Annotating images and image objects using a hierarchical Dirichlet process model

Oksana Yakhnenko
Computer Science Department
Iowa State University
Ames, IA, 50010
oksayakh@cs.iastate.edu

Vasant Honavar
Computer Science Department
Iowa State University
Ames, IA, 50010
honavar@cs.iastate.edu

ABSTRACT

Many applications call for learning to label individual objects in an image where the only information available to the learner is a dataset of images with their associated captions, i.e., words that describe the image content without specifically labeling the individual objects. We address this problem using a multi-modal hierarchical Dirichlet process model (MoM-HDP) - a nonparametric Bayesian model which provides a generalization for multi-model latent Dirichlet allocation model (MoM-LDA) used for similar problems in the past. We apply this model for predicting labels of objects in images containing multiple objects. 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. MoM-HDP generalizes a multi-modal latent Dirichlet allocation model in that it allows the number of components of the mixture model to adapt to the data. The model parameters are efficiently estimated using variational inference. Our experiments show that MoM-HDP performs just as well as or better than the MoM-LDA model (regardless the choice of the number of clusters in the MoM-LDA model).

1. INTRODUCTION

The traditional supervised classification task requires the use of labeled data. 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). For example, many web data sources e.g., social network communities such as Flickr, Facebook 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. 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.

Furthermore, in some cases we do not have objects labeled explicitly, but with more and more types of data available, the same information is contained in multiple data sources, and it is of interest to find correlations between the features of the different data sources.. Consider newspaper articles which contain pictures of the events and the description of the same events in text; or research articles which contain figures and captions which describe these figures. Consider the television news where in the video sequence voice-over describes what is happening. Similar situations appear in different domains - from the web, to multi-media, to biomedical domain and others.

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 caption for an entire image (the image *annotation* task), but specifically labels the individual objects (or regions of interest) in the image (the image object-label *correspondence* task)?

We first review the related work in the area of image annotation and the previous work in the area of the object-label correspondence.

Learning from multi-modal data, and annotation in particular has been cast as a multi-label, multiple instance learning problem [18]. In the context of image annotation 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 et al. [18] solved this problem by mapping multiple instances within a bag into a single meta instance and then solving the resulting single instance, multi-label learning task by training as many binary one-versus-all classifiers as there are class labels. It is unclear how this approach can be used to label each individual object in an image.

Hardoon et al. [8] have explored a kernelized version of canonical correlation analysis, for image retrieval and annotation.

Barnard et al. [2] has explored a range of models for solving the annotation and correspondence tasks. Of particular interest to us is a multi-modal latent Dirichlet allocation (MoM-LDA) model [5, 12] which was used in the study by Barnard et al. [2]. MoM-LDA is a model which combines as many latent Dirichlet allocation (LDA, [5]) models as there

MDM/KDD '08 August 24, 2008, Las Vegas, NV, USA

are modalities in the data. Latent Dirichlet allocation is a generative probabilistic model for independent *collections* of data where each collection is modeled by a mixture over latent factors. An advantage of LDA over other probabilistic mixture models (such as probabilistic latent semantic analysis, pLSA [9]) is that the distribution of the mixture components is not fixed, but is flexible for each document, thus allowing for multiple models. However a limitation of MoM-LDA is the need to specify the number of components of the mixture model. In practice, the model is trained for several choices of the mixture components and the optimal candidate for the number of clusters is chosen based on the validation on a hold-out set. This can be expensive because of the need to train several models. A somewhat similar model to classify events and scenes was used by Li and Fei-Fei [12].

Previous work [2] used global features for the image segments, such as shape, color, texture, etc. However, the reliance on global properties of image segments requires that the images be segmented prior to training the model. In contrast, by using local features of the image (obtained from small patches sampled from the image), it is possible to train the model on images without segmenting the images 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 [6]. In [2], the experiments were performed using the Corel dataset which only provides the captions for the image. In the absence of labels for individual objects or image segments, their study provided a limited assessment on the image object-label correspondence task on a small number of hand-annotated objects.

In this paper, we introduce MoM-HDP, a non-parametric generalization of the MoM-LDA model [5] using a hierarchical Dirichlet process. MoM-HDP, unlike MoM-LDA, adapts the number of clusters based on the training data. Also unlike previous work [5, 2] we use local features to represent the image to enable the model to find the needed correlations between the individual labels. We then apply this model to solve the problem of training a probabilistic model from a dataset of images and their associated captions to predict both the caption (the *image annotation* problem) as well as labels of individual regions of interest (the *image object-label correspondence* problem).

