Nonparametric Bayesian Upstream Supervised Multi-Modal Topic Models

Renjie Liao
Department of Computer Science & Engineering
The Chinese University of Hong Kong
rjliao@cse.cuhk.edu.hk

Jun Zhu
State Key Lab of Intell. Tech & Sys.; TNLIST Lab
Dept. of Comp. Sci & Tech
Tsinghua University, Beijing, 100084, China
dcszj@mail.tsinghua.edu.cn

Zengchang Qin
Intelligent Computing and Machine Learning Lab
School of ASEE
Beihang University, Beijing, 100191, China
zcqin@buaa.edu.cn

ABSTRACT

Learning with multi-modal data is at the core of many multimedia applications, such as cross-modal retrieval and image annotation. In this paper, we present a nonparametric Bayesian approach to learning upstream supervised topic models for analyzing multi-modal data. Our model develops a compound nonparametric Bayesian multi-modal prior to describe the correlation structure of data both within each individual modality and between different modalities. It extends the hierarchical Dirichlet process (HDP) through incorporating upstream supervised response variables and values of latent functions under Gaussian process (GP). Upstream responses shared by data from multiple modalities are beneficial for discriminatively training and GP allows flexible structure learning of correlations. Hence, our model inherits the automatic determination of the number of topics from HDP, structure learning from GP and enhanced predictive capacity from upstream supervision. We also provide efficient variational inference and prediction algorithms. Empirical studies demonstrate superior performances on several benchmark datasets compared with previous competitors.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Nonparametric statistics; H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.4 [**Information System Applications**]: Miscellaneous

Keywords

Multi-modal Learning; Nonparametric Bayesian; Topic Model; Cross-modal Retrieval

1. INTRODUCTION

Nowadays, large collections of data on the web consist of various modalities, such as images, texts, and audio or video clips. Extracting useful knowledge from these growing multi-modal data has become increasingly important in many application areas. Multi-modal learning, sometimes referred to as multi-view learning [8] or multi-field learning [26], aims at modeling collections of such kind of data and making predictions when data of some modalities is missing. One typical example is to model pairs of images and associated texts (e.g., captions, paragraphs or articles), which lays the foundation of many valuable applications, including cross-modal retrieval, image annotation and so on. However, this multi-modal learning problem is rather challenging, since it requires analyzing not only the characteristics of data in single modality but also the relevances of data across different modalities.

Previous works in this area have primarily focused on looking for latent representations shared by the multi-modal data. Based on the different techniques they used, these works can be roughly divided into three categories: subspace learning, undirected probabilistic graphical models (PGMs), and directed PGMs. Representative works of the first class include canonical correlation analysis (CCA) and its variants [31, 28]. By maximizing the correlation between different modalities, it aims to find a low dimensional subspace representation for multi-modal data. Another ones are based on distance metric learning [33, 34] and hash function learning [39], which try to discover optimal nonlinear subspace endowed with an expected similarity measure. For the second class, Markov random fields (MRF) based methods have commonly been used. For example, the dual-wing harmonium model [35] describes the latent representations shared by images and texts through tying the latent variables of two basic harmonium models. The above two types of models are capable of effectively discovering the desired latent representations, but often lack intuitive explanations and do not perform well in terms of prediction.

The main line of the third class of work are multi-modal topic models (mmTM), such as multi-modal latent Dirichlet allocation (mmLDA) [2], correspondence latent Dirichlet allocation (CorrLDA) [2] and the multi-modal aspect model [22]. Most of them describe the single-modal data via standard topic models, like latent Dirichlet allocation (LDA) [4]. Thereafter, [36] extends the Dirichlet prior in mmTM by using a hierarchical Dirichlet process (HDP) [29] which is able to automatically choose the number of topics. These models introduce shared latent variables which either indicate the topic proportions as in mmLDA or the

WSDM'14, February 24–28, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2556195.2556238.

indexes of topics as in CorrLDA, thus enforcing strong correlations between topics from different modalities. This kind of strong correlation is inappropriate since usually data in one modality contains a considerable amount of modality-private information that is unrelated to the other. Therefore, relying on the logistic normal distribution adopted in correlated topic model (CTM) [3], authors in [26] relax this strong correlation and allow structure learning of the correlation matrix. Beyond that, the discrete infinite logistic normal (DILN) [23] distribution is utilized in [30] to further keep the private topics inside each modality. All the above mmTMs learn the latent representations of the multi-modal data in an unsupervised manner and are able to offer intuitive probabilistic interpretations.

Recently, upstream supervised response variables, like categorical labels, which are available in many scenarios, are utilized for enhancing the performance of prediction. Upstream supervised models are common for understanding scenes in computer vision community [11, 41], since image categories are cheaply accessible on the web. In contrast to downstream supervised models, like supervised topic model and its variants [5, 40], upstream ones assume that response variables are directly or indirectly involved in generating latent variables. In the setting of multi-modal learning, these supervised response variables are usually shared by data from different modalities, thus help describing the relation between data from different modalities precisely. Moreover, with this kind of supervised guidance, latent representations learned by the model are presumably more discriminative. For example, it has been shown in [24] that applying logistic regression to class label, is helpful for improving the predictive capacity of CCA in cross-modal retrieval. At the same time, similar categorical information has been exploited in [8, 7] to extend the MRF-based multi-view models through a supervised max-margin approach and boosted predictive results are also demonstrated.

In this paper, we propose a novel nonparametric Bayesian upstream supervised (NPBUS) multi-modal topic model by combining the above advantages of previous works. Specifically, our model first inherits the two merits of having an explicit probabilistic explanation and automatic determination of the number of topics from the HDP-based mmTM. Then our model introduces a Gaussian process (GP) [25] to flexibly capture correlation structures both within each individual modality and between multiple different modalities. Moreover, upstream supervised response variables shared by data from different modalities are incorporated into a normalized gamma representation of HDP, thus making our method possess remarkable predictive ability. We also derive an efficient variational inference algorithm for training. The proposed NPBUS model demonstrates superior experimental results than various competitors in the predictive tasks of cross-modal retrieval and image annotation.

The rest of the paper is structured as follows. Sec. 2 introduces the background works related to our model. Sec. 3 elucidates our NPBUS multi-modal topic model. Sec. 4.1 presents our variational inference and prediction algorithms. Sec. 5 presents our empirical results on cross-modal retrieval and image annotation. Finally, Sec. 6 concludes.

2. BACKGROUND WORKS

In this section, we review the hierarchical Dirichlet process (HDP) [29] based topic model for single-modal data and the discrete infinite logistic normal (DILN) distribution [23], which lays the foundation of our model to be presented later. Terminologies from text modeling, e.g. "words", "documents" and "vocabulary", are used throughout the paper, since they can be well generalized in modeling data of other modalities. For example, in the context of bag-of-words model for image classification, "words", also referred to as "visual words", are clustering centers of some low level visual descriptors (e.g. SIFT [19]), and "documents" correspond to images.

