Deep Image Annotation and Classification by Fusing Multi-Modal Semantic Topics

YongHeng Chen[1], Fuquan Zhang[2], WanLi Zuo[3]
[1] College of Computer Science, Minnan Normal University, zhangzhou 363000, China
Key Laboratory of Data Science and Intelligence Application, Fujian Province University
[e-mail: yh_chen@mnnu.edu.cn]
[2]School of Software,Beijing Institute of Technology,Beijing,100081,China
Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang Univeristy, Fuzhou,350121,China
[e-mail: 8528750@qq.com]
[3]College of Computer Science and Technology, Jilin University, Changchun, China
[wanli@jlu.edu.cn]
*Corresponding author: YongHeng Chen

Received May 17, 2017; revised August 24, 2017; accepted September 11, 2017; published January 31, 2017

Abstract

Due to the semantic gap problem across different modalities, automatically retrieval from multimedia information still faces a main challenge. It is desirable to provide an effective joint model to bridge the gap and organize the relationships between them. In this work, we develop a deep image annotation and classification by fusing multi-modal semantic topics (DAC_mmst) model, which has the capacity for finding visual and non-visual topics by jointly modeling the image and loosely related text for deep image annotation while simultaneously learning and predicting the class label. More specifically, DAC_mmst depends on a non-parametric Bayesian model for estimating the best number of visual topics that can perfectly explain the image. To evaluate the effectiveness of our proposed algorithm, we collect a real-world dataset to conduct various experiments. The experimental results show our proposed DAC_mmst performs favorably in perplexity, image annotation and classification accuracy, comparing to several state-of-the-art methods.

Keywords: multi-modal topic model, image annotation, image classification, nonparametric Bayesian statistics, variational inference algorithm

YongHeng Chen's word was supported by the National Natural Science Foundation of China under Grant No. 60373099, No. 60973040, and No. 61303131;

1. Introduction

Traditional techniques of information retrieval are managed with the analysis of a document in a (high-dimensional) word-space at its core. As today's multimedia content becomes increasingly multi-modal with texts accompanying images and videos in the form of content description, transcribed text, or captions, current state-of-the-art multimedia search technology relies heavily on these collateral annotation texts to identify and retrieve images and video. Besides the fact that users often prefer the use of textual queries over examples, an important benefit of such an approach is the high-level semantic retrieval, e.g. retrieval of abstract concepts, that could not be achieved with low-level visual cues used in most query-by-example systems [7]. However, such task still faces a main challenge, which is how to build a joint model for bridging and revealing semantic information from information that spans multiple modalities [1]. This challenge has been referred to as the "semantic gap" problem across different modalities.

With annotation texts playing an increasingly vital role in modern multimedia retrieval systems, a relevant question one might ask is how to deal with numerous fast-growing user-generated content that often lacks descriptive annotation texts and classification label which would enable accurate semantic retrieval to be performed [7]. Image classification and annotation are two classical problems in modern multimedia information retrieval. Image annotation task aims to develop techniques to reliably describe the different objects depicted in the images by borrowing some annotation terms [2]. Matrix factorization [3], multi-label learning [4, 5] and probabilistic topic models [6, 7, 8, 9, 10, 11] have been developed. Given an image, image classification tells people what is the theme of the image according to its visual content from a high-level semantic meaning perspective. There has also been work on applying traditional supervised learning methods to perform classification, including support vector machine [12, 13], random forest [14, 15], and probabilistic topic models [16, 17, 18, 19, 20, 21, 22, 23]. It can be seen that probabilistic topic model has become popular in the computer vision community due to its solid theoretical foundation and promising performance. However, most existing probabilistic topic models for image annotation and classification are parametric probabilistic models, which require to determine the topic number beforehand. The selection of number of topics can have a significant impact on how well the model fits the data, and its ability to generalize on the training data. This is especially problem for applications where the topic space lacks a clear semantical interpretation, as often is the case in computer vision. Otherwise, there has been a lot of work on image classification and annotation separately, but few people focus on solving them jointly. However, because of the existing strongly relationship between the image classification and image annotation, this attracted much effort in combining both label and annotation information to design jointly model for image classification and annotation simultaneously [24, 25, 26, 27]. But the mentioned methods above typically are restricted to exploiting words associated to visible objects depicted in the images by only considering the keywords, captions or category labels. As a consequence, they are not able to exploit the entire information in which images with related loosely free-text descriptions. This joint model should be able to project multi-modal information to the same semantic space and make use of the full structure of documents that pair a body of text with a number of images, which are widely emerging on the internet, like newspaper articles, web-pages, and technical articles [2].

In this work, we concentrate on documents in which images with related loosely descriptions and class label, which are a natural way of providing rich information about the image, although many of the ideas would be applicable to other modalities. In this case, class

label and loosely related descriptions can be viewed as global and local description of image respectively, which can serve as a valuable complementary source of information for image annotation and classification. As shown in **Fig. 1**, objects irrelevant to the description of the image, such as suit and robe, are not present in the text, while non-visual words, such as energy, cooperation and terrorism, strongly help understanding the image at a high level. To address this issue, we propose a deep image annotation and classification by fusing multi-modal semantic topics (DAC_mmst) model, which is a joint topic model for image classification and annotation simultaneously. Conditioned on the shared latent semantic space, we leverage the information of the rich text loosely related to an image to reveal non-visual topics and visual topics. Visual annotation terms are supposed to extract from one of visual topics that is related with image regions and non-visual annotation terms from one of non-visual topics. The revealed non-visual words reduce the information loss of the traditional annotation and achieve deep annotation for images. At the same time, images classification can also be implemented according to revealed visual topics and non-visual topics. Moreover, DAC_mmst model provides the more flexible method to select the number of visual-topics by introducing Hierarchical Dirichlet Process (HDP) [28] to model each image as a Dirichlet process for topics discovery.

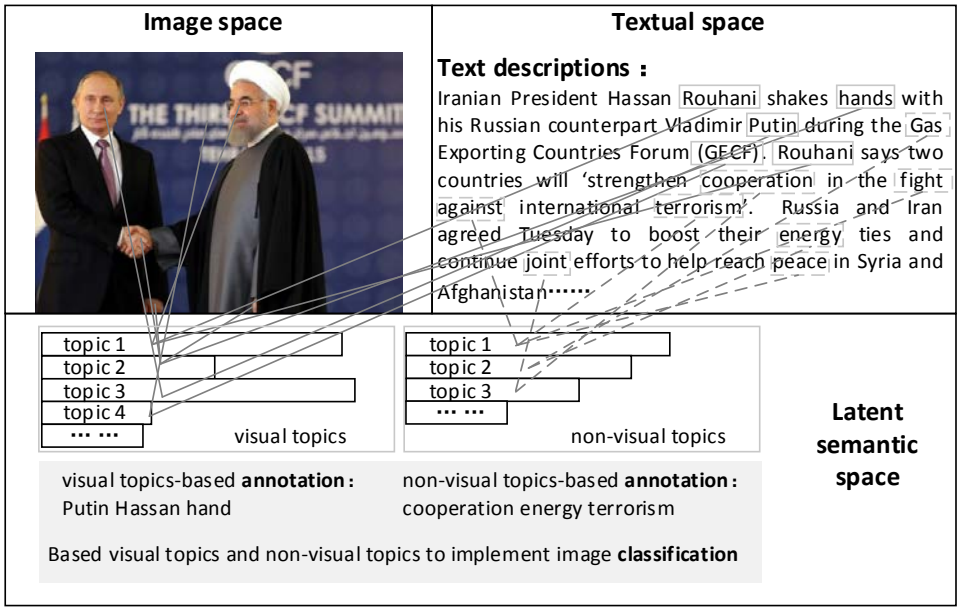

Fig. 1. Example of an image with a loosely related text description and class labels.

The outline of this paper is organized as follows. Related works are briefly summarized and discussed in Section 2. Section 3 introduces problem formulation and framework overview. In Section 4, approximate variation inference is given. Experimental results are reported and discussed in Section 5, followed by the conclusion in Section 6.

2. Related Work

Previous works in this area have primarily focused on looking for feature representations shared by the multi-modal data [30, 31, 32, 33, 34], which consider a kernelized discriminative canonical correlation analysis and casting the problem of image annotation to a classification problem based on feature extraction and representation. While

most of these approaches only consider the conditional distribution of image features under given annotation features rather than reveal the intrinsic semantic correlation of those features, which is not conducive to estimating the image annotation.

As previously mentioned, works on image classification and annotation are often modeled using a topic model, the most popular being Latent Dirichlet Allocation or LDA [29]. Latent Dirichlet allocation (LDA) is a Bayesian network that models a document by a latent low dimensional space spanned by a set of automatically revealed topical bases. Each topic defines a multinomial distribution over the original features and is assumed to have been drawn from a Dirichlet. To model multi-modal information, a wide variety of its extensions has been proposed in the area of image annotation, such as mm-LDA [6], Corr-LDA [6] and tr-mmLDA [7]. mm-LDA and Corr-LDA introduce a set of shared latent variables to represent the underlying causes of cross-correlations in the multi-modal data, thus enforcing strong correlations between textual and images. Therefore, relying on a latent variable regression approach to correlate latent variables of the two modalities, tr-mmLDA can linearly associate them to capture the correlated topic vectors, which allows the number of topics in the two data modalities to be different and relax the one-to-one correspondence between the topics of each modality. For image classification, Fei-fei Li [17] firstly used bag-of-word feature with the help of modified LDA model to learn and recognize natural scene categories automatically and without supervision. A. Bosch [20] uses probabilistic Latent Semantic Analysis(pLSA) [34] to achieve scene classification by a k-nearest neighbor classifier. Because there is a relationship between the image class label and image annotation terms, image annotation and classification should not be tackled separately. But these models only consider single task (image classification or image annotation), which cannot be used to achieve image annotation and classification simultaneously. MMLDA [24] integrates max-margin discriminative learning into generative topic models to achieve image annotation and classification. But, they also conduct these two tasks separately. Wang et al. [25] extended supervised topic modeling (sLDA) [36] to achieve image classification and annotation simultaneously, which leverage the relationship between the image annotation and class label by finding a latent topic space predictive of both.

