarchical Image Database. IEEE Compter Visual and Pattern Recognition (CVPR), 2009.

[15] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[16] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, Xiaohua Zhou: Exploiting Wikipedia as external knowledge for document clustering. KDD 2009: 389-396

[17] Hu, J., Fang, L., Cao, Y., et al. Enhancing Text Clustering by Leveraging Wikipedia Semantics. In Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (Singapore, July 20 – 24, 2008). ACM Press, New York, NY, 179-186.

[18] Wang, P. and Domeniconi, C. 2008. Building Semantic Kernels for text classification using Wikipedia. In Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (Nevada, Las Vegas, August 24 – 27, 2008). ACM Press, New York, NY, 713-721.

[19] F. Smadja, Retrieving collections from text: Xtract. Computational Linguistics, 1993, 19(1), pp. 143-177

[20] J. Yang, Y. G. Jiang, A. G. Hauptmann, C. W. Ngo, Evaluating Bag-of-Visual-Words Representations in Scene Classification. ACM SIGMM Int'l Workshop on Multimedia Information Retrieval (MIR'07), Augsburg, Germany, Sep. 2007.

[21] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. International Journal of Computer Vision, vol. 73, no. 2, June 2007, pp. 213-238

[22] Yu-Gang Jiang, Chong-Wah Ngo, Jun Yang: Towards optimal bag-of-features for object categorization and semantic video retrieval. CIVR 2007: 494-501

[23] Lowe, D. Distinctive Image Features from Scale-Invariant Key Points. International Journal of Computer Vision, 60(2): 91–110, 2004.

[24] Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. International Conference on Computer Vision. (2003) 1470– 1477

[25] J. Matas, O. Chum, U. M., T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In BMVC, 2002.

[26] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. IEEE T. PAMI, 27(10):1615–1630, 2005.

[27] L.-J. Li, R. Socher and L. Fei-Fei. Towards Total Scene Understanding:Classification, Annotation and Segmentation in an Automatic Framework. Computer Vision and Pattern Recognition (CVPR) 2009.

[28] Zhong Wu, Qifa Ke, M. Isard, Jian Sun,Bundling features for large scale partial-duplicate web image search Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (18 August 2009), pp. 25-32.

[29] Per-Erik Forssén and David G. Lowe, "Shape descriptors for maximally stable extremal regions," International Conference on Computer

[30] Van Rijsbergen, C.J., Information Retrieval, Butterworths, 1975.

[31] T. L. Griffiths, M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101:5228-5235, 2004.

[32] A. Gelman, J. Carlin, H. Stern, and D. Rubin. Bayesian Data Analysis 2nd edition. Chapman-Hall, 2003.