Appendix for Latent Dirichlet Allocation Models for Image Classification

Appendix I
Variational Inference in LDA models

Given an image $\mathcal{I} = \{w_1, \ldots, w_N\}, w_n \in \mathcal{V}$, inference consists of computing the posterior distribution of the unobserved variables, $P(\boldsymbol{\pi}, z_{1:N}|\mathcal{I})$[1]. Learning involves estimating the parameters $(\boldsymbol{\alpha}, \Lambda_{1:K})$, by maximizing the log likelihood, $l = \log P(\mathcal{D})$ of a training image dataset $\mathcal{D}$. Inference and learning are not tractable under LDA. A wide range of approximate inference methods have been proposed, such as Laplace or variational approximations, sampling methods, etc. We adopt variational inference. Variational methods approximate the posterior $P(\boldsymbol{\pi}, z_{1:N}|w_{1:N})$ by a mean-field variational distribution $q(\boldsymbol{\pi}, z_{1:N})$, indexed by free variational parameters, within some class of tractable probability distributions $\mathcal{F}$. These distributions usually assume independent factors,

$$q(\boldsymbol{\pi}, z_{1:N}) = q(\boldsymbol{\pi}; \boldsymbol{\gamma}) \prod_n q(z_n; \boldsymbol{\phi}_n) \tag{1}$$

where $q(y)$ and $q(z_n)$ are categorical models, and $q(\boldsymbol{\pi})$ a Dirichlet distribution. Given an observation $w_{1:N}$, the optimal variational approximation minimizes the Kullback-Leibler (KL) divergence between the two posteriors

$$q^* = \arg\min_{q \in \mathcal{F}} KL(q(\boldsymbol{\pi}, z_{1:N})||P(\boldsymbol{\pi}, z_{1:N}|w_{1:N})) \tag{2}$$

$$= \arg\min_{q \in \mathcal{F}} \log P(w_{1:N}) - \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \Lambda) \tag{3}$$

where

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \Lambda) = E_q[\log P(\boldsymbol{\pi}, z_{1:N}, w_{1:N}; \boldsymbol{\alpha}, \Lambda_{1:K})] - E_q[\log q(\boldsymbol{\pi}, z_{1:N}; \boldsymbol{\gamma}, \boldsymbol{\phi}_{1:N})] \tag{4}$$

is commonly known as the evidence lower bound (ELBO) [1]. This also lower bounds the true log likelihood of an image, for an arbitrary variational distribution $q(\boldsymbol{\pi}, z_{1:N})$ (see [2], Appendix A.3).

Since the data likelihood $P(w_{1:N})$ is constant for an observed image, the optimization problem of (3) is identical to the maximization of the ELBO,

$$q^*(\boldsymbol{\pi}, z_{1:N}) = \arg\max_{q \in \mathcal{F}} \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \Lambda) \tag{5}$$

From Appendix A.3 of [2], the update rule for coordinate descent of the variational parameters is

$$\gamma_k^* = \sum_n \phi_{nk} + \alpha_k \tag{6}$$

$$\phi_{nk}^* \propto \Lambda_{kw_n} \ \exp\left[\psi(\gamma_k)\right] \tag{7}$$

such that $\sum_k \phi_{nk} = 1$ and ψ is the Digamma function [3].

[1]Henceforth, wherever clear from context we shall omit the subscripts of probability distributions.

