Multi-Modal Hierarchical Dirichlet Process Model for Predicting Image Annotation and Image-Object Label Correspondence

Oksana Yakhnenko and Vasant Honavar
{oksayakh, honavar}@cs.iastate.edu
Iowa State University
Ames, IA, 50010

Abstract

Many real-world applications call for learning predictive relationships from multi-modal data. In particular, in multi-media and web applications, given a dataset of images and their associated captions, one might want to construct a predictive model that not only predicts a caption for the image but also labels the individual objects in the image. We address this problem using a *multi-modal hierarchical Dirichlet Process* model (MoM-HDP) - a stochastic process for modeling multi-modal data. MoM-HDP is an analog of a multi-modal Latent Dirichlet Allocation (MoM-LDA) with an infinite number of mixture components. Thus MoM-HDP allows circumventing the need for a priori choice of the number of mixture components or the computational expense of model selection. During training, the model has access to an un-segmented image and its caption, but not the labels for each object in the image. The trained model is used to predict the label for each region of interest in a segmented image. The model parameters are estimated efficiently using variational inference. We use two large benchmark datasets to compare the performance of the proposed MoM-HDP model with that of MoM-LDA model as well as some simple alternatives: Naive Bayes and Logistic Regression classifiers based on the formulation of the image annotation and image-label correspondence problems as one-against-all classification. Our experimental results show that unlike MoM-LDA, the performance of MoM-HDP is invariant to the number of mixture components. Furthermore, our experimental evaluation shows that the generalization performance of MoM-HDP is superior to that of MoM-HDP as well as the one-against-all Naive Bayes and Logistic Regression classifiers.

1 Introduction

Recent years have witnessed rapid advances in our ability to acquire and store massive amounts of data across different modalities (text, speech, images, etc.). The growth in the quantity of disparate types of data has far outstripped our ability to organize, analyze and extract useful knowledge from such data. *Multi-modal data mining* refers to the process of constructing predictive models from data that spans multiple modalities. Multi-modal data mining presents several challenges that are largely beyond the current state of the art in data mining. For example, many web data sources such as social network communities (i.e. Flickr, Facebook, etc) offer an abundant source of images with their associated captions i.e., words that describe the image content without specifically labeling the individual objects in the image. Newspaper articles or other types of digital media contain pictures of the events and the description of the same events in text; or research articles contain figures and captions which describe the figures. In many application scenarios [5], it is not enough to predict whether or not a particular object appears in the image; it is necessary to be able to label individual objects in the image.

Learning the relationships between image regions and words is an interesting example of multi-modal data mining. The available training data do not provide explicit labels for individual objects in an image. As more and more data becomes available, human annotation and labeling becomes prohibitively time consuming and expensive. This is especially true in the case of data that is derived from more than one modality (e.g., text and images; sound and images). More importantly, straightforward reductions of multi-modal data mining problems to standard supervised classification problems often fail to fully exploit the natural correlations that might exist among the basic entities within each modality and across modalities.

Given the expense of obtaining training datasets of images wherein each object in the image is labeled by a human annotator, there is a need for methods that can, given a dataset of images and their associated captions, learn to label individual objects in an image.

Against this background, this paper focuses on the following problem: Given a dataset of images and their associated captions, can we build a model that not only predicts a collection of labels for an entire image (the image *annotation* problem), but specifically labels the individual objects (or regions of interest) in the image (the image object-label *correspondence* problem)? Consequently, there is a growing interest in developing principled solutions to the image annotation problem and the image object-label correspondence problems [2, 5, 6, 16](see section 4 for details).

We describe an approach to solving the image annotation and image correspondence problems using a *multi-modal hierarchical Dirichlet Process* (MoM-HDP) model which is a natural generalization of the multi-modal latent Dirichlet Allocation model (MoM-LDA) [5]. Latent Dirichlet Allocation (LDA) is a generative probabilistic model for independent collections of data where each collection is modeled by a randomly generated mixture over latent factors. In topic modeling for text documents LDA assumes the following generative process: each document has its own distribution of topics, and given a specific topic, the words are generated. MoM-LDA is a generalization of LDA where the documents contain multiple types (modalities) of entities such as words, image regions (also called blobs). MoM-LDA describes the following generative process for the data: each document (consisting of both words and pictures) has a distribution for a fixed number of mixture components (topics), and given a specific mixture component the words and the image features are generated. However, selecting the number of mixture components to be used in a MoM-LDA model is difficult. In practice, several different MoM-LDA models corresponding to different choices of the number of mixture components are trained and evaluated using cross-validation and the best performing model is chosen.

The proposed MoM-HDP model is based on the hierarchical Dirichlet Process [25], a *stochastic process* that can be thought of as the analog of a mixture model, but with an *infinite* number of mixture components assumed in a mixture model. MoM-HDP thus allows us to circumvent the need for a priori (and hence potentially arbitrary) choice of the number of mixture components or the computational expense of training multiple MoM-LDA models before choosing one based on the results of cross-validation. Note however that in practice, the Dirichlet process is approximated by *truncating* it [14].

We compare the performance of the proposed MoM-HDP model with that of MoM-LDA model on the image annotation and image-label correspondence task on a dataset with variety of labels and objects using two datasets which provide the ground truth needed in order to evaluate the performance of the two approaches: Visual Object Classes (VOC) 2007 challenge data which has 20 possible labels and a subset of LabelMe, which has over 1700 possible labels. We also compare MoM-HDP and MoM-LDA with some simple alternatives: Naive Bayes and Logistic Regression classifiers based on the formulation of the image annotation and image-label correspondence problems as one-against-all classification problems. Our results show that the generalization performance of MoM-HDP is superior to that of MoM-HDP as well as the Naive Bayes and Logistic Regression classifiers. The results of our experiments show that the generalization performance of the MoM-LDA model is sensitive to the choice of the number of components that are assumed to exist in the mixture. In contrast, the performance of the MoM-HDP model is relatively insensitive to the specific choice of the cutoff used to truncate the Dirichlet Process.