Unfortunately, the existing multi-modal LDA models still need strong correlation between textual and images and rely on a limited textual representation, such as keywords and captions. However, in more realistic scenarios the images are related to richly and loosely related text. In such dataset, the image is not clearly representative of all the words in the text, and most of the text information is neglected. Otherwise, multi-modal models based on LDA inherit the defect of LDA model that the number of topics must be pre-determined. Considering all the problems mentioned above, in this paper, we propose a deep image annotation and classification by fusing multi-modal semantic topics (DAC_mmst) model to achieve image annotation and classification simultaneously. The main contributions of our work can be summarizes as follows:

(1) The proposed DAC_mmst model can effectively model multi-modal documents including long text with related images and learn the correlations between textual and visual modalities by capturing both types of topics, visual-topic and non-visual topic. The revealed non-visual topics reduce the information loss of the traditional multi-modal LDA models.

(2) In order to provide the more flexible method to select the number of visual-topics, DAC_mmst is the first joint model of image annotation and classification based on non-parametric HDP, which models each image as a Dirichlet process for topics discovery.

(3) We significantly boost the classification accuracy by considering the deeply non-visual topics.

3. MODEL

3.1 Data Representation

Without loss of generality, assume that we have a corpus with a collection of D multimedia documents, denoted by $D= \{d_1, d_2 \ldots, d_D\}$. In order to describe multi-modal information, we adopt the bag-of-word method for each modality. Each multimedia document consisting of an image and its corresponding descriptions text is thus summarized in our representation as a pair of vectors of bag-of-word counts corresponding to visual and textual information, respectively. The description text is a collection of M word occurrences denoted by $W = \{w_1, w_2 \ldots, w_M\}$, where each annotation word w_i defined as a unit-basis vector of T_w with exactly one ono-zero entry representing the membership to only one word in a vocabulary of T_w. To model visual information, following the work [17], we borrow a standard approach to obtain a bag of visual words by running the k-means algorithm on SIFT descriptors densely exacted from all training images. From that point on, any image can be represented a collection of N word occurrences denoted by $R = \{r_1, r_2 \ldots, r_N\}$, where each image region term r_i is similarity denoted for a vocabulary of size T_r and is the index of the closest K-means cluster to the i^{th} SIFT descriptor extracted from the image. Since we focus on classification tasks, a discrete response variable c is introduced from a label set $c \in C$, which is generated as a response to the visual topics along with the visual words

3.2 Sethuraman's stick-breaking construction for HDP

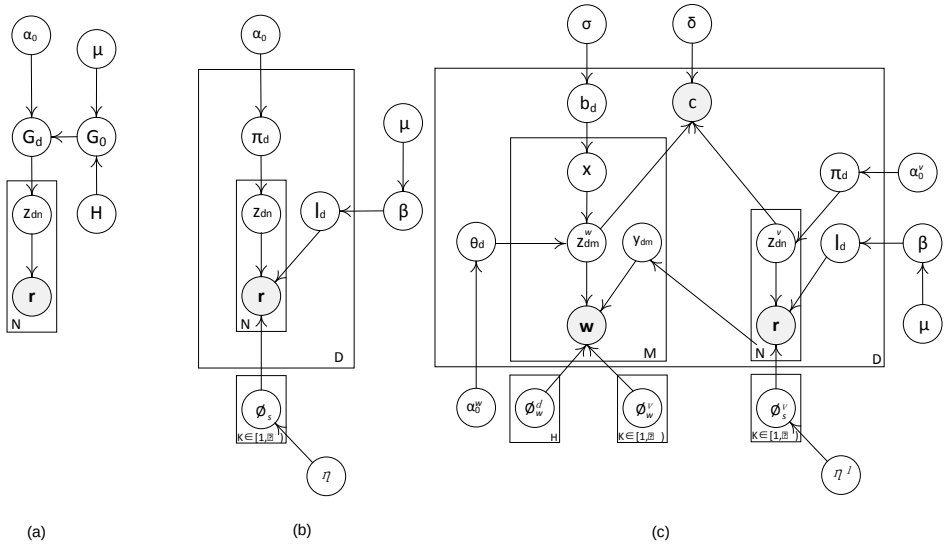

Fig. 2. Graphical model representations for (a) HDP (b) Sethuraman's stick-breaking construction for HDP (c) DAC_mmst model.

Before presenting the deep image annotation and classification by fusing multi-modal semantic topics (DAC_mmst) model, let us review the basic Hierarchical Dirichlet Process (HDP). The Dirichlet process (DP) is a distribution over distributions. It is a typical method for non-parametric Bayesian process, which is represented by DP(α_0, G_0), where α_0 is a concentration parameter and G_0 is a base measure parameter. Each document could be modeled as a DP, and each word in document d is a target object that is created by the distribution words over a topic sampled from the distribution of document based mixing vector over infinite number of topics. So DP allows the number of model parameters to grow

as more data is observed. However, in some applications, we may be interested in modeling groups of data with shared mixture components and prior over mixing measures.

To allow sharing data among the collection of topics across documents, Hierarchical Dirichlet Processes (HDP) [28] is proposed, which generates G_0 in the upper level DP from the mixture component space H as the common base measure to ensure that each group shares mixture components with a positive probability. Therefore, HDP uses multiple DPs to model multiple correlated documents sets, where each document is modeled as a DP mixture. As shown in **Fig.2 (a)** mathematically,

$$\begin{aligned} G_0 \mid \{\mu, H\} &\sim DP(\mu, H) \\ G_d \mid \{\alpha_0, G_0\} &\sim DP(\alpha_0, G_0) \\ z_{di} \mid G_d &\sim G_d \qquad d \in \{1, \ldots, D\} \\ r_{di} \mid z_{di} &\sim Mult(z_{di}) \qquad i \in \{1, \ldots, N_d\} \end{aligned} \tag{1}$$

where G_0 is a global topic distribution shared by all the documents collection, H is symmetric Dirichlet distribution parametrized by η, defined over a V-dimensional simplex, and each document d is generated based on local random measures $G_d \in DP(\alpha_0, G_0)$ that are also distributed as Dirichlet process and conditionally independent given G_0 with concentration parameter α_0 and base probability measure G_0.

The definition of the HDP in Eq. 1 is implicit. Wang [37] presented a more constructive representation of the HDP using two stick-breaking representations of a Dirichlet distribution called Sethuraman's stick-breaking construction as **Fig.2 (b)**. For the corpus-level DP draw, the base distribution $G_0 \mid \mu, H \sim DP(\mu, H)$ can be described using a stick-breaking representation as

$$\beta'_k \sim Beta(1, \mu) \quad \lambda_k \sim H \quad \beta_k = \beta'_k \prod_{s=1}^{k-1}(1 - \beta'_s) \quad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\lambda_k} \tag{2}$$

where the variables $\{\beta_k\}$ are viewed as the stick-breaking weights that sum to one, G_0 is discrete and has support at the topic λ_k with weights β_k, H is typically a symmetric Dirichlet distribution with parameter η. The per-document stick breaking representation, $G_d \mid \alpha_0, G_0 \sim DP(\alpha_0, G_0)$, can be descripted as

$$\pi'_{dt} \sim Beta(1, \alpha_0) \quad \psi_{dt} \sim G_0 \quad \pi_{dt} = \pi'_{dt} \prod_{s=1}^{t-1}(1 - \pi'_{dt}) \quad G_d = \sum_{t=1}^{\infty} \pi_{dt} \delta_{\psi_{dt}} \tag{3}$$

where the variables $\{\pi_{dt}\}$ are also the stick-breaking weights for topic ψ_{dt} that sum to one. The difference between corpus level and document level is how the topics themselves are drawn. The corpus level topics are drawn from H, but the per-document topics are drawn from G_0 as

$$l_{dt} \sim Mult(\beta) \quad \psi_{dt} = \lambda_{l_{dt}} \tag{4}$$

Where l_{dt} are indicator variables which index the corpus-level topic corresponding to ψ_{dt}. Given G_j, the generative process for i[th] words in the j[th] document is described as

$$z_{ji} \sim Mult(\pi_j), \theta_{ji} = \psi_{jz_{ji}} = \lambda_{l_{jz_{ji}}}, \quad r_{ji} \sim Mult(\theta_{ji}) \tag{5}$$

Since less and less of sick remains to partition when k or t becomes large, to make the parameter estimation feasible, we adopt a truncation technique employed by [38] to set the truncations of corpus and document levels to K and T, such that

$$\begin{aligned} \beta_K = 1 \quad \sum_{k=1}^{K} \beta_k = 1 \quad \beta_k = 0 \quad when \quad k > K \\ \pi_{jT} = 1 \quad \sum_{t=1}^{T} \pi_{jt} = 1 \quad \pi_{jt} = 0 \quad when \quad t > T \end{aligned} \tag{6}$$

where the number of T is far less than K, since in practice per-document G_d need far fewer topics than those required for the corpus.

