Latent Dirichlet Allocation Models for Image Classification

Nikhil Rasiwasia, *Member*, *IEEE*, and Nuno Vasconcelos, *Senior Member*, *IEEE*

Abstract—Two new extensions of latent Dirichlet allocation (LDA), denoted *topic-supervised* LDA (ts-LDA) and *class-specific-simplex* LDA (css-LDA), are proposed for image classification. An analysis of the supervised LDA models currently used for this task shows that the impact of class information on the topics discovered by these models is very weak in general. This implies that the discovered topics are driven by general image regularities, rather than the semantic regularities of interest for classification. To address this, ts-LDA models are introduced which replace the automated topic discovery of LDA with specified topics, identical to the classes of interest for classification. While this results in improvements in classification accuracy over existing LDA models, it compromises the ability of LDA to discover unanticipated structure of interest. This limitation is addressed by the introduction of css-LDA, an LDA model with class supervision at the level of image features. In css-LDA topics are discovered per class, i.e., a single set of topics shared across classes is replaced by multiple class-specific topic sets. The css-LDA model is shown to combine the labeling strength of topic-supervision with the flexibility of topic-discovery. Its effectiveness is demonstrated through an extensive experimental evaluation, involving multiple benchmark datasets, where it is shown to outperform existing LDA-based image classification approaches.

Index Terms—Image classification, graphical models, latent Dirichlet allocation, semantic classification, attributes

1 Introduction

Image classification is a topic of significant interest within computer vision. The goal is to classify an image into one of a prespecified set of image classes or categories. This is usually done by 1) designing a set of appearance features rich enough to describe the classes of interest, 2) adopting an architecture for the classification of these features, and 3) learning its parameters from training data. This strategy has been successful in the recent past, with advances in various aspects of the problem. With respect to features, a popular strategy is to model images as orderless collections of local descriptors, for example, edge orientation descriptors based on the scale invariant feature transform (SIFT). One successful orderless representation is the *bag-of-visual-words* (BoW). This consists of vector quantizing the space of local descriptors and using the means of the resulting clusters, commonly known as "visual-words,"[1] as their representatives. The set of visual-words forms a *codebook*, which is used for image quantization. Images are finally represented by histograms of visual word occurrence [9]. This representation mimics the time-tested *bag-of-words* of text-retrieval [21], with words replaced by visual-words [22]. It is the foundation of a number of methods for object recognition and image classification [14], [27], [6].

The simplest BoW image classification architecture is the equivalent of the *naive Bayes* approach to text classification [20]. It assumes that image *words*[2] are sampled independently from the BoW model, and relies on the Bayes decision rule for image classification. We refer to this as the *flat* model, due to its lack of hierarchical word groupings. Although capable of identifying sets of words discriminative for the classes of interest, it does not explicitly model the inter and intraclass structure of word distributions. To facilitate the *discovery* of this structure, various models have been recently ported from the text to the vision literature. Popular examples include hierarchical probabilistic models, commonly known as *topic models*, such as latent Dirichlet allocation (LDA) [4] and probabilistic latent semantic analysis (pLSA) [11]. Under these models, each document (or image) is represented as a finite mixture over an intermediate set of topics, which are expected to summarize the document semantics.

Since LDA and pLSA topics are discovered in an *unsupervised* fashion, these models have limited use for classification. Several LDA extensions have been proposed to address this limitation, in both the text and vision literatures. One popular extension is to apply a classifier, such as an SVM, to the topic representation [4], [5], [18]. We refer to these as discriminant extensions, and the combination of SVM with LDA as SVM-LDA. Such extensions are hampered by the inability of *unsupervised* LDA to latch onto semantic regularities of interest for classification [3], [28]. For example, it has been noted in the text literature [3] that, given a collection of movie reviews, LDA might discover as topics movie properties, for example, genres, which are not

1. In the literature the terms "textons," "keypoints," "visual-words," "visual-terms," "visual codewords," or "visterms" have also been used.

2. Henceforth, "word" refers to a semantic word in the context of text documents and visual word in the context of images.

- *N. Rasiwasia is with Yahoo! Labs Bangalore, Bengaluru 560071, Karnataka, India. E-mail: nikhil.rasiwasia@gmail.com.*
- *N. Vasconcelos is with the Department of Electrical and Computer Engineering, University of California San Diego. E-mail: nvasconcelos@ucsd.edu.*

Manuscript received 15 Dec. 2011; revised 28 July 2012; accepted 19 Mar. 2013; published online 4 Apr. 2013.
Recommended for acceptance by C. Lampert.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-12-0893.
Digital Object Identifier no. 10.1109/TPAMI.2013.69.

central to the classification task, for example, prediction of movie ratings. A second approach is to incorporate a class label variable in the generative model [10], [3], [25], [13], [28], [16]. These are denoted generative extensions. Two popular members of this family are the model of [10], here referred to as classLDA (cLDA), and the model of [25], commonly known as supervised LDA (sLDA). The latter was first proposed for supervised text prediction in [3]. Several other adaptations of LDA have been proposed for tasks other than classification: correspondence LDA [2] and topic-regression multimodal LDA [17] for image annotation, labeled LDA [19] (labLDA) for credit attribution in a multilabeled corpora, semiLDA [26] for human action recognition in videos, and so on.

In this paper, we focus on generative extensions of LDA for image classification. We start by showing that even the most popular supervised extensions, such as cLDA and sLDA, are unlikely to capture class semantics. This is shown by 1) a theoretical analysis of the learning algorithms, and 2) experimental evaluation on classification problems. Theoretically, it is shown that the impact of class information on the topics discovered by cLDA and sLDA is very weak in general and vanishes for large samples. Experiments show that the classification accuracies of cLDA and sLDA are not superior to those of unsupervised topic discovery. To address these limitations, we propose the family of *topic supervised (ts)* LDA models. Instead of relying on discovered topics, ts-LDA *equates topics to the classes of interest* for image classification, establishing a one-to-one mapping between topics and class labels. This forces LDA to pursue semantic regularities in the data. Topic supervision reduces the learning complexity of topic distributions and improves on the classification accuracy of existing sLDA extensions. This is demonstrated by the introduction of topic supervised versions of LDA, cLDA, and sLDA, denoted *ts-LDA*, *ts-cLDA*, and *ts-sLDA*, respectively. In all cases, the topic supervised models outperform the corresponding LDA models learned without topic-supervision.

However, topic-supervised LDA (ts-LDA) models do not outperform the flat model. This is partly due to limitations of LDA itself. We show that the bulk of the modeling power of LDA, in both existing and topic-supervised models, lies in dimensionality reduction, the mapping of images from a high-dimensional *word simplex* to a low-dimensional *topic simplex*, and not in the modeling of class specific distributions per se. Since all classes share a topic simplex of relatively low dimensionality, there is limited ability to simultaneously uncover rich intraclass structure and maintain discriminability. By obviating the discovery of this latent topic simplex, topic supervision increases discrimination, but sacrifices the model ability to discover unanticipated structure of interest for image classification. Hence, ts-LDA models are not fundamentally different from the flat model. To combine the *labeling strength* of topic-supervision with the *flexibility* of topic-discovery of LDA, we propose a novel classification architecture, denoted *class-specific simplex LDA* (css-LDA). Inspired by the flat model, css-LDA differs from the existing LDA extensions in that supervision is introduced directly at the level of image features. This induces the discovery of class-specific topic simplices and, consequently, *class-specific topic distributions*, enabling a much richer modeling of intraclass structure without compromise of discrimination ability. An extensive experimental evaluation shows that css-LDA outperforms both all existing extensions of LDA for image classification and the flat model on five benchmark datasets.

The paper is organized as follows: Section 2 briefly reviews the literature on generative models for image classification. The limitations of existing models are highlighted in Section 3 and the motivation for the proposed approaches is then given in Section 4. Next, Sections 5 and 6 introduce the topic-supervised and css-LDA models, respectively. An extensive experimental evaluation of all models is presented in Section 7. Finally, conclusions are drawn in Section 8.

