SUPERVISED MULTI-MODAL TOPIC MODEL FOR IMAGE ANNOTATION

Thu Hoai Tran[2] *and Seungjin Choi*[1,2]

[1] Department of Computer Science and Engineering, POSTECH, Korea
[2] Division of IT Convergence Engineering, POSTECH, Korea
thtlamson@gmail.com, seungjin@postech.ac.kr

ABSTRACT

Multi-modal topic models are probabilistic generative models where hidden topics are learned from data of different types. In this paper we present *supervised multi-modal latent Dirichlet allocation (smmLDA),* where we incorporate class label (global description) into the joint modeling of visual words and caption words (local description), for image annotation task. We derive variational inference algorithm to approximately compute posterior distribution over latent variables. Experiments on a subset of LabelMe dataset demonstrate the useful behavior of our model, compared to existing topic models.

***Index Terms*—** Image annotation, latent Dirichlet allocation, topic models

1. INTRODUCTION

Latent Dirichlet allocation (LDA) is a widely-used topic model, which was originally developed to model text corpora [5]. It is a hierarchical Bayesian model in which each observed item is modeled as a finite mixture over an underlying set of topics and each topic is characterized by a distribution over words. The basic idea of LDA, when it is applied to model a set of images, treating an image as a collection of visual words, is shown in Fig. 1. The same intuition, in the case of documents, can be found in [2].

Multi-modal extensions of LDA, referred to as *multi-modal topic models,* have been proposed to jointly model data of different types. These models were mainly applied to *image annotation*, where the goal is to assign a set of keywords to an image, learning underlying topics from a set of image-annotations pairs. Earlier work on this direction is correspondence LDA (cLDA) [3] which finds conditional relationships between latent variable representations of visual words and caption words. The conditional distribution of the annotation given visual descriptors is modeled for automatic image annotation. Topic regression multi-modal LDA (trmmLDA) [10] is an alternative method for capturing statistical association between image and text. Unlike cLDA, trmmLDA learns two separate sets of hidden topics and counts on a regression module to allow a set of caption topics to be linearly predicted from the set of image topics. It was motivated by the regression-based latent factor model [1], which was further elaborated in the hierarchical Bayesian framework [9]. It was shown in [10] that trmmLDA is more flexible than cLDA in the sense that the former allows the number of image topics to be different from the number of caption topics.

Class label is a global description of an image, while annotated keywords are local descriptions of image patches. Class label and annotations are related to each other. For instance, an image labeled as "highway" scene is more likely to be annotated with cars and road rather than apple and desk. In this paper we present *supervised multi-modal latent Dirichlet allocation (smmLDA)*, where we incorporate class label into trmmLDA so that two sets of hidden topics, which are related via linear regression, are learned from data of two types as well as from class label. Several extensions of LDA to incorporate supervision have been developed in the literature [4, 6, 7, 11, 13].

Most of these existing methods are limited to learning from data of single type. The model, trmmLDA, outperforms most of previous methods in the task of image annotation, but is an unsupervised method. Our model, smmLDA, extends the previous state of the arts in this domain, trmmLDA, by incorporating supervision of class label.

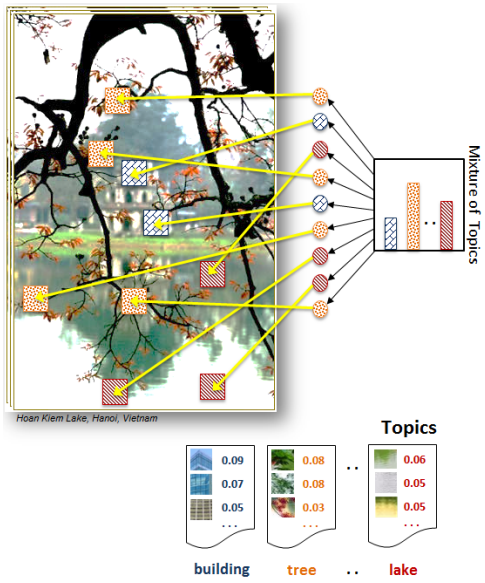

Fig. 1. A codeword is assigned to each image patch to represent an image as a collection of visual words. We assume that some number of topics, which are distributions over words, exist for the set of images. An illustration of how an image is generated by an LDA model is shown here. We first choose a distribution over topics (the histogram at right). Then for each visual word, choose a topic assignment (the circles with patterns filled in) and choose the visual word from the corresponding topic.

