Spatial-DiscLDA for Visual Recognition

Zhenxing Niu[1] Gang Hua[2] Xinbo Gao[1] Qi Tian[3]
[1]VIPS Lab, Xidian University
[2]IBM T. J. Watson Research Center
[3]University of Texas at San Antonio
zhenxingniu@gmail.com ghua@us.ibm.com xbgao.xidian@gmail.com qitian@cs.utsa.edu

Abstract

Topic models such as pLSA, LDA and their variants have been widely adopted for visual recognition. However, most of the adopted models, if not all, are unsupervised, which neglected the valuable supervised labels during model training. In this paper, we exploit recent advancement in supervised topic modeling, more particularly, the DiscLDA model for object recognition. We extend it to a part based visual representation to automatically identify and model different object parts. We call the proposed model as Spatial-DiscLDA (S-DiscLDA). It models the appearances and locations of the object parts simultaneously, which also takes the supervised labels into consideration. It can be directly used as a classifier to recognize the object. This is performed by an approximate inference algorithm based on Gibbs sampling and bridge sampling methods. We examine the performance of our model by comparing its performance with another supervised topic model on two scene category datasets, i.e., LabelMe and UIUC-sport dataset. We also compare our approach with other approaches which model spatial structures of visual features on the popular Caltech-4 dataset. The experimental results illustrate that it provides competitive performance.

1. Introduction

Originated from statistical natural language processing, topic model has been widely adopted for solving visual recognition problems. The representatives of them are the probabilistic Latent Semantic Analysis (pLSA) of Hofmann [1], and the Latent Dirichlet Allocation (LDA) of Blei [3]. Both are generative models for modeling the statistical relationships among documents, topics and vocabularies. Specifically, each document can be factorized into a probability distribution of topics, each of which is represented with a probability distribution of words. Hence they are also related to nonnegative matrix factorization and dimensionality reduction algorithms.

To employ topic models for image recognition, each image is represented by a set of quantized image features such as the SIFT features [12], namely visual words. Therefore, each image is considered to be a visual document composed of a bag of visual words [9], to which topic models such as pLSA [5], LDA [6], and corr-LDA [4] can be directly applied.

Nevertheless, most of adopted topic models for visual recognition in the past, if not all, are unsupervised. During training, important object category labels are neglected, which is undesired. This is the reason that they usually adopt the topic models for visual representation [2], and pose another layer of discriminative classifiers for recognition, *e.g.*, k-nearest neighbor classifiers or SVM classifiers. This treatment is less desirable because the two model training steps are performed in a separate fashion with different objectives, which may result in inferior results to a unified model.

Feifei and Perona [6] proposed a Bayesian hierarchical model based on LDA with the category label as a hidden variable. Fritz and Schiele [7] use LDA to learn a compact and low dimensional representation for multiple visual categories from multiple view points in an unsupervised fashion. Nevertheless, the training of the model is still unsupervised. Another issue of naively applying topic models to a bag-of-words representation is that important spatial structure of the object is discarded. As already manifested by several previous work [9][8][22][24][25], spatial structure modeling can not only improve recognition in static images, but also enable object discovery and localization in video sequences. For describing visual scenes, Sudderth *et al*. [8] proposed a Transformed Dirichlet Process to model the expected spatial locations of objects, and the appearance of visual features corresponding to each object.

To address the two issues discussed above, we extend a recent advancement of supervised topic models, namely the DiscLDA [21], to a part based visual representation for object category recognition. It explicitly introduces the discriminative category information in the generative topic model. In essence, it naturally defines multiple generative models pertaining to each specific visual category, which are all cast under a unified Bayesian model.

Our proposed representation also simultaneously models statistics of the appearances and locations of the different

parts of the object, which are automatically learnt without any manual specification. Comparing with previous part based representation such as the constellation model [10] and k-fans model [11], which all strive for modeling the spatial relationships among different parts, our part based representation only models the spatial distribution of each of the parts, which largely avoids the combinatory explosion of the hypothesis space. We call our proposed model to be Spatial-DiscLDA (*S-DiscLDA*).

In [9], another supervised topic model, namely sLDA has been examined for simultaneous image classification and annotation. It largely focuses on modeling the relationship between the class labels, *e.g.*, scene categories such as "outdoor", and image annotation, *e.g.*, object categories such as "tree", "flower", and "sky". So it is proper to model a scene, where there are several objects in a scene. However, the S-DiscLDA model explicitly exploits the probability distribution over vocabulary for different categories and hence may bear more capacity for object recognition then sLDA.

Our *S-DiscLDA* can be directly used as a classifier for visual recognition, where inference is performed by an approximate inference algorithm based on Gibbs sampling and bridge sampling methods [20]. Our contributions are hence three folds:

- We extended a recently developed supervised topical model, DiscLDA, to spatial modeling for solving visual recognition problems.
- We presented an effective part based visual representation, which simultaneously model the statistics of the appearances and locations of different parts of a visual object.
- Our model explicitly exploits the probability distribution over the vocabulary for different parts of different visual categories, and present more modeling capacity.

The remainder of the paper is organized as the following: Section 2 illustrates the part-based visual representation based on the proposed *S-DiscLDA* model. The inference and learning algorithms are presented in Section 3. Detailed experimental results are summarized and discussed in Section 4. Finally, we conclude with remarks on future work in Section 5.