2 Models for Image Classification

We start by reviewing LDA and its various extensions for classification. Images are observations from a random variable $\mathbf{X}$, defined on some feature space $\mathcal{X}$ of visual measurements. For example, $\mathcal{X}$ could be the space of discrete cosine transform (DCT), or SIFT descriptors. Each image is represented as a bag of N *feature vectors* $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\mathbf{x}_n \in \mathcal{X}$, assumed to be sampled independently. The feature space $\mathcal{X}$ is quantized into $|\mathcal{V}|$ bins, defined by a set of centroids, $\mathcal{V} = \{1, \ldots, |\mathcal{V}|\}$. Each feature vector $\mathbf{x}_n$, $n \in \{1, \ldots, N\}$, is mapped to its closest centroid. Images are represented as collections of visual words, $\mathcal{I} = \{w_1, \ldots, w_N\}$, $w_n \in \mathcal{V}$, where w_n is the bin containing $\mathbf{x}_n$. Each image in a *dataset*, $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_D\}$, is labeled with a class y, drawn from a random variable Y with values in $\mathcal{Y} = \{1, \ldots, C\}$. This is the set of classes that define the image classification problem, making $\mathcal{D} = \{(\mathcal{I}_1, y_1), \ldots, (\mathcal{I}_D, y_D)\}$.

A query image $\mathcal{I}_q$ is classified with the minimum probability of error criterion, where the optimal decision rule is to assign $\mathcal{I}_q$ to the class of maximum posterior probability, i.e.,

$$y^* = \arg\max_y P_{Y|W}(y|\mathcal{I}_q). \tag{1}$$

$P_{Y|W}(y|\mathcal{I}_q)$, the posterior probability of class y given $\mathcal{I}_q$, is computed with a combination of a probabilistic model for the joint distribution of words and classes and Bayes rule. We next review some popular models.

2.1 Flat Model

Fig. 1a presents the graphical form of the flat model. Visual words w_n are sampled independently conditioned on the class label. The generative process is

for each image **do**
  sample a class label $y \sim P_Y(y; \boldsymbol{\eta})$, $y \in \mathcal{Y}$
  for $i \in \{1, \ldots, N\}$ **do**
    sample a visual word $w_i \sim P_{W|Y}(w_i|y; \Lambda_y^{flat})$.
  end for
end for

Although any suitable distribution can be used as class prior $P_Y()$ and class-conditional distribution $P_{W|Y}()$, these are usually categorical distributions over $\mathcal{Y}$ and $\mathcal{V}$, respectively:

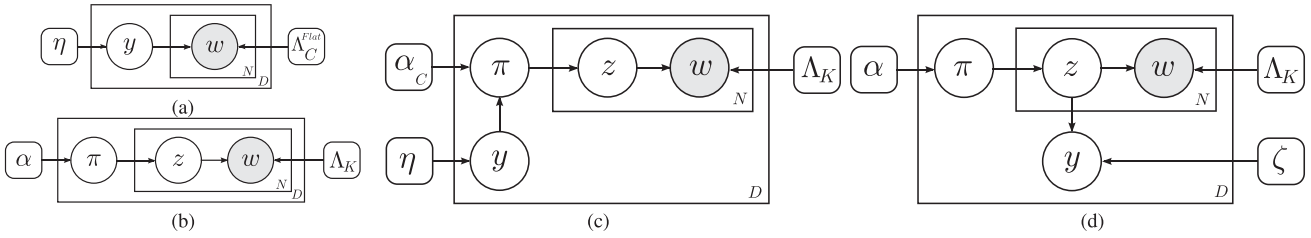


Fig. 1. Graphical models for (a) flat model, (b) LDA and ts-LDA, (c) cLDA and ts-cLDA, (d) sLDA and ts-sLDA. All models use the plate notation of [7], with parameters shown in rounded squares.

$$P_Y(y;\boldsymbol{\eta}) = \prod_{c=1}^{C} \eta_c^{\delta(c,y)}, \tag{2}$$

$$P_{W|Y}\big(w|y;\Lambda_y^{flat}\big) = \prod_{c=1}^{C}\prod_{v=1}^{|\mathcal{V}|} \Lambda_{cv}^{flat\,\delta(y,c)\delta(v,w)}, \tag{3}$$

where $\delta(x,y)$ is the Kronecker delta function, which takes the value of one if and only if $x=y$, zero otherwise, and $(\boldsymbol{\eta}, \Lambda_{1:C}^{flat})$ are model parameters such that $\sum_c \eta_c = 1$ and $\sum_v \Lambda_{cv}^{flat} = 1$. In this paper, wherever applicable, we assume a uniform class prior $\eta_y = \frac{1}{C},\ \forall y \in \mathcal{Y}$. However, for completeness, the prior is included in all equations. The parameters $\Lambda_{1:C}^{flat}$ can be learned by maximum likelihood estimation as

$$\Lambda_{yw}^{flat} = \frac{\sum_d \sum_n \delta\big(y^d, y\big)\delta\big(w_n^d, w\big)}{\sum_v \sum_d \sum_n \delta\big(y^d, y\big)\delta\big(w_n^d, v\big)}, \tag{4}$$

where d indexes the training images.

In general, images from a class may contain patches from others, for example, images of the "Street" class may contain patches of "Buildings" or "Highways." Hence, strictly supervised learning of the class-conditional distributions requires extensively labeled training sets, where each image patch is labeled with a class in $\mathcal{Y}$. Such supervision is not available in image classification problems, where labels are only provided for entire images. This is the *weakly supervised* learning problem also faced by image annotation [8]. As in that problem, class-conditional distributions are simply learned from all patches extracted from all images in the training set of the class. This type of learning has been shown effective through theoretical connections to multiple instance learning [24].

2.2 Unsupervised LDA Model

LDA is the generative model of Fig. 1b. The generative process is as follows:

for each image **do**
  sample $\boldsymbol{\pi} \sim P_\Pi(\boldsymbol{\pi};\boldsymbol{\alpha})$.
  for $i \in \{1,\ldots,N\}$ **do**
    sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{T} = \{1,\ldots,K\}$, where $\mathcal{T}$ is the set of topics.
    sample a visual word $w_i \sim P_{W|Z}(w_i|z_i;\Lambda_{z_i})$.
  end for
end for

where $P_\Pi()$ and $P_{W|Z}()$ are the prior and topic-conditional distributions, respectively. $P_\Pi()$ is a Dirichlet distribution on $\mathcal{T}$ with parameter $\boldsymbol{\alpha}$, and $P_{W|Z}()$ a categorical distribution on $\mathcal{V}$ with parameters $\Lambda_{1:K}$. Although the model parameters can be learned with the expectation maximization (EM) algorithm, the E-step yields intractable inference. To address this, a wide range of approximate inference methods have been proposed [1], such as Laplace or variational approximations, sampling methods, and so on. In this paper, we adopt variational inference for all models where exact inference is intractable. Variational inference for the LDA model is briefly discussed in Appendix I, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.69. In its original formulation, LDA does not incorporate class information and cannot be used for classification. We next discuss two models that address this limitation.

2.3 Class LDA

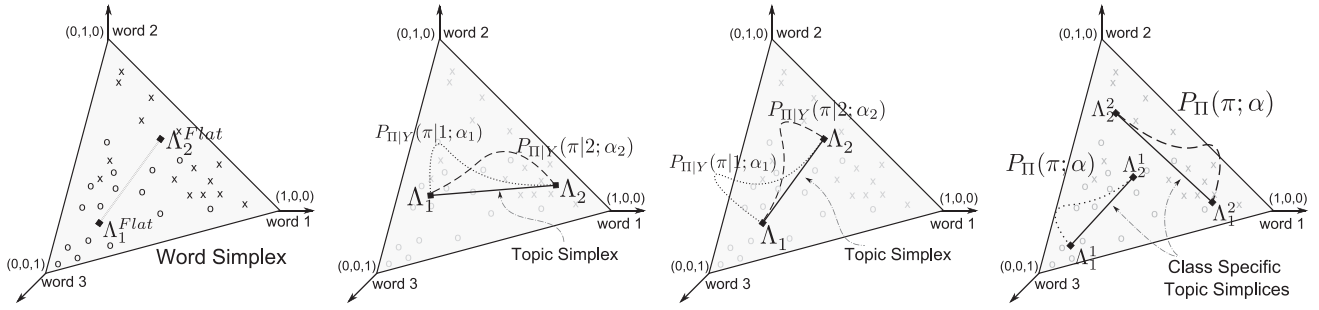
cLDA was introduced in [10] for image classification. In this model, as shown in Fig. 1c, a class variable Y is introduced as the parent of the topic prior Π. In this way, each class defines a prior distribution in topic space, conditioned on which the topic probability vector $\boldsymbol{\pi}$ is sampled. Images are sampled as follows:

for each image **do**
  sample a class label $y \sim P_Y(y;\boldsymbol{\eta})$, $y \in \mathcal{Y}$
  sample $\boldsymbol{\pi} \sim P_{\Pi|Y}(\boldsymbol{\pi}|y;\boldsymbol{\alpha}_y)$.
  for $i \in \{1,\ldots,N\}$ **do**
    sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{T} = \{1,\ldots,K\}$.
    sample a visual word $w_i \sim P_{W|Z}(w_i|z_i;\Lambda_{z_i})$
  end for
end for

where $\boldsymbol{\alpha}_y = \{\alpha_{y1},\ldots,\alpha_{yK}\}$. Parameter learning for cLDA is similar to that of LDA [10] and detailed in Appendix II, which is available in the online supplemental material. A query image $\mathcal{I}_q$ is classified with (1), using variational inference to approximate the posterior $P_{Y|W}(y|\mathcal{I}_q)$ [10].