3.3 DAC_mmst model

For the multi-modal documents containing images and related long descriptions, we suppose that the topic terms, image region terms, visual-annotation terms and non-visual-annotation terms, are constructed from two space, including visual topic space shared by images and texts, where topics are generated by visual and strongly related textual content jointly; and non-visual topics space only related to text space, where topics reflects weakly relationship between image and textual information and cannot be expressed by visual information. Image region terms and visual-annotation terms are constructed from visual topic space, and non-visual-annotation terms from non-visual topic space. Now we introduce the model of this paper, inspired by DAC_mmst, to simultaneously classify and annotate image data.

DAC_mmst model can be viewed in terms of a generative process that first generates the visual information, subsequently draws the discrete class label and lastly constructs the annotated information. More specifically, we first generate corpus and document level visual topics for images, and generate N image region terms r_n by introducing a HDP model. In order to achieve the classification task, we extend HDP to supervised HDP by learn an additional response variable to the document level visual topics. For each of the M annotation terms, we use binary variable to supervise whether the topic term is constructed from the visual topic space or the non-visual topic space. We force each visual-annotation term to directly share a hidden visual topic with a randomly selected image region term, which guarantees that the topics for visual-annotation terms are indeed a subset of the visual topics that occur in the corresponding image. Note that while visual-annotation term is restricted to association with one particular image region, this association allows the same image region to be associated with multiple visual-annotation terms. Meanwhile, in order to enable our model to take better advantage of deep annotation terms, we further leverage the non-visual topics to construct non-visual-annotation terms to reveal the latent semantic information for images. In order to reduce the complexity of our model, we adopt LDA model to recover the non-visual topics for simplicity.

More specifically, **Fig. 2 (c)** shows the graphical model representation of our DAC_mmst model. The generative process of DAC_mmst model for an image-text pair document with N image region terms, M annotation terms and their labels corresponding to the graphical model are given as follows:

(1)Generate the corpus level visual topic distribution according to Eq.(2)
(2)Generate the visual words distribution for each visual topic: $\phi_s^v \sim H$
(3)For each document d:
  (3.1)Generate document level visual topic distributionaccording to Eq.(3)
  (3.2)Generate a multinomial distribution over non-visual topics:$\theta_d \sim Dir(\alpha_0^w)$
  (3.3)Generate a binomial distribution over visual topics versus non-visual topics: $b_d \sim Beta(\sigma)$
  (3.4)For each Visual Word in document d:
    (3.4.1)Sample an visual topic assignment： $z_{dn}^v \sim Mult(\pi_d)$
    (3.4.2)Sample $r_{dn} \sim Mult(\phi_{sl_{dz_{dn}^v}}^v)$
  (3.5)Draw response variable: $c_d \sim softmax(z_d^v, z_d^w, \delta)$
  (3.6)For each textual term w_{dm}:
    (3.6.1)Draw a binary switch $x_{dm} \sim Binomial(b_d)$
    (3.6.2) If x_{dm}=0, then
      Sample a visual topic assignment: $y_{dm} \sim Unif(1, N_d)$
      Sample a visual-annotation term: $w_{dm} \sim Mult(\phi_{wz_{dydm}^v}^v)$
    (3.6.3) If x_{dm}=1, then
      Sample non-visual topic: $z_{dm}^w \sim Mult(\theta_d)$
      Sample non-visual-annotation term $w_{dm} \sim Mult(\phi_{wz_{dm}^w}^d)$

We can learn from the generative process that our DAC_mmst model is built on the traditional Corr-LDA and HDP models by introducing non-visual topics and image classification, which can effectively model multimedia document. Each document is jointed with two type topic distributions: π over visual topics owned for two modalities together, θ over non-visual topics exclusive to textual modality. Each kind of topics is probability distribution over textual terms M or visual terms N. Our DAC_mmst can make use of the supervised category label information to classify multiclass document directly. In sLDA, a response variable for each "document" (here, an image) is assumed drawn from a generalized linear model with input given by the empirical distribution of topics that generated the image patches. The response variable of sLDA is real valued and drawn from a linear regression, which simplified inference and estimation. However, a continuous response is not appropriate for our goal of building a classifier. Rather, we consider a class label response variable, drawn from a softmax regression for classification and incorporate it into HDP model. More specifically, conditioned on the distribution of visual topics and non-visual topics, classifiers (step 3.5) take the forms as follows.

$$p(c_d \mid z_d^w, z_d^v) = \frac{\exp(\delta_{c_d}^T \frac{1}{N+M}(\sum_{n=1}^{N} z_{dn}^v + \sum_{m=1}^{M} z_{dm}^w))}{\sum_{c=1}^{E} \exp(\delta_c^T \frac{1}{N+M}(\sum_{n=1}^{N} z_{dn}^v + \sum_{m=1}^{M} z_{dm}^w))} \tag{7}$$

4. Inference and parameter estimation

Input multimedia documents, the key inferential problem that we need to solve in order to use DAC_mmst model is that of computing the posterior distributions of two document-topic π and θ, and a set of distributions $\varphi v\ s$,$\varphi v\ w$,$\varphi d\ w$, and class label c. In this paper, we employ an approximate posterior inference algorithm to reveal the hidden posterior distributions in the model, which presents significant computational challenge in the face of massive data sets. Developing scalable approximate inference methods for topic models is an active area of research [38][39][40].The basic idea of variational method considers to employ Jensen's inequality to find an adjustable lower bound on the log-likelihood. Essentially, one considers a family of lower bounds, indexed by a set of variational parameters.

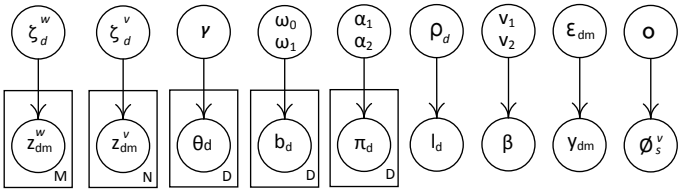

Fig. 3. Graphical model representation of the variational distribution used to approximate posterior.

For DAC_mmst model, the log-likelihood can be converted into a lower bound as follows:

$$\begin{aligned} &\log p(r, w, c \mid \alpha_0^w, \alpha_0^v, \mu, \eta, \sigma, \delta) \\ &\geq \underset{q}{E}[\log p(r, z^v, \pi, w, c, z^w, x, \beta, \theta, \phi_s^v, y, l \mid \alpha_0^w, \alpha_0^v, \mu, \eta, \sigma, \delta)] \\ &- \underset{q}{E}[\log q(z^w, z^v, \theta, b, \pi, l, \beta, y, \phi_s^v)] = \mathcal{L}(q) \end{aligned} \tag{8}$$

Where q denotes the variational distribution and Jensen's inequality is adopted to reach the variational lower bound of q. In order to find a tractable family of lower bounds, by removing some edges and nodes in the original graphical model, a simplified graphical

model with free variational parameters will is obtained, as shown in **Fig.3**. In particular, We define the following fully factorized variational distribution on the latent variables:

$$\begin{aligned} & q(z^{w}, z^{v}, \pi, l, \beta, \phi_{s}^{v}, \theta, y, b) \\ & = \prod_{j=1}^{D} \prod_{n=1}^{N} \prod_{t=1}^{T} q(z_{jn}^{w} \mid \zeta_{jn}^{w}) q(z_{jn}^{v} \mid \zeta_{jn}^{v}) q(l_{jt} \mid \rho_{jt}) q(\pi_{jt} \mid \alpha_{1jt}, \alpha_{2jt}) q(b_{j} \mid \varpi_{0}, \varpi_{1}) \\ & \cdot \prod_{k=1}^{K} q(\beta_{k} \mid v_{1k}, v_{2k}) q(\phi_{sk}^{v} \mid o_{k}) \cdot \prod_{m=1}^{M} q(y_{m} \mid \varepsilon_{m}) \cdot q(\theta_{j} \mid \gamma_{j}) \end{aligned} \tag{9}$$

where the free parameter ζ^{v} is a variational multinomial over the T visual-topics from the documents level topics, ζ^{w} is a H-dimensional multinomial free paremeter that governs the probability of non-visual topic assignment, ρ is a K-dimensional multinomial parameter which supervises the probability to determine corpus level topics, (α_1, α_2) and (v_1, v_2) are beta parameters managing the stick breaking for two levels, (ω_0, ω_1) is a variational parameter for binomial distribution, o is a variational Dirichlet, ε is a M N-dimensional multinomial over the image regions, and γ is a variational Dirichlet. The computation details of Equation (8) is given in Appendix.

Following the general equation for variational approximation, we minimize the KL-divergence between the factorized distribution and the true posterior by updating the parameters. To update the parameters, we adopt a coordinate ascent algorithm where we update one set of parameters while the rest remain unchanged. The coordinate ascent updates by computing the gradient of these parameters and setting them equal to zero. We sum up the parameter inference in **Algorithm 1**.