2. LATENT DIRICHLET ALLOCATION

We briefly give an overview of LDA [5]. LDA [5] is a generative probabilistic model of a corpus in which documents are represented

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

as random mixtures over latent topics where each topic is described by a distribution over words. Each document $\boldsymbol{w}_{d,1:N}$ is a sequence of N words, for $d = 1, \ldots, D$ (D is the size of a corpus) and each word $\boldsymbol{w}_{d,n} \in \mathbb{R}^V$ (V is the size of vocabulary) is a unit vector that has a single entry equal to one and all other entries equal to zero. For instance, if $\boldsymbol{w}_{d,n}$ is the vth word in the vocabulary, then $w_{d,n,v} = 1$ and $w_{d,n,j} = 0$ for $j \neq v$. The graphical model for LDA is shown in Fig. 2, where each document $\boldsymbol{w}_{d,1:N}$ is assumed to be generated as follows:

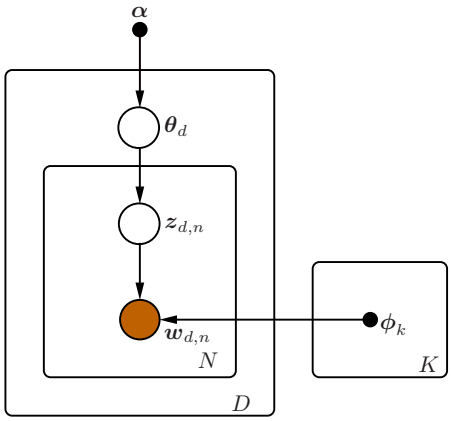

Fig. 2. Graphical model for LDA.

- Draw a vector of topic proportions, $\boldsymbol{\theta}_d \in \mathbb{R}^K$,

$$\boldsymbol{\theta}_d \sim \text{Dir}(\alpha_1, \ldots, \alpha_K).$$

- For each word n,
  - Draw a topic assignment $\boldsymbol{z}_{d,n} \in \mathbb{R}^K$ from multinomial distribution:

$$\boldsymbol{z}_{d,n} \,|\, \boldsymbol{\theta}_d \sim \text{Mult}(\boldsymbol{\theta}_d).$$

  - Draw a word $\boldsymbol{w}_{d,n} \in \mathbb{R}^V$:

$$\boldsymbol{w}_{d,n} \,|\, \boldsymbol{z}_{d,n}, \boldsymbol{\phi}_{1:K} \sim p(\boldsymbol{w}_{d,n}|\boldsymbol{z}_{d,n}, \boldsymbol{\phi}_{1:K}).$$

Given parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}_{1:K}$, the joint distribution of hidden and observed variables is given by

$$\begin{aligned} & p(\boldsymbol{\theta}_d, \boldsymbol{z}_{d,1:N}, \boldsymbol{w}_{d,1:N}|\boldsymbol{\alpha}, \boldsymbol{\phi}_{1:K}) \\ = \;& p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \left[\prod_{n=1}^{N} p(\boldsymbol{z}_{d,n}|\boldsymbol{\theta}_d) p(\boldsymbol{w}_{d,n}|\boldsymbol{z}_{d,n}, \boldsymbol{\phi}_{1:K}) \right]. \end{aligned}$$

Integrating over $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_{1:K}$, and summing over $\boldsymbol{z}_{d,1:N}$, the marginal distribution of a document is given by

$$\begin{aligned} & p(\boldsymbol{w}_{d,1:N}|\boldsymbol{\alpha}, \boldsymbol{\phi}_{1:K}) \\ = & \iint \sum_{\boldsymbol{z}_{d,1:N}} \left[\prod_{n=1}^{N} p(\boldsymbol{z}_{d,n}|\boldsymbol{\theta}_d) p(\boldsymbol{w}_{d,n}|\boldsymbol{z}_{d,n}, \boldsymbol{\phi}_{1:K}) \right] \\ & p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})\, d\boldsymbol{\theta}_d. \end{aligned}$$

Taking the product of marginal probabilities of single documents, the probability of a crops, the *marginal likelihood*, is given by

$$p(\boldsymbol{w}_{1:D,1:N}|\boldsymbol{\alpha}, \boldsymbol{\phi}_{1:K}) = \prod_{d=1}^{D} p(\boldsymbol{w}_{d,1:N}|\boldsymbol{\alpha}, \boldsymbol{\phi}_{1:K}). \quad (1)$$

Variational inference allows us to calculate approximate posterior distributions over hidden variables, $\{\boldsymbol{\theta}_d, \boldsymbol{z}_{d,n}\}$, by maximizing the variational lower-bound on the log marginal likelihood.