Thus, the main contributions of this paper are:

- Development of MoM-HDP, a HDP counterpart of MoM-LDA model for solving image annotation and image object-label correspondence problems under fairly general assumptions that circumvents the need for a priori (and hence potentially arbitrary) choice of the number of mixture components or the computational expense of training and evaluating multiple MoM-LDA models before choosing one based on the results of cross-validation.

- Experimental results that demonstrate that modeling the problem directly using MoM-LDA and MoM-HDP produces a better performance than one-against-all-learning scenario and that MoM-HDP outperforms MoM-LDA on image annotation and image-object label correspondence problems.

This paper is organized as follows: We briefly describe the MoM-LDA model and generalize it to a Dirichlet process in Section 2. We describe the dataset, experimental setup, evaluation procedure, and the results of our comparison of MoM-LDA and MoM-HDP models in Section 3. We conclude the paper with related work in Section 4 and a summary and a brief discussion of some directions for further research in Section 5.

2 Multi-modal Hierarchical Dirichlet Process model

We begin with describing Latent Dirichlet Allocation and multi-modal Dirichlet allocation, review the main principles behind the Dirichlet Processes and introduce our multi-modal hierarchical Dirichlet Process model.

2.1 Notation Let W be the vocabulary of all the possible words in the captions, and $\mathbf{w_i} = \{w_{1_i}...w_{N_i}\}, w_{j_i} \in W$ be the caption for image i. Let B be the vocabulary of all the possible visual words in the pictures, and $\mathbf{b_i} = \{b_{1_i}...b_{M_i}\}, b_{j_i} \in B$ be the "visual word" representation of the image. Let $\mathcal{D} = (\mathbf{w_i}, \mathbf{b_i})_{i=1}^{D}$ be the corpus of D images so that for each image the set of caption keywords is known.

2.2 Latent Dirichlet Allocation model for images and captions (MoM-LDA) We first describe a multi-modal Latent Dirichlet Allocation model (MoM-LDA) introduced by Blei and Jordan [5], and then generalize this model using a hierarchical Dirichlet Process (MoM-HDP). Informally, the following generative process is assumed for images and captions. The image topic (e.g. horseback riding) generates a distribution for intermediate level components (e.g. horse, person, grass, fence, sky, sun, building) and the intermediate level components generate specific words and image regions observed in the training data (e.g. the words "horse" and "person", and the image regions which correspond to horse's eyes, ears, person's face, arms and legs, etc). MoM-LDA assumes a pre-defined number of clusters which group the related entities in the modalities, and it groups the related visual words and the related words in the same clusters. In addition, the probability distribution of the clusters is different for each image-caption pair, which is achieved by introducing a Dirichlet prior for the distribution of clusters. Formally, the images and captions are described by the following generative process: For each image i, pick a distribution of topics $\pi_i \sim \text{Dirichlet}(\alpha)$. For each caption word j, pick a latent factor $t_{ij} \sim Mult(\pi_i)$ and then pick the word $w_{ij} \sim F(t_{ij})$. Similarly for each image feature j, pick a latent factor $s_{ij} \sim Mult(\pi_i)$ and then pick the feature $b_{ij} \sim F(s_{ij})$. The graphical model for this process is shown in Figure 1. Here $F(x)$ can be any appropriate distribution, such as Multinomial for words and discrete features, or Gaussian for continuous features. In our model and in our experiments, we use discrete-valued image features (visual words). Hence, we focus our discussion on the MoM-HDP model based on the multinomial distribution. However, the model described in this paper can be easily extended to other distributions.

2.3 Dirichlet Process A limitation of mixture models is that the need to specify a number of components to include in the mixture (namely K). The choice of number of the mixture components can have a major influence on how well the model fits the data, and its ability to generalize beyond the training data. Hence, we consider a model based on a hierarchical Dirichlet Process (HDP) [25], with *countably infinite* number of mixture components. For details on the HDP and their applications in probabilistic graphical models we refer the reader to [1], [25] or [4] and we only summarize the key aspects of DP, and then HDP in this paper.

The Dirichlet Process (DP) is a generalization of a finite mixture model, and it assumes countably infinite number mixture components. Unlike in the finite mixture models where the priors for the mixture components are assumed to be drawn from some distribution, the DP assumes that the priors are created according to some stochastic process.

DP is parametrized by a base distribution G_0 and a scaling parameter α and is denoted by $DP(\alpha, G_0)$. Let $z = \{z_1, z_2...\}$ be the mixture components, and let $X_1...X_N$ be a sample from the DP mixture. Then we can assume the following generative process for the data: draw mixture priors $\beta \sim DP(\alpha, G_0)$. For each mixture component $z = \{z_1, z_2...\}$ draw parameters $\phi_z \sim G_0$ which specify the distribution for the observations X. For each instance $i = 1...N$ draw parameters π_i which specify the distribution of the mixture components, draw a mixture component $z_i \sim Mult(\pi_i)$, and from the mixture component z_i draw $X_i \sim \phi_{z_i}$.

The two common approaches to constructing the DP are Chinese Restaurant Process [15], and stick-breaking construction [22]. In our work, we consider the latter. Intuitively, stick-breaking construction can be described as follows: the prior β is generated by taking a stick of length 1, and breaking off segments of the stick proportional to the remaining stick.

We use $\beta \sim \text{GEM}(\alpha)$ to denote that $\beta = (\beta_1, \beta_2...)$ is generated according to the stick-breaking distribution. Let $u_1, u_2...$ be countably infinite proportions that are generated according to the beta distribution. Then the weights β are defined in terms of $\beta_z = u_z \prod_{z'<z}(1 - u_{z'})$. Such construction ensures that β is countably infinite with each component drawn i.i.d.