2. Image representation for S-DiscLDA Model

It is common observation that an object is usually comprised of parts at different spatial locations. The appearances of the different parts are usually different. So an object can be naturally characterized by a set of parts with a certain spatial arrangement, while each part can also be represented by visual appearances of the corresponding set of image patches. Our model is intended to find where these parts are statistically located for an object category,

(a) (b)

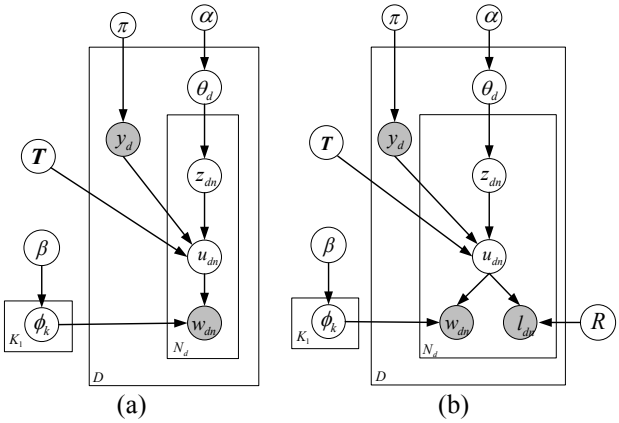
Fig.1 The spatial arrangement of (a) parts for an object and (b) elements for a scene. The appearances for different parts or elements are different.

and exploit the probability distribution of patch appearances for different parts.

The idea behind our model is that we use latent variables, *i.e.*, topics in S-DiscLDA model to characterize the object parts. The value of a latent variable for every image patch indicates which object part it comes from. By parameter estimation of our model, we can discover how the parts of an object are spatially arranged, and what the visual appearance is for each object part.

As shown in Fig.1, for an object recognition task (Fig.1(a)), the motorbike contains several parts, *i.e.*, wheel, seat, tail, *etc*. These parts present consistent spatial relationships, *e.g.*, seat is usually at the top, wheel is usually at the bottom, *etc*. For scene recognition task, we can consider the visual scene as an "object" in a general sense, and consider the different visual elements of the scene image as the "part". Take "coast" scene image as an example (Fig.1(b)), it contains several "parts", such as sky, beach, sea, *etc*. Regarding the spatial arrangement of parts, sky is usually on the top, and sea is usually at the bottom; regards the appearances of parts, the color of sea is usually blue, and the color of beach is usually white.

For each part, it consists of image patches with different visual appearances, which are described by the local



(a) (b)

Fig.2. The graphical model of (a) DiscLDA and (b) the proposed S-DiscLDA model. In S-DiscLDA, the location of image patch is modeled with variable l_{dn} and parameter set R .

Table 1. The notations of variables and parameters in S-DiscLDA, similar to those in [21]

variable/parameters	notations
$p_{dn}, (dn = 1,2,...,N_d)$	The *dn*-th patch in image I_d
$y_d \in \{1,2,...,C\}$	The supervised label of I_d
π	The parameter of the prior distribution of y_d
α, β	The parameters of Dirichlet distribution.
$\theta_d : \theta_d \sim Dir(\alpha)$	The instance of Dirichlet distribution for I_d, which is also the parameter of multinomial distribution of z_{dn}.
$z_{dn} \in \{1,2,...,K_0+1\}$ where $z_{dn} \sim Multi(\theta_d)$	The latent topic of p_{dn} without considering y_d, which obeys multinomial distribution.
$u_{dn} \in \{1,2,...,K_1+1\}$ where $(K_1 = CK_0)$, and $u_{dn} \sim Multi(T^{y_d}\theta_d)$	The latent topic of p_{dn} considering y_d, which obeys multinomial distribution.
$\mathbb{T} : \{T^c\}_{c=1}^{C}$	The mapping matrices between z_{dn} and u_{dn} for image category $y_d = c$.
$w_{dn} \in \{vw_i\}_{i=1}^{V}$ where $w_{dn} \sim Multi(\phi_k)$	The visual word of p_{dn}, which obeys multinomial distribution.
$\Phi : \{\phi_k\}_{k=1}^{K_1+1}$ where $\phi_k \sim Dir(\beta)$	The instance of Dirichlet distribution for $u_{dn} = k$, which is also the parameter of multinomial distribution of w_{dn}.
$l_{dn} = (x_{dn}, y_{dn})$	The location of p_{dn}, which obeys distribution as equation (2).
R: $\{\mu_k, \sigma_k\}_{k=1}^{K_1}$	The parameter of Gaussian distribution of l_{dn} for $u_{dn} = k$.

Table.2. The generative process of an image.

1.Draw topic proportions $\theta_d \sim Dir(\alpha)$;

2.For each patch p_{dn}, choose its topic z_{dn} drawn from the multinomial distribution, $z_{dn} \sim Multi(\theta_d)$;

3.For each patch p_{dn}, choose its topic $u_{dn} = T^{y_d} z_{dn}$ based on the image category $y_d \in \{1,2,...,C\}$. It indicates that topic u_{dn} drawn from the multinomial distribution, $u_{dn} \mid \theta_d, y_d, T^{y_d} \sim Multi(T^{y_d}\theta_d)$;

4.For each patch $p_{dn}, dn \in \{1,2,...,N_d\}$:

a) Draw its visual word $w_{dn} \sim Multi(\phi_k)$ for $u_{dn} = k$;

b) Draw its location $l_{dn} = (x_{dn}, y_{dn})$ from $prob(l_{dn} \mid u_{dn}, R)$

features of the patches. SIFT [12] is a local descriptor for image matching, in which it extracts a visual descriptor from each image patch surrounding a scale invariant keypoint for matching, which has been widely used for visual recognition problems [6][24][25][26].

The S-DiscLDA model can be illustrated by a graphical model as shown in Fig.2(b). We proceed to present the detailed mathematics of the proposed S-DiscLDA model. The notations of the variables and parameters in the model are presented in Table 1. The generative process of an image with our model can be described in Table 2.