We evaluate the performance of MoM-HDP and compare it with MoM-LDA on the image annotation and image-label correspondence task using a subset of VOC 2007 challenge data which has 20 possible labels. Our choice of datasets was motivated by the availability of labels for each region or object in the image which although not used in training the model, provides the ground truth needed to evaluate, unlike previous studies [5, 2], the performance of the trained model on the image object-label correspondence task. Our experiments show that MoM-HDP is invariant to the number of hidden factors (unlike MoM-LDA) and has a better generalization performance than MoM-LDA model on both the image annotation task and the the image object-label correspondence task.

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 in the results of our comparison of MoM-LDA and MoM-HDP models in Section 3. We conclude the paper with a summary and a brief discussion of some directions for further research in Section 4.

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 $\mathbf{w_i} = \{w_{i1}...w_{iN}\}, w_{ij} \in W$ be the caption for image i. Let B be the vocabulary of all the possible visual words in the pictures, and $\mathbf{b_i} = \{b_{i1}...b_{iM}\}, b_{ij} \in B$ be the 'visual words' 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 $\pi_i \sim \text{Dirichlet}(\alpha)$. 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 handle 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) [17], with *countably infinite* number of mixture

components. We summarize the key aspects of DP, and then HDP here. For details on the HDP and their applications in probabilistic graphical models we refer the reader to [1], [17] or [4].

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 [11], and stick-breaking construction [16]. 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 [17]. 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 probability distribution for the mixture components (π_z), 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 drawn from $DP(\alpha^\pi, \beta)$ were β 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.

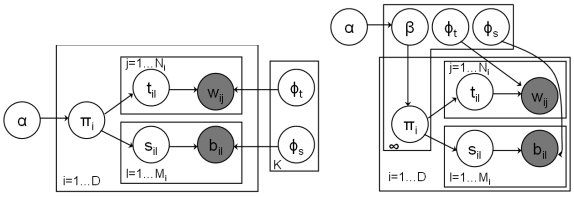

Figure 1: MoM-LDA model (left). Its MoM-HDP counterpart (right).

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

MoM-HDP	MoM-LDA
draw $\beta \sim \text{GEM}(\alpha)$	chose priors $(\alpha_1...\alpha_K)$
for each $z = 1, 2...$	for each $z = 1, ..., K$
draw $\phi_z^w \sim \text{Dirichlet}(\alpha_w)$	draw $\phi_z^w \sim G_0$
draw $\phi_z^b \sim \text{Dirichlet}(\alpha_b)$	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 $t_{ij} \sim \text{Mult}(\pi_i)$	draw $t_{ij} \sim \text{Mult}(\pi_i)$
draw $w_{ij} \sim \text{Mult}(\phi^w_{t_{ij}})$	draw $w_{ij} \sim \text{Mult}(\phi^w_{t_{ij}})$
for each word $j = 1, ..., M_i$	for each word $j = 1, ..., M_i$
draw $s_{ij} \sim \text{Mult}(\pi_i)$	draw $s_{ij} \sim \text{Mult}(\pi_i)$
draw $b_{ij} \sim \text{Mult}(\phi^b_{s_{ij}})$	draw $b_{ij} \sim \text{Mult}(\phi^b_{s_{ij}})$

To make the parameter estimation feasible, we assume a truncated DP [10], 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.

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 $q(\beta) \sim \text{GEM}(\alpha)$ 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.

Using mean-field approximation we get the following property: if $q(\mathbf{y}) = \prod_{i=1}^{n} q_i(y_i)$ is a factorized distribution for each of the factors y_i, then the solution for $q_i(y_i)$ has the form $q_i(y_i) \propto \exp(\mathbb{E}_{q-i} \log p(y_i|\mathbf{y}_{-i}))$ where $\mathbf{y}_{-i}$ is a set of all the factors which are not y_i (see [3] for details).

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: $q(\pi|\gamma) = \text{Dirichlet}(\alpha_\pi\beta + C_t(\cdot) + C_s(\cdot))$ so we use the prior γ of the form $\gamma = \alpha_\pi\beta + C_t(\cdot) + C_s(\cdot)$ as the update for the Dirichlet parameters, where $C_t(\cdot)$ is a vector of expected counts of the values that the factor t can take. Formally, define $C_t(\cdot) = C_t(t_1...t_k)$ where each of $C_t(t_k) = \mathbb{E}_q \sum_{i=1}^{N} \mathbb{1}_{(t_i,t_k)}$ ($\mathbb{1}_{(t_i,t_k)}$ is the indicator function). 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. The updates are very similar to [13]. In order to satisfy the constraint $\sum_{i=1}^{K} \beta_i = 1$ we use Quadratic Penalty method [15].

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:

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

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

3.1 Data

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 a subset of VOC 2007 challenge data [7] with 20 possible labels.