Appendix II
Parameter Estimation in cLDA

The parameters $(\boldsymbol{\eta}, \boldsymbol{\alpha}_{1:C}, \Lambda_{1:K})$ of cLDA are learned using variational expectation-maximization (EM). This iterates between:

a) Variational E-Step: approximates the posterior $P(\boldsymbol{\pi}^d, z_{1:N}^d|\mathcal{I}^d, y^d)$ given image $\mathcal{I}^d = \{w_1^d, \ldots, w_N^d\}$ by the variational distribution

$$q(\boldsymbol{\pi}^d, z_{1:N}^d) = q(\boldsymbol{\pi}^d; \boldsymbol{\gamma}^d) \prod_n q(z_n^d; \boldsymbol{\phi}_n^d). \tag{8}$$

Similar to the variational inference of LDA (see Appendix I), the variational parameters can be computed with the updates

$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_{y^d k} \tag{9}$$

$$\phi_{nk}^{d*} \propto \Lambda_{kw_n^d} \ \exp\left[\psi(\gamma_k^d)\right] \tag{10}$$

where $\sum_k \phi_{nk}^d = 1$. Note that in cLDA, where each class is associated with a separate prior over the topic simplex, (9) differs from (6), in that $\boldsymbol{\alpha}$ parameters are class specific.

b) M-Step: computes the values of the parameters $(\boldsymbol{\alpha}_{1:C}, \Lambda_{1:K})$. $\boldsymbol{\alpha}_y$ is obtained by maximizing,

$$\boldsymbol{\alpha}_y^* = \arg\max_{\boldsymbol{\alpha}_y} -\sum_d \delta(y^d, y) \log \mathcal{B}(\boldsymbol{\alpha}_y) + \sum_d \sum_k \delta(y^d, y)(\alpha_{y^d k} - 1) E_q[\log \pi_k^d] \tag{11}$$

with

$$E_q[\log \pi_k^d] = \psi(\gamma_k^d) - \psi(\sum_l \gamma_l^d) \tag{12}$$

$$\mathcal{B}(\boldsymbol{\alpha}_y) = \frac{\prod_k (\Gamma(\alpha_{yk})}{\Gamma(\sum_k \alpha_{yk})} \tag{13}$$

and $\Gamma()$ the Gamma function. This optimization can be carried out by the method of Newton-Raphson, as detailed in [3].

Λ_k is obtained by maximizing,

$$\Lambda_{kv}^* = \arg\max_{\Lambda_k} \sum_d \sum_n \delta(w_n^d, v) \phi_{nk}^d \log \Lambda_{kv} \tag{14}$$

such that $\sum_{v=1}^{|\mathcal{V}|} \Lambda_{kv} = 1$. This is done with the method of Lagrange multipliers, which results in the closed form update

$$\Lambda_{kv} \propto \sum_d \sum_n \delta(w_n^d, v) \phi_{nk}^d \tag{15}$$

where the proportionality symbol implies that Λ_k is normalized to sum to 1. Note that, as is common, we assume a uniform class prior $\boldsymbol{\eta}_y = \frac{1}{C}, \forall y \in \mathcal{Y}$.

APPENDIX III
PARAMETER ESTIMATION IN TOPIC-SUPERVISED LDA MODELS

In this appendix, we discuss parameter estimation for ts-cLDA. A similar approach can be used for the other topic-supervised models. Topic supervision decouples cLDA learning into two steps: 1) learning of the parameters $\Lambda_{1:K}$ of the topic-conditional distributions, and 2) learning of the parameters $\boldsymbol{\alpha}_{1:C}$ of the class-conditional distributions[2].

A. Learning Topic Conditional Distributions

As discussed in Section V, topics are defined over the class vocabulary $\mathcal{T} = \mathcal{V}$. In the absence of individual topic labels z_n^d for visual words w_n^d, it is assumed that all topic labels are equal to the image class y^d, i.e. $z_n^d = y^d \ \forall n, d$. Although this is not true, this assumption has been shown effective both through successful design of image labeling systems [5] and theoretical connections to multiple instance learning. In fact, this is an implicit assumption in learning the parameters of the flat model. The ML estimates of Λ_k are obtained from

$$\Lambda_{kv}^* = \arg\max_{\Lambda_k} \sum_d \sum_n \delta(y^d, k)\delta(w_n^d, v) \log \Lambda_{kv} \tag{16}$$

such that $\sum_{v=1}^{|\mathcal{V}|} \Lambda_{kv} = 1$. Using the method of Lagrange multipliers, the solution is

$$\Lambda_{kv} = \frac{\sum_d \sum_n \delta(y^d, k)\delta(w_n^d, v)}{\sum_j \sum_d \sum_n \delta(y^d, j)\delta(w_n^d, v)} \tag{17}$$