For ease of presentation, we employ the notations similar to those in [21]. There are C kinds of objects to be recognized, and it is assumed that each object consists of K_0 parts. Each image I_d is divided into N_d image patches denoted as $p_{dn}, (dn = 1,2,...,N_d)$. Each patch is described by its appearance and location. Each patch's appearance is represented by its corresponding visual word $w_{dn} \in \{vw_i\}_{i=1}^{V}$, where $\{vw_i\}_{i=1}^{V}$ is the codebook which is obtained by running the k-means algorithm on local features extracted from patches sampled from all images. The location of the patch is represented by its image coordinates $l_{dn} = (x_{dn}, y_{dn})$.

We use latent variables, *i.e.*, topic z_{dn} and u_{dn} in S-DiscLDA to characterize the object parts. Specifically, topic z_{dn} for the patch p_{dn} indicates which object part it comes from without considering the object category. Since every image patch must come from either background or part of an object, so the latent topic z_{dn} is used to indicate where p_{dn} comes from. More specifically, if p_{dn} comes from the *m*-th part of object (no matter which category of the object belongs to), we have $z_{dn} = m, (m = 1,2,...,K_0)$, and if it comes from background, we have $z_{dn} = K_0 + 1$. So z_{dn} takes a number of $(K_0 + 1)$ values, and it has no relationship with object category.

To discriminatively model different parts of different objects, we introduce another latent topic u_{dn}. Specifically, if p_{dn} comes from the *m*-th part of the *c*-th object category, we have $u_{dn} = (c-1) \cdot K_0 + m$, and if it comes from background, we have $u_{dn} = K_1 + 1$, where $K_1 = C \cdot K_0$. So u_{dn} has a number of $(K_1 + 1)$ values, which directly relates to the object categories.

The relationship between topic z_{dn} and topic u_{dn} can be described by mapping matrices $\mathbb{T} : \{T^c\}_{c=1}^{C}$, which maps the object parts for any object category to the corresponding object parts for a specific object category $y_d = c$, *i.e.*, $T^c : z_{dn} \to u_{dn}$. T^c is a $(CK_0 + 1) \times (K_0 + 1)$ matrix. Take $C = 2$ as an example, $T^c, (c = 1,2)$ is as

$$T^1 = \begin{pmatrix} I_{K_0} & 0 \\ 0_{K_0} & 0 \\ 0 & 1 \end{pmatrix}_{(2K_0+1)\times(K_0+1)}, T^2 = \begin{pmatrix} 0_{K_0} & 0 \\ I_{K_0} & 0 \\ 0 & 1 \end{pmatrix}_{(2K_0+1)\times(K_0+1)} \quad (1)$$

where I_{K_0} is a K_0 dimension identity matrix.

For example, if the patch p_{dn} comes from the m-th part, *i.e.*, the value of its topic z_{dn} is m, and the object belongs to category $y_d = c$, then the value of its topic u_{dn} is $(c-1)K_0 + m$. This also applies to the appearance and location of image patches from background for all object categories except that it is an one-to-one mapping because $z_{dn} = K_0 + 1 \rightarrow u_{dn} = CK_0 + 1$.

In our model, the w_{dn} is drawn from a distribution over the codebook, which is a multinomial distribution with parameters $\Phi: \{\phi_k\}_{k=1}^{K_1+1}$. The parameter ϕ_k describes what the appearance looks like for background ($k = K_1 + 1$) or for object parts ($k = 1, 2, ..., K_1$).

The l_{dn} is drawn from a distribution $prob(l_{dn} \mid u_{dn}, R)$, which depends on its topic u_{dn}. The location distribution of patches from background ($k = K_1 + 1$) is modeled by a uniform distribution, and the location distribution of patches from an object part ($k = 1, 2, ..., K_1$) is modeled by a Gaussian model with parameter set R: $\{\mu_k, \sigma_k\}_{k=1}^{K_1}$, *i.e.*,

$$prob(l_{dn} \mid u_{dn} = k, R) = \begin{cases} N(\mu_k, \sigma_k), & k = 1, 2, ..., K_1 \\ Uniform, & k = K_1 + 1 \end{cases} \quad (2)$$

The position and scale of each object part can be described by the Gaussian parameter R: $\{\mu_k, \sigma_k\}_{k=1}^{K_1}$. With parameter estimation algorithm (cf. Section 3.1), we can discover where the parts are spatially arranged for an object.

3. Learning and inference algorithms

3.1. Parameter learning

Given a corpus of image data with class labels $\mathrm{A} = \{(y_d, \boldsymbol{w}_d, \boldsymbol{l}_d)\}_{d=1}^{D}$, we find the maximum likelihood estimation for the distribution of visual word for each topic, *i.e.,* Φ; and the distribution of location for each topic, *i.e.*, R.

$$\{\Phi^*, R^*\} = \arg\max_{\Phi, R} \left(\prod_d P(y_d, \boldsymbol{w}_d, \boldsymbol{l}_d \mid \mathbb{T}, \Phi, R) \right) \quad (3)$$

We proposed an iterative learning algorithm using EM for parameter estimation of S-DsicLDA, as summarized in Algorithm-1. Specifically, given R^{i-1}, the posterior distribution over the assignments of words to topics for each image I_d, *i.e.*, $P(\boldsymbol{z}_d, \boldsymbol{u}_d \mid \boldsymbol{w}_d, \boldsymbol{l}_d, y_d, \mathbb{T}, R^{i-1})$ is evaluated firstly; then the Φ^i and R^i are estimated by examining this posterior distribution. After some iterative steps, the two parameter set will converge to $\{\Phi^*, R^*\}$.

We use the Rao-Blackwellized version of Gibbs sampling method (RB Gibbs sampling) [19] to sample from

Algorithm1: The training of S-DiscLDA

Input: Image data with class labels $\mathrm{A} = \{(y_d, \boldsymbol{w}_d, \boldsymbol{l}_d)\}_{d=1}^{D}$

Output: The estimated parameters of the S-DiscLDA $\{\Phi^L, R^L\}$.

