Max-Margin Latent Dirichlet Allocation for Image Classification and Annotation

Yang Wang
yangwang@uiuc.edu
Dept of Computer Science, University of Illinois at Urbana Champaign

Greg Mori
mori@cs.sfu.ca
School of Computing Science, Simon Fraser University

Abstract

We present the max-margin latent Dirichlet allocation, a max-margin variant of supervised topic models, for image classification and annotation. Our model for image classification (called MMLDA^c) integrates discriminative classification with generative topic models. Our model for image annotation (called MMLDA^a) extends MMLDA^c to the case of multi-label problems, where each image can be associated with more than one annotation terms. We derive efficient learning algorithms for both models and demonstrate experimentally the advantages of our proposed models over other baseline methods.

1 Introduction

We consider the problem of image classification and image annotation. In particular, we develop the *max-margin latent Dirichlet allocation* (MMLDA), a novel hierarchical model integrating max-margin discriminative learning and generative topic models to address these two tasks.

With the explosion of image data on the Internet and the availability of large-scale image databases in the vision community (e.g. LabelMe [18], ImageNet [9]), automatically classifying and annotating these large collections of images is becoming an important challenge. The work in [2] learn to recognize people's faces from face images and their associated captions. There is also work [1] that treats object recognition as a machine translation task and learns a model for the correspondence between image segments and annotation terms. The work in [4] uses a similar idea and develops a probabilistic model for images and texts. There is also work [13, 20] that combine image classification and annotation together.

Much work [1, 3, 4, 12, 13, 20, 21] in image classification and labeling uses topic models, which are a class of powerful tools originally proposed in text modeling and have gained much popularity in computer vision recently. Examples of topic models include the probabilistic latent semantic analysis (pLSI) [10], latent Dirichlet allocation (LDA) [7], correlated topic models (CTM) [5], etc. Most topic models (e.g. LDA) are unsupervised, i.e. only the words in the document collection are modeled. LDA assumes that each document is a mixture of latent topics, and each topic defines a multinomial distribution over a given vocabulary. The goal of topic models is to discover those topics underlying the document collection to facilitate tasks like browsing, searching, etc. Unsupervised topic models have also been used in many computer vision applications. One example is to automatically discover object classes from image collections [17].

BMVC 2011 http://dx.doi.org/10.5244/C.25.112

There has been work on applying topic models to construct features for classification. The hope is that those topics discovered by topic models can be fed to a classifier like SVM. Unfortunately, the SVM trained in this two-stage fashion typically performs worse than the one trained directly based on the original features (e.g. histogram of word counts in text analysis) [7]. This is mainly because the topic discovery and classification are disconnected in this approach. The topics discovered by topic models are not necessarily the ones useful for classification. To address this issue, several supervised variants of topic models have been proposed, including the *supervised LDA* (sLDA) [6] and the *discriminative LDA* (DiscLDA) [11]. The goal of these models is to learn topic models by considering the class label in the training process, so the latent topics discovered by the models are directly tied to classification tasks. Recently, Zhu et al. [22] propose a max-margin variant of supervised topic models called the *maximum entropy discriminative latent Dirichlet allocation* (MedLDA). MedLDA integrates the max-margin learning principle with topic models, which results in topic representations arguably more suitable for classification tasks.

Despite the success of topic models in visual recognition, we believe there is something important missing. Almost all the above-mentioned topic models in computer vision assume the "bag-of-words" image representation, i.e. an image is represented by a collection of un-ordered feature descriptors computed from small local patches. Although the "bag-of-words" representation has been proven successful, other more holistic image representations (e.g. GIST [16]) have been shown to be powerful in many applications too. As we will demonstrate in the experiments, models that exploit both types of feature representations work better than the ones based only on bag-of-words.

In this paper, we propose the *max-margin latent Dirichlet allocation* (MMLDA), a variant of MedLDA. We introduce two different versions of MMLDA, called MMLDA^c for image classification, and MMLDA^a for image annotation. MMLDA^c is based on MedLDA. The main difference is that MedLDA only uses the latent topics as the feature vector for classification, while MMLDA^c uses latent topics together with any other image features. This extension allows MMLDA^c to make use of image features (e.g. GIST) that cannot be easily represented as bag-of-words. MMLDA^a is an extension of MMLDA^c for image annotation. In image annotation, the goal is to choose a set of annotation terms (also called *tags*) to describe an image. Since an image can be associated with more than one tag, image classification is a multi-label classification. In MMLDA^a, various tags are implicitly coupled by the latent topics defined in the model. Training MMLDA^a results in topic representations that are suitable for predicting those tags.