B. Learning Class Conditional Distribution with known z_n^d

Under the weak learning assumption, where z_n^d are equated to the class of the image, i.e. $z_n^d = y^d, \forall n, d$, the class-conditional distributions for cLDA can be learned using standard EM algorithm. This iterates between two steps:

c) E-Step: computes

$$E_{\boldsymbol{\pi}^d|y^d,z_{1:N}^d}[\log \pi_k^d] = \psi(\alpha_{y^d k} + n_k^d) - \psi(\sum_l \alpha_{y^d l} + n_l^d) \tag{18}$$

where $n_k^d = \sum_n \delta(z_n^d, k)$.

d) M-Step: computes the values of the parameters $\boldsymbol{\alpha}_y$ by maximizing,

$$\boldsymbol{\alpha}_y^* = \arg\max_{\boldsymbol{\alpha}_y} -\sum_d \delta(y^d, y) \log \mathcal{B}(\boldsymbol{\alpha}_y) + \sum_d \sum_k \delta(y^d, y)(\alpha_{y^d k} - 1) E_{\boldsymbol{\pi}^d|y^d,z_{1:N}^d}[\log \pi_k^d] \tag{19}$$

similar to that of standard cLDA (see Appendix II).

C. Learning Class Conditional Distribution with unknown z_n^d

As discussed in Section V-C learning the class-conditional distributions under weak supervision leads to degenerate solutions. Instead, they are learned assuming unknown patch labels z_n^d. This is done by maximizing the data likelihood, $P(y^d, w_{1:N}^d)$, using the variational EM algorithm, iterating between two steps:

[2]Note that $\boldsymbol{\eta}$ is again assumed to follow a uniform distribution.

e) Variational E-Step: computes

$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_{y^d k} \tag{20}$$

$$\phi_{nk}^{d*} \propto \Lambda_{k w_n^d} \ \exp\left[\psi(\gamma_k^d)\right] \tag{21}$$

where the proportionality symbols implies that ϕ_n^d is normalized to sum to 1.

f) M-Step: computes the values of parameters $\boldsymbol{\alpha}_{1:C}$ (note that $\Lambda_{1:K}$ is already computed), using (11).

APPENDIX IV
EXPERIMENTAL SETUP

In this appendix, we describe the experimental setup used to evaluate the performance of both css-LDA and topic-supervised LDA models.

A. Datasets

All experiments were based on datasets from the scene classification literature.

1) Natural Scene Categories (N15, N13, N8): This dataset consists of images of natural scenes. The initial version consisted of 8 different scene categories, viz. "coast", "forest", "highway", "insidecity", "mountain", "opencountry", "street", "tallbuilding" and has been used in [4], [10], [6]. Five more categories, viz. "bedroom", "suburb", "kitchen", "livingroom", "office" were added in [8]. This version has been further used in [6], [10]. Two more categories, viz. "store","industrial" were added in [7], for a total 15 categories. We refer to the complete 15-category version as $N15$, the 13-category subset as $N13$ and the original 8-category subset as $N8$. Each category contains 200 to 400 images, of average size 300×250 pixels. One hundred images per scene were used to learn models, the remaining being used as test set. The final experiments were repeated six times, with random train/test splits.