Algorithm 1: Variational Inference for tm_MMC

Initialize all the variational parameters
While Not converged or within MAX iteration **do**
E Step:
For each documents
Update per documents stick
$\alpha_{1jt} = 1 + \sum_{n=1}^{N} \zeta_{jnt}^{v}$ $\alpha_{2jt} = \sum_{n=1}^{N} \sum_{i=t+1}^{T} \zeta_{jni}^{v} + \alpha_{0}^{v}$
Update per visual word topic indices ζ^{v} using Eq.(10) and (11)
Update per documents topic indices ρ using Equation(12) and δ
Update per textual word topic indices ζ^{w}
$\phi_{wi,w_{jn}}^{d} \exp(\psi(\gamma_{i}) - \psi(\sum_{j=1}^{H} \gamma_{j}))$
Update beta parameters of binomial distribution
$\omega_{0} = n_{jx_{0}} + \sigma_{0}$ $\omega_{1} = n_{jx_{1}} + \sigma_{1}$
Update the approximate posterior distribution over regions
$\varepsilon_{mn} \propto \exp(\sum_{i=1}^{T} \zeta_{\mathrm{ni}}^{v} \log \phi_{wi,w_{m}}^{v})$
Update the posterior Dirichlet parameters:
$\gamma_{i} = \sum_{n=1}^{M} \zeta_{n,i}^{w} + \alpha_{0i}^{w}$
M Step:
Update corpus level stick
$v_{1k} = 1 + \sum_{d=1}^{D} \sum_{t=1}^{T} \rho_{dtk}$ $v_{2k} = \sum_{d=1}^{D} \sum_{t=1}^{T} \sum_{l=k+1}^{K} \rho_{dtl}$
Update topic mixute
$o_{ki} = \sum_{j=1}^{D} \sum_{t=1}^{T} \rho_{jtk} (\sum_{n=1}^{N} \zeta_{jnt}^{v} [r_{jn} = i] + \eta$
Update the label parameter c using Eq.(7)
Finds maximum likelihood estimates of the model parameters

More specifically, We turn to the update for the variational multinomial ζ^{v}. We observer that $\sum_{l=1}^{C}\prod_{n=1}^{N}(\sum_{e=1}^{K}\sum_{i=1}^{T}\rho_{jie}\zeta_{jni}^{v}\ \exp(\frac{1}{N+M}\delta_{le}))$ is only a linear function of ζ^{v} for fixed j and n. We follow the approach of [41] to derive the fixed point update. The main idea is that considering a previous estimation of value $\zeta^{v old}$ to maximize the lower bound of $\ell'_{|\zeta^{v}|}$ so that this lower bound is tight on $\zeta^{v old}$. We know the inequality $\log(y) \le \tau^{-1}y + \log(\tau) - 1$, where equality holds if and only if $x = \tau$. Thus, set $y=x^{T}\zeta_{jn}^{v}$ and $\tau = x^{T}\zeta_{jn}^{v\ old}$. The new bound can be written as follows.

$$\begin{aligned}\ell \ge & \sum_{m=1}^{D}\sum_{n=1}^{N}\sum_{t=1}^{T}\varepsilon_{jmn}\zeta_{jnt}^{v}\log(\phi_{wt,w_{jm}}^{v}) + \sum_{j=1}^{D}\sum_{n=1}^{N}\sum_{t=1}^{T}\zeta_{jnt}^{v}((\sum_{k=1}^{K}\rho_{jtk}\\ & \sum_{i=1}^{Tr}(\psi(o_{ki}) - \psi(\sum_{p}o_{kp}))[r_{jn} = i]) + \zeta_{jnt}^{v}((\psi(\alpha_{1jt}) - \psi(\sum_{i=1}^{2}\alpha_{ijt})) + \\ & \sum_{t=1}^{t-1}(\psi(\alpha_{2jt}) - \psi(\sum_{i=1}^{2}\alpha_{ijt}))) - \zeta_{jnt}^{v}\log\zeta_{jnt}^{v}) + \sum_{j=1}^{M}(\delta_{ej}^{T}(\frac{1}{N+M}\sum_{n=1}^{N}\sum_{t=1}^{T}\zeta_{jnt}^{v}))\\ & -(x^{T}\zeta_{jnt}^{v\ old})^{-1}x^{T}\zeta_{jn}^{v} - \log(x^{T}\zeta_{jn}^{v\ old}) + 1) = \ell'_{|\zeta^{v}|}.\end{aligned} \tag{10}$$

This lower bound of $\ell'_{|\zeta^{v}|}$ is tight when $\zeta^{v}=\zeta^{v old}$. We compute the derivative for the new bound as $\partial\ell'_{|\zeta|}/\partial\zeta_{jnt}^{v}=0$ under the constraint $\sum_{t=1}^{T}\zeta_{jnt}^{v} = 1$, and determine the fixed-point update as follows.

$$\begin{aligned}\zeta_{jnt}^{v} \propto & \sum_{m=1}^{D}\sum_{n=1}^{N}\sum_{t=1}^{T}\varepsilon_{jmn}\log(\phi_{wt,w_{jm}}^{v}) +\\ & \exp(\sum_{k=1}^{K}\rho_{jtk}\ \underset{q}{E}[\log p(r_{jn} \mid \phi_{sk}^{v})] + \underset{q}{E}[\log\pi_{jt}] + \frac{1}{N+M}\delta_{yj}^{T} - (x^{T}\zeta_{jn}^{v\ old})^{-1}x_{t})\end{aligned} \tag{11}$$

In order to re-estimate the variational parameter ρ, we set $\partial\ell/\partial\rho=0$, however, this will not lead to a closed-form solution and make the gradient vanish. This is a non-linear optimization problem. There are many algorithms and libraries for solving this. For example, the conjugate gradient method, which only requires the computation of partial derivatives, can be used. Thus, in order to update ρ, the conjugate gradient will be adopted to solve this optimization problem. We set $\xi = \sum_{l=1}^{C}\prod_{n=1}^{N}(\sum_{k=1}^{K}\sum_{t=1}^{T}\rho_{jtk}\zeta_{jnt}^{v}\ \exp(\frac{1}{N+M}\delta_{lk}))$, and conjugate gradient only requires the derivatives as follows.

$$\begin{aligned}\frac{\partial\ell}{\partial\rho_{jtk}} \approx & \sum_{n=1}^{N}\zeta_{jnt}^{v}\ \underset{q}{E}[\log p(r_{jn}\mid\phi_{sk}^{v})] + \underset{q}{E}[\log\beta_{k}] - 1 - \log\rho_{jtk}\\ & +\delta_{yjk}\exp(\frac{1}{N}\sum_{n=1}^{N}\zeta_{jnt}^{v}) - \xi^{-1}\left\{\sum_{l=1}^{C}(\prod_{m=1}^{N}(\sum_{e=1}^{K}\sum_{i=1}^{T}\rho_{jie}\zeta_{jmi}^{v}\exp(\frac{1}{N+M}\delta_{le})))\right.\\ & \left.\cdot\sum_{n=1}^{N}\left(\frac{\zeta_{jni}^{v}\cdot\exp(\frac{1}{N+M}\delta_{lk})}{\sum_{e=1}^{K}\sum_{i=1}^{T}\rho_{jie}\zeta_{jni}^{v}\cdot\exp(\frac{1}{N+M}\delta_{le})}\right)\right\}\end{aligned} \tag{12}$$

The derivation of δ is similar to Equation (12), and the detailed derivation is shown in Appendix A.2.

5. Experiments

To evaluate the performance of the newly proposed model, in this section, we present our experiments on two predictive tasks: image classification and annotation tasks. We first discuss details of the dataset used and the experimental settings, and then show experimental

results using the different models.

5.1 Dataset and Preprocessing

The evaluation of a cross-modal retrieval system requires a document corpus with images, loose text descriptions and category label. Most existing multi-modal datasets are limited to annotations that describe visible object names only, in this paper, we gathered the dataset with paired images and a brief and loose descriptions from Flickr. The dataset consists of eight different social events taken place in the last several years as demonstrated in **Table 1**. We manually adopted the site's API and keywords to crawl and parse related textual information and its corresponding images in the timeline of each social event category. The eight different social events span across economy, war, politics, and science.

Table 1. Illustration of dataset.

Car.Id	Category Identification	Start Time	End Time	# Doc.
1	Occupy wall street(ows)	2011.09	2012.08	7206
2	South China Sea arbitration(scsd)	2013.01	2016.07	8040
3	Syrian civil war(scw)	2011.04	2013.06	7942
4	War in Afghanistan(wia)	2001.10	2009.04	6503
5	Mars Reconnaissance Orbiter(mro)	2005.04	2012.06	6823
6	Apple Jobs(ajo)	2008.09	2013.10	4987
7	Financial crisis(fcr)	2007.04	2012.08	4297
8	Greek Recession(gre)	2011.05	2012.04	5738

For each image, we rescaled the image for the maximum height of 256 pixels, and then obtained 20×20 image patches by grid sampling with a 20-pixel interval, resulting in 144 patches per image. We adopted SIFT descriptors of each gray-scale patch to extract 128-dim features for all images. We then ran k-means clustering algorithm on 128-dim descriptors to create a visual codebook of 260 discrete code words. We tried other codebook sizes and the results were similar. The words of related text information were reprocessed by (1) Deleting stop-words and non-English alphabets; (2) Deleting low frequencies of annotation terms that appeared less than three times, and (3) Adopting the snowball algorithm to stem from for words.