2 Background

Our proposed models are based on the *supervised latent Dirichlet allocation* (sLDA) [6] and the *maximum entropy discrimination latent Dirichlet allocation* (MedLDA) [22]. In this section, we give a brief introduction to these two models.

Suppose we are given a collection $\mathscr{D}$ of M documents $\mathscr{D} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$. Each document $\mathbf{w}$ is a collection of words $\mathbf{w} = (w_1, w_2, ..., w_N)$. For each of presentation, we also write $\mathbf{w}$ as $w_{1:N}$ from now on. A document is also associated with a response variable y. Since we focus on classification, the response variable y is a discrete value from a finite label set $y \in \mathscr{Y}$.

Let K be the number of topics, and V be the size of the vocabulary. Let β be a $K \times V$ matrix where each row β_k is a distribution over the V words. For classification problems, sLDA assumes the following generative process of a document $\mathbf{w}$ and its response variable y:

1. Draw topic proportions $\theta|\alpha \sim \mathrm{Dir}(\alpha)$

2. For each word w_n
   (a) Draw topic assignment $z_n|\theta \sim \text{Mult}(\theta)$
   (b) Draw word $w_n|z_n,\beta_{1:K} \sim \text{Mult}(\beta_{z_n})$
3. Draw response variable $y|z_{1:N},\eta,\sigma^2 \sim \text{GLM}(\eta^\top \bar{z},\sigma^2)$, where $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$

where $\text{GLM}(\bar{z},\eta,\sigma)$ denotes a generalize linear model.

In sLDA, the parameters $(\alpha,\beta,\eta,\sigma^2)$ are estimated by maximizing the joint likelihood $p(\mathbf{y},\mathscr{D}|\alpha,\beta,\eta,\sigma^2)$, where $\mathbf{y}$ is the vector of response variables for all the documents in $\mathscr{D}$. Since directly maximizing the joint likelihood is intractable, sLDA maximizes its lower bound. Given a document $w_{1:N}$ and its response variable y, it can be shown that:

$$\log p(\mathbf{w},y|\alpha,\beta,\eta,\sigma^2) \geq \mathscr{L}(q) = \mathbb{E}[\log p(\theta,\mathbf{z},\eta,y,\mathbf{w})] + \mathscr{H}(q) \tag{1}$$

The expectation $\mathbb{E}[\cdot]$ in Eq. 1 is taken with respect to a variational distribution $q(\theta,\mathbf{z}|\gamma,\phi)$, which is used to approximate the posterior $p(\theta,\mathbf{z}|\alpha,\beta,\sigma^2,y,\mathbf{w})$. $\mathscr{H}(q)$ is the entropy of the distribution q. sLDA maximizes this lower bound $\mathscr{L}(q)$.

In sLDA, $q(\theta,\mathbf{z}|\gamma,\phi)$ is assumed to has a fully factorized form $q(\theta,z_{1:N}|\gamma,\phi_{1:N}) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n)$, where γ is a K-dimensional Dirichlet parameter vector and each ϕ_n parametrizes a multinomial distribution over K elements.

Recently, max-margin based learning methods have gained much popularity in the computer vision community due to their superior performance in a variety of tasks. Examples include support vector machines (SVM) for standard classification problems, and structural SVMs for structured output problems. It is therefore desirable to combine topic models with max-margin learning.