2.4 Hierarchical Dirichlet Process We described a simple Dirichlet Process which we will use as a basis for a more complicated model. DP assumes one model with infinitely many mixture components for all documents, and it is a non-parametric equivalent of the probabilistic latent semantic analysis (p-LSA). We would like a learning algorithm which creates a model for each document (just like LDA) and therefore we assume a hierarchical Dirichlet Process (HDP) to provide a non-parametric generalization of the LDA model [25]. HDP assumes a separate generative model for each document $j = 1...J$, and that each model shares a collection of the mixture components. Each model provides a probabil-

ity distribution for the mixture components for a document i (π_i), and these distributions are tied between the models via the prior β.

2.5 Hierarchical Dirichlet Process multi-modal model (MoM-HDP) We now apply the stick-breaking construction of the priors for the hierarchical Dirichlet Process to the multi-modal generative model. Like in the case of MoM-LDA, we assume that each observable modality is clustered by the mixture components, so that each word w is generated by a cluster t, each image component b is generated by a cluster s. The clusters for image-caption pair $\mathbf{w}_i$, $\mathbf{b}_i$ have multinomial distribution parametrized by π_i ($p(s_i) = p(t_i) = \pi_i$) drawn from $DP(\alpha^\pi, \beta)$ were $\beta \sim GEM(\alpha)$ is constructed using a stick-breaking distribution. Furthermore, the parameters for observations given their clusters $\phi_t^w = p(w|t)$ and $\phi_s^b = p(b|s)$ are generated from some base distribution G_0 (such as a Dirichlet distribution).

We show MoM-LDA and MoM-HDP in graphical notation in Figure 1. We also note that if the prior β is assumed to be drawn from finite Dirichlet instead of a stick-breaking distribution, this model becomes a Dirichlet-smoothed version of the MoM-LDA.

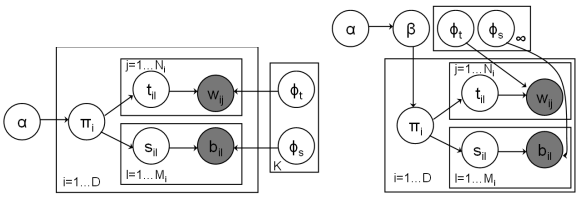

Figure 1: MoM-LDA model (left). Its MoM-HDP counterpart (right).

We summarize the generative processes modeled by MoM-HDP and MoM-LDA below.

MoM-HDP	MoM-LDA
draw $\beta \sim$ GEM(α)	chose priors ($\alpha_1 ... \alpha_K$)
for each $z = 1, 2...$	for each $z = 1, ..., K$
draw $\phi_z^w \sim$ Dirichlet(α_w)	draw $\phi_z^w \sim G_0$
draw $\phi_z^b \sim$ Dirichlet(α_b)	draw $\phi_z^b \sim G_0$
for each image $i = 1, ..., D$	for each image $i = 1, ..., D$
draw $\pi_i \sim \mathbf{DP}(\alpha^\pi, \beta)$	draw $\pi_i \sim \mathbf{Dir}(\alpha_1 ... \alpha_K)$
for each word $j = 1...N_i$	for each word $j = 1...N_i$
draw $t_{ij} \sim$ Mult(π_i)	draw $t_{ij} \sim$ Mult(π_i)
draw $w_{ij} \sim$ Mult($\phi_{t_{ij}}^w$)	draw $w_{ij} \sim$ Mult($\phi_{t_{ij}}^w$)
for each word $j = 1, ..., M_i$	for each word $j = 1, ..., M_i$
draw $s_{ij} \sim$ Mult(π_i)	draw $s_{ij} \sim$ Mult(π_i)
draw $b_{ij} \sim$ Mult($\phi_{s_{ij}}^b$)	draw $b_{ij} \sim$ Mult($\phi_{s_{ij}}^b$)

To make the parameter estimation feasible, we assume a truncated DP [14], and truncate β at K, so that $\beta_z = 0$ for all $z > K$. In this case, $\pi_i \sim \text{DP}(\alpha^\pi, \beta)$ simply becomes $\pi_i \sim \text{Dirichlet}(\alpha^\pi, \beta_1 ... \beta_K)$. While the model has infinite number of states, the density of the process is determined by the first several states, and as the cut-off K increases, the approximation improves.

Next we describe the parameter estimation procedure for the hierarchical Dirichlet Process model using variational inference.

2.6 Parameter Estimation via Variational Inference Let θ be the model parameters and z be all the hidden variables and x be the observations. The goal of fully Bayesian inference is to estimate parameters θ which maximize the probability $p(\theta, z|x)$. Such estimation puts hidden variables and the model parameters on equal footing. Because the exact inference is intractable we use variational inference. The probability $p(\theta, z|x)$ can be approximated by some distribution $q^*(\theta, z)$, such that

$$q^*(\theta, z) = \text{argmin}_{q \in \mathcal{Q}} KL(q(\theta, z)||p(\theta, z|x)$$

where $\mathcal{Q}$ is a tractable subset of distributions. In particular, if $\mathcal{Q}$ is a fully factorized distribution, then each of the factors will have a closed form solution which depends on other factors, and the solution which minimizes the original problem is obtained in the iterative fashion, similar to the expectation maximization procedure.

Using mean-field approximation we get the following property: if $q(\theta) = \prod_{i=1}^n q_i(\theta_i)$ is a factorized distribution for each of the factors θ_i, then the solution for $q_i(\theta_i)$ has the form $q_i(\theta_i) \propto \exp(\mathbb{E}_{q-i} \log p(\theta_i|\theta_{-i}))$ where θ_{-i} is a set of all the factors which are not θ_i (see [3] for details). Thus the model parameters are fitted using an iterative procedure: fix all factors but θ_i, solve for $q_i(\theta_i) \propto \exp(\mathbb{E}_{q-i} \log p(\theta_i|\theta_{-i}))$, and move on to the next factor θ_{i+1}.