3. SUPERVISED MULTI-MODAL LDA

In this section, we present the main result, *discriminative multi-modal LDA (smmLDA),* where we incorporate class label into the joint modeling of visual words $\{\boldsymbol{r}_{d,n}\}$ and caption words $\{\boldsymbol{w}_{d,m}\}$, whose latent variable representations are related via linear regression. The graphical model for smmLDA is shown in Fig. 3.

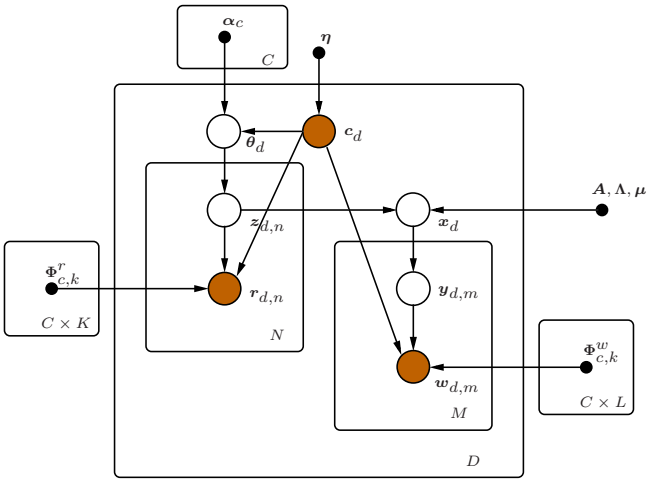

Fig. 3. Graphical model for discriminative multi-modal LDA (smmLDA).

3.1. Model

The generation process for each visual word $\{\boldsymbol{r}_{d,n}\}$ and caption word $\{\boldsymbol{w}_{d,m}\}$ is as follows.

- Choose a category label:

$$\boldsymbol{c}_d \in \mathbb{R}^C \sim \text{Mult}(\boldsymbol{\eta}) = \prod_{j=1}^{C} \eta_j^{c_{d,j}},$$

  where $\boldsymbol{c}_d$ is the C-dimensional unit vector. If $\boldsymbol{c}_d$ is the class label j, then $c_{d,j} = 1$ and $c_{d,i} = 0$ for $i \neq j$.

- Draw a vector of image topic proportions:

$$\boldsymbol{\theta}_d \in \mathbb{R}^K \sim \prod_{j=1}^{C} \text{Dir}(\boldsymbol{\theta}_d|\boldsymbol{\alpha}_j)^{c_{d,j}},$$

- For each visual word $\boldsymbol{r}_{d,n}$,
  - Draw an image topic assignment:

$$\boldsymbol{z}_{d,n} \in \mathbb{R}^K \sim \text{Mult}(\boldsymbol{\theta}_d) = \prod_{k=1}^{K} \theta_{d,k}^{z_{d,n,k}}.$$

– Draw a visual word:

$$\begin{aligned}\boldsymbol{r}_{d,n}|\boldsymbol{z}_{d,n},\boldsymbol{c}_d &\sim \text{Mult}(\boldsymbol{\Phi}^r)\\ &= \prod_{i=1}^{C}\prod_{k=1}^{K}\prod_{j=1}^{V_r}\left[\Phi^r_{i,k,j}\right]^{c_{d,i}z_{d,n,k}r_{d,n,j}},\end{aligned}$$

where V_r is the size of visual word vocabulary.