The first attempt of integrating max-margin learning and topic models is the work of *maximum entropy discrimination latent Dirichlet application* (MedLDA) [22]. For classification with $|\mathscr{Y}|$ possible classes, given the latent topic assignment $z_{1:N}$, MedLDA assumes a linear discriminative function in the form of $F(y,z_{1:N},\eta) = \eta_y^\top \bar{z}$, where $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$ as in sLDA, η_y is a class-specific K-dimensional vector associated with class y and η is the concatenation of η_y for all $y \in \mathscr{Y}$. If we assume a normal prior $\mathscr{N}(0,I)$ on the parameter η, MedLDA can be written in the following form (please refer to [22] for the details):

$$\min_{q,\alpha,\beta,\eta,\xi} \quad -\mathscr{L}(q) + \frac{1}{2}||\eta||^2 + C\sum_{d=1}^{D}\xi_d, \quad \text{s.t.} \quad \forall d,y: \quad \xi_d \geq 0 \tag{2a}$$

$$\eta_{y_d}^\top \mathbb{E}[\bar{Z}_d] - \eta_y^\top \mathbb{E}[\bar{Z}_d] \geq \Delta(y,y_d) - \xi_d \tag{2b}$$

where $\bar{Z}_d$ denotes the random variable corresponding to $\bar{z}$ in the d-th document, and the expectation $\mathbb{E}[\bar{Z}_d]$ is taken with respect to the variational distribution $q(\cdot)$. Without the term $-\mathscr{L}(q)$ in Eq. 2, the optimization problem in Eq. 2 simply defines a multi-class SVM [8] with $\mathbb{E}(\bar{Z}_d)$ being the feature vector, ξ_d being the slack variable associated with each document. $\Delta(y,y_d)$ is a loss function indicating the cost of misclassifying y_d to be y. In classification problem, we typically use the 0-1 lose, i.e. $\Delta(y,y_d) = 1$ if $y \neq y_d$, and $\Delta(y,y_d) = 0$ otherwise.

The rationale of MedLDA is that we want to find a distribution on latent topic representation Z_d for the d-th document and a model distribution q which satisfy: (1) the expectation of latent topic representation $\mathbb{E}[\bar{Z}_d]$ tends to produce a good classifier (in the typical max-margin sense) when used as feature vectors; (2) the model distribution explains the data well, i.e. by minimizing $-\mathscr{L}(q)$. MedLDA uses the generative part (i.e. $-\mathscr{L}(q)$) of the model as a regularization to the max-margin discriminative learning. Without this regularization, the max-margin learning is not sensible since we can arbitrarily assign latent topics to perfectly separate the training data.

3 Max-Margin LDA

In this section, we introduce the max-margin LDA. Our model extends MedLDA to make it more suitable for vision tasks.

First, the feature vector used for classification in MedLDA is constructed only from the latent topic representation $\mathbb{E}(\bar{Z}_d)$ of the document. However, for many vision applications, the raw feature representations (e.g. bag-of-words histogram on vector-quantized local patches) usually already give very good performance (see the results in Sec. 4), there is no reason to ignore the raw features when learning a classifier. More importantly, the feature vector based on latent topics implicitly assumes a bag-of-words image representation. When local features (e.g. SIFT [15] descriptors computed from interest points) are used, it is relatively straightforward to represent an image as a bag of words. But for many image classification problems, global image features (e.g. GIST [16]) can be effective too. It is not clear how to use GIST features in the bag-of-words representation, hence we cannot use MedLDA together with GIST features. In Sec. 3.1, we present a variant of MMLDA for image classification (we call it MMLDA^c). The advantage of MMLDA^c is that it can be used together with any feature representations.

Second, MedLDA is only for standard classification problems, where each data instance is associated with a single class label. When it comes to image annotation, MedLDA cannot be directly applied, since each image can be associated with multiple annotation terms. In Sec. 3.2, we propose a model called MMLDA^a to address those multi-label classification problems arising in image annotation.

3.1 MMLDA^c

In this section, we present the MMLDA^c model for image classification problems. We use $\mathbf{x}$ to denote an image. We use $\mathbf{w}$ to denote the bag-of-words representation of $\mathbf{x}$, e.g. $\mathbf{w}$ can be obtained by vector-quantization of SIFT descriptors. The topic assignment of the words in the document is denoted by $\mathbf{z}$. We assume a linear discriminative function of the form $F(y,\mathbf{z},\mathbf{w},\mathbf{x},\eta) = \eta_y^\top f(\mathbf{z},\mathbf{w},\mathbf{x})$. Note the definition of $F(\cdot)$ is similar to that in MedLDA. In fact, if we assume $f(\mathbf{z},\mathbf{w},\mathbf{x}) = \bar{z} = \frac{1}{N}\sum_{n=1}^N z_n$, we can recover $F(\cdot)$ in MedLDA. So the definition of $F(\cdot)$ in MMLDA is a strict generalization of that in MedLDA. One important thing to remember is that since $\mathbf{z}$ is not observed, $f(\mathbf{z},\mathbf{w},\mathbf{x})$ is actually a random vector implicitly defined by the distribution on Z.