5.2 Qualitative Evaluation

In this section, we will qualitatively demonstrate the effectiveness of our proposed model for deep image annotation and classification. For simplicity, we first visualize the learned visual and non-visual topics in the **Fig. 4**, which can validate the effectiveness of our proposed DAC_mmst model. By providing a multi-modal information of the representative textual and visual terms, it is very intuitive to interpret the associated topic. For simplicity, we only visualize the discovered eight topics most including four visual topics and four non-visual topics related respectively to two categories. We provides five top-ranked visual-annotation terms and five most related images for each visual topic, and five top-ranked non-visual terms for each non-visual topic. The non-visual words are ranked by their probability of being drawn from the corresponding non-visual topic $p(w|zw\ h)$, the visual-annotated words from $p(r|zv\ k)$, while the images are sorted by counting the number of visual words and textual visual-annotation words with the corresponding topic for a document as follows.

$$p(z_k^v \mid w_d, r_d) = \frac{n_{d,k}^v + n_{d,k}^w}{\sum_{k=1}^{K} (n_{d,k}^v + n_{d,k}^w)} \tag{13}$$

where n^v and n^w indicate the numbers of visual topic k assigned to the visual-annotation terms and the visual terms of document d, respectively.

As shown in **Fig. 4**, the results are impressive and satisfy our expectation, where the visual topics represented by textual words and well relevant visual patches are meaningful and show high consistency between semantic concepts and visual content. For example, visual topic #55 illustrates the five top-ranked visual-annotated words, such as fighters, opposition, rebels, soldiers, and SDF, and five most related images according to *p(w|zv k)* and Eq.(13), respectively. Otherwise, in each non-visual topic, the discovered topic words represented by textual information are difficult to be described by visual content. For example, the non-visual words of non-visual topic #60, such as casualties, violence and bloodshed, are difficult to be expressed by image, which are extracted according to the rank of their probability values calculated by *p(w|zw h)*. For visual topic #43, visual topic #55, non-visual topic #49 and non-visual topic #60, the most related category is Syrian Civil War. Based on the results, we can confirm that our proposed model can effectively mine the topics of multi-modal dataset.

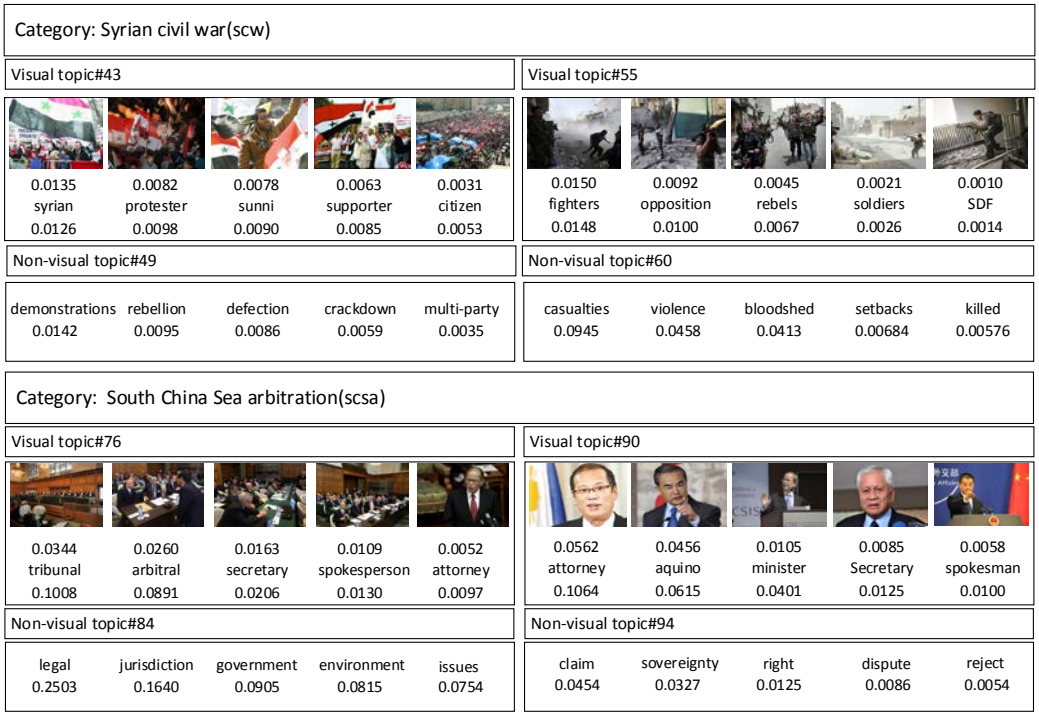

Fig. 4. The discovered sample topics.

5.3 Quantitative Evaluation

A. Perplexity

The objective of document modeling is a density estimation that describes the underlying structure of data. The most common way to achieve this is by estimating the model's generalization performance on previously unobserved documents. In this subsection, we employ perplexity, which is defined as the geometric mean of the inverse marginal probability of each word in the held-out test data, to measure the quality of model generalization. The

higher the likelihood is, the lower the perplexity will be, and a smaller value of perplexity represents a better effect of clustering and generalization capability. Formally, given the test set, we computed the perplexity of the given long text under p (w | r) for each image as follows.

$$Perp = exp(-\frac{\sum_{d=1}^{D}\sum_{m=1}^{M_d}\log p(w_{d,m}|r_{d,1:N})}{\sum_d M_d}) \\ = exp(-\frac{\sum_{d=1}^{D}\sum_{m=1}^{M_d}\{\log p(w_{d,m}|r_{d,1:N}, x_{d,m}=0) + \log p(w_{d,m}|r_{d,1:N}, x_{d,m}=1)\}}{\sum_d M_d}) \quad (14)$$

where $p(w_{d,m}|r_{d,1:N}, x_{d,m})$ is the conditional probability of textual words, given an image $r_{d,1:N}$, x is a binary variable denoting the type of topic generating this word, and M_d is the number of textual words in document d. The higher conditional likelihood leads to the lower perplexity.

To demonstrate the perplexity performance, we compare our proposed model with two existing baseline multi-modal topic models, corrLDA and mmLDA. But both model based on LDA model assume a one-to-one correspondence between the topics of each modality, they need to specify the topic number beforehand. As our DAC_mmst model introduces non-parametric HDP to estimate the best number of visual topics that can perfectly explain the image, we only regard the number of non-visual topics as an adjustable parameter. **Fig. 5** shows the perplexity under the maximum likelihood estimates of each model for different number of topics. We can find that mm_LDA does not present good conditional distributions, mainly because this model permits annotation terms to be constructed by topic that did not contribute to constructing the image regions. Although mm_LDA has the capability of modeling the joint distribution of texts and image regions, it cannot model the relationship between them. Furthermore, corrLDA presents as flexible a joint distribution by allowing annotation terms to be allocated to different visual topics, so it can achieve superior generalization performance to the mm_LDA.

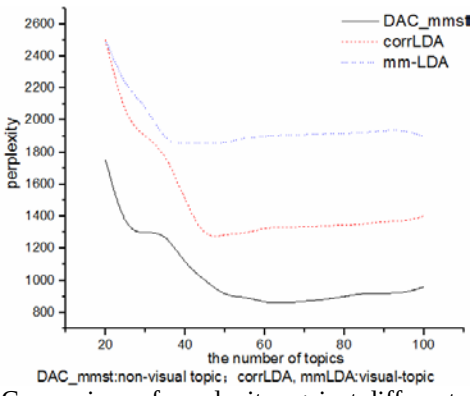

Fig. 5. Comparison of perplexity against different models.

Most notably, we found DAC_mmst model can provide better generation quality than corr-LDA and mm_LDA models for two reasons. The main reason is that DAC_mmst is more suited to loosely relationship than one-to-one mapping used by corrLDA model. By separating the visual and non-visual topics, DAC_mmst not only generates strong relativity between two modalities, but also can model textual information without the corresponding visual information. Second, DAC_mmst determines the proper number of visual-topics by non-parametric Bayesian, which avoids under and over fitting for visual information. Thus,

with DAC_mmst, we can achieve a competitive fit of the joint distribution and find superior conditional distributions given images.

B. Image Annotation and Region Labeling

Table 2. Image annotation and region labeling results in terms of precision, recall and F-measure.

Image annotation												
mm-LDA			corrLDA			tr-mmLDA			DAC_mmst			
P	R	F	P	R	F	P	R	F	P	R	F	
15	23	18	20	37	26	32	71	44	40	74	51	

Region labeling												
mm-LDA			corrLDA			tr-mmLDA			DAC_mmst			
P	R	F	P	R	F	P	R	F	P	R	F	
29	34	31	31	59	41	46	51	48	60	53	56	

The performance measure of annotation is compared with mm-LDA, corr-LDA and tr-mmLDA. Annotation task usually involves two objectives: image annotation and image region labeling. Given a segmented image without its annotation, we can use compute two distributions over visual-annotation terms and non-visual annotation terms conditioned on the image according to Equation (15).

$$\begin{aligned} p(w \mid r, x_w=0) &\approx \sum_{n=1}^{N} \sum_{z_n} q(z_{jn}^v \mid \zeta_{jn}^v) p(w \mid z_{jn}^v, \phi_w^v, x_w=0) \\ p(w \mid r, x_w=1) &\approx \sum_{n=1}^{N} \sum_{z_n} q(z_{jn}^w \mid \zeta_{jn}^w) p(w \mid z_{jn}^w, \phi_w^d, x_w=1) \end{aligned} \tag{15}$$

In addition, we can use our model to make prediction for the region annotation. The label assigned to an image region is then the one which gives the highest probability for following region-based conditional distribution over visual annotation terms

$$p(w \mid r, r_n) \approx \sum_{z_n} q(z_{jn}^v \mid \zeta_{jn}^v) p(w \mid z_{jn}^v, \phi_w^v) \tag{16}$$