Initialization:

Initializing the $R^0\{\mu_k^0, \sigma_k^0\}$ manually.

Training:

For $i = 1, 2, ..., L$

1. **Sample from the posterior distribution.** Given R^{i-1}, Sample N RB Gibbs steps for each image I_d from $p(\boldsymbol{z}_d, \boldsymbol{u}_d \mid \boldsymbol{w}_d, \boldsymbol{l}_d, y_d, \mathbb{T}, R^{i-1})$ with Eq.(4)
2. **Estimate Φ^i and R^i**
   a) Estimate Φ^i with last sample $\boldsymbol{u}^{(N)}$ with Eq.(5)
   b) For each I_d, estimate $p(u_{dn} \mid \boldsymbol{w}_d, \boldsymbol{l}_d, y_d, \mathbb{T}, R^{i-1})$ with N samples $\boldsymbol{u}_d^{(t)}, (t = 1, 2, ..., N)$, and Estimate R^i with Eq.(6)

the distribution $p(\boldsymbol{z}_d, \boldsymbol{u}_d \mid \boldsymbol{w}_d, \boldsymbol{l}_d, y_d, \mathbb{T}, R)$. To sample from it, we need to compute

$$p(z_{dn} = h, u_{dn} = k \mid \boldsymbol{z}_{-dn}, \boldsymbol{u}_{-dn}, \boldsymbol{w}_d, \boldsymbol{l}_d, y_d, \mathbb{T}, R) \propto \frac{(\beta_{w_{dn}h} + m_{-dn,h}^{(w_{dn})})}{(\beta_{(\cdot)h} + m_{-dn,h}^{(\cdot)})} T_{hk}^{y} (\alpha_k + n_{-dn,k}) p(l_{dn} \mid u_{dn}, R) \quad (4)$$

where $m_{-dn,h}^{(w_{dn})}$ is the number of patches with w_{dn} and $z_{dn} = h$ in image I_d except for p_{dn}, and $n_{-dn,k}$ is the number of patches with $u_{dn} = k$ in image I_d except for p_{dn}. Here we omit certain details of the derivation due to the space. The reader may refer to the Appendix 6.1 for details.

Given the samples from the posterior distribution, we can estimate parameter set $\{\Phi, R\}$. Since the posterior is

$$p(\phi_k \mid \boldsymbol{u}, \boldsymbol{w}) = Dir(\{\beta_{vk} + m_k^{(v)}\}_{v=1}^{V}) \quad (5)$$

where $\boldsymbol{u} = \{\boldsymbol{u}_d\}_{d=1}^{D}$, $\boldsymbol{w} = \{\boldsymbol{w}_d\}_{d=1}^{D}$ and $m_k^{(v)}$ is the number of patches with $w_{dn} = v$ and $u_{dn} = k$ in all images $\{I_d\}_{d=1}^{D}$. So the $\Phi = \{\phi_k\}_{k=1}^{K_1+1}$ can be estimated as the posterior mean of $p(\phi_k \mid \boldsymbol{u}, \boldsymbol{w})$, which is simply the normalized Dirichlet parameters.

The spatial arrangement and the scale of different object parts are characterized by the mean values and variance of a Gaussian distribution, respectively. These form the parameter set R: $\{\mu_k, \sigma_k\}_{k=1}^{K_1}$. With the samples generated from the posterior distribution $p(\boldsymbol{z}_d, \boldsymbol{u}_d \mid \boldsymbol{w}_d, \boldsymbol{l}_d, y_d)$, we evaluate the probability $p(\boldsymbol{u}_d \mid \boldsymbol{w}_d, \boldsymbol{l}_d, y_d)$ by marginalizing out the topic variable $\boldsymbol{z}_d$. The probability $\omega_{dn}(k) \triangleq p(u_{dn} = k \mid \boldsymbol{w}_d, \boldsymbol{l}_d, y_d)$ indicates how likely the patch that comes from the object part k. In the learning algorithm, we take the weighted means and weighted variances as the robust estimation of the locations and scales, *i.e.*,

$$\begin{cases} \mu_k = \varpi(k)\cdot\sum_{d=1}^{D}\sum_{dn=1}^{N_d}\omega_{dn}(k)l_{dn} \\ \sigma_k^2 = \varpi(k)\cdot\sum_{d=1}^{D}\sum_{dn=1}^{N_d}\omega_{dn}(k)(l_{dn}-\mu_k)^2 \end{cases}, \quad k=1,...,K_1 \tag{6}$$

where $\varpi(k) = 1/\sum_{d=1}^{D}\sum_{dn=1}^{N_d}\omega_{dn}(k)$

3.2. Approximate Inference

To recognize an object in a new image, we first divide the new image into patches just as the training process, and extract the local features for each patch. For each patch p_{dn}, we get its corresponding image coordinate l_{dn}, and quantify its local feature into visual words w_{dn} with the codebook we obtained in the training process. With the known model parameters $\{\mathbb{T},\Phi,R\}$ in the training process, we can predict the image's label $\hat{y}_d$ by maximum a posteriori estimation,

$$\hat{y}_d = \arg\max_{y_d\in\{1,2,...,C\}} \left(P(y_d \mid \boldsymbol{w}_d,\boldsymbol{l}_d,\mathbb{T},\Phi,R)\right). \tag{7}$$