We assume $f(\mathbf{z},\mathbf{w},\mathbf{x})$ is a concatenation of two sub-vectors $f(\mathbf{z},\mathbf{w},\mathbf{x}) = \text{cat}(\bar{z}; g(\mathbf{w},\mathbf{x}))$, where $g(\mathbf{w},\mathbf{x})$ is a vector defined on $\mathbf{w}$ and $\mathbf{x}$, $\bar{z}$ is defined as $\bar{z} = \frac{1}{N}\sum_{n=1}^N z_n$ similar to sLDA and MedLDA, $\text{cat}(\mathbf{a};\mathbf{b})$ denotes the concatenation of two vectors $\mathbf{a}$ and $\mathbf{b}$. Notice that we do not have any assumption on the form of $g(\mathbf{w},\mathbf{x})$, it can be any feature vector extracted from the image, e.g. histogram of words, GIST descriptors, or both. Similarly, we assume η_y is also a concatenation of two sub-vectors $\eta_y = \text{cat}(\zeta_y; \nu_y)$, so that $\eta_y^\top f(\mathbf{z},\mathbf{w},\mathbf{x}) = \zeta_y{}^\top \bar{z} + \nu_y{}^\top g(\mathbf{w},\mathbf{x})$. Fig. 1 (a) shows a graphical illustration of MMLDA^c. Similar to MedLDA, we learn the model parameter by solving an optimization problem as follows:

$$\min_{q,\alpha,\beta,\eta,\xi} -\mathscr{L}(q) + \frac{1}{2}\lambda||\eta||^2 + \tau\sum_{d=1}^D \xi_d, \quad \text{s.t.}\forall d, y: \quad \xi_d \geq 0 \tag{3a}$$

$$(\eta_{y_d}^\top - \eta_y^\top)\mathbb{E}[f(\bar{Z}_d, \mathbf{w}_d, \mathbf{x}_d)] \geq \Delta(y, y_d) - \xi_d \tag{3b}$$

A minor difference from MedLDA is that, we have used two regularization parameters λ and τ (instead of one parameter C in MedLDA) in Eq.3 to allow more flexibility in terms of the relative contribution of each term.

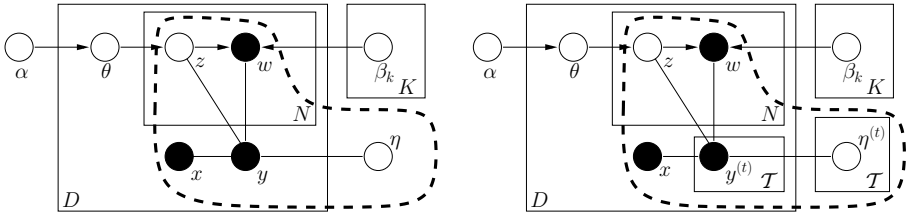


Figure 1: Graphical illustrations of (left) MMLDA^c for image classification; (right) MMLDA^a for image annotation. The variables enclosed by the dashed line are involved in the max-margin component of the model.

The optimization in Eq. 3 is generally intractable. But we can use a co-ordinate descent algorithm that iteratively optimizes over γ, ϕ, η, α and β. By writing the Lagrange of Eq. 3 and setting its derivative with respect to γ, ϕ, η, α and β to zero, we get a set of updating rules similar to those in MedLDA.

Optimize over γ: since the constraints in Eq. 3 do not involve γ, the updating rules similar to MedLDA: $\gamma \leftarrow \alpha + \sum_{n=1}^{N} \phi_{dn}$.