2.1 Hierarchical Dirichlet Process Topic Models

We first introduce latent Dirichlet allocation (LDA) [4] which assumes that documents are represented as mixtures of latent topics η and words appearing in a document are drawn independently from their corresponding topics. Specifically, a topic mixing proportion θ for each document is first sampled from a Dirichlet prior and thereafter the topic for each word is chosen through sampling a topic index $z \sim Mult(\theta)$, where $Mult(\cdot)$ is a multinomial distribution. At last, word x is drawn from its corresponding topic η_z which is a multinomial distribution over the words in the vocabulary. Note that, in a fully Bayesian treatment, we may place a prior for topics η. To further model the topic mixing proportion, hierarchical Dirichlet process (HDP) [29], a nonparametric Bayesian prior, is widely adopted. Formally, HDP is a Dirichlet process (DP) [13] that has another Dirichlet process as its base probability measure. In this paper, we focus on two-level HDPs, though it may be extended to arbitrary levels. Its hierarchical representation is

$$\begin{aligned} G &\sim DP(\alpha G_0) \\ \tilde{G} &\sim DP(\beta G), \end{aligned} \tag{1}$$

where G_0 is the aforementioned Dirichlet prior for η, and α and β are the first and second level concentration parameters respectively. The main advantage of HDP formulation is the automatical determination of the number of topics owing to its nonparametric nature. Moreover, due to the almost sure discreteness property of DP [1], the HDP prior enables the implicit sharing of atomic probability measures between the repeatedly sampled topic proportions θ. And it will largely ease our work of specifying multiple priors, since there will be multiple different sets of topics in the multi-modal setting.

2.2 Discrete Infinite Logistic Normal (DILN) Distribution

Though HDP is a popular prior for modeling topic proportions, it is insufficient to capture the correlation structure between different topics since its construction is a random measure which means all random variables as well as all summations of subsets are independent. To overcome this weak point, discrete infinite logistic normal (DILN) model [23] is proposed. It takes advantages from both HDP and CTM, and offers an effective prior for structure learning of topics.

Specifically, DILN first introduces an auxiliary variable ℓ for each topic η. It can be interpreted as the latent location of a topic in some latent space. As ℓ and η are associated, the base distribution used in the first level HDP is augmented as a product measure $G_0 \times L_0$, where G_0 is still the prior distribution for topics and L_0 is a distribution over latent locations. To construct DILN, a hierarchical sampling process is

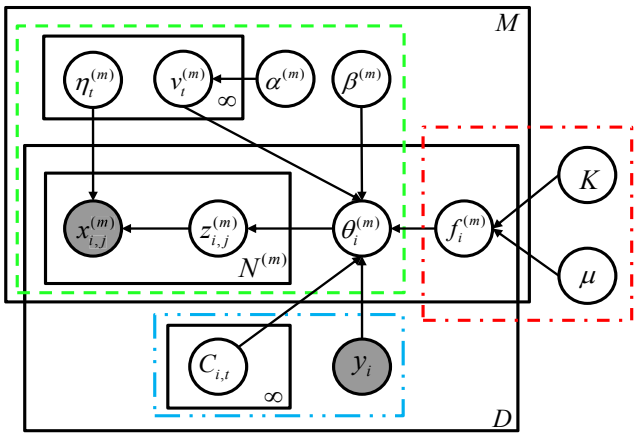

Figure 1: A graphical illustration of our NPBUS model. The red rectangle (single dot dashed lines) indicates the part of GP, the green rectangle (dashed lines) indicates the part of HDP topic model and the blue rectangle (double dots dashed lines) indicates the part of upstream supervised response variable.

implemented as HDP. In the first level, a random measure G is drawn from Dirichlet process, $G \sim DP(\alpha G_0 \times L_0)$. And in the second level, a random measure $\tilde{G}$ is drawn from another DP, $\tilde{G} \sim DP(\beta G)$. Meanwhile, a random function $f(\ell)$ is drawn from a Gaussian process, $f(\ell) \sim GP(\mu(\ell), k(\ell, \ell'))$, where $f(\cdot)$ and GP are defined on the latent location ℓ, μ and k are the mean function and kernel function respectively. Finally, a sample from DILN, i.e., the topic mixing proportion, is constructed via multiplying the random measure $\tilde{G}$ by the exponentiated value of the random function $f(\ell)$. Hence, a topic assignment z could be drawn as,

$$z \sim \exp(f(\ell))\tilde{G}. \tag{2}$$

Since ℓ is explicitly correlated via a GP, topic proportions drawn from DILN are also correlated. And, due to the discreteness of $\tilde{G}$, the sampled topic mixing proportion is discrete.

3. NPBUS MULTI-MODAL TOPIC MODEL

We now develop the nonparametric Bayesian upstream supervised (NPBUS) multi-modal topic model and elaborate its merits in modeling multi-modal data. Without loss of generality, we focus on the two-modal dataset which contains collection of images and texts. We emphasize that the generalization to more and different modalities is straightforward. For clarity, we denote the two-modal observable dataset by $X = \{x_i^{(m)}, y_i | i = 1, 2, ..., D, m = 1, 2\}$, where $x_i^{(m)}$ and y_i are the ith word of mth modality and ith response variable respectively, and D is the size of the dataset. Note that in our problem setting, each pair of image and text shares an extra response variable which plays an supervision role. Here we highlight the key challenges dealing with such a type of dataset: (i) modeling correlations of topic proportions both within modality and between different modalities; and (ii) exploiting upstream supervising information of response variable. Then we will explain in detail our NPBUS model and how it tackles the above challenges.

3.1 Compound Nonparametric Bayesian Multi-Modal Prior

In the context of multi-modal learning, words in each modality are drawn independently from the modality-specific topic given their corresponding topic assignments. For the ith document of mth modality, we denote the vector of topic mixing proportion as $\theta_i^{(m)}$ and the probability of choosing the tth topic is thus $\theta_{i,t}^{(m)}$. The topic assignment of the jth word $x_{i,j}^{(m)}$ is defined as $z_{i,j}^{(m)}$ and the auxiliary variable of tth topic appeared in DILN is $f_{i,t}^{(m)}$. Note that, $f_{i,t}^{(m)}$ is the value of a random function modeled by GP and we omit its input latent location ℓ_t here and clarify the reason later.