We propose an approximate inference algorithm to predict the label $\hat{y}_d$, as shown in Algorithm-2. By employing the notations as in [22], we define a function as $q_c(\boldsymbol{z}_d) \triangleq p(\boldsymbol{w}_d,\boldsymbol{l}_d \mid \boldsymbol{z}_d, y_d=c,\mathbb{T},\Phi,R)p(\boldsymbol{z}_d)$, and we denote its normalization constant as $Z_c \triangleq \int q_c(\boldsymbol{z}_d)d\boldsymbol{z}_d$. It is obvious that we have the relationship between the posterior probability $P(y_d \mid \boldsymbol{w}_d,\boldsymbol{l}_d,\mathbb{T},\Phi,R)$ and Z_c as

$$\frac{p(y_d=c \mid \boldsymbol{w}_d,\boldsymbol{l}_d,\mathbb{T},\Phi,R)}{p(y_d=1 \mid \boldsymbol{w}_d,\boldsymbol{l}_d,\mathbb{T},\Phi,R)} = \frac{p(y_d=c)}{p(y_d=1)}\frac{Z_c}{Z_1}, c=2,3,...,C \tag{8}$$

To estimate the posterior $P(y_d \mid \boldsymbol{w}_d,\boldsymbol{l}_d,\mathbb{T},\Phi,R)$, we need to estimate the ratio $Z_c/Z_1, c=2,3,...,C$. As an extension of importance sampling to estimate the ratio between two normalization factors, we leverage bridge sampling [19] to estimate it.

Define $\Delta^{y_d}_{w_{dn},z_{dn},l_{dn}} \triangleq p(w_{dn},l_{dn} \mid z_{dn},y_d,\mathbb{T},\Phi,R)$, we have

$$\frac{Z_c}{Z_1} \approx \frac{\sum_{i=1}^{M} h_{c1}(\boldsymbol{z}_d^{(i)1})}{\sum_{i=1}^{M} h_{1c}(\boldsymbol{z}_d^{(i)c})}, c=2,3,...,C \tag{9}$$

where $h_{ab} = \sqrt{\prod_{dn}\frac{\Delta^{y_d=a}_{w_{dn},z_{dn},l_{dn}}}{\Delta^{y_d=b}_{w_{dn},z_{dn},l_{dn}}}}$. The reader may refer to the Appendix 6.2 for detailed derivation.

To sample from $q_c(\boldsymbol{z}_d)$, we need to compute

$$\begin{aligned} &p(z_{dn}=h \mid \boldsymbol{z}_{-dn},\boldsymbol{w}_d,\boldsymbol{l}_d,y_d,\mathbb{T},\Phi,R) \\ &\propto (\alpha_h + n_{-dn,h})\Phi^{y_d}_{wh}\sum_k T^{y_d}_{kh}\, p(l_{dn} \mid u_{dn}=k,R) \end{aligned} \tag{10}$$

Again, detailed derivation is available in Appendix 6.3.

Algorithm2: The inference of S-DiscLDA

Input: The visual words and corresponding locations for the new image $(\boldsymbol{w}_d,\boldsymbol{l}_d)$

Output: The posterior probability $P(y_d \mid \boldsymbol{w}_d,\boldsymbol{l}_d,\mathbb{T},\Phi,R)$

1. Sample $\boldsymbol{z}_d^{(i)c}, i=1,2,...,M$ from $p(\boldsymbol{z}_d \mid y_c,\boldsymbol{w}_d,\boldsymbol{l}_d,\mathbb{T},\Phi,R)$ with Eq.(10).
2. Compute the ratio $Z_c/Z_1, c=2,3,..,C$ with Eq.(9).
3. Estimate the $P(y_d \mid \boldsymbol{w}_d,\boldsymbol{l}_d,\mathbb{T},\Phi,R)$ with Eq.(8).

4. Experiments

We evaluate our method with two experiments using three datasets, *i.e.*, the LabelMe dataset [11], the UIUC-Sport dataset [11], and the Caltech-4 dataset [16], followed by detailed discussions. The LabelMe dataset in [11] is obtained by an on-line tool with the following 8 scene categories, namely "highway", "inside city", "tall building", "street", "forest", "coast", "mountain", and "open country". The UIUC-Sport dataset has 8 scene categories, *i.e.*, "badminton", "bocce", "croquet", "polo", "rockclimbing", "rowing", "sailing", and "snowboarding". The Caltech-4 dataset has 5 categories, *i.e.*, "face", "motorbike", "airplane", "car" and "background".

4.1. Object recognition

In the first experiment on Caltech-4, we compare our method with other object recognition methods which also conduct spatial modeling over either a set of object parts [10], or a set of local features. The constellation model [10] attempts to represent an object by a set of parts under mutual geometric constraints, which simultaneously learns shape, appearance and relative scale represented by Gaussian densities using EM. It is an unsupervised model with high computation cost, which neglects the valuable labels during training.

In [16], the authors model the spatial relationship between visual words by extracting higher-order spatial features. It focuses on how to select lower order features and how to build higher order features. The authors use spatial histogram with distance approximately in log scale to build second order feature, and illustrate that the algorithm can avoid exhaustive computation. However, the spatial histogram is relatively coarse, and it cannot explicitly indicate the spatial arrangement of object parts.

We extract SIFT descriptors from densely sampled image patches for the S-DiscLDA model for this recognition task. Our S-DiscLDA uses the latent variables to discover the spatial arrangement of object parts, and it is a supervised model by exploiting the object category information. It can model the spatial relationship of all image patches, while the

higher-order spatial features [16] can only model the relative relationships among several image patches.

For comparison, we followed the same experimental setting as those adopted in [16] and [12]. The data of each category was randomly splitted into two subsets of equal size. The model was then trained on the first subset and tested on the second subset. For the S-DiscLDA model, β and α are set to 0.01 and $50/K_0$ respectively, $\mathrm{T}:\{T^y\}_{y=1}^{C}$ is set according to Equation (1), where $C=2$ for this task and we set $K_0=15$. These parameter settings will be further discussed in more detail in Section 4.3.