Optimize over ϕ: For a document d and each word i, we set $\partial L / \partial \phi_{di} = 0$ and get the following updating rule:

$$\phi_{di} \propto \exp\left(\mathbb{E}[\log\theta|\gamma] + \mathbb{E}[\log p(w_{di}|\beta)] + \frac{\tau}{N}\sum_{y\neq y_d}\mu_d(y)(\zeta_{y_d} - \zeta_y)\right) \tag{4}$$

Note that each term in Eq. 4 is a K-dimensional vector. Eq. 4 without the last term exactly corresponds to the updating rules of ϕ_{di} in unsupervised LDA [7]. The updating rule is also similar to that in MedLDA. The only difference is that for each η_y, we only need a subset of its elements (i.e. ζ_y) in the updating equation. The Lagrange variables $\mu_d(y)$ are obtained when optimizing L over η (see below).

Optimize over β: This optimization can be done via the following updating rules

$$\beta_{k,w} \propto \sum_{d=1}^{D}\sum_{n=1}^{N} 1(w_{dn} = w)\phi_{dhk} \tag{5}$$

Optimize over α: The optimization over α can be done using a Newton-Raphson iterative method identical to that in unsupervised LDA [7].

Optimize over η: When fixing all the other parameters, the optimization over η amounts to solve the following optimization problem similar to a multi-class SVM [8]:

$$\min_{\eta,\xi} \frac{1}{2}||\eta||^2 + \frac{\tau}{\lambda}\sum_{d=1}^{D}\xi_d, \quad \text{s.t.} \forall d, y: \quad \xi_d \geq 0 \tag{6a}$$

$$(\eta_{y_d}^\top - \eta_y^\top)\mathbb{E}[f(\bar{Z}_d, \mathbf{w}_d, \mathbf{x}_d)] \geq \Delta(y, y_d) - \xi_d \tag{6b}$$

When fixing the remaining parameters other than η, we can easily get $\mathbb{E}[\bar{Z}_d] = \frac{1}{N}\sum_{n=1}^{N}\phi_{dn}$. Combining with the fact that $g(\mathbf{w}, \mathbf{x})$ is a fixed vector that does not depend on $q(\cdot)$, we can get the following:

$$\mathbb{E}[f(\bar{Z}_d, \mathbf{w}_d, \mathbf{x}_d)] = \text{cat}\Big(\mathbb{E}[\bar{Z}_d]; g(\mathbf{w}, \mathbf{x})\Big) = \text{cat}\Big(\frac{1}{N}\sum_{n=1}^{N}\phi_{dn}; g(\mathbf{w}, \mathbf{x})\Big) \tag{7}$$

In summary, we simply compute a feature vector according to Eq. 7, then plug it into the optimization problem in Eq. 6. When optimizing Eq. 6, we also need to keep track of the Lagrange dual variable $\mu_d(y)$ associated with each constraint, which are needed for computing ϕ for the next iteration. We use the SVM implementation in [8] to solve Eq. 6.

The algorithm of MMLDA^c is very similar to MedLDA. The main difference is that the SVM involved in the optimization over η can use richer feature representations, while the SVM in MedLDA only uses features constructed from latent topics. As a result, the dual variables $\mu_d(y)$ obtained from solving Eq. 6 are different from those in MedLDA as well. Since those dual variables are involved in the optimization of latent topics (see Eq. 4), the latent topics discovered in MMLDA^c will be different from those in MedLDA.

3.2 MMLDA^a

Both MedLDA and MMLDA^c are for standard classification problems, where each datum $\mathbf{x}$ is associated with a single label $y \in \mathscr{Y}$. In this section, we introduce another model called MMLDA^a for the scenario where each $\mathbf{x}$ is associated with more than one label. An important application of this scenario is image annotation. In image annotation, the goal is to use a set of tags to describe a given image. This is not a standard classification problem because an image can be associated with more than one tag, see examples in Fig. 2.

Let us assume an image annotation task with $\mathscr{T}$ possible tags. For a given image $\mathbf{x}$, our goal is to predict a $\mathscr{T}$ dimensional binary vector $\mathbf{y} = (y^{(1)}, y^{(2)}, ..., y^{(\mathscr{T})})$, where $y^{(t)}$ is 1 or 0 indicating the presence/absence of the t-th tag for this image. One simple solution is to formulate image annotation as $\mathscr{T}$ binary classification problems, where a binary classifier is learned separately to predict the presence/absence of each tag based on the image features. During testing, we run $\mathscr{T}$ binary classifiers to independently predict whether each tag should be chosen to describe an unseen image. The disadvantage of this approach is that those $\mathscr{T}$ are learned independently of each other, but the annotations of an image are usually correlated, e.g. annotation terms such as "car", "road", "sky" tend to appear together. There are different ways to exploit the correlation among annotations. Our approach is inspired by the hypothesis speculated in [14] that there exists a latent low-dimensional feature space that are shared by classifiers for different tags. In [14], a low-rank matrix factorization approach is used to exploit this low-dimensional space. In our work, we directly use the latent topics as the low-dimensional space shared by tag classifiers.