Similar with HDP topic models, the first level Dirichlet process of our model is represented via a stick breaking process [27],

$$\begin{aligned} \eta_t^{(m)} &\sim G_0 \\ \tilde{v}_t^{(m)} &\sim Beta(1, \alpha^{(m)}) \\ v_t^{(m)} &= \tilde{v}_t^{(m)} \prod_{j=1}^{t-1} (1 - \tilde{v}_j^{(m)}) \\ G &= \sum_{t=1}^{\infty} v_t^{(m)} \delta_{\eta_t^{(m)}}, \end{aligned} \tag{3}$$

where G_0 is the Dirichlet prior for topics, $v_t^{(m)}$ is the stick proportion and $Beta(\cdot)$ is a beta distribution.

In the second level, we first draw a random measure $\tilde{G}$ from another Dirichlet process, $DP(\beta^{(m)} G)$. Note that we here focus on presenting our prior and leave the detailed construction of this DP in next section. As aforementioned, different topics are just weakly correlated owing to the normalization property of probability measure $\tilde{G}$. To capture flexible correlation structures, we then concatenate the auxiliary variables from each modality into an single vector, like below,

$$f_i = [f_{i,1}^{(1)}, ...f_{i,\infty}^{(1)}, f_{i,1}^{(2)}, ...f_{i,\infty}^{(2)}, ..., f_{i,1}^{(M)}, ..., f_{i,\infty}^{(M)}]. \tag{4}$$

Then we use a GP to model the whole vector f_i, which is similar with what [26] and [30] have done. Finally, as in DILN, we could multiply the samples from G by the exponentiated value of $f_{it}^{(m)}$, which fulfills the introduction of correlation to topic proportions. Specifically, in the covariance matrix of GP, the diagonal blocks describe the correlation of topic proportions within the same modality, and off-diagonal ones describe the correlation of topic proportions between different modalities. Note that, to sample $f_{it}^{(m)}$, we merely need to sample f_i and obtain $f_{it}^{(m)}$ immediately according to its position in the whole vector.

We now turn to model the upstream supervised response variable y which can be categorical labels, rating scores and so on. Here we only consider the discrete case, i.e., $y \in \{1, 2, ..., \mathbb{K}\}$. Since these responses induce a supervised grouping of multi-modal data, we could conduct correlation structure learning within each group. Therefore, in this way, the whole multi-modal dataset would be modeled more precisely and the generated topic proportions would be more discriminative. Specifically, we introduce another random scaling factor $C_{i,t}$ which is drawn from a gamma distribution $Gamma(a_{t,y_i}, b_{t,y_i})$. Then we divide exponentiated value $\exp(f_{i,t}^{(m)})$ by $C_{i,t}$ to obtain a response dependent random scaling factor. By doing so, the compound nonparametric

Bayesian multi-modal prior distribution of topic proportions is formulated as,

$$\theta_i^{(m)} \propto \sum_{t=1}^{\infty} \frac{\exp(f_{i,t}^{(m)})}{C_{i,t}} \tilde{G}. \tag{5}$$

Now the correlation structures of multi-modal data is captured in the weight term $\exp(f_{i,t}^{(m)})$. And, multi-modal data within the same group, i.e., having the same value of y_i, would share the same hyperparameters a_{t,y_i} and b_{t,y_i}. Thus the group specific random weight $C_{i,t}$ has introduced the upstream supervising information. As a result, with this prior at hand, we have solved the two challenges mentioned above.

3.2 Normalized Gamma Representation

Following the normalized gamma process construction of HDP in [23], we now develop a representation of our NPBUS multi-modal topic model. It will be clear that this kind of representation incorporates the correlation structure via the second parameter of gamma distribution and simplifies our posterior inference.

First of all, a normalized gamma process construction of the second level HDP could be expressed as,

$$\begin{aligned} \tilde{\theta}_{i,t}^{(m)} &\sim Gamma(\beta^{(m)} v_t^{(m)}, 1) \\ \theta_i^{(m)} &= \sum_{t=1}^{\infty} \frac{\tilde{\theta}_{i,t}^{(m)}}{\sum_{j=1}^{\infty} \tilde{\theta}_{i,j}^{(m)}} \delta_{\eta_t^{(m)}}. \end{aligned} \tag{6}$$

where the normalizing constant $\sum_{j=1}^{\infty} \tilde{\theta}_{i,j}^{(m)}$ is almost surely finite [23]. Building upon Eq. (6), a topic mixing proportion $\theta_i^{(m)}$ drawn from the second level of our prior could be constructed below,

$$\begin{aligned} f_{i,t}^{(m)} &\sim GP(\mu, K), \\ C_{i,t} &\sim Gamma(a_{y_i,t}, b_{y_i,t}), \\ \tilde{\theta}_{i,t}^{(m)} &\sim Gamma(\beta^{(m)} v_t^{(m)}, \exp(-f_{i,t}^{(m)})/C_{i,t}), \\ \theta_i^{(m)} &= \sum_{t=1}^{\infty} \frac{\tilde{\theta}_{i,t}^{(m)}}{\sum_{j=1}^{\infty} \tilde{\theta}_{i,j}^{(m)}} \delta_{\eta_t^{(m)}}. \end{aligned} \tag{7}$$

where the third line is derived by absorbing the scaling factor $\exp(f_{i,t}^{(m)})/C_{i,t}$ of Eq. (5) into the second parameter of gamma distribution.

Once the topic mixing proportion $\theta_i^{(m)}$ is obtained, we are capable to generate multi-modal data. The overall generative process of our model is summarized below:

- For each modality $m = 1, ..., M$
  - For each latent topic $t = 1, 2, ..., \infty$

    draw $\{v_t^{(m)}, \eta_t^{(m)}\}$ according to Eq. (3)
  - For each document $i = 1, 2, ..., D$

    draw $\theta_i^{(m)}$ according to Eq. (7)
    - For each word $j = 1, 2, ..., N^{(m)}$

      draw $z_{i,j}^{(m)} \sim Mult(\theta_i^{(m)})$

      draw $x_{i,j}^{(m)} \sim Mult(\eta_{z_{i,j}^{(m)}}^{(m)})$

Graphically, the PGM of our model is illustrated as in Fig. 1 in which we separately annotate the part of HDP topic model, the part of GP prior and the part of upstream supervised response variable. Here we omit the prior G_0 of topics and hyperparameters of $C_{i,t}$ in the figure for simplicity. As can be seen from the figure, both within modality and between modality correlation of topic proportions are captured and the upstream supervised response variable is also exploited.