Similar to [12] and [16], we use ROC equal error rate (EER) to evaluate the classification performance, as shown in Table. 3. The experimental results demonstrate that the proposed model significantly outperforms the method [12], and also outperforms the method [16], except for the car category, where we obtain only slightly worse result. The slightly inferior result in the car category may be attributed to the large variation of the car size, which presents large spatial variation among the different parts.

Table 3. Comparison to method [12] and method [16]. The table gives ROC equal error rate (EER) on the Caltech-4 dataset.

Class	Method [12]	Method [16]	Our method
Face	3.6%	0.92%	**0.5%**
Motorbike	7.5%	1.0%	**0.9%**
Airplane	9.8%	1.75%	**1.0%**
Car	11.5%	**0.5%**	0.9%

4.2. Scene classification

Our model can also be used for scene classification. The scene image can be regarded as containing scene elements in different spatial locations, and the spatial arrangement of these elements is usually relatively consistent for images from same scene category and is diverse for images from different scene categories. In [11], Wang *et al.* exploited a supervised topic mode, *i.e.*, the sLDA for scene image classification. It focuses on modeling the relationships between the scene categories and image annotations. The image annotation in [11] is similar to the concept of scene element or part in S-DiscLDA. However, it does not explicitly exploit the spatial relationship of the different image annotations.

Our S-DiscLDA model is also a supervised model which models the appearance and location of image patches simultaneously. To compare with this sLDA based scene classification method, we followed the same experimental protocol adopted in [11], and performed the scene classification experiment on the LabelMe dataset and UIUC-Sport dataset. Similar to [11], we use the SIFT region descriptors extracted from a sliding grid (5×5). We report results on a codebook of 800 codewords (larger codebook sizes gave similar performance). For the S-DiscLDA model, we set $K_0=15$, and $C=8$ for this task, other parameters are similar to the setting in the previous section.

This task is a multi-class classification problem. The results are illustrated in the confusion matrix of Fig. 3, where Fig. 3(a) and Fig. 3(c) are the classification results from the sLDA method quoted from [11] on LabelMe and UIUC-Sport dataset respectively. And Fig. 3(b) and Fig. 3(d) presented the results of S-DiscLDA on corresponding datasets respectively. On LabelMe dataset, S-DiscLDA can reduce the error of sLDA [11] by at least 4% and achieve better results in 7 out of the 8 scene categories. On UIUC-Sport dataset, our models can reduce the error of sLDA [11] by at least 3% and also achieve better results in 7 out of the 8 scene categories.

4.3. Discussion on S-DiscLDA

The original DiscLDA model can also be directly employed for object recognition. More specifically, we utilize the DiscLDA as a classifier, and use the bag of visual words as the visual representation for an image, and use the object category label as the class label. Given a corpus of image data with class labels, $B=\{(y_d,\boldsymbol{w}_d)\}_{d=1}^{D}$, we train the DiscLDA with the learning algorithm proposed in [21] and obtain the model parameter set $\{\Phi\}$. For a test image, we

	highw.	insid.	tallb.	street	forest	coast	mount.	openc.
highw.	**77**	02	01	04	00	05	04	07
insid.	01	**75**	04	17	00	01	00	02
tallb.	01	06	**84**	04	01	00	03	01
street	05	17	04	**71**	00	00	01	02
forest	00	00	00	00	**92**	00	04	04
coast	10	00	00	00	01	**71**	01	17
mount.	03	00	01	01	02	00	**83**	10
openc.	07	00	01	01	08	12	10	**61**

(a) sLDA [11]. LabelMe: avg. accuracy: 76%.

	highw.	insid.	tallb.	street	forest	coast	mount.	openc.
highw.	**80**	00	00	02	00	08	03	07
insid.	00	**89**	00	03	03	00	03	02
tallb.	00	10	**86**	01	00	00	02	01
street	00	09	00	**74**	09	00	07	01
forest	00	01	00	01	**93**	00	05	00
coast	06	00	00	00	01	**73**	06	14
mount.	03	01	03	00	07	01	**77**	08
openc.	01	02	00	03	02	14	12	**66**

(b) S-DiscLDA. LabelMe: avg. accuracy: 80%

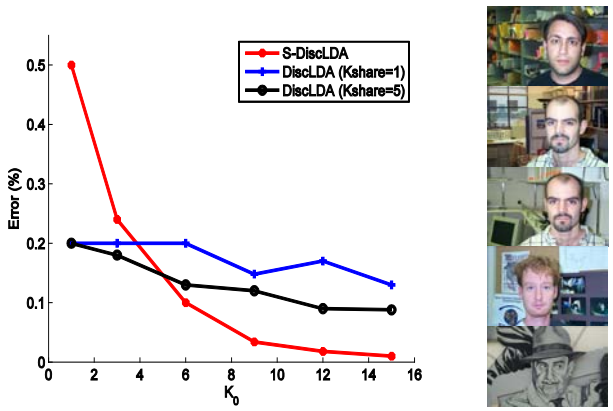
	badmi.	bocce	croquet	polo	rockc	rowing	sailing	snowb
badmi.	**82**	02	09	03	00	01	02	01
bocce	07	**26**	43	05	04	03	03	08
croquet	02	08	**70**	08	02	04	04	02
polo	10	03	22	**51**	01	06	02	05
rockc	00	02	01	03	**85**	01	01	07
rowing	03	02	10	07	01	**71**	03	03
sailing	03	00	11	03	01	07	**66**	09
snowb.	02	05	07	05	12	04	05	**60**

(c) sLDA [11]. UIUC-Sport: avg. accuracy: 65%

	badmi.	bocce	croquet	polo	rockc	rowing	sailing	snowb
badmi.	**85**	01	09	00	02	02	00	01
bocce	09	**30**	46	00	04	08	01	02
croquet	02	05	**79**	00	02	06	04	02
polo	09	03	25	**50**	06	06	00	01
rockc	01	00	04	00	**87**	04	00	04
rowing	01	00	09	07	01	**74**	03	05
sailing	00	02	09	03	02	09	**71**	04
snowb.	05	02	11	01	10	05	02	**64**

(d) S-DiscLDA. UIUC-Sport: avg. accuracy: 68%

Fig 3. Comparison with method [11] using confusion matrices. The rows denote true label and the columns denote the estimated label. All the numbers stand for percentage (%).