2.4 Supervised LDA

sLDA was proposed in [3]. As shown in Fig. 1d, the class variable Y is conditioned by topics Z. In its full generality, sLDA uses a generalized linear model of Y, which can be either discrete or continuous. Wang et al. [25] applied this generic framework to the task of image classification, where Y takes on discrete responses, by making use of the generalized linear model exponential family relevant to a categorical response. An alternative extension to binary image annotation was proposed in [16], using a multivariate Bernoulli variable for Y. In [28], the max-margin principle is used to train sLDA, which is denoted maximum



(a) flat (b) cLDA (c) ts-cLDA (d) css-LDA

Fig. 2. Representation of various models on a three word simplex. Also shown are sample images from two classes: "o" from class-1 and "x" from class-2. (a) The flat model. (b) The cLDA model with two topics. The line segment depicts a 1D topic simplex whose vertices are topic-conditional word distributions. Classes define smooth distributions on the topic simplex, denoted by dashed and dotted lines. (c) The ts-cLDA model. Topic-conditional word distributions are learned with supervision and aligned with the class-conditional distributions of the flat model. (d) The css-LDA model. Each class defines its own topic simplex.

entropy discrimination LDA (medLDA). In this paper, sLDA refers to the formulation of [25] since this was the one previously used for image classification. Images are sampled as follows:

for each image **do**
  sample $\boldsymbol{\pi} \sim P_{\Pi}(\boldsymbol{\pi}; \boldsymbol{\alpha})$.
  for $i \in \{1, \ldots, N\}$ **do**
    sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{T} = \{1, \ldots, K\}$
    sample a visual word $w_i \sim P_{W|Z}(w_i|z_i; \Lambda_{z_i})$.
  end for
  sample a class label $y \sim P_{Y|Z}(y|\bar{\mathbf{z}}; \boldsymbol{\zeta}_{1:C})$, $y \in \mathcal{Y}$
end for

where $\bar{\mathbf{z}}$ is the mean topic assignment vector $\bar{\mathbf{z}}_k = \frac{1}{N}\sum_{n=1}^{N} \delta(z_n, k)$, and

$$P_{Y|Z}(y|\bar{\mathbf{z}}; \boldsymbol{\zeta}) = \frac{\exp\left(\boldsymbol{\zeta}_y^T \bar{\mathbf{z}}\right)}{\sum_{l=1}^{C} \exp\left(\boldsymbol{\zeta}_l^T \bar{\mathbf{z}}\right)}, \tag{5}$$

a softmax activation function with parameter $\boldsymbol{\zeta}_c \in \mathbb{R}^K$. Variational inference is used to learn all model parameters and to approximate the posterior $P_{Y|W}(y|\mathcal{I}_q)$, used in (1) for classifying an image $\mathcal{I}_q$ [25].

2.5 Geometric Interpretation

The models discussed above have an elegant geometric interpretation [4], [23]. Associated with a vocabulary of $|\mathcal{V}|$ words, there is a $|\mathcal{V}|$-dimensional space, where each axis represents the occurrence of a particular word. A $|\mathcal{V}| - 1$-simplex in this space, here referred to as *word simplex*, represents all probability distributions over words. Each image (when represented as a word histogram) is a point on this space. Fig. 2a illustrates the 2D simplex of distributions over three words. Also shown are sample images from two classes, "o" from class-1 and "x" from class-2, and a schematic of the flat model. Under this model, each class is modeled by a class-conditional word distribution, i.e., a point on the word simplex. In Fig. 2a, Λ_1^{flat} and Λ_2^{flat} are the distributions of class-1 and class-2, respectively.

Fig. 2b shows a schematic of cLDA with two topics. Each topic in an LDA model defines a probability distribution over words and is represented as a point on the word simplex. Since topic probabilities are mixing probabilities for word distributions, a set of K topics defines a $K-1$ simplex in the word simplex, here denoted the *topic simplex*. If the number of topics K is strictly smaller than the number of words $|\mathcal{V}|$, the topic simplex is a low-dimensional subsimplex of the word simplex. The projection of images on the topic simplex can be thought of as dimensionality reduction. In Fig. 2b, the two topics are represented by Λ_1 and Λ_2, and span a 1D simplex, shown as a connecting line segment. In cLDA, each class defines a distribution (parameterized by $\boldsymbol{\alpha}_y$) on the topic simplex. The distributions of class-1 and class-2 are depicted in the figure as dotted and dashed lines, respectively. Similar to cLDA, sLDA can be represented on the topic simplex, where each class defines a softmax function.[3]

3 Limitations of Existing Models

In this section, we present theoretical and experimental evidence that, contrary to popular belief, topics discovered by sLDA and cLDA are not more suitable for discrimination than those of standard LDA. We start by showing, through an analysis of the variational learning equations of sLDA and cLDA, that class labels have very weak influence in the learning of the topic distributions of these models.

In both sLDA and cLDA, the parameters $\Lambda_{1:K}$ of the topic distributions are obtained via the variational M-step as

$$\Lambda_{kv} \propto \sum_d \sum_n \delta\left(w_n^d, v\right) \phi_{nk}^d, \tag{6}$$

where d indexes the images, $\sum_v \Lambda_{kv} = 1$, $\delta()$ is a Kronecker delta function, and ϕ_{nk} is the parameter of the variational distribution $q(z)$. This parameter is computed in the E-step:

$$\text{For cLDA:} \qquad \gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_{y^d k}, \tag{7}$$

$$\phi_{nk}^{d*} \propto \Lambda_{kw_n^d} \exp\left[\psi\left(\gamma_k^d\right)\right], \tag{8}$$

3. Strictly speaking, the softmax function is defined on the average of the sampled topic assignment labels $\bar{\mathbf{z}}$. However, when the number of features N is sufficiently large, $\bar{\mathbf{z}}$ is proportional to the topic distribution $\boldsymbol{\pi}$. Thus, the softmax function can be thought of as defined on the topic simplex.

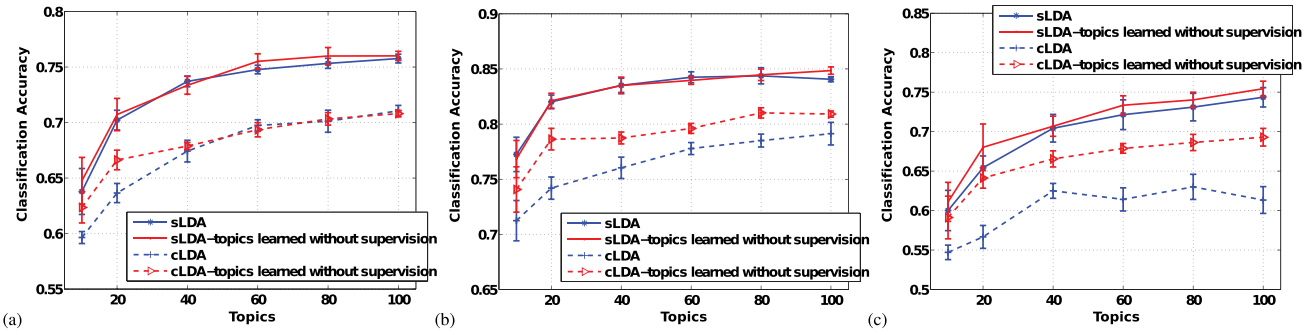


Fig. 3. Classification accuracy as a function of the number of topics for sLDA and cLDA, using topics learned with and without class influence and codebooks of size 1,024, on (a) N13, (b) N8, and (c) S8. Similar behavior was observed for codebooks of different sizes.

$$\text{For sLDA:} \qquad \gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha_k, \tag{9}$$

$$\phi_{nk}^{d*} \propto \Lambda_{kw_n^d} \exp\left[\psi(\gamma_k^d) + \frac{\zeta_{y^d k}}{N} - \frac{\sum_c \exp\frac{\zeta_{ck}}{N} \prod_{m\neq n} \sum_j \phi_{mj}^d \exp\frac{\zeta_{cj}}{N}}{\sum_c \prod_m \sum_j \phi_{mj}^d \exp\frac{\zeta_{cj}}{N}}\right], \tag{10}$$

where $\boldsymbol{\gamma}$ is the parameter of the variational distribution $q(\boldsymbol{\pi})$ (see [4] for the details of variational inference in LDA). The important point is that the class label y^d only influences the topic distributions through (7) for cLDA (where $\boldsymbol{\alpha}_{y^d}$ is used to compute the parameter $\boldsymbol{\gamma}^d$) and (10) for sLDA (where the variational parameter ϕ_{nk}^d depends on the class label y^d through $\zeta_{y^d k}/N$).