3.3 The Correlation Structure

We now exploit the correlation structure of topic proportions in our model. Owing to the normalized gamma process representation above, we could analytically calculate the first two order central moments and the covariance of the unnormalized topic proportions (i.e., $\tilde{\theta}_{i,t}^{(m)}$). These moments essentially contain the desired correlation structures, since the correlation matrix can be obtained by merely normalizing the covariance matrix. The mean function μ of GP is assumed to be zero in deriving the following equations. After some integration procedures, we could obtain the results below,

$$\begin{aligned} &\mathbb{E}[\tilde{\theta}_{i,t}^{(m)} | \Theta] = \frac{\beta^{(m)} v_t^{(m)} a_{y_i,t}}{b_{y_i,t}} \mathrm{e}^{K_{t,t}^{(m,m)}/2} \\ &\mathbb{V}[\tilde{\theta}_{i,t}^{(m)} | \Theta] = \\ &\frac{\beta^{(m)} v_t^{(m)} a_{y_i,t}}{b_{y_i,t}} \mathrm{e}^{2K_{t,t}^{(m,m)}} \left[1 + \frac{\beta^{(m)} v_t^{(m)} a_{y_i,t}}{b_{y_i,t}} (1 - \frac{1}{\mathrm{e}^{K_{t,t}^{(m,m)}}}) \right] \\ &Cov[\tilde{\theta}_{i,t}^{(m)}, \tilde{\theta}_{i,s}^{(n)} | \Theta] = \\ &\frac{\beta^{(m)} \beta^{(n)} v_t^{(m)} v_s^{(n)} a_{y_i,t} a_{y_i,s}}{b_{y_i,t} b_{y_i,s}} \mathrm{e}^{(K_{t,t}^{(m,m)} + K_{s,s}^{(n,n)})/2} (\mathrm{e}^{K_{t,s}^{(m,n)}} - 1) \end{aligned} \tag{8}$$

where the conditional set is $\Theta = \{\beta^{(m)}, v_t^{(m)}, a_{t,y_i}, b_{t,y_i}, K | t = 1, 2, ..., \infty\}$. Recall that the diagonal block sub-matrices of K denote the correlation of topic proportions within each modality and off-diagonal ones denote the correlation between different modalities. To make expressions clear, we use the superscripts of K to denote the row and column indexes of block sub-matrices in K. For example, $K^{(1,1)}$ may denote the covariance matrix of topics from image modality, and $K^{(1,2)}$ may denote the covariance matrix between topics of image modality and text modality. Then we use the subscripts of K to denote the row and column indexes in the corresponding sub-matrix.

Note that the correlation structures shown in the above equations are distinct from ones in [26] and [30]. As discussed in [23], the term $v_t^{(m)}$ indicates how sparsity is enforced in the first level DP. Moreover, in our model, it is clear that the covariance depends both on the kernel matrix K of GP and hyperparameters a_{t,y_i} and b_{t,y_i} of gamma distribution. We could control the correlation structure flexibly through learning these parameters.

4. INFERENCE AND PREDICTION

In this section, we focus on the computational problems of our NPBUS model and present an efficient variational posterior inference algorithm and a corresponding prediction algorithm.

4.1 Variational Inference

To infer the posterior distributions of the latent variables, especially the unnormalized topic mixing proportions $\tilde{\theta}$, and to learn the model parameters, we employ a truncated mean-field variational inference algorithm, which has been shown to be effective in dealing with nonparametric Bayesian models [29]. First of all, we denote the previous set of observed variables as X and the set of all latent variables as V. By introducing a variational distribution $q(V)$, we can write the general variational lower bound for the log evidence $\log p(X)$ as

$$\mathcal{L} = \mathbb{E}_q[\log p(V, X)] - \mathbb{E}_q[\log q(V)], \tag{9}$$

where $\mathbb{E}_q[\cdot]$ means the expectation is calculated with respect to the distribution q. According to the PGM in Fig. 1,

$$V = \{\mathrm{C}_{i,t}, \alpha^{(m)}, \beta^{(m)}, v_t^{(m)}, \eta_t^{(m)}, z_{i,j}^{(m)}, \tilde{\theta}_{i,t}^{(m)}, f_{i,t}, \mu, K | \\ m = 1, 2, i = 1, ..., D, j = 1, ..., N^{(m)}, t = 1, ..., T\}, \tag{10}$$

where $N^{(m)}$ is the maximum number of words per document in mth modality. T is the truncation level, which means $q(\tilde{v}_T^{(m)} = 1) = 1$, and $\tilde{v}_T^{(m)}$ is defined as in Eq. (3). The joint probability distribution $p(V, X)$ can be obtained according to the generative process mentioned before. As for the variational distribution $q(V)$, it can be factorized according to the mean-field assumption,

$$q(V) = \prod_{m=1}^{M} \prod_{i=1}^{D} \prod_{j=1}^{N^{(m)}} \prod_{t=1}^{T} q(\alpha^{(m)}) q(\beta^{(m)}) q(v_t^{(m)}) \\ q(\eta_t^{(m)}) q(\mathrm{C}_{i,t}) q(z_{i,j}^{(m)}) q(\tilde{\theta}_{i,t}^{(m)}) q(f_i) q(\mu) q(K). \tag{11}$$

Note that here we exploit the variational distribution of f_i for approximation. And $q(f_{i,t}^{(m)})$ thus can be obtained straightforwardly via marginalization of $q(f_i)$, due to the construction in Eq. (4). Next, distributions in the right hand side of Eq. (11) are further defined as,

$$\begin{aligned} &q(\eta_t^{(m)}) \sim Dir(\pi_t^{(m)}), \\ &q(z_{i,j}^{(m)}) \sim Mult(\phi_{i,j}^{(m)}), \\ &q(f_i) \sim Normal(\tilde{\mu}_i, diag(\tilde{\sigma}_i)), \\ &q(C_{i,t}) \sim Gamma(\tilde{a}_{y_i,t}, \tilde{b}_{y_i,t}), \\ &q(\tilde{\theta}_{i,t}^{(m)}) \sim Gamma(\hat{a}_{i,t}^{(m)}, \hat{b}_{i,t}^{(m)}), \\ &q(\alpha^{(m)}) q(\beta^{(m)}) q(v_t^{(m)}) = \delta_{\alpha^{(m)}} \cdot \delta_{\beta^{(m)}} \cdot \delta_{v_t^{(m)}}, \\ &q(\mu) q(K) = \delta_\mu \cdot \delta_K, \end{aligned} \tag{12}$$

where $\delta(\cdot)$ is the Kronecker delta function and is used for tractability of inference as in [17]. Moreover, $Dir(\cdot)$ and $Normal(\cdot)$ are the Dirichlet distribution and multivariate normal distribution respectively. Here $\tilde{\sigma}_i$ is a vector and $q(f_i)$ thus has a diagonal covariance matrix.