(a) (b)

Fig.4 (a) The performance comparison of object recognition between DiscLDA and S-DiscLDA on face dataset in Caltech-4. (b) Some samples that can be correctly recognized with S-DiscLDA but incorrectly recognized with DiscLDA.

get its visual words as the feature and predict the image label with the trained DiscLDA. In this way, the latent topic variables z_d and u_d have no spatial meaning, and the image patch location l_d is not modeled.

In this experiment, we provide a comparison between DiscLDA and S-DiscLDA model in object recognition. It can evaluate how much the spatial information improves the performance of object recognition. As shown in Fig.4(a), we use the face dataset in Caltech-4 to illustrate the performance difference. The experiments on the other three categories have similar results. In this experiment, we also explore the different configuration of K_{share} for DiscLDA.

The performance is evaluated by the error rate, *i.e.*, $Error = \frac{FP+FN}{Total}$. From Fig.4(a), we can clearly observe that the classification error of both models will decrease with the increase of K_0. This is because the dimension of parameters $\Phi: \{\phi_k\}_{k=1}^{K_1+1}$ and $R: \{\mu_k, \sigma_k\}_{k=1}^{K_1}$ depends on K_0. Hence the appearance and location of object parts can be characterized more precisely with large K_0 than that with small K_0. However, the classification error of S-DsicLDA will decrease faster than that of DiscLDA. So spatial modeling in S-DiscLDA can significantly improve its capacity for object recognition. From Fig.4(a), we can also observe that the performance only improves slightly after the number of parts increased beyond 15, *i.e.*, after $K_0 > 15$. So in our experiments presented in Section 4.1 and 4.2, we set $K_0 = 15$. Fig. 4(b) shows some samples that can be correctly recognized with S-DiscLDA but incorrectly recognized with DiscLDA. In these images, the background is cluttered, and it contains many patches with various appearances. So it is difficult to recognize the face without modeling the spatial structure.

The parameter β and α of Dirichlet distribution are the hyper parameters of both DiscLDA and S-DiscLDA. They can be interpreted as the virtual samples contributing to the smoothing of ϕ_k and θ respectively, which are set according to the setting in [19].

For the computational cost of S-DiscLDA, since the appearances and locations for different parts are modeled with conditional independent latent variables, S-DiscLDA is very efficient to model a large number of object parts. So it can be used for objects with many parts and complex structure.

5. Conclusions

In this paper, we propose a novel model S-DiscLDA for visual recognition. Our model aims to bridge the gap between bag of words model and visual modeling by taking spatial structure modeling into consideration. By considering the spatial relationships among local image features, we extend a recent supervised topic model, the DiscLDA to spatial modeling. Our model captures the appearance and location of the different object parts simultaneously, and leverages the supervised label information to facilitate automatic learning of different parts. Our experiments on both object recognition and scene recognition demonstrated the effectiveness of the proposed model. As we have noticed, there are some prior constraints on the spatial relationships of object parts. In our future work, we plan to further explore to augment the proposed model with this kind of prior knowledge, and hence improve the richness of the proposed model in knowledge representation.

6. Acknowledgement

This research was supported by the Ph.D. Programs Foundation of Ministry of Education of China under Grant 20090203110002, and the Research Projector Funding of Microsoft Research Asia. This work was supported in part to Dr. Qi Tian by NSF IIS 1052851, Google Faculty Research Award and FXPAL Research Award, respectively

7. Appendix

7.1. The derivation of Equation (4)

It is obvious that we have

$$p(z_{dn}, u_{dn} \mid \boldsymbol{z}_{-dn}, \boldsymbol{u}_{-dn}, \boldsymbol{w}_d, \boldsymbol{l}_d, y_d, \mathbb{T}, R) = \frac{p(z_d, \boldsymbol{u}_d, \boldsymbol{w}_d, \boldsymbol{l}_d \mid y_d, \mathbb{T}, R)}{p(\boldsymbol{z}_{-dn}, \boldsymbol{u}_{-dn}, \boldsymbol{w}_{-dn}, \boldsymbol{l}_{-dn} \mid y_d, \mathbb{T}, R)} . \quad (11)$$

Based on S-DiscLDA, the joint probability can be factorized as

$$p(\boldsymbol{z}_d, \boldsymbol{u}_d, \boldsymbol{w}_d, \boldsymbol{l}_d \mid \alpha, \beta, y_d, \mathbb{T}, R) = p(\boldsymbol{z}_d \mid \alpha) p(\boldsymbol{u}_d \mid \boldsymbol{z}_d, y_d, \mathbb{T}) p(\boldsymbol{w}_d \mid \boldsymbol{u}_d, \beta) p(\boldsymbol{l}_d \mid \boldsymbol{u}_d, R) . \quad (12)$$

The distributions of image patch locations are assumed

independent to each other in S-DiscLDA, so we have

$$p(\boldsymbol{l}_d \mid \boldsymbol{u}_d, R) = \prod_{dn} p(l_{dn} \mid u_{dn}, R) \tag{13}$$

Combining these equations together we get Equation (4).