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. This requirement influenced our choice of the datasets used in our experiments. 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. The Corel dataset used in previous studies [2] does not include labels for objects within each image, and hence cannot be used to assess the performance of the model on the image object-label correspondence task. In addition this dataset is also no longer publicly available.

The VOC 2007 database contains 2501 training images in 20 categories, from which we selected images in 7 categories that contain roughly the same number of images/labeled regions: 'boat', 'cat', 'cow', 'motorbike', 'sofa', 'sheep', 'train', resulting in a training set of 714 images (at training we also include all the caption that came in the images, so the actual number of captions is 20).

We rescaled the images for the maximum height of 256 pixels. We then used SIFT detector [14] 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 descriptors 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 then selected images from the test set in the same categories (resulting in 1414 test images). 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 are not segmented, we can use standard segmentation algorithms 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. Each test image has two to three objects, resulting in a total of 3726 test objects.

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

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

3.2 Experiments and results

3.2.1 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 is not especially critical).

3.2.2 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).

VOC2007.

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

We summarize the performance of MoM-LDA 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.

While an accuracy of 34% 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

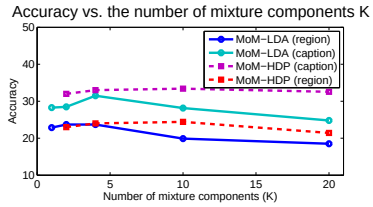

Figure 3: Performance of MoM-LDA (represented by the solid line) and performance of MoM-HDP (represented by the dotted line) vs the number of mixture components. The accuracy on the region labeling is denoted by label (region), and the overall accuracy on the captions constructed from the region labels is denoted by label (caption).

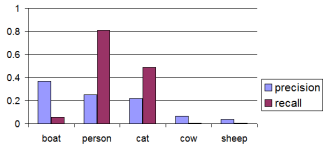

Figure 4: Region annotation result: per-label precision recall on all predicted region labels. The labels not shown in the plot had precision/recall 0.

we also note that they used a probability threshold, thus allowing for several label to be predicted for a given region).

We also examine in detail the performance on the per-label basis, and we show the precision/recall plots in Figure 4 for MoM-HDP.

Note that the label 'person' has a very high recall and low prevision, which indicates that it was often predicted as a possible label. While we constructed the dataset by attempting to include the regions which had similar frequency distributions, we discovered that in the training data 456 captions included '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 additional training examples. Another possible source of improvement is better quality captions, i.e., captions that are descriptive of all objects in the image.

4. CONCLUSION AND DISCUSSION

In this paper we considered a problem of semi-supervised object recognition: Given a dataset of images and their associated captions, can we build a model that not only predicts a caption for an entire image (the image annotation task), but specifically labels the individual objects (or regions of interest) in the image (the image object-label correspondence task)? The need to find such correlations between the text and images is important for many problems in domains in which labeled data is expensive or is not readily available.

Specifically, we have introduced MoM-HDP, a hierarchical Dirichlet process which generalizes the MoM-LDA model [5]. MoM-HDP, unlike MoM-MDP, adapts the number of clusters based on the training data. We also used local features (visual words) to represent the image, to enable the model to find the needed correlations between the individual labels, and the 'visual words' that represent the image segments.

We compared the performance of MoM-LDA and MoM-HDP on the image annotation and image-label correspondence task on a dataset with variety of labels and objects. Our experiments show that MoM-HDP performs just as well as or better than the MoM-LDA model (regardless the choice of the number of clusters in the MoM-LDA model) on both the image annotation task and the the image object-label correspondence task.

It is only relatively recently that probabilistic models for learning from multi-modal data and associated learning algorithms have begun to receive significant attention in the literature. For the purposes of evaluation, we restricted our work and application to images and text, however it is of interest to apply such models in domains with other data types (such as sound, video sequences, and others). The model we presented in this paper considers two modalities, however it can be easily extended to three or more, and it is also of interest to access the performance of this model when more data modalities are present.

Other interesting questions remain to be answered in the future. In this model we considered bag-of-words representation for each modality. It will be interesting to extend the model to consider dependencies between the features into account. In our experiments we encountered the problem of the sparsity of the captions. It is of interest to develop models which overcome this problem. The model we considered is a generative model. In light of the superior performance of discriminative models on classification tasks, it would be interesting to investigate discriminative correspondence models.

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

[7] Mark Everingham, Luc Van-Gool, Chris Williams, John Winn, and Andrew Zisserman. The PASCAL

Visual Object Classes Challenge 2007 (VOC2007) Results.

[8] 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.

[9] Thomas Hofman. Probabilistic latent semantic analysis. In *UAI*, 1999.

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

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

[12] 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.

[13] 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.

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

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

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

[17] 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.

[18] 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.