2) UIUC Sports Dataset (S8): This dataset contains images from eight sports categories, viz. "badminton", "bocce", "croquet", "polo", "rock climbing", "rowing", "sailing", "snowboarding". It was first proposed in [9] for LDA based classification, and subsequently used by [14] to evaluate sLDA. Each category has 137 to 250 large size images. In our experiments, the images were resized to a maximum of 256 pixels along the larger border. In all, there are 1579 images. As in [9], 70 images per scene were used to learn the models, and 60 images as test set. Again, the final experiments were repeated 6 times with random train/test splits.

3) Corel Image Collection (C50): This dataset consists of images from 50 Corel Stock Photo CDs, where each CD contains 100 images of a common scene. The annotated version of this dataset (where each image is further annotated with 1-5 concepts) is commonly used for the evaluation of image annotation systems [11], [12], [13]. In this work, we used the 50 scene classes, each corresponding to one CD in the collection, as the ground truth for classification. For each CD, 90 images were used to learn class models and the remaining for testing. All images were normalized to size 181×117 or 117×181 and converted from RGB to the YBR color space.

B. Appearance features

Two feature transforms were used for appearance representation, viz. scale invariant feature transform (SIFT) and discrete cosine transform (DCT). SIFT features were used for monochrome images and DCT features for color images. For SIFT, as is common in the literature, 128-dimensional SIFT descriptors were computed over 16×16 pixel patches, sampled densely over a grid with a regular spacing of 8 pixels in both the horizontal and vertical directions. On average, 1000 SIFT[3] features were computed per image. DCT features were also computed on a dense regular grid, with a step of 8 pixels. 8×8 image patches were extracted around each grid point, and DCT coefficients computed per patch and color channel. This resulted in a 64 dimensional space per channel, of which we used the first 43 DCT coefficients.

Codebooks of visual words were obtained with K-means clustering, for K ranging from 128 to 4096. For each dataset, codebooks were generated from a random collection of 300 examples per training image. K-means initialization was performed with a vector quantizer designed by the Linde-Buzo-Gray (LBG) algorithm, using a variation of the cell splitting method described in [15]. For experiments using LDA and sLDA, we used the code available online[4]. This code was modified for cLDA, topic-supervised LDA and css-LDA (which will be made available online). The number of topics was varied from 10 to 100 for topic discovery approaches. For topic-supervised models, the number of topics was equal to the number of classes. The α_k parameter was set to 1 in all experiments except cLDA and ts-cLDA, where an asymmetric $\boldsymbol{\alpha}_y$ parameter was learned per class. Although not explicitly shown in Figure 1, we used the "smoothed" version of various models. The flat model is regularized using Laplace smoothing with a hyper-parameter of 0.1 and LDA models are regularized using a Dirichlet prior on the topic-distributions [2], using a symmetric hyper-parameter of 0.001. The performance of the various models was not very sensitive to the choice of both α_k and the smoothing parameter.

REFERENCES

[1] D. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128, 2008.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] T. Minka. Estimating a dirichlet distribution. *http://research.microsoft.com/ minka/papers/dirichlet/*, 1:3, 2000.

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

[5] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):394–410, 2007.

[6] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(4):712–727, 2008.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Conference on*, pages 2169–2178. IEEE, 2006.

[3]We use the LEAR implementation of SIFT to compute the descriptors, http://lear.inrialpes.fr/people/dorko/downloads.html.

[4]http://www.cs.princeton.edu/~blei/lda-c/ and http://www.cs.princeton.edu/~chongw/slda/ respectively.

[8] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005 IEEE Conference on*, pages 524–531. IEEE, 2005.

[9] L. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *Computer Vision and Pattern Recognition, 2007 IEEE Conference on*. IEEE, 2007.

[10] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(9):1575–1589, 2007.

[11] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, Copenhagen, Denmark, 2002.

[12] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Computer Vision and Pattern Recognition, 2004 IEEE Conference on*. IEEE, 2004.

[13] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Advances in Neural Information Processing Systems, Vancouver*, 2003.

[14] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009 IEEE Conference on*, pages 1903–1910. IEEE, 2009.

[15] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1):84–95, 1980.