Define

$$\begin{aligned}\mathcal{Q} =& q(\beta, \pi, \mathbf{s}, \mathbf{t}, \phi_\mathbf{s}, \phi_\mathbf{t}) \\ =& q(\beta)q(\pi)\prod_{i=1}^{M} q(\mathbf{s})\prod_{i=1}^{N} q(\mathbf{t})\prod_{z=1}^{K}\left(q(\phi_\mathbf{z}^\mathbf{b})q(\phi_\mathbf{z}{}^w)\right)\end{aligned}$$

where $q(\beta) \sim$ GEM(α) is drawn from the stick-breaking distribution, $q(\pi) \sim \text{DP}(\alpha_\pi, \beta)$ is drawn from the Dirichlet Process, $q(\phi_z)$'s are drawn from the Dirichlet distributions, and $q(s), q(t)$ are Multinomial.

The variational mean-field for the hierarchical Dirichlet Process can be viewed as a three-step process: the expectation step involves optimizing hidden multinomial factors $q(s)$ and $q(t)$ (equivalent E-step in

the EM). The maximization step involves parameter estimation to optimize $q(\phi)$ and $q(\pi)$ (equivalent to the M-step in the EM). The last step is optimizing the top-level distribution $q(\beta)$ (this step has no counterpart in the standard EM).

2.6.1 Updating Dirichlet distribution factors $q(\pi)$, $q(\phi_z^w)$, $q(\phi_z^b)$ (M-step) Since we have truncated β at a finite K, the Dirichlet Process reduces to a finite Dirichlet distribution. Using mean-field $q(\pi) \propto \mathbb{E}_q \log(p(\pi|t,s)) \propto \mathbb{E}_q \log(p(t,s,\pi))$. The optimal $q(\pi)$ parametrized by γ is given by standard update for a Dirichlet distribution. Computing the expectation we get the following expression:

$$\begin{aligned}
q(\pi) &= \exp \mathbb{E}_q \log[\prod_{z\in Z} \pi_z^{\alpha_\pi \beta} \prod_{z\in Z} \pi_z^{\sum_{i=1}^{M} \mathbb{1}_{(s_i,z)}} \prod_{z\in Z} \pi_z^{\sum_{i=1}^{N} \mathbb{1}_{(ti,z)}}] \\
&= \exp \mathbb{E}_q(\alpha_\pi \beta + \sum_{i=1}^{M} \mathbb{1}_{(s_i,z)} + \sum_{i=1}^{N} \mathbb{1}_{(t_i,z)}) \sum_{z\in Z} \log \pi_z \\
&= \exp \sum_{z\in Z} \log \pi_z^{\mathbb{E}_q \alpha_\pi \beta + \mathbb{E}_q \sum_{i=1}^{M} \mathbb{1}_{(s_i,z)} + \mathbb{E}_q \sum_{i=1}^{N} \mathbb{1}_{(ti,z)}} \\
&= \prod_{z\in Z} \pi_z^{\alpha_\pi \beta + C_s(z) + C_t(z)} \\
&= \text{Dirichlet}\left(\alpha_\pi \beta + C_s(\cdot) + C_t(\cdot)\right)
\end{aligned}$$

Therefore the solution to factor π of the form $\gamma = \alpha_\pi \beta + C_t(\cdot) + C_s(\cdot)$ as the update for the Dirichlet parameters, where $C_t(\cdot) = C_t(t_1 ... t_k)$ is a vector of expected counts of the values that the factor t can take. Similarly $C_s(\cdot) = C_s(s_1 ... s_k)$ is the vector of expected counts that the factor s can take. These expected counts are computed using $q(s)$ and $q(t)$ that we describe below (E-step).

The updates for the $q(\phi)$ are obtained similarly, and are $q(\phi_z^w|\lambda_z^w) = \text{Dirichlet}(\alpha_w + C^w(z,\cdot))$ where $\lambda_z^w = \alpha_w + C^w(z,\cdot)$ and $q(\phi_z^b|\lambda_z^b) = \text{Dirichlet}(\alpha_b + C^b(z,\cdot))$ where $\lambda_z^b = \alpha_b + C^b(z,\cdot)$. Here $C^w(z,\cdot) = C(z, w_1 ... w_W)$ is the vector of expected counts of words of the image in cluster z and $C^b(z,\cdot) = C(z, b_1 ... b_B)$ is the vector of expected counts for visual words in cluster z that describe the image.

2.6.2 Updating multinomial distribution factors $q(t)$, $q(s)$ (E-step) In order to introduce dependency of the data, we first define $q(t_j|w_i) \propto q(t_j, w_i)$ and $q(t_j)$ can be recovered by marginalizing over the words w. Using mean-field approximation,

$$\begin{aligned}
q(t_j|w_i) &= \exp\left(\mathbb{E}_q \log(p(t_j|w_i))\right) \\
&\propto \exp\left(\mathbb{E}_q \log(p(t_j, w_i))\right) \\
&\propto \exp\left(\mathbb{E}_q \log \pi(j)\right) \exp\left(\mathbb{E}_q \log \phi_{t_j}^w(w_i)\right)
\end{aligned}$$

Define multinomial weights as $W(t_j) = \exp\left(\mathbb{E}_q \log \pi(j)\right)$ and $W_{t_j}(w_i) = \exp\left(\mathbb{E}_q \log \phi_{t_j}^w(w_i)\right)$. The weights W can be computed efficiently, namely $W_{t_j}(w_i) = \frac{\exp(\Psi(\lambda_t(w_i)))}{\exp \Psi(\sum_i \lambda_t(w_i))}$ and $W_t(t_j) = \frac{\exp(\Psi(\gamma_j))}{\exp \Psi(\sum_i \gamma_i)}$ where $\Psi(x) = \frac{\partial}{\partial x} \log \Gamma(x)$ is the Digamma function (which can be computed using Taylor-series approximation). The Dirichlet priors λ and γ are used after updating the Dirichlet distribution factors (which was described in the previous step).