Now we have set up the variational lower bound, which up to a constant is equivalent to the negative Kullback-Leibler (KL) divergence of the true posterior of latent variables $p(V|X)$ from the approximated variational distribution $q(V)$ [15]. We now derive a coordinate ascent algorithm for maximizing the lower bound by taking derivatives of Eq. (9) with respect to variational parameters. Since our variational inference algorithm is partly related to the one given by [23], here we only list the different updates. Specifically, the update equation or gradients of variational parameters are listed as below:

$$\begin{aligned} &\frac{\partial L}{\partial \tilde{\mu}_i} = \lambda_i - \gamma - K^{-1}(\tilde{\mu}_i - \mu) \\ &\frac{\partial L}{\partial \tilde{\sigma}_{i,t}} = -\frac{1}{2}(\lambda_{i,t} + K_{t,t}^{-1} - \frac{1}{\tilde{\sigma}_{i,t}}) \\ &\frac{\partial L}{\partial a_{i,t}} = \sum_{j=1}^{D} \delta(y_j = i)\{\log \frac{b_{y_j,t}}{\tilde{b}_{y_j,t}} + \psi(\tilde{a}_{y_j,t}) - \psi(a_{y_j,t})\} \\ &b_{i,t} = \frac{a_{i,t}\tilde{b}_{i,t}}{\tilde{a}_{i,t}} \\ &\frac{\partial L}{\partial \tilde{a}_{i,t}} = \sum_{j=1}^{D} \delta(y_j = i) \left\{1 - (\gamma_t + \tilde{a}_{y_j,t})\psi'(\tilde{a}_{y_j,t}) - \frac{b_{y_j,t}}{\tilde{b}_{y_j,t}} + \frac{\lambda_{y_j,t}}{\tilde{a}_{y_j,t}}\right\} \\ &\frac{\partial L}{\partial \tilde{b}_{i,t}} = \sum_{j=1}^{D} \delta(y_j = i) \left\{\frac{\gamma_t - \tilde{a}_{y_j,t} - \lambda_{y_j,t}}{\tilde{b}_{y_j,t}} + \frac{\tilde{a}_{y_j,t}}{\tilde{b}_{y_j,t}^2} b_{y_j,t}\right\}, \end{aligned} \tag{13}$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the digamma function and trigamma function respectively. λ_i is a vector which has the same size as f_i. And its tth element is given as,

$$\lambda_{i,t} = \exp(-\tilde{\mu}_{i,t} + \frac{1}{2}\tilde{\sigma}_{i,t}) \frac{\hat{a}_{y_i,t}\tilde{b}_{y_i,t}}{\hat{b}_{y_i,t}\tilde{a}_{y_i,t}}. \tag{14}$$

γ is a vector which is also of the same size as f_i and similarly constructed as in Eq. (4),

$$\gamma = [\beta^{(1)} v_1^{(1)}, \ldots, \beta^{(1)} v_T^{(1)}, \ldots, \beta^{(M)} v_1^{(M)}, \ldots, \beta^{(M)} v_T^{(M)}]. \tag{15}$$

Based on above derivations, we thus are able to perform efficient gradient-based techniques, e.g. Newton-Raphson, stochastic gradient descent, to find the optimized variational parameters.

To update the parameters of GP, we maximize the marginal likehood, which provides equations of μ and K:

$$\begin{aligned} \mu &= \frac{1}{D}\sum_{i=1}^{D} \tilde{\mu}_i \\ K &= \frac{1}{D}\sum_{i=1}^{D} \{(\tilde{\mu}_i - \mu)(\tilde{\mu}_i - \mu)^T + diag(\tilde{\sigma}_i)\}, \end{aligned} \tag{16}$$

We update the kernel matrix directly rather than the kernel function, since the latter one will increase the computational burden by inferring latent locations ℓ which are as many as latent topics of both image and text modality. Moreover, in this way, low rank approximations of the gram matrix K, like Nystrom approximation [25], could be applied for speeding up the inference.

4.2 Prediction

Prediction tasks under the upstream supervised multi-modal setting usually involve two objectives, (i) given testing multi-modal samples $\{\bar{x}_j^{(m)} | m = 1, 2, j = 1, ..., \mathbb{N}\}$, predicting their corresponding response variables $\{\bar{y}_j\}$; (ii) given testing samples of one modality $\{\bar{x}_j^{(1)}\}$, predicting samples of the missing modality $\{\bar{x}_j^{(2)}\}$ and the corresponding response variable $\{\bar{y}_j\}$. Here we focus on achieving the second objective, since (ii) is generally harder than (i) and is more closely related to the scenarios of cross-modal retrieval and image annotation.

Specifically, the aim of (ii) is to obtain the predictive posterior distribution $p(\bar{y}_j, \bar{x}_j^{(2)}|\bar{x}_j^{(1)}, X)$, where X is the aforementioned set of training dataset. Relying on the Bayes theorem, the desired probability could be expressed as,

$$p(\bar{y}_j, \bar{x}_j^{(2)}|\bar{x}_j^{(1)}, X) = \int p(\bar{x}_j^{(2)}|\bar{x}_j^{(1)}, \bar{y}_j, V)p(\bar{y}_j)dp(V|X), \tag{17}$$

where V is the set of all latent variables as in Sec. 4.1 and $p(\bar{y}_j)$ is the prior of response variable. Note that, in the training process, response variable is observable, whereas it is a latent variable to be inferred during testing. In our experiments, we just use the empirical distribution of y in the training dataset. Analogous to the empirical approximation of MCMC sampling, we calculate the predictive distribution using variational posterior as,

$$p(\bar{y}_j, \bar{x}_j^{(2)}|\bar{x}_j^{(1)}, X) \approx E_{q^*}[p(\bar{x}_j^{(2)}|\bar{x}_j^{(1)}, \bar{y}_j, V)]p(\bar{y}_j), \tag{18}$$

where $q^*(V)$ stands for the approximated variational distribution inferred during training process according to Sec. 4.1.

To compute $E_{q^*}[p(\bar{x}_j^{(2)}|\bar{x}_j^{(1)}, \bar{y}_j, V)]$, we could first infer the topic mixing proportion vector $\theta_j^{(1)}$ and the vector of auxiliary latent variables $f_j^{(1)}$ through maximizing a lower bound as in Eq. (9). Then, based on the conditional expectation of multivariate Gaussian distribution, we can obtain the auxiliary variables for the missing modality as,

$$f_j^{(2)} = \mu^{(2)} + K^{(2,1)}(K^{(1,1)})^{-1}(f_j^{(1)} - \mu^{(1)}), \tag{19}$$

where $\mu^{(1)}$ and $\mu^{(2)}$ are factorized from μ according to the construction in Eq. (4), and superscripts of K denote the row and column indexes of its block sub-matrices as in Sec. 3.3. With $f_j^{(2)}$ at hand, we now are capable to calculate the expectation.