- Given the empirical image topic frequency,

$$\overline{\boldsymbol{z}}_d = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{z}_{d,n},$$

sample a real-valued topic proportion variable for caption text:

$$\boldsymbol{x}_d|\overline{\boldsymbol{z}}_d,\boldsymbol{A},\boldsymbol{\mu},\boldsymbol{\Lambda}\sim\mathcal{N}(\boldsymbol{x}_d|\boldsymbol{A}\overline{\boldsymbol{z}}_d+\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}).$$

- Compute caption topic proportions:

$$v_{d,l} = \frac{e^{x_{d,l}}}{\sum_{l=1}^{L}e^{x_{d,l}}}.$$

- For each caption word $\boldsymbol{w}_{d,m}$,

– Draw a caption topic assignment:

$$\boldsymbol{y}_{d,m}\sim\text{Mult}(\boldsymbol{v}_d)=\prod_{l=1}^{L}v_{d,l}^{y_{d,m,l}}.$$

– Draw a caption word:

$$\begin{aligned}\boldsymbol{w}_{d,m}|\boldsymbol{y}_{d,m},\boldsymbol{c}_d &\sim \text{Mult}(\boldsymbol{\Phi}^w)\\ &= \prod_{i=1}^{C}\prod_{l=1}^{L}\prod_{j=1}^{V_w}\left[\Phi^w_{i,l,j}\right]^{c_{d,i}y_{d,m,l}w_{d,m,j}},\end{aligned}$$

where V_w is the size of caption word vocabulary.

We define sets of variables as

$$\begin{aligned}\mathcal{R} &= \{\boldsymbol{r}_{d,n}\},\ \mathcal{W}=\{\boldsymbol{w}_{d,m}\},\ \mathcal{Z}=\{\boldsymbol{z}_{d,n}\},\ \mathcal{Y}=\{\boldsymbol{y}_{d,m}\},\\ \boldsymbol{C} &= \{\boldsymbol{c}_d\},\ \boldsymbol{\Theta}=\{\boldsymbol{\theta}_d\},\ \boldsymbol{X}=\{\boldsymbol{x}_d\}.\end{aligned}$$

Then the joint distribution over these variables obeys the following factorization:

$$\begin{aligned}&p(\mathcal{R},\mathcal{W},\boldsymbol{C},\boldsymbol{\Theta},\boldsymbol{X},\mathcal{Z},\mathcal{Y})\\ &= p(\boldsymbol{C}|\boldsymbol{\eta})p(\boldsymbol{\Theta}|\boldsymbol{C},\boldsymbol{\alpha})p(\mathcal{Z}|\boldsymbol{\Theta})p(\mathcal{R}|\mathcal{Z},\boldsymbol{\Phi}^r,\boldsymbol{C})\\ &\quad p(\boldsymbol{X}|\mathcal{Z},\boldsymbol{A},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\mathcal{Y}|\boldsymbol{X})p(\mathcal{W}|\mathcal{Y},\boldsymbol{\Phi}^w,\boldsymbol{C}),\end{aligned}\tag{2}$$

where

$$\begin{aligned}p(\boldsymbol{C}|\boldsymbol{\eta}) &= \frac{1}{C},\\ p(\boldsymbol{\Theta}|\boldsymbol{C},\boldsymbol{\alpha}) &= \prod_{d=1}^{D}\prod_{j=1}^{C}\text{Dir}(\boldsymbol{\theta}_d|\alpha_j)^{c_{d,j}},\\ p(\mathcal{Z}|\boldsymbol{\Theta}) &= \prod_{d=1}^{D}\prod_{n=1}^{N}\prod_{k=1}^{K}\theta_{d,k}^{z_{d,n,k}},\\ p(\mathcal{R}|\mathcal{Z},\boldsymbol{\Phi}^r,\boldsymbol{C}) &= \prod_{d=1}^{D}\prod_{n=1}^{N}\left[\prod_{i=1}^{C}\prod_{k=1}^{K}\prod_{j=1}^{V_r}\left(\Phi^r_{i,k,j}\right)^{c_{d,i}z_{d,n,k}r_{d,n,j}}\right],\\ p(\boldsymbol{X}|\mathcal{Z},\boldsymbol{A},\boldsymbol{\mu},\boldsymbol{\Lambda}) &= \prod_{d=1}^{D}\mathcal{N}(\boldsymbol{x}_d|\boldsymbol{A}\overline{\boldsymbol{z}}_d+\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}),\\ p(\mathcal{Y}|\boldsymbol{X}) &= \prod_{d=1}^{D}\prod_{m=1}^{M}p(\boldsymbol{y}_{d,m}|\boldsymbol{x}_d)=\prod_{d=1}^{D}\prod_{m=1}^{M}\prod_{l=1}^{L}v_{d,l}^{y_{d,m,l}},\\ p(\mathcal{W}|\mathcal{Y},\boldsymbol{\Phi}^w,\boldsymbol{C}) &= \prod_{d=1}^{D}\prod_{m=1}^{M}\left[\prod_{i=1}^{C}\prod_{l=1}^{L}\prod_{j=1}^{V_w}\left(\Phi^w_{i,l,j}\right)^{c_{d,i}y_{d,m,l}w_{d,m,j}}\right],\end{aligned}$$