We assume the classifier for the t-th tag is a binary linear SVM taking the feature vector defined in Eq. 7 as its input feature. We use $\eta^{(t)}$ to denote the parameter of this SVM classifier. Accordingly, $\eta^{(t)}$ has two sub-parts corresponding to vector obtained via latent topics $\mathbb{E}(\bar{Z}_d)$, and the vector obtained via $g(\mathbf{w}, \mathbf{x})$. The training data are in the form of $\{\mathbf{x}_d, \mathbf{y}_d\}_{d=1}^D$, where the label $\mathbf{y}_d$ is a $\mathscr{T}$ dimensional binary vector $\mathbf{y}_d = (y_d^{(1)}, y_d^{(2)}, ..., y_d^{(\mathscr{T})})$, and $y_d^{(t)} = 0$ or 1. We use $\Delta^a(\mathbf{y}, \mathbf{y}_d)$ to denote the loss incurred by predicting $\mathbf{y}$ while the ground-truth annotation is $\mathbf{y}_d$. We assume $\Delta^a(\mathbf{y}, \mathbf{y}_d)$ decomposes into the summation of $\mathscr{T}$ per-tag losses as $\Delta^a(\mathbf{y}, \mathbf{y}_d) = \sum_{t=1}^{\mathscr{T}} \Delta(y^{(t)}, y_d^{(t)})$, where $y_d^{(t)}$ denotes the ground-truth label for the t-th tag for document d, $y^{(t)}$ denotes an arbitrary label for the t-th tag. The loss $\Delta(y^{(t)}, y_d^{(t)})$ is the 0-1 loss.

Then we can formulate the multi-label image annotation using the following optimization problem:

$$\min_{q,\alpha,\beta,\eta,\xi} -\mathscr{L}(q) + \frac{1}{2}\lambda \sum_{t=1}^{\mathscr{T}} ||\eta^{(t)}||^2 + \tau \sum_{d=1}^{D} \sum_{t=1}^{\mathscr{T}} \xi_d^{(t)}, \quad \text{s.t.} \quad \forall d, t, y^{(t)}: \quad \xi_d^{(t)} \geq 0 \tag{8a}$$

$$\eta_{y_d^{(t)}}^{(t)\top} \mathbb{E}[f(\bar{Z}_d, \mathbf{w}_d, \mathbf{x}_d)] - \eta_{y^{(t)}}^{(t)\top} \mathbb{E}[f(\bar{Z}_d, \mathbf{w}_d, \mathbf{x}_d)] \geq \Delta(y^{(t)}, y_d^{(t)}) - \xi_d^{(t)} \tag{8b}$$

Fig. 1 (b) shows a graphical illustration of MMLDA^a. We use the same iterative method in MMLDA^c to optimize Eq. 8. Optimizing Eq. 8 with respect to γ, β, α leads to updating rules identical to those in MMLDA^c. The only differences are in the optimization over ϕ and η.

Optimize over η: when fixing the remaining parameters, the optimization over η can be performed separately for each tag $t \in \{1,2,...,\mathscr{T}\}$. For a fixed t, we need to solve the following optimization problem:

$$\min_{q,\alpha,\beta,\eta,\xi} \frac{1}{2}\lambda||\eta^{(t)}||^2 + \tau\sum_{d=1}^{D}\xi_d^{(t)}, \quad \text{s.t.} \quad \forall d, y^{(t)}: \quad \xi_d^{(t)} \geq 0 \tag{9a}$$

$$\eta_{y_d^{(t)}}^{(t)\top}\mathbb{E}[f(\bar{Z}_d, \mathbf{w}_d, \mathbf{x}_d)] - \eta_{y^{(t)}}^{(t)\top}\mathbb{E}[f(\bar{Z}_d, \mathbf{w}_d, \mathbf{x}_d)] \geq \Delta(y^{(t)}, y_d^{(t)}) - \xi_d^{(t)} \tag{9b}$$

Eq. 9 is equivalent to a binary SVM and can be solved using the same technique in [8].