Further more, if we are only interested in estimating $\bar{y}_j$, e.g. in the situation of cross-modal retrieval, we can integrate $\bar{x}_j^{(2)}$ from the predictive posterior distribution Eq. (18) which provides the maximum a posterior (MAP) estimation. When the prior for $\bar{y}_j$ is a uniform distribution, MAP is equivalent to the maximum likelihood estimation which is what common upstream models [11, 41] do. Therefore, we could make our prediction algorithm flexible and powerful by imposing a proper application dependent prior for response variable. In turn, $\bar{x}_j^{(2)}$ could be obtained in a similar manner for situations where $\bar{y}_j$ is less cared, like image annotation.

5. EXPERIMENTS

In this section, we present in detail the experiments for evaluating the performance of our NPBUS model in capturing correlation structures of multi-modal data and its predictive capacity.

5.1 Data & Experimental Settings

We evaluate the performance of our model on two predictive tasks—cross-modal retrieval and image annotation. For the first one, we use the public wiki dataset[1] contributed by N. Rasiwasia et al. [24]. The dataset consists of 2,866 image-text pairs which are collected from 2,700 articles selected and reviewed by Wikipedia, of which 2,173 pairs are randomly chosen to be the training set and the other 693 image-text pairs are chosen to be the test set. Moreover, a response variable of category label is offered for each training pair. These labels cover 10 semantic categories, like art, biology, music and so forth. For the experiments on image annotation we use the public Corel5K [10] as the benchmark dataset, which contains around 5,000 images that are only accompanied by 1 to 5 annotations of keywords. We use a fixed set of 499 images for testing and the rest for training, following the setup in [14]. Since there is no human labeled category information for training data, we cluster the tags through hierarchical Dirichlet process (HDP) topic model and automatically find 21 clusterings in total. Then the obtained clustering labels are regarded as response variables. Note that tags for each image are represented as a vector of words distribution on the vocabulary and response variable adopts 1-of-W vector representation. Besides, all visual features and ground-truth annotated text tags are available on the web[2].

5.2 Correlation Structure

We first investigate how our NPBUS model chooses topics and how topics learned by our model are correlated on the wiki dataset. First, we demonstrate the stick proportions $v_t^{(m)}$ obtained through inference. We notice that, in Fig. 2, the stick proportions of image vary differently with ones of text, which motivates our incorporation of separate HDPs for the modalities. Also, from Fig. 2, we see that the last stick proportions are close to zero which justifies the truncation level $T = 100$ is sufficient.

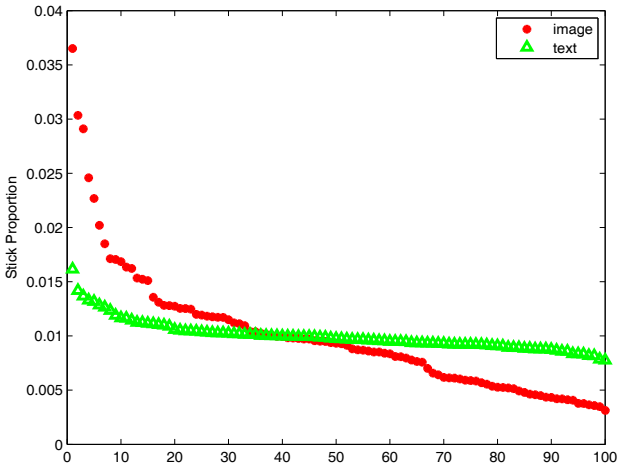

Figure 2: Visualization of stick proportions learnt from our model for image modality (red ball) and text modality (green triangle).

Next, we turn to illustrate the learned correlation structures of topics. For saving space, we only show the correlation matrix between topics of image modality and topics of text modality in Fig. 3. We can see that some correlation values are negative which may suggest absence of specific topics from the other modality. Moreover, only a few topics of image are strongly correlated with topics of text. This is as expected, since every pair of image and text in this wiki dataset is extracted owning to their close positions on the webpage which does not imply close similarity of their contents. To further inspect whether the correlation learned by

[1] http://www.svcl.ucsd.edu/projects/crossmodal/

[2] http://lear.inrialpes.fr/people/guillaumin/data.php

our model matches our own intuition, we are supposed to present the content of a pair of correlated topics. However, it is not intuitive for visualize the topics of image, since they are distributions over SIFT vocabulary. In order to overcome this problem, we adopt the same visual relevance measurement ρ as in [30]. For each topic of text modality, ρ is defined to be the mean value of absolute correlations with all image topics. Note that, this measurement captures the co-occurrences of text topics and combined image topics. Then we can rank topics according to its value of visual relevance. In Table. 1, ranked topics and their text contents are summarized. We see that topics with strong visual relevance, i.e., large values of ρ, contain clear and concrete visual counterparts. For example, the first topic is about birds and their living environments, which are easily depicted by pictures. Nevertheless, topics with weak visual relevance are difficult to be described by visual contents, like the fifth topic in the last row. Therefore, our NPBUS model does capture the correlation structures of multi-modal data.

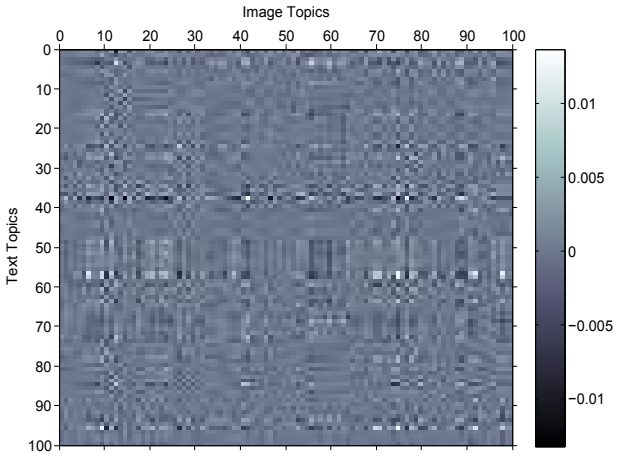

Figure 3: Visualization of correlation structure between topics of image modality (columns) and topics of text modality (rows).

5.3 Cross-Modal Retrieval

Table 2: MAP scores of cross-modal retrieval

Method	Image Query	Text Query	Average
SCM [24]	0.277	0.226	0.252
GMA [28]	0.272	0.232	0.253
CMTC [38]	0.293	0.232	0.266
MLBE [39]	0.381	0.496	0.439
NPBUS	**0.408**	**0.544**	**0.476**

We then conduct cross-modal retrieval which contains two sub-tasks—text retrieval through image queries and image retrieval through text queries. We first execute our variational inference algorithm for the training dataset. Then, given a query in one modality, we predict its corresponding response variable y by the prediction algorithm mentioned in Sec. 4.2. Finally, we use the probability vector $p(y)$ to calculate the ranking of all retrieved data. More specifically, the response variables of training data are encoded in 1-of-$\mathbb{K}$ vector representation and here $\mathbb{K}$ equals 10

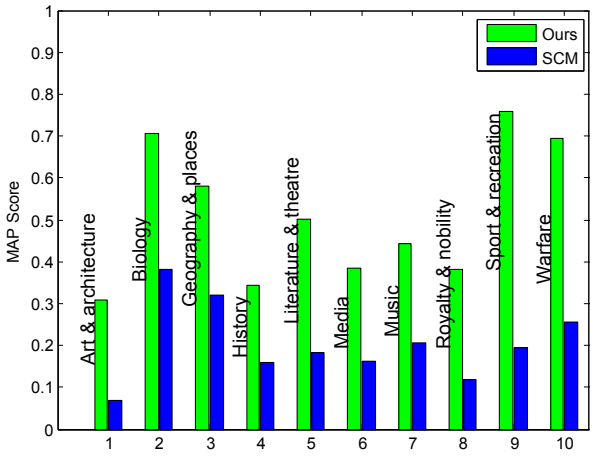

Figure 4: MAP scores per category with text query.