It should be noted that our model is not directly comparable to that of the state-of-the-art methods, as we are the first to achieve annotation by introducing non-visual terms. In order to assess the annotation performance, compared with the other models, we only consider the visual-annotation performance using precision, recall and F-measure as the performance measure. An annotated word is considered to be correct if and only if it appears in the ground truth annotation of the target image. Let A be the number of images automatically annotated with a given word, B the number of images correctly annotated with that word, and C the number of images having that word in ground-truth annotation. Then precision(P), recall(R) and F-measure(F) are computed as R = B/C ,P = B/A and F=2×P×R/(P + R). When there are repeated words in the ground-truth annotations, the repeated terms were removed to calculate the F. Otherwise, We used 5 random train/test splits to estimate the average P, R and F.The experimental comparisons are listed in **Table 2**. From this table, we can find that our method can consistently beat the other methods evaluated by the F-measure metric for more than 7 percent. The causes for more F-measure than the state-of-art may be the reduced-dimension visual-annotation terms, which equals the textual information minus non-annotation terms, used in our model. In other words, non-visual topics remove some non-visual words, which reduce the noise information for visual-annotation and purify the visual terms. Also, the best fit number of visual topics determined by non-parametric HDP make a contribution to improve the annotation performance

C. Classification Accuracy

The goal in this experiment is to classify an unknown image as one of the ten learned scene classes. To measure the performance of different learning approaches for classfication, we implement four approaches (sLDA, mmLDA-SVM, mmLDA-SG, and our proposed DAC_mmst) that are popularly employed in social media analysis to train the models. sLDA is

the supervised model from [36], which only consider textual feature. For mmLDA-SVM, based on unsupervised mmLDA model and SIFT to denote image features, we represent each multi-modal document as a vector adopting text and image content. Then, classifier is learned adopting SVM, which is used to predict and attach two type of labels to text data. mmLDA-SG is similar to the training process of mmLDA-SVM. But we adopt softmax regression method to train classifier instead of SVM.

We report 8-category discriminations of 87% by adopting DAC_mmst model. The results are illustrated in the confusion matrices of **Fig. 6**. Note that we observe that SLDA is better than mmLDA-SVM and mmLDA-SG, which shows the textual information is much more helpful than the visual information for classification of multi-modal information with richly and loosely descriptions. We can also observe that our DAC_mmst can reduce the error of SLDA by at least 10%. It may be because SLDA only uses the supervised information. Different from these methods, DAC_mmst can jointly exploit the multi-modal property and the multiclass property, which can reasonably model the correlations between two modalities more accurately. Through separating the visual topics and non-visual-topics, our model makes better document representations in the latent semantic space and boosts the classification accuracy. Otherwise, the combination with non-parametric Bayesian HDP model effectively prevent over or under fitting due to the number of topics, which contributes to improving the classification accuracy.

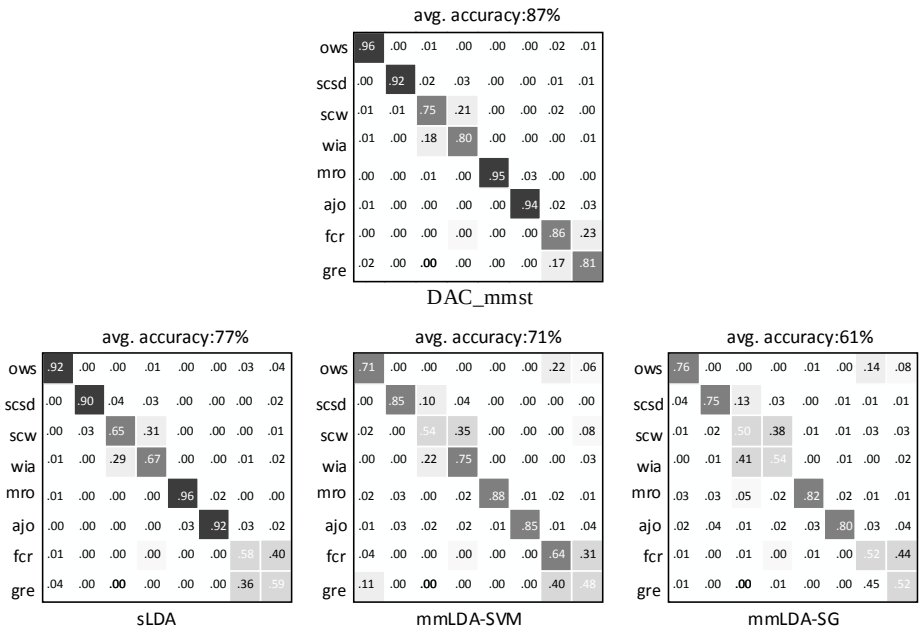

Fig. 6. Image classification accuracy analysis.

Conclusions and Future Work

In this paper, we have developed DAC_mmst, a supervised topic model for the task of image annotation and classification in a unified framework. While existing methods require strong correspondence between the modalities, our DAC_mmst can flexibly bridge and reveal semantic information from information across different modalities by separating the visual topics and non-visual topics, where the non-visual topics reduce the information loss of the traditional annotation and achieve deeper annotation for images. At the same time, our DAC_mmst can effectively integrate the revealed topics to implement classification. Last, it is

capable to automatically determine the number of latent topics for visual modality. Extensive experiments demonstrate the above advantages in terms of perplexity, image annotation, and classification.

In the future, we intend to investigate more tasks based on the expansion of our proposed model, such as multi-modal sentiment analysis. Otherwise, one of the limitations of our model is that the computational cost is high in terms of both time and space complexity. In the future, we will develop more efficient model to optimize this problem.

Reference

[1] J. Deng, W. Dong and et al., "ImageNet: A large-scale hierarchical image database," *Computer Vision and Pattern Recognition*, pp. 248-255, Jun., 2009. Article (CrossRef Link)

[2] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the Gap: Query by Semantic Example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923-938, July, 2007. Article (CrossRef Link)

[3] Q. Liu and Z. Li, "Projective nonnegative matrix factorization for social image retrieval," *Neurocomputing*, vol. 172, pp. 19-26, Jan., 2016. Article (CrossRef Link)

[4] B. Wu, S. L. yu, B.G. Hu and Q. Ji, "Multi-label learning with missing labels for image annotation and facial action unit recognition," *Pattern Recognition*, vol. 48, no. 7, pp. 2279-2289, July, 2015. Article (CrossRef Link)

[5] X. Y. Jing and et al, "Multi-label Dictionary Learning for Image Annotation," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp.2712-2725, June, 2016. Article (CrossRef Link)

[6] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. of International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 127-134, July, 2003. Article (CrossRef Link)

[7] D. Putthividhy, H. T. Attias and S. S. Nagarajan, "Topic regression multi-modal Latent Dirichlet Allocation for image annotation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 238, no. 6, pp. 3408-3415, June, 2010. Article (CrossRef Link)

[8] Y. Q. Jia and M. Salzmann and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. of IEEE International Conference on Computer Vision*, vol. 32, no. 14, pp. 2407-2414, Nov., 2011. Article (CrossRef Link)

[9] Y. Wang and G. Mori, "Human Action Recognition by Semilatent Topic Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762-1774, Oct. , 2009. Article (CrossRef Link)

[10] S. Chonglin and et al., "Efficient Methods for Multi-label Classification," in *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 164-175, Apr., 2015. Article (CrossRef Link)

[11] S. H. Amiri and M. Jamzad, "Automatic image annotation using semi-supervised generative modeling," *Pattern Recognition*, vol. 48, no. 1, pp. 174-188, Jan, 2015. Article (CrossRef Link)

[12] Y. Lin and et al, "Large-scale image classification: Fast feature extraction and SVM training," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1689-1696, June, 2011. Article (CrossRef Link)

[13] B. F. Guo and et al., "Customizing Kernel Functions for SVM-Based Hyperspectral Image Classification," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 622-629, Apr., 2008. Article (CrossRef Link)

[14] A. Bosch, A. Zisserman and X. Munoz, "Image Classification using Random Forests and Ferns," in *Proc. of International Conference on Computer Vision IEEE*, pp. 1-8, Oct., 2007. Article (CrossRef Link)

[15] B. Xu, Y. Ye and L. Nie, "An improved random forest classifier for image classification," in *Proc. of IEEE International Conference on Information and Automation*, pp. 795-800, Jun., 2012. Article (CrossRef Link)