We next consider the case of cLDA. Given that $q(\boldsymbol{\pi})$ is a posterior Dirichlet distribution (and omitting the dependence on d for simplicity), the estimate of γ_k has two components: $\hat{l}_k = \sum_{n=1}^N \phi_{nk}$, which acts as a vector of counts, and α_{yk}, which is the parameter from the prior distribution. As the number of visual words N increases, the amplitude of the count vector, $\hat{\mathbf{l}}$, increases proportionally, while the prior $\boldsymbol{\alpha}_y$ remains constant. Hence, for a sufficiently large sample size N, the prior $\boldsymbol{\alpha}_y$ has a very weak influence on the estimate of $\boldsymbol{\gamma}$. This is a hallmark of Bayesian parameter estimation, where the prior only has impact on the posterior estimates for small sample sizes. It follows that the connection between class label Y and the learned topics Γ_k is extremely weak. This is not a fallacy of the variational approximation. In cLDA (see Fig. 1b), the class label distribution is simply a prior for the remaining random variables. This prior is easily overwhelmed by the evidence collected at the feature-leve, whenever the sample is large.

A similar effect holds for sLDA, where the only dependence of the parameter estimates on the class label is through the term $\zeta_{y^d k}/N$. This clearly diminishes as the sample size N increases.[4] In summary, topics learned with either cLDA or sLDA are very unlikely to be informative of semantic regularities of interest for classification, and much more likely to capture generic regularities common to all classes.

To confirm these observations, we performed experiments with topics learned under two approaches. The first used the original learning equations, i.e., (7) and (8) for cLDA and (9) and (10) for sLDA. In the second, we severed all connections with the class label variable *during topic learning* by reducing the variational E-step (of both cLDA and sLDA) to

$$\gamma_k^{d*} = \sum_n \phi_{nk}^d + \alpha, \tag{11}$$

$$\phi_{nk}^{d*} \propto \Lambda_{kw_n^d} \exp\left[\psi(\gamma_k^d)\right], \tag{12}$$

with $\alpha = 1$. This guarantees that the topic-conditional distributions are learned without any class influence. The remaining parameters ($\boldsymbol{\alpha}_y$ for cLDA, $\boldsymbol{\zeta}_y$ for sLDA) are still learned using the original equations. The rationale for these experiments is that if supervision makes any difference, models learned with the original algorithms should perform better.

Fig. 3 shows the image classification performance of cLDA and sLDA under the two learning approaches on the N13, N8, and S8 data sets (see Appendix IV, which is available in the online supplemental material, for details on the experimental setup). The plots were obtained with a 1,024 words codebook, and between 10 and 100 topics. Clearly, the classification performance of the original models *is not* superior to that of the ones learned without class supervision. The sLDA model has almost identical performance under the two approaches on the three datasets. For cLDA, unsupervised topic discovery is in fact *superior* on the N8 and S8 dataset. This can be explained by poor regularization of the original cLDA algorithm. We have observed small values of α_{yk}, which probably led to poor estimates of the topic distributions in (7). For example, the maximum, median, and minimum values of α_{yk} learned with 10 topics on S8 were 0.61, 0.12, and 0.04, respectively. In contrast, the corresponding values for unsupervised topic discovery were 7.09, 1.09, and 0.55. Similar effects were observed in experiments with codebooks of different size. These results show that the performance of cLDA and sLDA is similar (if not inferior) to that of topic learning without class supervision. In both cases, the class variable has very weak impact on the learning of topic distributions.

4. This discussion refers to the sLDA formulation of [25], which was proposed specifically for image classification. Note that the original sLDA formulation of [3] includes a dispersion parameter δ which provides additional flexibility in modeling the variance of Y. Inclusion of δ makes the parameter estimates dependent on the class label via $\zeta_{y^d k}/(N\delta)$. In this case, the importance of the class label y^d can be controlled by scaling δ appropriately. However, the basic argument still holds in the sense that, for any $\delta > 0$, this importance vanishes as the sample increases.

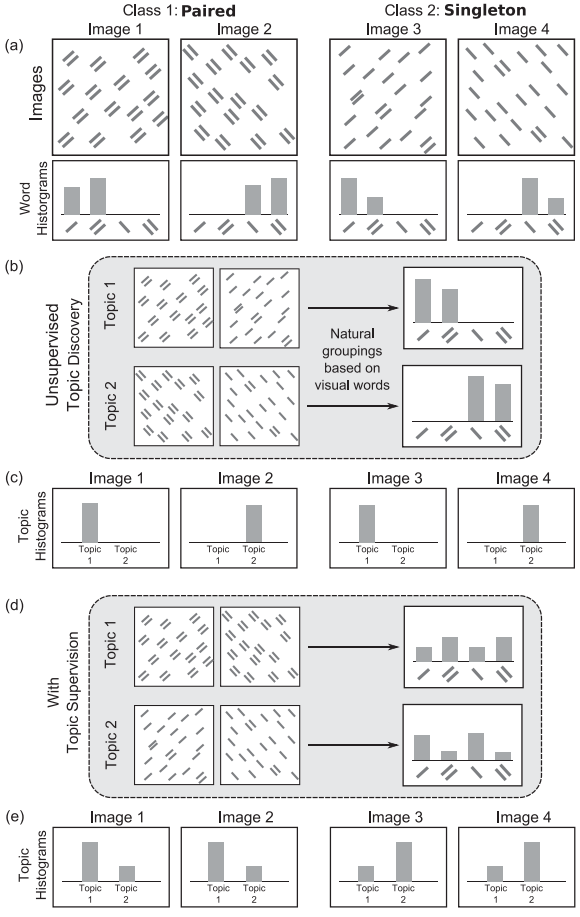


Fig. 4. An example of the need for supervision in learning topic distributions. (a) Images from the class *paired*—composed of paired line segments tilted left or right, and *singleton*—composed of singleton segments tilted left or right. Shown under each image is the associated word histogram, based on a set of four natural visual words for this problem—singleton-right, paired-right, singleton-left, and paired-left segments. (b) Image groupings based on word histograms uncover "generic" regularities of the data. (c) Expected topic vectors for the images in (a) based on topics discovered in (b). (d) Topics discovered with topic supervision. (e) Expected topic vectors for images in (a) based on topics discovered in (d).

4 Generic versus Semantic Regularities

Since all models discussed so far effectively implement unsupervised topic discovery, it is worth considering the limitations of this paradigm.

4.1 Generic versus Semantic Regularities

In this section, we show that unsupervised topic discovery is prone to latching onto generic image regularities, rather than the semantic regularities of interest for classification. Consider a dataset composed by the four types of images of Fig. 4a. These are images from a toy world of line segments. The segments have a *single* orientation, and the world has two *states*. Under the first state, segments are placed on the world in pairs, referred to as the *"paired"* state. Under the second, the segments are placed individually, referred to as the *"singleton"* state. Inferring the state of the world is not trivial because, as is common in computer vision, it can be imaged from two camera angles. Under the first, segments retain their orientation. Under the second, this orientation is reversed. Assuming a natural set of visual words for this problem (singleton-right, paired-right, singleton-left, and paired-left segments), each image produces the visual word histogram shown under it. Close inspection of these histograms shows that the largest overlap occurs between those of images shot from the same camera angle as these contain segments with the same orientation. Indeed, as can be seen in Fig. 4b, grouping the images by camera angle produces two topics of perfectly disjoint word histograms. On the other hand, as shown in Fig. 4d, grouping images by state of the world generates two topics of highly overlapping histograms because it requires grouping segments of mixed orientation.