We now show how to compute the expectation of the multinomial weight which depends on the Dirichlet prior. For a variable ϕ drawn from a Dirichlet distribution parametrized by γ:

$$p(\phi|\gamma) = e^{\left(\sum_i \gamma_i \log \phi_i - \sum_i \log \Gamma(\gamma_i) + \log \Gamma(\sum_i \gamma_i)\right)}$$

where $\log \phi_i$ is the sufficient statistic and $\log \Gamma(\sum_i \gamma_i) - \sum_i \log \Gamma(\gamma_i)$ is the log-normalization factor. Using the general fact that the expectation of the sufficient statistic is the first moment of the log-normalization factor w.r.t to its natural parameter, we get $\mathbb{E}_q \log \phi_i = \Psi(\gamma_i) - \Psi(\sum_j \gamma_j)$.

2.6.3 Updating top-level component $q(\beta)$ Finally we summarize the updates for the stick-breaking parameters β. Again, using mean-field it is easy to show that $q(\beta) \propto \mathbb{E}_q p(\beta|\alpha) + \mathbb{E}_q p(\pi|\beta)$, and so $q(\beta) = \mathbb{E}_q \text{GEM}(\beta;\alpha) + \mathbb{E}_q \text{DP}(\pi;\alpha^\pi \beta)$, however since we truncated β at K, it becomes $q(\beta) = \mathbb{E}_q \text{GEM}(\beta;\alpha) + \mathbb{E}_q \text{Dirichlet}(\pi;\alpha^\pi \beta)$. There are no closed-form solutions for β, however it is possible to use maximize $q(\beta)$ using gradient ascent and update the components of β with $\eta \frac{\partial q(\beta)}{\partial \beta_k}$ iteratively (where η is the learning rate). The updates are very similar to [17]. In order to satisfy the constraint $\sum_{i=1}^{K} \beta_i = 1$ we use Quadratic Penalty method [19].

2.7 Making predictions Given the model, we can now use it to make predictions for the region annotation. To predict the label for the described by $\mathbf{b} = b_1 ... b_T$, we can use the word which has the highest probability given all the visual words in the region: $p(w|\mathbf{b})$. This probability can be computed using:

$$p(w|\mathbf{b}) = \sum_{m=1}^{T} \sum_{z_m} p(w|z_m) \int p(z_m|\pi_s) p(\pi_s|b_m) d\pi_s$$
$$\approx \sum_{m=1}^{T} \sum_{z_m} p(w|z_m) q(z_m|b_m)$$

Note that the integral can be computed efficiently using variational inference for the test region.

The label assigned to a region is then the one which gives the highest probability $w_{pred} = \arg\max_{w_i \in W} p(w_i|\mathbf{b})$.

3 Experiments and Results

We describe the datasets used in our evaluation, and the experimental set-up.

3.1 Data In order to evaluate the performance of the model on image object label correspondence, we need to assume that the image to be labeled is segmented into regions or objects and need to have labels for each region or object in each test image. The images can be segmented using one of the magnitude of available segmentation algorithms (such as normalized cuts [23] or superpixels[20]). Note that we do not use object-level labels in training the model. A major goal of this work is to explore the feasibility of using models trained on a dataset of images and their associated annotations to perform both image annotation as well as labeling of individual objects in each images. We proceed with describing the image data and its representation.

3.1.1 PASCAL Visual Objects Classes We compare both MoM-LDA and MoM-HDP on the image annotation and image-label correspondence tasks on Visual Object Classes 2007 challenge data [10].

The VOC 2007 database contains 2501 training images in 20 categories and 4952 images in the test set. We resized the images for the maximum height of 256 pixels. We use grid sampling to extract patches of 13x13 pixels from each image. We then use SIFT representation of each patch [18] to extract 128 features for all images in the training set. These features are invariant to rotation and occlusion, which is often present in the images. The 150,000 descriptors (extracted randomly from the training images) were clustered into 1500 clusters using k-means clustering to create a codebook of "visual words". Each *image* was then represented as a bag of visual words, and a bag of caption words (labels). The codebook created from the training images was used to represent the test objects.

We assume that the test images are segmented and extract the SIFT features from the regions, and use the codebook created at training to represent the *test objects*. If the images were not segmented, we could have used segmentation algorithms (such as normalized cuts or superpixels) to segment each image into regions before processing them further. However, the results of such segmentation may or may not coincide with the segmentation that forms the basis of object-level labels used as reference to evaluate the performance of the model on the image object-label correspondence task. Hence we assume here that segmented images are provided during the test phase. There are 14,976 objects in the test set.

Figure 2: Sample from the VOC 2007 training and test images.

We show some representative training and test images in Figure 2 to demonstrate the variety of the images and complexity of the task.

3.1.2 LabelME To evaluate the performance of the models on the large scale data set with many possibilities for captions we use LabelMe [21] database. LabelMe is a web-based image database and an annotation tool which allows users to annotate images and objects in the images in the database. The annotators select the regions which correspond to the objects in the image, and label these regions with the keywords. The database contains a great variety of image categories and themes, and it continues to grow over-time as more and more people contribute the new images and annotate the existing ones.

For our experiments we selected a subset using 9 keywords to query for images and used the union of these images as the data ("building", "car", "tree", "cat", "dog", "person", "plant", "water" and "sky" were the keywords). We then selected images which have between 4 and 19 objects. From the resulting subset we used 80% of the images as the training data set (resulting in 7373 images), and the rest as the test set (1513 images). The test set contains ∼14000 regions, and so on average each image has 10 captions. All

images were rescaled for the maximum height and width of 256 pixels.

The captions were lower cased and stemmed, resulting in ≈ 1700 distinct caption words in the vocabulary. As before, we extract SIFT features in order to create a codebook of 1500 visual words from the training data from 15,000 image patches randomly sampled from the training images, train the model on the image and caption information only, and test the model on the regions.