3.2. Variational Inference

The log marginal likelihood is given by

$$\begin{aligned}&\log p(\mathcal{R},\mathcal{W},\boldsymbol{C})\\ &= \log\int_{\boldsymbol{\Theta}}\int_{\boldsymbol{X}}\sum_{\mathcal{Z}}\sum_{\mathcal{Y}}p(\mathcal{R},\mathcal{W},\boldsymbol{C},\boldsymbol{\Theta},\boldsymbol{X},\mathcal{Z},\mathcal{Y})d\boldsymbol{\Theta}d\boldsymbol{X}\\ &\geq \int_{\boldsymbol{\Theta}}\int_{\boldsymbol{X}}\sum_{\mathcal{Z}}\sum_{\mathcal{Y}}q(\boldsymbol{\Theta},\boldsymbol{X},\mathcal{Z},\mathcal{Y})\\ &\quad \log\left(\frac{p(\mathcal{R},\mathcal{W},\boldsymbol{C},\boldsymbol{\Theta},\boldsymbol{X},\mathcal{Z},\mathcal{Y})}{q(\boldsymbol{\Theta},\boldsymbol{X},\mathcal{Z},\mathcal{Y})}\right)d\boldsymbol{\Theta}d\boldsymbol{X}\\ &= \mathcal{F}(q),\end{aligned}\tag{3}$$

where $q(\boldsymbol{\Theta},\boldsymbol{X},\mathcal{Z},\mathcal{Y})$ denotes the *variational distribution* and Jensen's inequality is used to reach the *variational lower-bound* $\mathcal{F}(q)$.

We assume that the variational distribution factorizes as

$$q(\boldsymbol{\Theta},\boldsymbol{X},\mathcal{Z},\mathcal{Y}) = q(\boldsymbol{\Theta})q(\boldsymbol{X})q(\mathcal{Z})q(\mathcal{Y}),\tag{4}$$

where each distribution is assumed to be of the form in Table 1. Variational parameters, $\left\{\{\overline{\alpha}_{d,k}\},\{\overline{\boldsymbol{x}}_d,\boldsymbol{\Gamma}_d^{-1}\},\{\tau_{d,n,k}\},\{\rho_{d,m,l}\}\right\}$, are determined by maximizing the variational lower-bound

$$\begin{aligned}\mathcal{F}(q) &= \mathbb{E}_q\Big[\log p(\boldsymbol{C}|\boldsymbol{\eta})+\log\log p(\boldsymbol{\Theta}|\boldsymbol{C},\boldsymbol{\alpha})+\log p(\mathcal{Z}|\boldsymbol{\Theta})\\ &+ \log p(\mathcal{R}|\mathcal{Z},\boldsymbol{\Phi}^r,\boldsymbol{C})+\log p(\boldsymbol{X}|\mathcal{Z},\boldsymbol{A},\boldsymbol{\mu},\boldsymbol{\Lambda})\\ &+ \log p(\mathcal{Y}|\boldsymbol{X})+\log p(\mathcal{W}|\mathcal{Y},\boldsymbol{\Phi}^w,\boldsymbol{C})\Big]\\ &- \mathbb{E}_q\Big[\log q(\boldsymbol{\Theta})+\log q(\boldsymbol{X})+\log q(\mathcal{Z})+\log q(\mathcal{Y})\Big],\end{aligned}$$

where $\mathbb{E}_q[\cdot]$ denotes the statistical expectation with respect to the variational distribution $q(\cdot)$. Detailed derivations for variational inference are omitted here due to the space limitation. In fact, derivations can be done in a similar manner to [10]. Especially,

Table 1. Updating equations for variational parameters.