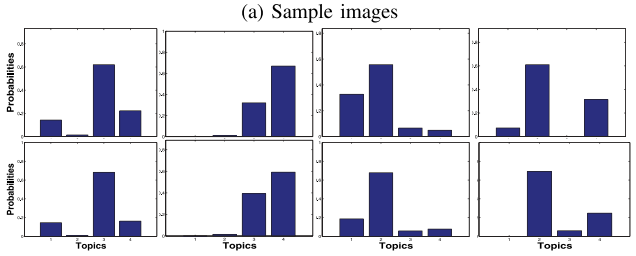
Hence, for this dataset, the prevalent regularity is the grouping of images by segment orientation. Unsupervised topic discovery will latch onto this regularity, producing the expected topic vectors shown in Fig. 4c, i.e., grouping image 1 with image 3 and image 2 with image 4. However, this is a *generic* regularity, totally unrelated to the semantic regularities of the problem, which are the *singleton versus paired states of the world*. To reflect these semantic regularities the model must be *forced* to have the topics of Fig. 4d, i.e., topic learning has to be *supervised*. Only this will guarantee the grouping of images 1and 2 and images 3 and 4, and the expected topic vectors of Fig. 4e. In summary, supervised topic discovery is usually required for LDA models to equate topics to states of the world. Unsupervised learning can easily latch onto generic regularities of the data (in this case the segment orientation), ignoring the semantic regularities of interest for classification.

4.2 Limitation of Current Models

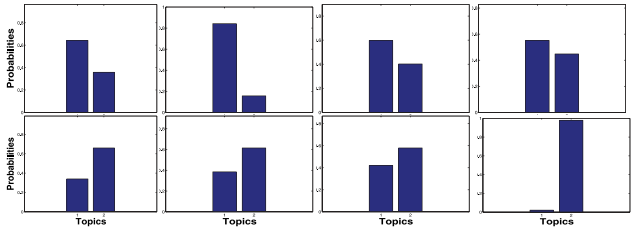
In Section 3, we have seen that models such as sLDA or cLDA *effectively* learn topics without supervision. This makes them prone to latching onto generic regularities. An illustration of this problem, on real data, is given in Fig. 5, which shows some example images from classes "Sailing" (top row) and "Rowing" (bottom row) of the S8 dataset (see Appendix IV-A, which is available in the online supplemental material). Fig. 5b shows the topic histograms corresponding to the images in Fig. 5a, for a four-topic cLDA model trained on 100 images per class. Note that, although belonging to different classes, the two images in each column have nearly identical topic distributions. On the other hand, the topic distributions vary significantly within each class (image row). In fact, the topic of largest probability varies from column to column.

The inability of these topics to reflect semantic regularities is particularly problematic for image classification with LDA-like models, whose distinctive feature is precisely to isolate the topic probability vector Π from the features W. Consider, for example, the generative model of cLDA (see Fig. 1c). When conditioned on the topic-probability vector Π, the class label Y is independent of the image features W, i.e., $P_{Y|\Pi,W}(y|\boldsymbol{\pi},\mathcal{I}) = P_{Y|\Pi}(y|\boldsymbol{\pi})$. Although this independence is usually desirable—as topics introduce abstraction over visual features—it becomes a problem for classification, where it implies that images of similar topic distribution are assigned to the same class. In the case of Fig. 5, this means that images in the same column will receive the same class label, while images in

(a) Sample images



(b) Topics vectors produced by cLDA



(c) Topic vectors produced by ts-cLDA

Fig. 5. (a) Images of "Sailing" (top row) and "Rowing" (bottom row). (b) Expected topic vectors produced, for the images in (a), by cLDA with four topics. (c) Expected topics vectors produced, for the images in (a), by ts-cLDA.

the same row will likely not. This makes the class assignments of cLDA close to the worst possible.[5]

5 Topic Supervision

In this section, we discuss topic supervision for LDA models and its impact in learning and inference.

5.1 Topics Supervision in LDA Models

The simplest solution to the problem of Fig. 5 is to *force* topics to reflect the semantic regularities of interest. This consists of equating topics to class labels, and is denoted ts-LDA. For example, in Fig. 4, topics would be defined as *paired* and *singleton*, leading to the image grouping of Fig. 4d and the topic histograms of Fig. 4e. Images 1 and 2 and images 3 and 4 would then have identical topic distributions, and the topic distributions of the paired and singleton classes would be different. This is unlike Fig. 4c, where the two classes can have identical topic distributions.

Topic supervision was previously proposed in semi-LDA [26] and labLDA [19] for action and text classification, respectively. However, its impact on classification performance is difficult to ascertain from these works for several reasons. First, none of them performed a systematic comparison to existing LDA methods, preventing any analysis of the benefits of topic-supervision over standard LDA. Second, both semi-LDA and labLDA are topic-supervised versions of LDA. None of the existing works present topic-supervised versions of the classification models, such as cLDA and sLDA (which as we shall see outperform ts-LDA). Third, semi-LDA adopts an unconventional inference process, which assumes that $p(z_n|w_1, w_2, \ldots, w_n) \propto p(z_n|\boldsymbol{\pi})p(z_n|w_n)$. It is unclear how this affects the performance of the topic-supervised model. Finally, the goal of labLDA is to assign multiple labels per document. This is somewhat different from image classification, although it reduces to a topic-supervised model for classification if there is a single label per item.

5. A similar outcome occurs for sLDA, for which a similar independence property holds. As can be seen in Fig. 1d, when conditioned on the hidden topic assignment variables Z, the class label Y becomes independent of the image features W, i.e., $P_{Y|Z,W}(y|\bar{\mathbf{z}}, \mathcal{I}) = P_{Y|Z}(y|\bar{\mathbf{z}})$.

5.2 Models and Geometric Interpretation

To analyze the impact of topic-supervision on the various LDA models, we start by noting that the graphical model of the topic supervised extension of any LDA model is *exactly* the same as that of the model without topic supervision. The only subtle yet significant difference is that the topics are no longer discovered, but specified. It is thus possible to introduce topic-supervised versions of all models in the literature. In this paper, we consider three such versions, viz. "ts-LDA," "topic-supervised class LDA (ts-cLDA)," and "topic-supervised supervised LDA (ts-sLDA)." These are the topic-supervised versions of LDA, cLDA, and sLDA, respectively, with the following three distinguishing properties:

- The set of topics $\mathcal{T}$ *is* the set of class labels $\mathcal{Y}$.
- The samples from the topic variables Z_i are class labels.
- The topic conditional distributions $P_{W|Z}()$ are learned in a supervised manner.

We will see that this has the added advantage of simpler learning. It should be noted that ts-LDA is structurally equivalent to semi-LDA [26], but uses more principled inference.

Fig. 2c shows the schematic of ts-cLDA for a two class problem on a three word simplex. As with cLDA, Fig. 2b, Λ_1 and Λ_2 are topic-distributions. There is, however, a significant difference. While the topic distributions of cLDA, learned by topic discovery, can be positioned anywhere on the word simplex, those of ts-cLDA are specified and identical to the image classes. This makes the topic-conditional distributions of ts-cLDA identical to the class-conditional distributions of the flat model.

5.3 Learning and Inference with Topic-Supervision

In this section, we discuss learning and inference procedures for the ts-LDA models. The introduction of topic-level supervision decouples the learning of the topic-conditional distribution $P_{W|Z}()$ from that of the other model parameters, reducing learning complexity. In general, learning would require a strongly supervised training set, where each image patch is labeled with one of the topics in $\mathcal{Y}$ (recall that $\mathcal{T} = \mathcal{Y}$), i.e., known z_i^d for all i and d. However, such a set is usually not available. As in flat model learning, the parameters of ts-LDA models are learned with weak supervision, relying on multiple instance learning. Under

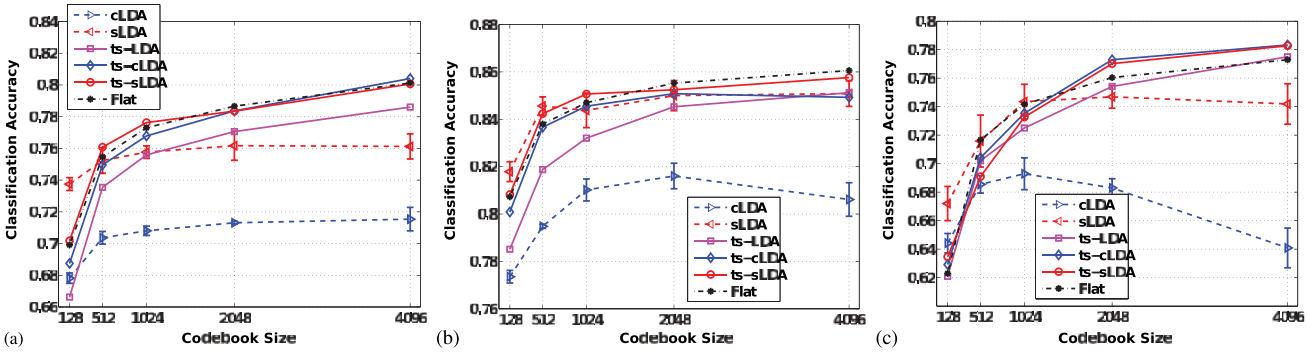


Fig. 6. Classification accuracy versus codebook size for ts-sLDA, ts-cLDA, sLDA, cLDA, and flat model on (a) N13, (b) N8, and (c) S8. For ts-sLDA and ts-cLDA the number of topics is equal to the number of classes. For sLDA and cLDA, results are presented for the number of topics of best performance.