3.2 Experiments and results

3.2.1 Multiple label learning as one-against-all classification To establish a baseline, we first consider reducing the multiple label problem to one-against-all learning scenario, similar to the set-up in [26]. Given the dataset $\mathcal{D} = \{\mathbf{b}_j, \mathbf{w}_j\}_{j=1}^{D}$, vocabulary of caption words W of size T, we train T binary classifiers. Each classifier h_{w_i} is trained on a new dataset were all the target words were kept and considered as one class, and all the words that are not the target words where considered the second class: $\mathcal{D}^{'} = \{\mathbf{b}_j, w_j^{'}\}_{j=1}^{D}$ where $\mathbf{b}_j$ is that as in $\mathcal{D}$ and $w_j^{'} = 1$ if $\mathbf{w}_j$ of $\mathcal{D}$ contains word $w_i \in W$ and 0 otherwise. Given a test object $\mathbf{b}_{test}$ each of the classifiers $h_{w_i}(\mathbf{b}_{test})$ assigns a score r_{w_i} and the word with the highest score $w_{new} = \arg\max_{w_i}[r_{w_1}, ..., r_{w_T}]$ is used as a prediction. We considered Naive Bayes and Logistic Regressions as the classifiers.

3.2.2 Initialization for parameter estimation Variational inference is susceptible to local minima. Since one of the local minima corresponds to the setting where all factors are equally likely, we initialize the model by randomly assigning several image/caption pairs to a factor. We set the hyperparameters $\alpha = \{\alpha, \alpha_\pi, \alpha_b, \alpha_w\}$ to 1. Given the large size of the training dataset, we believe that the choice of hyperparameters for priors is not especially critical.

3.2.3 Image annotation and region labeling In order to assess the performance of the models on the image annotation task, we used accuracy of annotation as the performance measure. Let C be the predicted set of words in a caption. Let R be the actual caption (the actual set of words that appear in the caption for a particular image). To avoid the complication of having to deal with multiple objects with the same name, we binarize C and R. To measure how close C is to R we count how many elements are in common in C and R; In other words, we are interested in the cardinality of the intersection $|C \cap R|$. We can now define accuracy as $Acc = P(R|C) = \frac{|C \cap R|}{|C|}$.

Since we have the ground truth or object-level labels for the regions, we can also evaluate the performance of the model on the object recognition task on the per-label basis using standard performance measures such as precision (the fraction of the actual objects with a given label out of all the objects classified as such), recall (the fraction of the objects that were assigned a particular label out of all the existing objects with that label), and accuracy (the fraction of correctly labeled objects in the entire set of test images).

	per region	per caption
NB OneVsAll	30.56	38.03
LR OneVsAll	20.19	20.79
MoM-LDA	31.67	40.82
MoM-HDP	**34.5**	**41.92**
Chance prediction	5	5

Table 1: Comparison of accuracies (in %) of various algorithms for per-region and per-caption annotation task for VOC 2007 dataset

VOC2007 In the VOC 2007 dataset, the number of labels is 20, and so predicting a label at random results in 5% accuracy.

Table 1 shows the comparison of one-against-all learning scenario and the combined LDA model.

Notice that Logistic Regression has the worst performance. We believe that this could be due to overfitting on the training data. Since one Logistic Regression is trained to maximize accuracy for each keyword, since the distribution of the target word and its compliment is very unbalanced, it is possible that Logistic Regression overfits on the compliment of the keyword, thus assigning low scores to the words.

Statistical Significance test In order to test significance of the results on region labeling we use a simple statistical test for difference in two error proportions [24]. Let the null hypothesis be that two algorithms f_1 and f_2 have the same error on the same test dataset T of size N. Let $e_1 = \frac{N_{f1}}{N}$ be the fraction of the test examples that f_1 predicted incorrectly and let $e_2 = \frac{N_{f2}}{N}$ be the fraction of the test examples that f_2 predicted incorrectly. Then the quantity $e_1 - e_2$ can be viewed as a random variable with 0 mean and standard deviation $s_e = \sqrt{\frac{2p(1-p)}{N}}$ where $p = \frac{e_1+e_2}{2}$ is the average of two errors. From this, we use the statistic $z = \frac{e_1-e_2}{s_e}$ and if $|z| > Z_{0.975} = 1.96$ then the null hypothesis is rejected. We compute the z-value between the MoM-HDP and the other algorithms considered, and the improvement on the test set is statistically significant.[1]

[1]The z-values for the difference between errors of various algorithms are: z(MoM-HDP,NB)=7.1, z(MoM-HDP,LR)=27.6,

Since we use region labeling to construct the full caption, we believe that the significance tests on the region labeling are enough for the caption reconstruction. We also note that to the best of our knowledge there is no well-defined statistical significance test for a learning algorithm which predicts multiple labels to a test instance, and that it is of interest to develop such test.

A reason for using a simple statistical test instead of a k-fold cross validation test [9] is that the VOC 2007 challenge dataset that we used consists of a prespecified training set and a test set; with test set being much larger and more "difficult" that the training set [10].[2]

We next take a closer look the performance of the combined mixture models MoM-LDA and MoM-HDP on the region annotation and overall image annotation as a function of the number of the mixture components K, in Figure 3. The best precision of MoM-LDA in terms labels assigned to objects in the image and in terms of the caption assigned to the image was obtained at $K = 5$. The performance of MoM-HDP is less sensitive to the choice of K used to truncate the HDP model. We also observe that the performance of Mom-LDA degraded when the number of mixture components exceeded the optimum value ($K = 5$ whereas the performance of MoM-HDP was more robust with respect to K.