7.2. The derivation of Equation (9)

With the definition of the normalization constant $Z_c \triangleq \int q_c(\boldsymbol{z}_d) d\boldsymbol{z}_d$, we have

$$\frac{Z_c}{Z_1} \approx \frac{\sum_{i=1}^{M} \frac{q_b}{q_1}(\boldsymbol{z}_d^{(i)1})}{\sum_{i=1}^{M} \frac{q_b}{q_c}(\boldsymbol{z}_d^{(i)c})}, c = 2,3,...,C \tag{14}$$

where $\boldsymbol{z}_d^{(i)c}, i = 1,2,...,M$ are M independent distribution samples from the distribution $q_c(\boldsymbol{z}_d)$, and the $q_b(\boldsymbol{z}_d)$ is the bridge distribution. In our algorithm, we use a geometric bridge [20], *i.e.*, $q_b(\boldsymbol{z}_d) = \sqrt{q_1(\boldsymbol{z}_d) q_c(\boldsymbol{z}_d)}$.

The joint distribution $p(\boldsymbol{z}_d, \boldsymbol{u}_d, \boldsymbol{w}_d, \boldsymbol{l}_d \mid \alpha, \beta, y_d, \mathbb{T}, R)$ can be factorized as shown Equation (12). By marginalizing out variable $\boldsymbol{u}_d$, we have

$$\begin{aligned} &\Delta_{w_{dn}, z_{dn}, l_{dn}}^{y_d} \triangleq p(w_{dn}, l_{dn} \mid z_{dn}, y_d, \mathbb{T}, \Phi, R) \\ &= \sum_k p(u_{dn} = k \mid z_{dn}, y_d, \mathbb{T}) p(w_{dn} \mid u_{dn} = k, \Phi) p(l_{dn} \mid u_{dn} = k, R) \end{aligned} \tag{15}$$

Given $\boldsymbol{z}_d$, the $\boldsymbol{w}_d$ and $\boldsymbol{l}_d$ are conditional independence, we have

$$\begin{aligned} q_c(\boldsymbol{z}_d) &\triangleq p(\boldsymbol{z}_d) p(\boldsymbol{w}_d, \boldsymbol{l}_d \mid \boldsymbol{z}_d, y_d, \mathbb{T}, \Phi, R) \\ &= p(\boldsymbol{z}_d) \prod_{dn} \Delta_{w_{dn}, z_{dn}, l_{dn}}^{y_d} \end{aligned} \tag{16}$$

Using geometric bridge $q_b(\boldsymbol{z}_d) = \sqrt{q_1(\boldsymbol{z}_d) q_c(\boldsymbol{z}_d)}$, we have

$$\frac{Z_c}{Z_1} \approx \frac{\sum_{i=1}^{M} \frac{q_b}{q_1}(\boldsymbol{z}_d^{(i)1})}{\sum_{i=1}^{M} \frac{q_b}{q_c}(\boldsymbol{z}_d^{(i)c})} = \frac{\sum_{i=1}^{M} \sqrt{\frac{q_c}{q_1}(\boldsymbol{z}_d^{(i)1})}}{\sum_{i=1}^{M} \sqrt{\frac{q_1}{q_c}(\boldsymbol{z}_d^{(i)c})}} \tag{17}$$

Combining Equation (16) and (17), we have Equation (9).

7.3. The derivation of Equation (10)

It is obvious that we have

$$\begin{aligned} &p(z_{dn} \mid \boldsymbol{z}_{-dn}, \boldsymbol{w}_d, \boldsymbol{l}_d, y_d, \mathbb{T}, \Phi, R) \\ &= \frac{p(\boldsymbol{z}_d, \boldsymbol{w}_d, \boldsymbol{l}_d \mid y_d, \mathbb{T}, \Phi, R)}{p(\boldsymbol{z}_{-dn}, \boldsymbol{w}_{-dn}, \boldsymbol{l}_{-dn} \mid y_d, \mathbb{T}, \Phi, R)} \end{aligned} \tag{18}$$

Expand Equation (12), we have

$$\begin{aligned} &p(\boldsymbol{z}_d, \boldsymbol{w}_d, \boldsymbol{l}_d \mid y_d, \mathbb{T}, \Phi, R) \\ &= p(\boldsymbol{z}_d) \prod_{dn} \Phi_{w_{dn} z_{dn}}^{y_d} \prod_{dn} \sum_k p(u_{dn} = k \mid z_{dn}, y_d, \mathbb{T}) p(l_{dn} \mid u_{dn} = k, R) \end{aligned} \tag{19}$$

Combining Equation (18) and (19), we have Equation (10).

8. References

[1] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[2] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001.

[3] D. M. Blei, A. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[4] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR*, 2003.

[5] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *ECCV*, 2006.

[6] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[7] M. Fritz, and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *CVPR*, 2008.

[8] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed dirichlet processes. In *NIPS*, 2005.

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[10] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *ICCV*, 2007.

[11] C. Wang, D. Blei and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

[12] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[13] D. J. Crandall, P. F. Felzenszwalb, and D. P. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, 2005.

[14] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[15] Z. Lin, G. Hua, and L. Davis, Multiple instance feature for robust part-based object detection. In *CVPR*, 2009.

[16] D. Liu, G. Hua, P. A. Viola, and T. Chen, Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[17] E. Nowak, F. Jurie, and B. Triggs, Sampling strategies for bag-of-features image classification. In *ECCV*, 2006.

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

[19] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*, 2004.

[20] X. L. Meng, and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica*, 1996.

[21] S. L. Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.

[22] S. L. Julien. *Discriminative machine learning with structure*. PhD Thesis, University of California, Berkeley, 2009.

[23] D. Liu and T. Chen. A topic-motion model for unsupervised video object discovery. In *ICCV*, 2007.

[24] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian. Spatial coding for large scale partial-duplicate web image search. In *ACM Multimedia*, 2010.

[25] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *ACM Multimedia*, 2009.

[26] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.