TABLE 1
Classification Results

	Dataset				
model	**N15**	**N13**	**N8**	**S8**	**C50**
css-LDA	76.62 ± 0.32	81.03 ± 0.74	87.97 ± 0.84	80.37 ± 1.36	46.04
flat	74.91 ± 0.38	79.60 ± 0.38	86.80 ± 0.51	77.87 ± 1.18	43.20
ts-sLDA	74.82 ± 0.68	79.70 ± 0.48	86.33 ± 0.69	78.37 ± 0.80	42.33
ts-cLDA	74.38 ± 0.78	78.92 ± 0.68	86.25 ± 1.23	77.43 ± 0.97	40.80
ts-LDA	72.60 ± 0.51	78.10 ± 0.31	85.53 ± 0.41	77.77 ± 1.02	39.20
medLDA [28]	72.08 ± 0.59	77.58 ± 0.58	85.16 ± 0.57	78.19 ± 1.05	41.89
sLDA [25]	70.87 ± 0.48	76.17 ± 0.92	84.95 ± 0.51	74.95 ± 1.03	39.22
cLDA [10]	65.50 ± 0.32	72.02 ± 0.58	81.30 ± 0.55	70.33 ± 0.86	34.33
LDA-SVM	73.19 ± 0.51	78.45 ± 0.34	86.82 ± 0.93	76.32 ± 0.71	45.46

this form of learning, all patch labels in an image are made equal to its class label, i.e., $z_n^d = y^d \; \forall n, d$.

Given z_n^d, the parameters Λ_k of the topic-conditional distribution are learned by ML estimation:

$$\Lambda_{kv}^* = \arg\max_{\Lambda_k} \sum_d \sum_n \delta(y^d, k)\delta(w_n^d, v) \log \Lambda_{kv}, \tag{13}$$

under the constraint $\sum_{v=1}^{|\mathcal{V}|} \Lambda_{kv} = 1$. The ML solution is

$$\Lambda_{kv} = \frac{\sum_d \sum_n \delta(y^d, k)\delta(w_n^d, v)}{\sum_j \sum_d \sum_n \delta(y^d, j)\delta(w_n^d, v)}. \tag{14}$$

Comparing to (4), it is clear that, when $\mathcal{T} = \mathcal{Y}$, the topic distributions of topic-supervised models are equivalent to the class-conditional distributions of the flat model.[6]

The remaining model parameters could also be learned under the assumption of known z_n^d (see Appendix III-B, which is available in the online supplemental material, for the case of ts-cLDA). However, under weak supervision, all patches of a given class would be assigned to the same topic (that of the class), leading to degenerate parameters for the class conditional distributions. For example, the ts-cLDA parameters $\boldsymbol{\alpha}_c$ would be zero for all topics other than the class namesake. Hence, although useful for learning topic-conditional distributions, multiple instance learning only produces trivial class-conditional distributions. A better solution, which we adopt, is to learn these distributions under the assumption of unknown patch labels, as is done in the original algorithms. In summary, weakly supervised learning is only used to learn topic-conditional distributions. Parameter estimation for ts-cLDA is detailed in Appendix III, which is available in the online supplemental material.

5.4 Experimental Analysis

Fig. 5c shows the topic vectors produced by ts-cLDA for the images of Fig. 5a. Note how these vectors are different across columns but similar across rows, i.e., have the opposite behavior of those produced by cLDA (see Fig. 5b). Clearly, ts-cLDA latches onto the semantic regularities of interest for classification. It is thus expected to outperform cLDA in image classification. This is confirmed by Fig. 6, which presents classification results of ts-LDA, ts-cLDA, and ts-sLDA as a function of codebook size under the experimental conditions of Fig. 3. Also shown are the accuracies of cLDA, sLDA, and the flat model. In all cases, each image is assigned to the class of highest posterior probability.

All three topic supervised approaches outperform sLDA and cLDA. This holds for all datasets and codebook sizes when compared to cLDA, and for all datasets and codebooks with over 1,024 codewords when compared to sLDA. The best performance across codebook and topic cardinalities is reported in Table 1. On average, topic-supervision improves the accuracy of cLDA and sLDA by 12 and

6. Note that the extension to the case where topics are supervised but different from the image classes, i.e., $\mathcal{T} \neq \mathcal{Y}$, would be trivial. For example, for classes {"Beach," "Lake"}, topics could be defined as {"sand," "water," "sky," "trees"}, promoting a natural hierarchy of concepts. In other applications, topics could be class attributes. All of these would, however, require additional labeling of training sets with respect to the desired topics.

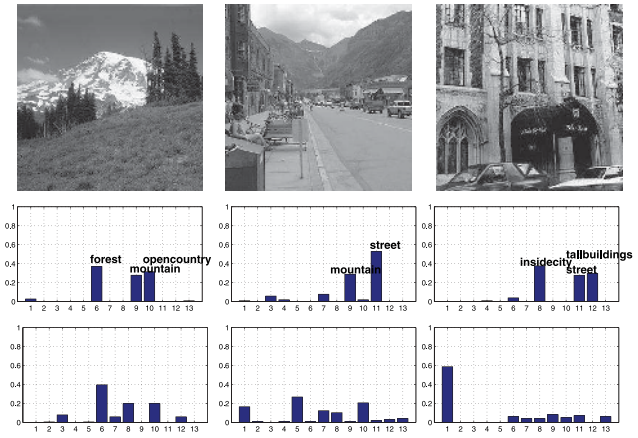


Fig. 7. Top: Images misclassified by cLDA but not ts-cLDA. Bottom: Expected topic distributions of ts-cLDA (middle) and cLDA (bottom), using 13 topics. ts-cLDA topics are class labels, and high probability topics capture the image semantics. cLDA topics lack clear semantics.

5 percent, respectively. This happens despite the fact that the number of topics of cLDA and sLDA is usually much larger than those of the topic-supervised models (number of classes). Among the topic-supervised models, ts-cLDA and ts-sLDA achieve comparable performance, superior to that of ts-LDA. Fig. 7 shows images incorrectly classified by cLDA but correctly classified by ts-cLDA on N13. Also shown are the topic histograms obtained in each case, with ts-cLDA in the middle and cLDA in the bottom row. Again, the figures illustrate the effectiveness of ts-cLDA at capturing semantic regularities. Topics with high probability are indeed representative of the image semantics. This interpretation is only possible as the topic labels in ts-cLDA have a one-to-one correspondence with the class labels. For cLDA, topic histograms merely represent visual clusters.

5.5 Topic Supervision versus the Flat Model

In this section, we discuss the similarities and differences between ts-LDA models and the flat model.

5.5.1 ts-LDA versus Flat Model

The closest topic-supervised model to the flat model is ts-LDA. Although the graphical models are different (cf. Figs. 1a and 1b), the weakly supervised learning procedure is similar. In fact, since the learned topic-conditional distributions of ts-LDA are the class-conditional distributions of the flat model, the training procedure is essentially the same. There are, however, significant differences between the inference procedures of the two models. For the flat model, assuming a uniform class prior, the class assignment is

$$y^* = \arg\max_y \prod_n P_{W|Y}(w_n|y) P_Y(y) \tag{15}$$

$$= \arg\max_y \prod_n \Lambda_{yw_n}. \tag{16}$$

For ts-LDA, the class assignment is

$$y^* = \arg\max_y \gamma_y, \tag{17}$$

where γ_y is obtained through the iteration (see Appendix I, which is available in the online supplemental material):

$$\gamma_y^* = \sum_n \phi_{ny} + \alpha, \tag{18}$$

$$\phi_{ny}^* \propto \Lambda_{yw_n} \exp[\psi(\gamma_y)]. \tag{19}$$

While in (16) the word probabilities Λ_{yw_n} are aggregated via multiplication, in (18) and (19) they are effectively aggregated via summation. In summary, while the class assignments of the flat model are primarily driven by the product of word probabilities, those of ts-LDA are driven by their sum. The two quantities can be quite different. Fig. 6 shows that this difference is of consequence as ts-LDA frequently underperforms the flat model.