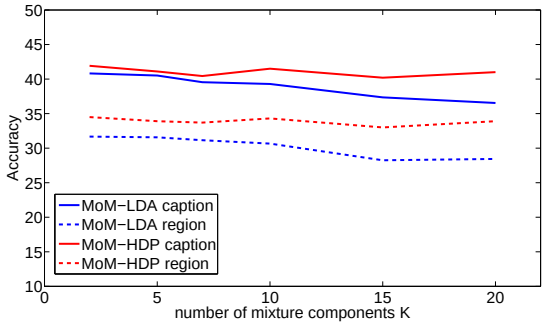

Figure 3: Performance of MoM-LDA (represented by the red line) and performance of MoM-HDP (represented by the blue line) vs the number of mixture components. The accuracy on the region labeling is shown using solid line, and the overall accuracy on the captions constructed from the region labels is shown using dashed line.

While an accuracy of 41% may be viewed as poor in the standard supervised learning setting, it is worth noting that the more general multi-modal learning setting considered in this paper is far more challenging (see for example, the results reported in [2] where a similar performance measure was used to evaluate the performance of MoM-LDA, however we also note that they used a probability threshold, thus allowing for several label to be predicted for a given region or [11] for performance on image annotation).

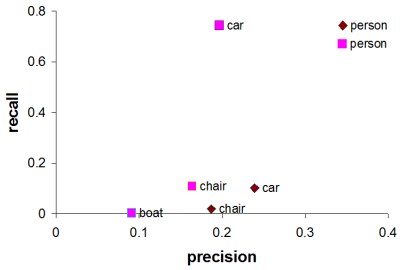

Figure 4: Region annotation result: per-label precision recall on all predicted region labels for VOC 2007. Square: HDP precision/recall, diamond: LDA precision/recall.

We now take a closer look at the performance of the models on the per-region task, and examine in detail the performance measured by precision/recall on the per-label basis. The results are presented in Figure 4 for MoM-HDP and LDA for a cut-off $K = 5$.

Only several labels have relatively high precision/recall measures ([2] reported similar trends, however the precision/recall was calculated for whether a word was present or absent in the caption, not for the labels directly). While both models have similar performance on the precision measure, the recall is much higher for HDP model. In addition, HDP was able to assign a relatively high precision/recall for the label "boat", while LDA did not predict any boats correctly.

Note that the label "person" has a very high recall and low prevision, which indicates that it was often predicted as a possible label. We discovered that in the training data about half of the captions included the word "person". Consider an image which has many objects of which only a few have corresponding labels in the caption. In such a scenario, the visual words associated with the image (which could be very diverse) are likely to get assigned to the clusters associated with the few labels that appear in the image caption, thereby biasing the predictions towards those labels. We conjecture that the sparsity of captions relative to the number of objects in the image biases the model towards the labels that are overrepresented in captions. One possible approach to correcting this bias is to use partially supervised training data and to add region/caption pairs as

$\overline{z(\text{MoM-HDP, MoM-LDA})}$=5.02

[2]Moreover, because the words that appear in a given caption as well as the objects that appear in an image are unlikely to be independent which presents challenges in devising reliable tests for comparing different models - see Section 5

additional training examples. Another possible source of improvement is better quality captions, i.e., captions that are descriptive of all objects in the image.

LabelMe Lastly, we show the performance of the model on region labeling on the large data set derived from LabelMe. Here we set K to 20, and train MoM-LDA and MoM-HDP and we report the accuracy on the per-region basis and on the entire caption set in Table 2. A random classifier would have 1 in 1700 chance to predict a caption correctly (0.06% accuracy).

LabelME	image annot.	region annot.
MoM-LDA	15.56	10.5
MoM-HDP	34.84	**28.45**
NB OneVsAll	**38.2**	24.21

Table 2: Performance (accuracy in %) of MoM-LDA and MoM-HDP on image annotation and region recognition for LabelME

While Naive Bayes one-against-all training scenario produces a better result on the caption prediction than the MoM-HDP model, MoM-HDP model has a much better result on the region labeling, and the improvement is statistically significant.[3] We do not present result for Logistic Regression due to the training time demand. It takes about 1 hour to train one Logistic Regression for one target word. Therefore, to train all the needed classifiers would take approximately 1700 hr = 70 days (as opposed to several minutes for Naive Bayes and about an hour for MoM-LDA and MoM-HDP). The training time is one major drawback of having one-against-all training scenarios.

4 Related Work

We briefly summarize work on the image annotation and image object-label correspondence or closely related problems. Learning from multi-modal data, and in particular, learning to annotate images, has been cast as a multiple label multiple instance learning problem [26]. In this formulation, each image is represented by a bag of objects (instances), and the corresponding image caption is represented by a bag of words (set of labels). Zhou and Zhang [26] proposed to use a multiple-instance learning for each label using one-against all multiple-instance learners. However, this work did not address the problem of labeling each individual object within an image.

Hardoon et al. [11] have explored a kernelized version of canonical correlation analysis for image retrieval and annotation. Specifically they show how a semantic representation of images and their associated text can be learned and how the resulting representation in a common semantic space can be used to compare data from the text and image modalities. However, the primary focus of this work was not on solving the image object-label correspondence problem.

Barnard et al. [2] have examined several solutions for the image annotation and image-object label correspondence problems. They developed several models for the joint distribution of image regions and words, including those that explicitly learn the correspondence between image regions and words. They studied a multi-modal and correspondence extensions to hierarchical mixture models [13], and probabilistic latent semantic indexing (pLSI) [12] for text, a translation model adapted from statistical machine translation [8], and a multi-modal extension to mixture of Latent Dirichlet Allocation (MoM-LDA) [5, 16] which generalizes LDA [6] to the setting where the data combines multiple modalities (e.g., image, text).

Selecting the number of mixture components to be used in a MoM-LDA model is difficult. In practice a model is trained for several numbers of mixture components, evaluated on a held-out set, and the best performing model is chosen. The need to train and evaluate several models makes this approach computationally expensive, especially in the case of large datasets consisting of large numbers of image and text features. In contrast, the proposed MoM-HDP model allows us to circumvent the need for a priori choice of the number of mixture components. It addressed the computational expense associated with the model selection for MoM-LDA since in practice multiple MoM-LDA models need to be trained before choosing one based on the results of cross-validation.