[16] T. N. Hong, C. Barat and C. Ducottet, "Approximate image matching using strings of bag-of-visual words representation," in *Proc. of International Conference on Computer Vision Theory and Applications*, vol. 2, pp. 345-353, Jan., 2014. Article (CrossRef Link)
[17] F. F. Li and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 524-531, June, 2005. Article (CrossRef Link)
[18] L. J. Li and F. F. Li, "What, where and who? Classifying events by scene and object recognition," in *Proc. of IEEE International Conference on Computer Vision*, pp. 1-8, Oct., 2007. Article (CrossRef Link)
[19] X. Liu and et al., "Boosting image classification with LDA-based feature combination for digital photograph management," *Pattern Recognition*, vol. 38, no.6, pp. 887-901, Jun., 2005. Article (CrossRef Link)
[20] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Proc. of European Conference on Computer Vision*, vol. 3954, pp. 517-530, May, 2006. Article (CrossRef Link)
[21] L. J. Li, R. Socher and F. F. Li, "Towards total scene understanding: classification, annotation and sementation in an automatic framework," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Jun., 2009. Article (CrossRef Link)
[22] K. Xu and et al., "Unsupervised Satellite Image Classification Using Markov Field Topic Model," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 1, pp. 130-134, Jan., 2013. Article (CrossRef Link)
[23] N. Rasiwasia and N. Vasconcelos, "Latent Dirichlet allocation models for image classification," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 11, pp. 2665-2679, Nov., 2013. Article (CrossRef Link)
[24] Y. Wang and G. Mori, "Max-margin latent dirichlet allocation for image classification and annotation," *Lecture Notes in Computer Science*, vol. 1674, no. 1, pp. 39-48, Sep., 2011. Article (CrossRef Link)
[25] C. Wang, D. Blei and F. F. Li, "Simultaneous image classification and annotation," in *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, vol. 19, no. 2, pp. 1903-1910, Jun., 2009. Article (CrossRef Link)
[26] W. Hua, H. Heng and C. Ding, "Image annotation using bi-relational graph of images and semantic labels," in *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, vol. 42, no. 7, pp. 793-800, Jun., 2011. Article (CrossRef Link)
[27] X. Cai and et al. "Joint stage recognition and anatomical annotation of drosophila gene expression patterns," *Bioinformatics*, vol. 28, no. 12, pp. 16-24, Jun., 2012. Article (CrossRef Link)
[28] J. Paisley and et al., "Nested Hierarchical Dirichlet Processes," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 2, pp. 256-270, Feb., 2015. Article (CrossRef Link)
[29] D. M Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, May, 2003. Article (CrossRef Link)
[30] J. Kandola, T. Graepel and J. Shawetaylor, "Reducing Kernel Matrix Diagonal Dominance Using Semi-definite Programming," *Lecture Notes in Computer Science*, vol. 2777, pp. 288-302, 2003. Article (CrossRef Link)
[31] Z. Wei, X. Luo and F. Zhou, "Ontology Based Automatic Image Annotation Using Multi-class SVM," in *Proc. of International Conference on Image and Graphics*, pp. 434-438, Jul., 2013. Article (CrossRef Link)
[32] G. Carneiro and N. Vasconcelos, "Formulating Semantic Image Annotation as a Supervised Learning Problem," in *Proc. of IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, vol. 2, pp. 163-168, 2005. Article (CrossRef Link)
[33] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 25, no. 9, pp. 1075-1088, Sep., 2003. Article (CrossRef Link)
[34] D. R. Hardoon and et al., "A Correlation Approach for Automatic Image Annotation," in *Proc. of Conference on Advanced Data Mining and Applications*, pp. 681-692, Aug., 2006. Article (CrossRef Link)

[35] X. Li, Q. Lv and W. Huang, "Learning Similarity with Probabilistic Latent Semantic Analysis for Image Retrieval," *Ksii Transactions on Internet & Information Systems*, vol. 9, no. 4, pp. 424-440, Apr., 2015. Article (CrossRef Link)
[36] J. Zhu, "MedLDA: Max-Margin Supervised Topic Models," *Journal of Machine Learning Research*, vol. 13, no. 4, pp. 2237−2278, 2009. Article (CrossRef Link)
[37] W. Fan, N. Bouguila, "Online Data Clustering Using Variational Learning of a Hierarchical Dirichlet Process Mixture of Dirichlet Distributions," in *Proc. of International Conference on Database Systems for Advanced Applications*, pp. 18-32, July, 2014. Article (CrossRef Link)
[38] X. Liu, J. Zeng and et al., "Scalable Parallel EM Algorithms for Latent Dirichlet Allocation in Multi-Core Systems," in *Proc. of International Conference on World Wide Web*, pp. 669-679, May, 2015. Article (CrossRef Link)
[39] D. M. Blei, A. Kucukelbir and J. D. Mcauliffe "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859-877, Feb., 2017. Article (CrossRef Link)
[40] J. Taghia, Z. Ma and A. Leijon, "Bayesian Estimation of the von-Mises Fisher Mixture Model with Variational Inference," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 9, pp. 1701-1715, Sep., 2014. Article (CrossRef Link)
[41] J. Huang, "Maximum Likelihood Estimation of Dirichlet Distribution Parameters," *Distribution Cmu Technique Report*, vol. 44, no. 5, pp. 1049-1050, 2005. Article (CrossRef Link)
[42] N. Rasiwasia and et al., "A new approach to cross-modal multimedia retrieval," *ACM International Conference on Multimedia*, pp. 251-260, Oct., 2010. Article (CrossRef Link)
[43] O. Yakhnenko, V. Honavar, "Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies," in *Proc. of the British Machine Vision Conference*, pp. 1-12, Sep., 2011. Article (CrossRef Link)

Yongheng cheng was born in Heilongjiang of China in Dec 1980 and received the Ph.D. degree at the Department of Computer Science and technology, Jilin University. His current main research interests include Data Mining, Web Intelligence and Ontology Engineering and Information integration. He is a member of System Software Committee of China's Computer Federation. More than 20 papers of him were published in key Chinese journals or international conferences, 10 of which are cited by SCI/EI.

Yaojin Lin received the Ph.D. at Hefei University of Technology, and a Professor in the Department of Computer and Engineering, Minnan Normal University. His research interests include data mining, granular computing.

Wan-Li Zuo was born in Jilin of China in Dec 1957. He is a professor and doctoral supervisor at Department of Computer Science and technology, Jilin University. Main research area covers Database Theory, Machine Learning, Data Mining and Web Mining, Web Search Engines, Web Intelligence.

A. Appendix

A.1. Computation of every expectation from the lower bound:

$$\log p(r, w, c \mid \alpha_0^w, \alpha_0^v, \mu, \eta, \sigma, \delta)$$

$$\geq \underset{q}{E}[\log p(r, z^v, \pi, w, c, z^w, x, \beta, \theta, \phi_s^v, y, l \mid \alpha_0^w, \alpha_0^v, \mu, \eta, \sigma, \delta)] - \underset{q}{E}[\log q(z^w, z^v, \theta, b, \pi, l, \beta, y, \phi_s^v)]$$

$$= \ell(q) = \sum_j^D (\underset{q}{E}[\log p(r_j \mid z_j^v, \phi_s^v, l_j)] + \underset{q}{E}[\log p(z_j^v \mid \pi_j)] + \underset{q}{E}[\log p(\pi_j \mid \alpha_0^v)] + \underset{q}{E}[\log p(\phi_s^v \mid \eta)]$$

$$+ \underset{q}{E}[\log p(l_j \mid \beta)] + \underset{q}{E}[\log p(\beta \mid \mu)] + \underset{q}{E}[\log p(c_j \mid z_j^v, l_j, \delta)] + \underset{q}{E}[\log p(w_j \mid z_j^w, y_j, \phi_w^v, \phi_w^d)]$$

$$+ \underset{q}{E}[\log p(x \mid b_j) p(b_j \mid \sigma)] + \underset{q}{E}[\log p(z_j^w \mid \theta_j)] + \underset{q}{E}[\log p(\theta_j \mid \alpha_0^w)])$$

$$- \sum_{j=1}^D (\underset{q}{E}[\log q(l_j)] + \underset{q}{E}[\log q(z_j^w) + \log q(z_j^v)] + \underset{q}{E}[\log q(\pi_j)] + \underset{q}{E}[\log q(b) + \underset{q}{E}[\log q(y_j)])$$

$$- \underset{q}{E}[\log q(\beta)] - \underset{q}{E}[\log q(\phi_s^v)] - \underset{q}{E}[\log q(\theta)]$$

1st term:

$$\underset{q}{E}[\log p(r_j \mid z_j^v, \phi_s^v, l_j) = \sum_{n=1}^N \underset{q}{E}[\log p(r_{jn} \mid z_{jn}^v, \phi_s^v, l_j)$$

$$= \sum_{n=1}^N \sum_{t=1}^T \sum_{k=1}^K \underset{q}{E}[[z_{jn}^v = t][l_{jt} = k] \log p(r_{jn} \mid \phi_{sk}^v)]$$

$$= \sum_{n=1}^N \sum_{i=1}^{T_r} \sum_{t=1}^T \sum_{k=1}^K \zeta_{jnt}^v \rho_{jtk} (\psi(o_{ki}) - \psi(\sum_{p=1}^V o_{kp})) [r_{jn} = i]$$

2nd term:

$$\underset{q}{E}[\log p(z_j^v \mid \pi_j)] = \sum_{n=1}^N \underset{q}{E}[\log p(z_{jn}^v \mid \pi_j)]$$

$$= \sum_{n=1}^N \sum_{t=1}^T [z_{jn}^v = t](\underset{q}{E}[\log(\pi_{jt})] + \sum_{l=1}^{k-1} \underset{q}{E}[\log(1 - \pi_{jt})])$$

$$= \sum_{n=1}^N \sum_{t=1}^T (\sum_{i=t+1}^T \zeta_{jni}(\psi(\alpha_{2jt}) - \psi(\alpha_{1jt} + \alpha_{2jt})) + \zeta_{jnt}(\psi(\alpha_{1jt}) - \psi(\alpha_{1jt} + \alpha_{2jt})))$$