5.5.2 ts-{c,s}LDA versus Flat Model

Like ts-LDA, the topic-conditional distributions of ts-cLDA and ts-sLDA are identical to the class-conditional distributions of the flat model. However, since the class-conditional distributions of ts-cLDA and ts-sLDA have additional parameters ($\boldsymbol{\alpha}_y$ for ts-cLDA and $\boldsymbol{\zeta}_y$ for ts-sLDA), their learning algorithms are different from that of the flat model. During inference, the class posterior $P_{Y|W}(y|\mathcal{I})$ critically depends on the class specific parameters $\boldsymbol{\alpha}_y/\boldsymbol{\zeta}_y$. These are quite different from the class specific parameters of the flat model Λ_y. Thus, for ts-cLDA and ts-sLDA, both training and testing have significant differences from those of the flat model.

Fig. 6 shows that, although outperforming their unsupervised counterparts, topic-supervised models cannot beat the flat model. This shows that the differences between these models are of little consequence. In particular, it suggests that modeling class distributions on the topic simplex (see Fig. 2c) does not necessarily improve recognition performance. This is troubling since such modeling increases the complexity of both LDA learning and inference, which are certainly larger than those of the flat model. It places in question the usefulness of the whole LDA approach to image classification. This problem has simply been ignored in the literature, where comparisons with the flat model are usually not presented.

6 Class-Specific-Simplex Latent Dirichlet Allocation

To overcome this limitation, we introduce a new LDA model for image classification, denoted css-LDA.

6.1 Motivation

The inability of the LDA variants to outperform the flat model is perhaps best understood by returning to Fig. 2. Note that both cLDA and ts-cLDA map images from a high-dimensional word simplex to a low-dimensional topic simplex, which is common to all classes. This restricts the scope of the class models, which are simple Dirichlet distributions over the topic simplex. Similar pictures hold for sLDA and ts-sLDA, where the classes define a softmax function in the simplex. In fact, even SVM-LDA learns an SVM classifier on this space. Since the topic simplex is

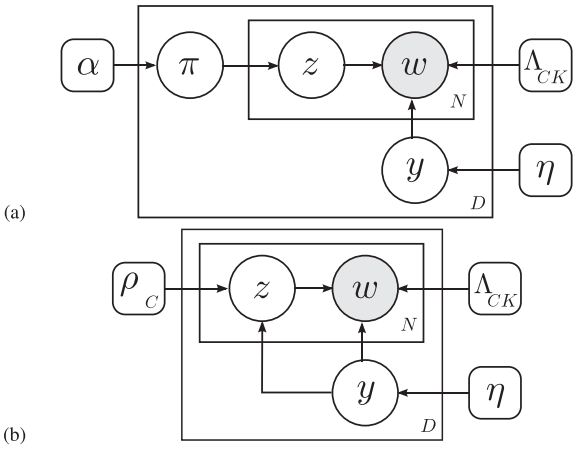


Fig. 8. Graphical model of (a) css-LDA and (b) MFM.

common and low-dimensional, too few degrees of freedom are available to characterize intraclass structure, preventing a very detailed discrimination of the different classes. In fact, the main conclusion of the previous sections is that the bulk of the modeling power of LDA lies in the selection of the topic simplex, and not in the modeling of the data distribution in it. Since to capture the semantic regularities of the data, the simplex has to be aligned with the class labels—as is done under topic-supervision—there is little room to outperform the flat model.

This limitation is common to any model that constrains the class-conditional distributions to lie on a common topic simplex. This is the case whenever the class label Y is connected to either the prior Π or topic Z variables, as in the graphical models of Fig. 1. Since the topic simplex is smaller than the word simplex, it has limited ability to simultaneously model rich intraclass structure and keep the classes separated. For this, it is necessary that the class label Y affect the word distributions *directly*, freeing these to distribute themselves across the word simplex in the most discriminant manner. This implies that Y must be connected to the word variable W, as in the flat model. The result is the graphical model of Fig. 8a, which turns out to have a number of other properties of interest.

The first follows from the fact that it makes the topic conditional distributions dependent on the class. Returning to Fig. 2, this implies that the vertices of the topic simplex are class-dependent, as shown in Fig. 2d. Note that there are two 1D topic simplices, one for each class defined by the parameters Λ_1^y and Λ_2^y, $y \in \{1, 2\}$. The dotted and dashed lines denote the prior distribution on the topic simplices, which is controlled by the $\boldsymbol{\alpha}$ parameter. Hence, each class is endowed with its own topic simplex justifying the denomination of the model as css-LDA.

Second, css-LDA extends both the flat and the LDA model, simultaneously addressing the main limitations of these two models. With respect to LDA, it is a supervised extension that, unlike cLDA or sLDA, relies on the most expressive component of the model (topic simplex) to achieve class discrimination. Because there are *multiple topic simplices*, the class-conditional distributions can have little overlap in word-simplex even when topic simplices are low-dimensional. In particular, topic distributions no longer need to have the consistency of Fig. 5c. On the contrary, the distributions from images of a given class are now free to be all over *its* topic simplex. Since the simplex is different from those of other classes, this does not compromise discrimination. On the other hand, because a much larger set of topic distributions is now possible per class, the model has much greater ability to model intraclass structure. In summary, when compared to LDA, cLDA, or sLDA, css-LDA combines improved class separation with improved capacity to model intraclass structure.

With respect to the flat model, css-LDA inherits the advantages of LDA over bag-of-words. Consider the collection of images from the class "Bocce" of the S8 dataset (see Appendix IV-A, which is available in the online supplemental material), shown on the left side of Fig. 9. Note that the game of Bocce can be played indoors or outdoors, on beach or on grass, on an overcast day or a sunny day, and so on. Each of these conditions leads to drastically different scene appearances and thus a great diversity of word distributions. Under the flat model, the "Bocce" class is modeled by a single point in the word simplex, the average of all these distributions, as shown in Fig. 2a. This is usually insufficient to capture the richness of each class. Rather than this, css-LDA devotes to each class a topic simplex, as shown in Fig. 2d. This increases the

Fig. 9. Two-dimensional embedding of the topic vectors discovered by css-LDA (marked #1-#10), and class-conditional distribution of flat model (marked flat model), for left: "Bocce" (S8) and right: "Highway" (N13) classes. Also shown are the nearest neighbor images of sample topic conditional distributions.

expressive power of the model because there are now many topic-conditional word distributions per class.

Thus, css-LDA can account for much more complex class structure than the flat counterpart. In the example of Fig. 2, while the flat model approximates all the images of each class by a point in word simplex, css-LDA relies on a line segment. In higher dimensions, the difference can be much more substantial since each topic simplex is a subspace of dimension $K-1$ (K the number of topics), while the approximation of the flat model is always a point. In summary, when compared to the flat model, css-LDA has a substantially improved ability to model intraclass structure. In fact, it is able to harness the benefits of both topic-supervision-as each topic learned is learned under class supervision, and topic-discovery-as several topics are discovered per class.

6.2 The css-LDA Model

The generative process of css-LDA is

for each image **do**
  sample $\boldsymbol{\pi} \sim P_{\Pi}(\boldsymbol{\pi}; \boldsymbol{\alpha})$.
  sample a class label $y \sim P_Y(y; \boldsymbol{\eta})$, $y \in \mathcal{Y} = \{1, \ldots, C\}$.
  for $i \in \{1, \ldots, N\}$ **do**
    sample a topic, $z_i \sim P_{Z|\Pi}(z_i|\boldsymbol{\pi})$, $z_i \in \mathcal{T} = \{1, \ldots, K\}$,
    where $\mathcal{T}$ is the set of topics.
    sample a visual word $w_i \sim P_{W|Z,Y}(w_i|z_i, y; \Lambda^y_{z_i})$.
  end for
end for

Similar to the earlier LDA extensions, $P_Y()$ is a categorical distribution over $\mathcal{Y}$ with parameter $\boldsymbol{\eta}$, $P_{\Pi}()$ a Dirichlet distribution on the topic simplex with parameter $\boldsymbol{\alpha}$, $P_{Z|\Pi}()$ a categorical distribution over $\mathcal{T}$ with parameter $\boldsymbol{\pi}$, and $P_{W|Z,Y}()$ a categorical distribution over $\mathcal{V}$ with a class dependent parameter Λ^y_z.