Optimize over ϕ: the main difference between MMLDA^a and MMLDA^c lies in the optimization over ϕ.

$$\phi_{di} \propto \exp\left(\mathbb{E}[\log\theta|\gamma] + \mathbb{E}[\log p(w_{di}|\beta)] + \frac{\tau}{N}\sum_{t=1}^{\mathscr{T}}\sum_{y^{(t)}}\mu_d^{(t)}(y^{(t)})(\zeta_{y_d^{(t)}}^{(t)} - \zeta_{y^{(t)}}^{(t)})\right) \tag{10}$$

We can compare Eq. 10 with Eq. 4 to see the difference. The first two terms in Eq. 4 and Eq. 10 come from the unsupervised LDA. The third term in Eq. 4 biases ϕ_{di} towards a distribution that favors a more accurate classification. The third term in Eq. 10, on the other hand, biases ϕ_{di} towards more accurate annotations for *all* the possible tags (notice the summation over all t in Eq. 10).

4 Experiments

We test our models on two real-world datasets containing both class labels and annotations: a subset from LabelMe [18], and the UIUC sport dataset [12]. Both datasets have been used in [20]. The LabelMe dataset contains images of 8 different scene categories: "coast", "forest", "highway", "inside city", "mountain", "open country", "street" and "tall building". Similar to [20], we remove annotation terms occurring fewer than 3 times. On average there are about six annotation terms per image in the LabelMe dataset. We randomly choose half of the data as the test set. From the other half, we randomly select 50 images from each class to form the validation set. The remaining data are used as the training set. The UIUC sport dataset contains 8 sport classes: "badminton", "bocce", "croquet", "polo", "rock climbing", "rowing", "sailing" and "snowboarding". On average there are about seven annotation terms per image in this dataset. We split the dataset into training, validation, and test sets in a way similar to LabelMe.

Following [20], we extract the 128-dimensional SIFT [15] descriptors densely selected on a sliding grid. Those SIFT descriptors are clustered to form the codebook. We report results using the codebook of size 250. We have tried other codebook sizes and the results are similar.

Image classification: We compare the overall classification accuracies of our proposed models with several baseline methods in Table 4. We have tried several different ways of using our models, denoted as *MMLDA^c*, *MMLDA^c + GIST*, *MMLDA^c + GIST* and *MMLDA^c + SIFT + GIST* in Table 4. *MMLDA^c* only uses the latent topic representation as the feature vector. *MMLDA^c+SIFT* uses the concatenation of latent topic representation and

method	LabelMe	UIUC sport
sLDA	81.66	72.23
SVM+SIFT	79.54	72.74
SVM+GIST	79.63	72.61
SVM+SIFT+GIST	79.46	72.73
MMLDA^c	**81.74**	**74.65**
MMLDA^c+SIFT	**84.53**	**76.05**
MMLDA^c+GIST	**86.05**	**82.17**
MMLDA^c+SIFT+GIST	**86.73**	**83.06**

Table 1: Image classification accuracies (%) on the LabelMe and UIUC sport datasets.

method	LabelMe	UIUC sport
sLDA*	38.7	35.0
SVM+SIFT	46.25	45.75
SVM+GIST	45.98	45.35
SVM+SIFT+GIST	46.24	45.78
MMLDA^a	**46.64**	**44.51**
MMLDA^a+SIFT	**47.71**	**47.51**
MMLDA^a+GIST	**47.19**	**50.61**
MMLDA^a+SIFT+GIST	**47.94**	**52.49**

Table 2: Image annotation results in terms of F-measure (%) on the LabelMe and UIUC sport datasets. *The performance measurement of sLDA is taken from [20] and is not directly comparable to others in the table.

the histogram of visual words as the feature vector. *MMLDA^c+GIST* uses the concatenation of latent topic representation and the GIST descriptor as the feature vector. *MMLDA^c+SIFT+GIST* uses the concatenation of all three vectors: latent topics, histogram of visual words, GIST descriptors. For baseline methods, we compare with linear SVM trained with various combinations of feature vectors (histogram of visual words, GIST). We also compare with the sLDA model by running the code[1] of [20] on our dataset. We use the validation set to set the free parameters in each method (e.g. the C parameter in SVM, or the number of topics K in sLDA).