3rd term:

$$\underset{q}{E}[\log p(\pi_j \mid \alpha_0^v)] = \sum_{t=1}^T \underset{q}{E}[\log p(\pi_{jt} \mid \alpha_0^v)]$$

$$= \sum_{t=1}^T \underset{q}{E}[\log \frac{\Gamma(\alpha_0^v + 1)}{\Gamma(\alpha_0^v)\Gamma(1)} (1 - \pi_{jt})^{\alpha_0^v - 1}]$$

$$= \sum_{t=1}^T [\log \alpha_0^v + (\alpha_0^v - 1)(\psi(\alpha_{2jt}) - \psi(\alpha_{1jt} + \alpha_{2jt}))]$$

4th term:

$$\underset{q}{E}[\log p(\phi_s^v \mid \eta)$$

$$= \sum_{k=1}^K \underset{q}{E}[\log \frac{\Gamma(T_r \eta)}{\prod_{i=1}^R \Gamma(\eta)} \prod_{w=1}^{T_r} \phi_{s_{kw}}^{v\eta - 1}]$$

$$= \sum_{k=1}^K (\log \Gamma(T_r \eta) - \sum_{i=1}^{T_r} \Gamma(\eta) + \sum_{w=1}^{T_r} (\eta - 1)(\psi(o_{kw}) - \psi(\sum_{p=1}^{T_r} o_{kp})))$$

5th term in analog to the 2nd term

$$\underset{q}{E}[\log p(l_j \mid \beta)]$$

$$= \sum_{t=1}^T \sum_{k=1}^K (\sum_{i=k+1}^K \rho_{jti}(\psi(v_{2k}) - \psi(v_{1k} + v_{2k})) + \rho_{jtk}(\psi(v_{1k}) - \psi(v_{1k} + v_{2k})))$$

6th term:

$$E_q[\log p(\beta \mid r)]$$

$$= \sum_{k=1}^{K}[\log r + (r-1)(\psi(v_{2k}) - \psi(v_{1k} + v_{2k}))]$$

7th term:

$$E_q[\log p(c_j \mid z_j^v, z_j^w)]$$

$$\geq \delta_{ej}^T(\frac{1}{M+N}\sum_{m=1}^{M}\sum_{t=1}^{H}\zeta_{jmt}^w) - (\sum_{c=1}^{L}\prod_{m=1}^{M}E_q[\sum_{t=1}^{H}\zeta_{jmt}^w(\exp(\frac{1}{N+M}\delta_{ct})]) +$$

$$\delta_{ej}^T(\frac{1}{M+N}\sum_{n=1}^{N}\sum_{t=1}^{T}\rho_{jt}\zeta_{jnt}^v) - (\sum_{c=1}^{L}\prod_{n=1}^{N}E_q[\sum_{t=1}^{T}\rho_{jie}\zeta_{jnt}^v(\exp(\frac{1}{N+M}\delta_{ct})])$$

8th term:

$$E_q[\log p(w_j \mid z_j^w, z_j^v, y_j, \phi_w^v, \phi_w^d)]$$

$$= \sum_{m=1}^{M}\sum_{n=1}^{N}[\sum_{t=1}^{T}(\varepsilon_{jmn}\zeta_{jnt}^v E_q[\log(w_{jm} \mid \phi_{wt}^v)] + \sum_{i=1}^{H}\zeta_{jmi}^w E_q[\log p(w_{jm} \mid \phi_{wi}^d)])]$$

$$= \sum_{m=1}^{M}\sum_{n=1}^{N}[\sum_{t=1}^{T}\varepsilon_{jmn}\zeta_{jnt}^v \log(\phi_{wt,w_{jm}}^V) + \sum_{i=1}^{H}\zeta_{jmi}^w \log(\phi_{wi,w_{jm}}^d)]$$

9th term:

$$E_q[\log p(x_j \mid b_j)p(b_j \mid \sigma)]$$

$$= \log\frac{\Gamma(\sigma_0 + \sigma_1)}{\Gamma(\sigma_0)\Gamma(\sigma_1)} + (n_{jx_0} + \sigma_0 - 1)\psi(\omega_0) + (n_{jx_1} + \sigma_1 - 1)\psi(\omega_1)$$

$$-(n_{jx_0} + \sigma_0 + n_{jx_1} + \sigma_1 - 2)\psi(\omega_0 + \omega_1)$$

10th term:

$$E_q[\log p(z_j^w \mid \theta_j)]$$

$$= \sum_{n=1}^{M}\sum_{i=1}^{H}\zeta_{jni}^w(\psi(\gamma_i) - \psi(\sum_{j=1}^{H}\gamma_j))$$

11th term:

$$E_q[\log p(\theta_j \mid \alpha_0^w)]$$

$$= (\sum_{i=1}^{H}(\alpha_{0i}^w - 1)(\psi(\gamma_i) - \psi(\sum_{i=1}^{H}\gamma_i))) + \log\Gamma(\sum_{i=1}^{H}\alpha_{0i}^w) - \sum_{i=1}^{H}\log(\alpha_{0i}^w)$$

12th term:

$$-E_q[\log q(l_j) = -E_q[\log\prod_{t=1}^{T}\rho_{jt}^{[l_j=t]}] = -\sum_{t=1}^{T}\rho_{jt}\log\rho_{jt}$$

13th term:

$$-E_q[\log q(z_j^w) + \log q(z_j^v)] = -\sum_{n=1}^{M}\sum_{i=1}^{H}\zeta_{jni}^w \log\zeta_{jni}^w - \sum_{n=1}^{N}\zeta_{jn}^v \log\zeta_{jn}^v$$

14th term:

$$-E_q[\log q(\pi_j)] = -\sum_{t=1}^{T}E_q[\log\frac{\Gamma(\alpha_{1jt} + \alpha_{2jt})}{\Gamma(\alpha_{1jt})\Gamma(\alpha_{2jt})}\pi_{jt}^{\alpha_{1jt}-1}(1-\pi_{jt})^{\alpha_{2jt}-1}]$$

$$= \sum_{t=1}^{T}(\log\frac{\Gamma(\alpha_{1jt} + \alpha_{2jt})}{\Gamma(\alpha_{1jt})\Gamma(\alpha_{2jt})} - (\alpha_{1jt} - 1)\psi(\alpha_{1jt}) - (\alpha_{2jt} - 1)\psi(\alpha_{2jt})$$

$$+(\alpha_{1jt} + \alpha_{2jt} - 2)\psi(\alpha_{1jt} + \alpha_{2jt}))$$

15th term:

$$-\underset{q}{E}[\log q(b_j)]$$

$$= \log \frac{\Gamma(\omega_0 + \omega_1)}{\Gamma(\omega_0)\Gamma(\omega_1)} - (\omega_0 - 1)\psi(\omega_0) - (\omega_1 - 1)\psi(\omega_1) + (\omega_0 + \omega_1 - 2)\psi(\omega_0 + \omega_1))$$

16th term:

$$-\underset{q}{E}[\log q(y_j)] = -\sum_{m=1}^{M}\sum_{n=1}^{N}\varepsilon_{jmn}\log\prod_{n=1}^{N}\varepsilon_{jmn}$$

17th term:

$$-\underset{q}{E}[\log q(\beta)]$$

$$= \sum_{k=1}^{K}(\log\frac{\Gamma(v_{1k} + v_{2k})}{\Gamma(v_{1k})\Gamma(v_{2k})} - (v_{1k} - 1)\psi(v_{1k}) - (v_{2k} - 1)\psi(v_{2k}) + (v_{1k} + v_{2k} - 2)\psi(v_{1k} + v_{2k}))$$

18th term:

$$-\underset{q}{E}[\log q(\phi_s^v)]$$

$$= \sum_{k=1}^{K}(-\log\Gamma(\sum_{p=1}^{V}o_{kp}) + \sum_{i=1}^{V}\log\Gamma(o_{kpi}) - \sum_{i=1}^{V}(o_{ki} - 1)(\psi(o_{ki}) - \psi(\sum_{p=1}^{V}o_{kp})))$$

19th term:

$$\underset{q}{E}[\log q(\theta)]$$

$$= (\sum_{i=1}^{H}(\gamma_i - 1)(\psi(\gamma_i) - \psi(\sum_{i=1}^{H}\gamma_i))) + \log\Gamma(\sum_{i=1}^{H}\gamma_i) - \sum_{i=1}^{H}\log(\gamma_i)$$

A.2. We also use the conjugate gradient method to derivate this parameter as δ in analog to ρ as follows.

$$\xi = \sum_{l=1}^{C}\prod_{n=1}^{N}(\sum_{e=1}^{K}\sum_{i=1}^{T}\rho_{jie}\zeta_{jni}^{v}\exp(\tfrac{1}{N+M}\delta_{le}))$$

$$\frac{\partial\ell}{\partial\delta_{yk}} = \sum_{j=1}^{D}\Bigg([c_j = c]\tfrac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\rho_{jtk}\zeta_{jni}^{v} - \xi^{-1}\prod_{m=1}^{N}(\sum_{e=1}^{K}$$

$$\sum_{i=1}^{T}\rho_{jie}\zeta_{jmi}^{v}\exp(\tfrac{1}{N+M}\delta_{ye}))\cdot\sum_{n=1}^{N}\left(\frac{\sum_{i=1}^{T}\rho_{jik}\zeta_{jni}^{v}\cdot\exp(\tfrac{1}{N+M}\delta_{yk})}{\sum_{e=1}^{K}\sum_{i=1}^{T}\rho_{jie}\zeta_{jni}^{v}\cdot\exp(\tfrac{1}{N+M}\delta_{ye})}\right)\Bigg)$$