Variational posterior distributions	Updating equations for variational parameters
$q(\boldsymbol{\Theta}) = \prod_{d=1}^{D} \text{Dir}(\boldsymbol{\theta}_d \vert \overline{\boldsymbol{\alpha}}_d)$	$\overline{\alpha}_{d,k} = \sum_{i=1}^{C} c_{d,i}\alpha_{i,k} + \sum_{n=1}^{N} \tau_{d,n,k}$
$q(\mathcal{Z}) = \prod_{d=1}^{D}\prod_{n=1}^{N}\prod_{k=1}^{K} \tau_{d,n,k}^{z_{d,n,k}}$	$\log \tau_{d,n,k} \propto \left[\psi(\overline{\alpha}_{d,k}) - \psi(\overline{\alpha}_{d,1} + \cdots + \overline{\alpha}_{d,K})\right] + \sum_{i=1}^{C}\sum_{j=1}^{V_r} c_{d,i} r_{d,n,j} \log \Phi^r_{i,k,j}$ $+ \frac{1}{N}\left[\boldsymbol{A}^\top \boldsymbol{\Lambda}(\overline{\boldsymbol{x}}_d - \boldsymbol{\mu})\right]_k - \frac{1}{2N^2}\left[\text{diag}(\boldsymbol{A}^\top \boldsymbol{\Lambda} \boldsymbol{A})\right]_k - \frac{1}{N^2}\left[\boldsymbol{A}^\top \boldsymbol{\Lambda} \boldsymbol{A} \sum_{m \neq n} \boldsymbol{\phi}_{d,m}\right]_k$
$q(\mathcal{Y}) = \prod_{d=1}^{D}\prod_{m=1}^{M}\prod_{l=1}^{L} \rho_{d,m,l}^{y_{d,m,l}}$	$\log \rho_{d,m,l} \propto \sum_{i=1}^{C}\sum_{j=1}^{V_w} c_{d,i} w_{d,m,j} \log \Phi^w_{i,l,j} + \overline{x}_{d,l}$
$q(\boldsymbol{X}) = \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{x}_d \vert \overline{\boldsymbol{x}}_d, \boldsymbol{\Gamma}_d^{-1})$	$\xi_d = \sum_{l=1}^{L} e^{\overline{x}_{d,l} + \frac{0.5}{\gamma_{d,l}}}$ Determine $\overline{\boldsymbol{x}}_d$ and $\gamma_{d,l}$ by Newton-Raphson method

$\mathbb{E}_q\left[\log v_{d,l}\right] = \mathbb{E}_q\left[\frac{e^{x_{d,l}}}{\sum_{l=1}^{L} e^{x_{d,l}}}\right]$ is not directly maximized. Instead, as in [10], its convex lower-bound is maximized:

$$\begin{aligned}\mathbb{E}_q\left[\log v_{d,l}\right] &\geq \mathbb{E}_q\left[x_{d,l} - \log \xi_d - \frac{1}{\xi_d}\sum_{l=1}^{L} e^{x_{d,l}} + 1\right] \\ &= \overline{x}_{d,l} - \log \xi_d - \frac{1}{\xi_d}\sum_{l=1}^{L} e^{\overline{x}_{d,l} + \frac{.5}{\gamma_{d,l}}} + 1.\end{aligned}$$

3.3. Parameter Estimation

Coordinate ascent algorithms for updating variational parameters are summarized in Table 1. Regression parameters $\{\boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\}$ are updated:

$$\boldsymbol{A} = \left(\frac{1}{N}\sum_{d=1}^{D}(\overline{\boldsymbol{x}}_d - \boldsymbol{\mu})\sum_{n=1}^{N}\boldsymbol{\phi}_{d,n}^\top\right) \left(\frac{1}{N^2}\sum_{d=1}^{D}\text{tr}\left(\sum_{n=1}^{N}\text{diag}(\boldsymbol{\phi}_{d,n}) + \sum_{n=1}^{N}\boldsymbol{\phi}_{d,n}\sum_{m\neq n}\boldsymbol{\phi}_{d,m}^\top\right)\right)^{-1},$$

$$\boldsymbol{\mu} = \frac{1}{D}\sum_{d=1}^{D}\left(\overline{\boldsymbol{x}}_d - \frac{1}{N}\boldsymbol{A}\sum_{n=1}^{N}\boldsymbol{\phi}_{d,n}\right),$$

$$\boldsymbol{\Lambda}^{-1} = \frac{1}{D}\left((\overline{\boldsymbol{x}}_d - \boldsymbol{\mu})(\overline{\boldsymbol{x}}_d - \boldsymbol{\mu})^\top + \boldsymbol{\Gamma}_d^{-1} - \frac{1}{N}\boldsymbol{A}\left(\sum_{n=1}^{N}\boldsymbol{\phi}_{d,n}\right)\overline{\boldsymbol{x}}_d^\top\right).$$