We can see that sLDA performs similarly with SVM. MMLDA^c alone only performs comparably to sLDA and SVM. But when we combine MMLDA^c with various raw features (e.g. SIFT,GIST), the performance is much better.

Image Annotation: For image annotation, we compare various methods using the F-measure in Table 2. The baseline method is a set of linear SVMs separately trained for predicting the presence/absence of each annotation term. Since the code for image annotation from [20] is not publicly available, we cannot run their method on our dataset. The performance measurements of sLDA shown in Table 2 are the published results reported in [20], it is important to remember that they are not directly comparable to other numbers in the tables, since they use different subsets of the datasets.

We can see that our proposed methods outperform other baseline methods on both datasets. Fig. 2 shows some examples of the annotations generated by MMLDA^a+SIFT+GIST.

[1] Available at `http://www.cs.princeton.edu/~chongw/slda`


sea water, sky


mountain, trees, sky, snowy mountain


car, mountain, road, sky, trees


building, door, road, sky, sidewalk, trees, window


mountain, trees, sky

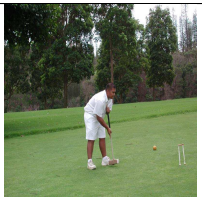

field, mountain, sky, trees


building, car, road, sidewalk, sky

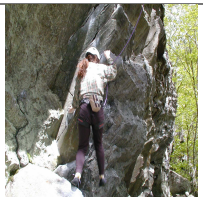

building occluded, skyscraper occluded, building, sky,


athlete, audience, floor, net, wall


athlete, ball, grass, mallet, plant, wicket


athlete, ball, grass, lawn, mallet, plant, tree, wicket


athlete, grass, horse, mallet, tree


climber, plant, rock, rope


athlete, oar, rowboat, tree, water


athlete, sky, sailing boat, water

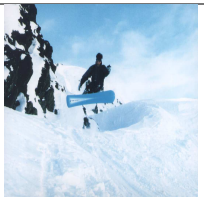

ski, skier, sky

Figure 2: Examples of image annotation. The first two rows are examples from the LabelMe dataset. The last two rows are examples from the UIUC sport dataset.

5 Conclusion and Future Work

We have presented the max-margin latent Dirichlet allocation (MMLDA) that uses the max-margin criterion to train topic models for image classification and annotation. Our experimental results on two benchmark datasets show the promise of MMLDA. As future work, we like to extend our model to perform image classification and annotation jointly. We also like to improve our model by adapting more advanced inference algorithms recently proposed for topic models, e.g. the collapsed variational inference [19].

Acknowledgement: This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] 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.

[2] Tamara L. Berg, Alexander C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee Whye Teh, Erik Learned-Miller, and David A. Forsyth. Names and faces in the news. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[3] Alessandro Bissacco, Ming-Hsuan Yang, and Stefano Soatto. Detecting humans via their pose. In *Advances in Neural Information Processing Systems*, volume 19, pages 169–176. MIT Press, 2007.

[4] David M. Blei and Michael I. Jordan. Modeling annotated data. In *Proc. of 26th International Conference on Research and Development in Information Retrieval(SIGIR)*, 2003.

[5] David M. Blei and John D. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems*, volume 18, 2006.

[6] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20. MIT Press, 2008.

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

[8] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[10] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of Twenty-Second Annual International Conference on Research and Development in Information Retrieval(SIGIR)*, pages 50–57, 1999.

[11] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems*, volume 21. MIT Press, 2008.

[12] Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[13] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[14] Nicolas Loeff and Ali Farhadi. Scene discovery by matrix factorization. In *European Conference on Computer Vision*, 2008.

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

[16] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3): 145–175, 2001.

[17] Bryan C. Russell, Alexei A. Efros, Josef Sivic, William T. Freeman, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[18] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

[19] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2006.

[20] Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[21] Yang Wang and Greg Mori. Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1762–1774, 2009.

[22] Jun Zhu, Amr Ahmed, and Eric P. Xing. MedLDA: Maximum margin supervised topic models for regression and classification. In *International Conference on Machine Learning*, 2009.