In contrast to previous work [2] which relied on representation of image segments using *global* features such as shape, color, texture, etc. we have chosen to use *local* features [7]. A consequence of reliance on global properties of image segments is that the images must be segmented prior to training the model. In contrast, representation of image segments (blobs) using local image features makes it possible to train the model on images without segmenting them prior to training. Furthermore, recent work in the image processing community has shown that local representation of the image can substantially improve the performance of the resulting models [7]. In [2], the experiments were performed using the Corel dataset which only provides the captions for the image, and this dataset is also no longer publicly available. In the absence of labels for individual objects or image segments, their study provided a

[3] The z-values for the difference between errors of various algorithms are: z(MoM-LDA,MoM-HDP)=37.92, z(MoM-HDP, NB)=8.05

limited assessment on the image object-label correspondence task on a small number of hand-annotated objects. In contrast, in this paper, we used two datasets which provide the ground truth needed evaluating the performance of alternative solutions image annotation and image object-label correspondence tasks: Visual Object Classes (VOC) 2007 challenge data which has 20 possible labels and a subset of LabelMe, which has over 1700 possible labels.

5 Summary

In this paper we considered an interesting example of multi-media data mining: Given a dataset of images and their associated captions, can we build a model that not only predicts a caption i.e., a collection of labels for an entire image (the image annotation task), but specifically labels the individual objects (or regions of interest) in the image with a collection of labels (the image object-label correspondence task)? We have described a solution to this problem based on a *multi-modal hierarchical Dirichlet Process* (MoM-HDP). MoM-HDP generalizes the hierarchical Dirichlet Process (HDP) model (that can be thought of as the analog of a mixture model, but with an infinite number of mixture components assumed in a mixture model) to deal with multi-modal data (e.g., images, text). MoM-HDP thus allows us to circumvent the need, in the case of alternatives such as the multi-modal latent Dirichlet Allocation (MoM-LDA), for a priori and hence potentially arbitrary choice of the number of mixture components or the computational expense of choosing the best performing model from among multiple models corresponding to different choices of the number of mixture components. During training, the model has access to an un-segmented image and its caption, but not the labels for each object in the image. The trained model is used to predict the label for each region of interest in a segmented image. We use variational inference to efficiently estimate model parameters. Our experiments using two large-scale datasets show that the generalization performance of MoM-HDP is superior to that of MoM-HDP as well as the Naive Bayes and Logistic Regression classifiers (in one-against-all learning scenario).

Although our experiments with the MoM-HDP model have been limited to data consisting of images and text, the underlying probabilistic model and the algorithm for training the model readily generalize to data that include multiple modalities (e.g., text, image, speech, etc.). MoM-HDP model can be extended along several interesting directions: The current model is based on a simple bag of features (words, visual features or visual words) representation of the data from each modality. It would be interesting to consider more sophisticated models of interaction among features within and across modalities.

Sparsity of captions in data extracted from online image collections presents a significant challenge. It would be interesting to augment the dataset with some fully labeled data. It also would be interesting to explore *active learning* strategies by which the system could seek labels of specific objects in specific images that would be most beneficial for reducing the ambiguities that exist in its current model of the data.

In our experiments, we have used rather simple metrics (adapted from the more extensively studied supervised learning scenario) for evaluating the performance of a learned model and for comparing alternative models (or algorithms for multi-modal data mining). Rigorous experimental comparison of the models requires the development of performance measures that reflect the inherent complexities of learning predictive models from multi-modal data. For example, in the case of image annotation or image object-label correspondence have to take into account the correlations among labels within an annotation or collection of labels associated with an object in an image. Performance measures for standard classification tasks are defined in terms of the contingency matrix of true positives, false positives, true negatives, and false negatives. It is not immediately apparent how to generalize such measures to cope with the complexities of predictive models for data across multiple, typically not independent modalities. Statistical tests for comparison of different predictive models need to take into account the complications arising from multiple hypotheses testing on the same data. It would also be interesting to consider generalizations of the problems considered in this paper that take into account different costs or risks associated with mislabeling different image objects or inclusion or exclusion of different labels. Our ongoing research is aimed at addressing some of these problems.

Acknowledgements We thank the reviewers for providing insightful comments. This research was supported in part by a grant (IIS 0711356) from the National Science Foundation.

References

[1] Charles Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

[2] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] David Blei and Michael I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2004.

[5] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2006.

[8] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

[9] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[10] Mark Everingham, Luc Van-Gool, Chris Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[11] David Hardoon, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. A correlation approach for automatic image annotation. In Xue Li, Osmar Zaiane, and Zahnhuai Li, editors, *Second International Conference on Advanced Data Mining and Applications, ADMA 2006*, volume 4093, pages 681–692. Springer, 2006.

[12] Thomas Hofman. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[13] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *Proc. of IJCAI*, 1999.

[14] Hermant Ishwaran and Lancelot James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161 – 174, 2001.

[15] Hermant Ishwaran and Lancelot James. Generalized weighted chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13:1211 – 1235, 2003.

[16] Li-Jia Li and Li Fei-Fei. What, where and who? classifying event by scene and object recognition. In *IEEE International Conference in Computer Vision (ICCV)*, 2007.

[17] Percy Liang, Slav Petrov, Michael Jordan, and Dan Klein. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 688–697, 2007.

[18] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[19] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer, 2000.

[20] X. Ren and J. Malik. Learning a classification model for segmentation. In *In Proc. 9th Int. Conf. Computer Vision*, 2003.

[21] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: a database and web-based tool for image annotation. Technical report, MIT AI Lab Memo AIM-2005-025, 2005.

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

[23] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[24] G. W. Snedecor and W. G. Cochran. *Statistical Methods*. Iowa State University Press, 1989.

[25] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

[26] Zhi-Hua Zhou and Min-Ling Zhang. Multi-instance multi-label learning with application to scene classification. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1609–1616. MIT Press, Cambridge, MA, 2007.