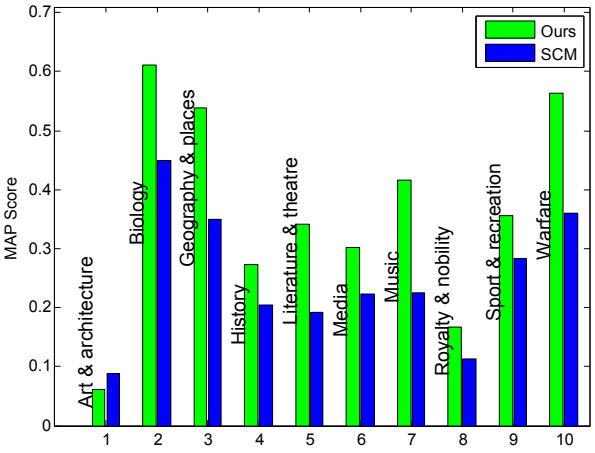

Figure 5: MAP scores per category with image query.

since there are 10 categories. Therefore, the distance between two response variables is established as distance of two vectors in vector space. Note that we experimented with different distance measures, including L_1, L_2, normalized correlation (NC) and chi-squared χ^2 distance, and we found that the L_1 distance outperforms all others. To make the comparison fair, we take the same feature settings as in [24], that is, we use topic mixing proportions to represent text documents and use bag of SIFT features to represent images. In the training stage, we set the initial values of two concentration parameters in HDP as: $\alpha = 15$ and $\beta = 5$, and the truncation level as $T = 100$ for each modality. 34 and 66 topics are learnt automatically for training texts and images, respectively. We compare our NPBUS model with semantic correlation matching (SCM) [24], generalized multi-view analysis (GMA) [28], cross-modal topic correlations (CMTC) [38] and multi-modal latent binary embedding (MLBE) [39]. The performance is measured with mean average precision (MAP) which has good discrimination and stability and is widely used in the literature of information retrieval [9].

The overall MAP scores are reported in Table 2. It is clear from this table that our NPBUS outperforms other models

Topics with top-5 visual relevances
population species north birds forest area males females found breeding trees year largest high conservation
film music production movie scene play musical studio role american sound hollywood pictures performance
ship navy fleet naval guns war sea british battleship royal squadron world hitler enemy iowa coast aircraft
king family george prince life wife duke earl father died royal princess marriage lord son daughter married
book works published novel story life writing critics stories popular characters literary history author volume
Topics with least-5 visual relevances
up base out hit second single two followed third cardinals left home inning pearl fly field down yankees
through wall sculpture neilston placed although models agricultural model turrets flight entire having computer
gold richard california order bangladesh nails amp due entire jews raised hotel videos 1976 far observer hours
canadian adams states united tournament international revolutionary gold winter world people old historical
people female eggs number american african minnesota many census 2001 chicks days rate around home life

Table 1: Visualization of topic contents. Each row of words indicates a topic. From top row to bottom row, the corresponding visual relevances are decreasing. Words from left to right in each topic are ranked in descending order according to their probabilities.

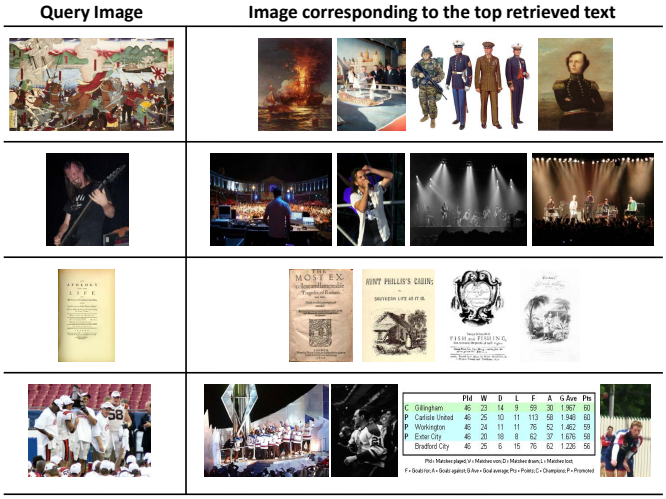

Figure 6: Examples of cross-modal retrieval with image query. Images corresponding to retrieved texts with top-4 rank scores are shown.

in both sub-tasks on the wiki dataset. This improvement of performance validates the benefits of incorporating supervising information of semantic abstracts. Note that the boost of performance is larger in sub-task with text query than the one with image query. It is perhaps because that the SIFT feature is not expressive enough for generally describing visual objects with semantic meaning. Since only SCM published their per category MAP scores, we compare the histograms of both text query and image query with SCM in Fig. 4 and Fig. 5 respectively. And our NPBUS achieves better results for almost all categories. More instances of cross-modal retrieval are demonstrated in Fig. 6 and Fig. 7. We consider the top-4 ranked retrieved texts for image queries in Fig. 6. And in Fig. 7, we show the top-5 ranked images for text queries. Note that, for the ease of display, we only show the images corresponding to the retrieved texts in Fig. 6 and contents of text queries have been fragmented in Fig. 7.

5.4 Image Annotation

In experiments of image annotation, variational inference is first conducted to obtain the approximated variational distribution for training dataset. For testing images, we compute the desired conditional probability of tags given response variables and image feature according to the prediction algorithm. Following the aforementioned convention, captions with first 5 highest conditional probabilities are drawn as the final annotation results. As for image features, we exploit the same types of ones as in [14], and normalize them by their L_1 norm separately before the combination step. Since the overall features they used are of more than 30,000 dimensions, we reduce the dimension to 500 by principle component analysis (PCA) for computational efficiency. Moreover, 3 standard measures are adopted as in [6]: the mean precision per word (i.e., P), the mean recall per word (i.e., R) and number of keywords with non -zero recall value (i.e., $N+$). We compare our NPBUS with 10 other methods which publicly report their results on this dataset. The experimental comparisons are listed in Table 3. From this table, we can find that NPBUS ranks 2nd with respect to P and achieves the best results in terms of R and $N+$. The causes for less precision than the state-of-art may be the reduced-dimension visual features used in our experiments. Also, since some training images have too few attached tags, the topic model may not perform better compared to some simple model, like logistic regression in [14]. Finally, the supervising information obtained via unsupervised HDP clustering is somewhat limited.