Multinomial parameters $\{\boldsymbol{\Phi}^r, \boldsymbol{\Phi}^w\}$ are updated:

$$\Phi^r_{i,k,j} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N} c_{d,i}\tau_{d,n,k} r_{d,n,j}}{\sum_{j=1}^{V_r}\sum_{d=1}^{D}\sum_{n=1}^{N} c_{d,i}\tau_{d,n,k} r_{d,n,j}},$$

$$\Phi^w_{i,l,j} = \frac{\sum_{d=1}^{D}\sum_{m=1}^{M} c_{d,i}\rho_{d,m,l} w_{d,m,j}}{\sum_{j=1}^{V_w}\sum_{d=1}^{D}\sum_{m=1}^{M} c_{d,i}\rho_{d,m,l} w_{d,m,j}}.$$

Dirichlet parameters $\{\boldsymbol{\alpha}_c\}$ are updated using Newton-Raphson method, as in LDA [5].

Given a test image $\boldsymbol{r}_{*,1:N}$, class label and annotations are determined by choosing the most probable ones among conditional probabilities $p(\boldsymbol{c}_d \vert \boldsymbol{r}_{*,1:N})$ and $p(\boldsymbol{w}_{d,m} \vert \boldsymbol{r}_{*,1:N})$.

4. EXPERIMENTS

We use the 8-category subset of LabelMe dataset [12] to perform image annotation experiments. Categories include 'coast', 'forest', 'highway', 'inside city', 'mountain', 'open country', 'street', and 'tall building'. This subset has 2686 images of size 256×256 with complete annotations.

We use 128-dimensional SIFT descriptors [8] computed on 20×20 image patches where each image patch is obtained by sliding a window with a 20-pixel interval. Then we run k-means clustering on 128-dimensional descriptors to learn a 256-word visual codebook. For the annotation words, we remove the words appearing less than 3 times in the whole data. Finally, we have a complete set of triples (visual words, caption words, class label). The whole data is separated into the training set of size 2000, and the test set of size 686.

We evaluate the performance in terms of *caption perplexity*, defined as

$$\text{Perplexity} = \exp\left\{-\frac{\sum_{d=1}^{D}\sum_{m=1}^{M_d}\log p(\boldsymbol{w}_{d,m} \vert \boldsymbol{r}_{d,1:N})}{\sum_{d=1}^{D} M_d}\right\},$$

where $p(\boldsymbol{w}_{d,m} \vert \boldsymbol{r}_{d,1:N})$ is the conditional probability of caption words given an image $\boldsymbol{r}_{d,1:N}$ and M_d is the number of caption words in document d. The higher conditional likelihood leads to the lower perplexity. The performance comparison to the previous state of the arts, trmmLDA, is summarized in Table 2, where our smmLDA outperforms trmmLDA.

Table 2. Perplexity comparison.

Method	$K = 25$	$K = 30$
trmmLDA [10]	35	36
smmLDA (our method)	**28.5**	**30.4**

5. CONCLUSIONS

In this paper we have presented a multi-modal extension of LDA with supervision, leading to smmLDA. We have developed variational inference algorithms to approximately compute posterior distributions over variables of interest in smmLDA. Applications to image annotation demonstrated the high performance of smmLDA compared to the previous state of the arts.

Acknowledgments: This work was supported by National Research Foundation (NRF) of Korea (NRF-2013R1A2A2A01067464) and POSTECH Rising Star Program.

6. REFERENCES

[1] D. Agarwal and B.-C. Chen, "Regression-based latent factor models," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Paris, France, 2009.

[2] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[3] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Toronto, Canada, 2003.

[4] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 20. MIT Press, 2008.

[5] D. M. Blei, A. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[6] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, 2005.

[7] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 21, 2009.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] S. Park, Y.-D. Kim, and S. Choi, "Hierarchical Bayesian matrix factorization with side information," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Beijing, China, 2013.

[10] D. Putthividhya, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent Dirichlet allocation for image annotation," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, 2010.

[11] D. Ramage, D. Hall, H. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2009.

[12] B. C. Russel, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, pp. 157–173, 2008.

[13] C. Wang, D. M. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009.