Table 3: Comparison of performance for image annotation

Methods	P	R	$N+$
CorrLDA [2]	6	9	59
CRM [16]	16	19	107
InfNet [21]	17	24	112
NPDE [37]	18	21	114
SML [6]	23	29	137
MBRM [12]	24	25	122
TGLM [18]	25	29	131
MSC [32]	25	32	136
JEC [20]	27	32	139
TagProp [14]	**33**	42	160
NPBUS	29	**44**	**187**

On 31 January, the effort to retake the city began anew. The attack was launched at 08:30 hours, and was met by inaccurate Iraqi fire which knocked-out two Saudi V-150 wheeled vehicles.Stanton, p. 9, claims that two vehicles were destroyed, while Westermeyer, p. 31, claims that three were knocked-out. The 8th battalion of the Saudi brigade was ordered to deploy to the city by 10:00 hours, while 5th Battalion to the north engaged another column of Iraqi tanks attempting to reach the city. The latter engagement led to the destruction of around 13 Iraqi tanks and armored personnel carriers, and the capture of 6 more vehicles and 116 Iraqi soldiers, costing the Saudi battalion two dead and two wounded……

"Honoured members: the Hockey Hall of Fame", p. 91. On March 30, 1993, it was announced that Gil Stein, who at the time was the president of the National Hockey League, would be inducted into the Hall of Fame. There were immediate allegations that he had engineered his election through manipulation of the hall's board of directors. Due to these allegations, NHL commissioner Gary Bettman hired two independent lawyers, Arnold Burns and Yves Fortier, to lead an investigation. They concluded that Stein had "improperly manipulated the process" and "created the false appearance and illusion" that his nomination was the idea of Bruce McNall……

Figure 7: Two examples of cross-modal retrieval with text query. Left parts are fragmented queries and right parts are corresponding retrieved images with top-5 rank scores. The content of the top text describes a war and the below one is about hockey.

6. CONCLUSIONS

In this paper, we have presented a nonparametric Bayesian upstream supervised (NPBUS) multi-modal topic model. Our NPBUS model allows flexible learning of correlation structures of topics within individual modality and between different modalities. And it becomes more discriminative via incorporating upstream supervising information shared by multi-modal data. Last, it is capable to automatically determine the number of latent topics in each modality. We also devise efficient variational inference and prediction algorithms. Extensive experiments demonstrate the above advantages in terms of cross-modal retrieval and image annotation.

In future work, we intend to develop truncation free algorithms to improve our approximated inference, or utilize sampling methods, like MCMC. And we also will exploit low rank approximations to accelerate the kernel learning. Moreover, we will apply our model for mining more different kinds of multi-modal data, e.g., multi-lingual webpages, multi-source perception data of robotics.

7. ACKNOWLEDGMENTS

Jun Zhu is supported by the National Basic Research Program (973 Program) of China (Nos. 2013CB329403, 2012CB316301), and National Natural Science Foundation of China No. 61322308. Zengchang Qin is supported by the National Science Foundation of China No. 6130504.

8. REFERENCES

[1] D. Blackwell and J. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[2] D. Blei and M. Jordan. Modeling annotated data. In *ACM SIGIR*, pages 127–134, 2003.

[3] D. Blei and J. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[5] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[6] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410, 2007.

[7] N. Chen, J. Zhu, F. Sun, and E. Xing. Large-margin predictive latent subspace learning for multiview data

analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(12):2365–2378, 2012.

[8] N. Chen, J. Zhu, and E. Xing. Predictive subspace learning for multi-view data: A large margin approach. In *NIPS*, 2010.

[9] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

[10] P. Duygulu, K. Barnard, J. De Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, pages 349–354. Springer, 2002.

[11] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, pages 524–531. IEEE, 2005.

[12] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *CVPR*, volume 2, pages II–1002. IEEE, 2004.

[13] T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

[14] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, pages 309–316. IEEE, 2009.

[15] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[16] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[17] P. Liang, S. Petrov, M. I. Jordan, and D. Klein. The infinite PCFG using hierarchical Dirichlet processes. In *EMNLP/CoNLL*, 2007.

[18] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[20] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, volume 8, pages 316–329, 2008.

[21] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *CIVR*, 2004.

[22] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1802–1817, 2007.

[23] J. Paisley, C. Wang, and D. M. Blei. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(4):997–1034, 2012.

[24] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, pages 251–260, 2010.

[25] C. Rasmussen and C. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

[26] K. Salomatin, Y. Yang, and A. Lad. Multi-field correlated topic modeling. *SIAM SDM*, 2009.

[27] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[28] A. Sharma, A. Kumar, H. Daume, and D. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, pages 2160–2167. IEEE, 2012.

[29] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[30] S. Virtanen, Y. Jia, A. Klami, and T. Darrell. Factorized multi-modal topic model. In *UAI*, 2012.

[31] S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In *ICML*, 2011.

[32] C. Wang, S. Yan, L. Zhang, and H. Zhang. Multi-label sparse coding for automatic image annotation. In *CVPR*, pages 1643–1650. IEEE, 2009.

[33] P. Wu, S. C.-H. Hoi, P. Zhao, and Y. He. Mining social images with distance metric learning for automated image tagging. In *WSDM*, pages 197–206. ACM, 2011.

[34] H. Xia, P. Wu, and S. C. Hoi. Online multi-modal distance learning for scalable multimedia retrieval. In *WSDM*, pages 455–464. ACM, 2013.

[35] E. Xing, R. Yan, and A. Hauptmann. Mining associated text and images with dual-wing harmoniums. In *UAI*, 2005.

[36] O. Yakhnenko and V. Honavar. Multi-modal hierarchical Dirichlet process model for predicting image annotation and image-object label correspondence. In *SIAM SDM*, 2009.

[37] A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *CIVR*, 2005.

[38] J. Yu, Y. Cong, Z. Qin, and T. Wan. Cross-modal topic correlations for multimedia retrieval. In *ICPR*, 2012.

[39] Y. Zhen and D. Yeung. A probabilistic model for multimodal hash function learning. In *ACM SIGKDD*, 2012.

[40] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research*, 13:2237–2278, 2012.

[41] J. Zhu, L.-J. Li, L. Fei-Fei, and E. P. Xing. Large margin learning of upstream scene understanding models. In *NIPS*, pages 2586–2594, 2010.