Like previous models, learning and inference are intractable. Given an image $\mathcal{I} = \{w_1, \ldots, w_N\}$, $w_n \in \mathcal{V}$, inference consists of computing the posterior distribution

$$P_{Y,\Pi,Z|W}(y, \boldsymbol{\pi}, z_{1:N}|\mathcal{I}) = P_{\Pi,Z|W,Y}(\boldsymbol{\pi}, z_{1:N}|\mathcal{I}, y) P_{Y|W}(y|\mathcal{I}), \tag{20}$$

where

$$P_{Y|W}(y|\mathcal{I}) = \frac{P_{Y,W}(y, \mathcal{I})}{\sum_c P_{Y,W}(c, \mathcal{I})}. \tag{21}$$

Both $P_{Y,W}(y, \mathcal{I})$ and $P_{\Pi,Z|W,Y}(\boldsymbol{\pi}, z_{1:N}|\mathcal{I}, y)$ are intractable and approximated using variational methods. The posterior $P_{\Pi,Z|W,Y}(\boldsymbol{\pi}, z_{1:N}|\mathcal{I}, y)$ is approximated by the variational distribution

$$q(\boldsymbol{\pi}, z_{1:N}) = q(\boldsymbol{\pi}; \boldsymbol{\gamma}) \prod_n q(z_n; \boldsymbol{\phi}_n). \tag{22}$$

The marginal likelihood $P_{Y,W}(y, \mathcal{I})$ is approximated by maximizing the evidence lower bound $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda)$ for different values of $y \in \{1, \ldots, C\}$, i.e.,

$$P_{W,Y}(\mathcal{I}, y) \sim \max_{\boldsymbol{\gamma}, \boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda), \tag{23}$$

$$\text{where} \quad \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda) = E_q[\log P(y, \boldsymbol{\pi}, z_{1:N}, w_{1:N})] - E_q[\log q(\boldsymbol{\pi}, z_{1:N})]. \tag{24}$$

Solving (23) for a given y results in updates similar to the standard LDA inference equations (see Appendix I, which is available in the online supplemental material):

$$\gamma^*_k = \sum_n \phi_{nk} + \alpha_k, \tag{25}$$

$$\phi^*_{nk} \propto \Lambda^y_{kw_n} \exp[\psi(\gamma_k)]. \tag{26}$$

Note that for css-LDA, where each class is associated with a separate topic simplex, (26) differs from standard LDA in that the Λ parameters are class specific.

Learning involves estimating the parameters $(\boldsymbol{\eta}, \boldsymbol{\alpha}, \Lambda^{1:C}_{1:K})$ by maximizing the log likelihood, $l = \log P_W(\mathcal{D})$, of a training image dataset $\mathcal{D}$. This is done with a variational EM algorithm that iterates between.

Variational E-step approximates the posterior $P_{\Pi,Z}(\boldsymbol{\pi}^d, z^d_{1:N}| \mathcal{I}^d, y^d)$ by a variational distribution $q(\boldsymbol{\pi}^d, z^d_{1:N})$ parameterized by $(\boldsymbol{\gamma}^d, \boldsymbol{\phi}^d_n)$, where d indexes the images in the training set. This leads to the update rules of (25) and (26).

M-Step computes the values of parameters $(\boldsymbol{\alpha}, \Lambda^{1:C}_{1:K})$. $\boldsymbol{\alpha}$ is obtained from

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \sum_d \left[-\log \mathcal{B}(\boldsymbol{\alpha}) + \sum_k (\alpha_k - 1) E_q\big[\log \pi^d_k\big] \right], \tag{27}$$

with

$$E_q\big[\log \pi^d_k\big] = \psi\big(\gamma^d_k\big) - \psi\Big(\sum_l \gamma^d_l\Big), \tag{28}$$

$$\mathcal{B}(\boldsymbol{\alpha}) = \frac{\prod_k (\Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}, \tag{29}$$

and $\Gamma()$ the Gamma function. This optimization can be carried out by the method of Newton-Raphson, as detailed in [15].

The parameters $\Lambda^{1:C}_{1:K}$ are obtained from

$$\Lambda^{y\ *}_{kv} = \arg\max_{\Lambda^y_{kv}} \sum_d \sum_n \delta\big(y^d, y\big) \delta\big(w^d_n, v\big) \phi^d_{nk} \log \Lambda^y_{kv}, \tag{30}$$

such that $\sum_{v=1}^{|\mathcal{V}|} \Lambda_{kv} = 1$. Using the method of Lagrange multipliers this results in the closed form update:

$$\Lambda^y_{kv} = \frac{\sum_d \sum_n \delta\big(y^d, y\big) \delta\big(w^d_n, v\big) \phi^d_{nk}}{\sum_u \sum_d \sum_n \delta\big(y^d, y\big) \delta\big(w^d_n, u\big) \phi^d_{nk}}. \tag{31}$$

Similarly to previous models, we assume a uniform class prior, i.e., $\eta_y = \frac{1}{C}$, $\forall y$. Note that in (27) learning $\boldsymbol{\alpha}$ requires computing $E_q[\log \pi^d_k]$ for $d \in \{1, \ldots, |\mathcal{D}|\}$, i.e., for all images in the training set. Furthermore, in (31) Λ^y_k has an indirect dependence on $\boldsymbol{\alpha}$ through (25) and (26), in effect coupling the learning of the topic distributions to the entire training dataset. However, as we shall see, the performance of css-LDA is not very sensitive to the $\boldsymbol{\alpha}$ parameter. Hence, instead of learning it, we set it to a fixed value. This in turn simplifies the learning of the topic distributions in (31) since Λ^y_k depends only on the

images of class y, i.e., the topic distributions can be learned independently for each class, in parallel.

Given an image $\mathcal{I}_q$, css-LDA based classification is performed with the minimum probability of error rule of (1).

6.3 Comparison with Mixture Models

Css-LDA can be interpreted as a mixture of LDA models, each with its topic simplex. An interesting question is whether this is the simplest mixture model with the properties of the previous section. A natural alternative is the *mixture of flat models* (MFM) with the graphical model of Fig. 8b. It models each class as a mixture of K multinomial distributions:

$$P_{W|Y}(w|y) = \sum_z P_{W|Z,Y}(w|z,y)P_{Z|Y}(z|y) \tag{32}$$

$$= \sum_z \Lambda^y_{zw}\rho^y_z, \tag{33}$$

where $\Lambda^y_{1:K}$ are the parameters of the mixture components of class y ($\sum_v \Lambda^y_{zv} = 1$) and $\rho^y_{1:K}$ the mixing probabilities ($\sum_z \rho^y_z = 1$). While most mixture models are quite effective at modeling multimodal probability distributions, this is not the case for the mixture of multinomials. In this case,

$$P_{W|Y}(\mathcal{I}|y) = \sum_{z_1}\cdots\sum_{z_n}\prod_n P_{W,Z|Y}(w_n, z_n|y) \tag{34}$$

$$= \prod_n \sum_z P_{W|Y,Z}(w_n|y,z)P_{Z|Y}(z|y) \tag{35}$$

$$= \prod_n \bar{\Lambda}^y_{w_n}, \tag{36}$$

with $\bar{\Lambda}^y_w = \sum_z \Lambda^y_{zw}\rho^y_z$. Hence, the mixture model collapses to a multinomial distribution whose parameters are an average of those of the mixture components. It follows that the MFM is equivalent to the flat model of Fig. 2a.

This suggests that the properties of css-LDA, Fig. 2d, may not be trivial to achieve with simpler mixture models. In fact, the image specific mixing prior Π of css-LDA appears to be critical since it makes the class conditional distributions:

$$P_{W|Y}(\mathcal{I}|y) = \int d\boldsymbol{\pi} \sum_{z_1}\cdots\sum_{z_n}\prod_n P_{W,Z,\Pi|Y}(w_n, z_n, \boldsymbol{\pi}|y) \tag{37}$$

$$= \int d\boldsymbol{\pi} \prod_n \sum_z P_{W|Y,Z}(w_n|y,z)P_{Z|\Pi}(z|\boldsymbol{\pi})P_\Pi(\boldsymbol{\pi}) \tag{38}$$

$$= E_{\boldsymbol{\pi}}\left[\prod_n \sum_z \Lambda^y_{zw_n}\pi_z\right], \tag{39}$$

where $E_{\boldsymbol{\pi}}[.]$ is the expected value under $P_\Pi()$. Unlike ρ_z in (33), π_z is a random variable. The expectation of (39) prevents the collapse to a unimodal distribution, as in MFM. In summary, css-LDA is one of the simplest ways to endow the flat model framework with multimodal class distributions. In fact, css-LDA with one topic reduces to the flat model.

7 Results

Several experiments were performed to evaluate css-LDA, using the experimental setup discussed in Appendix IV, which is available in the online supplemental material.

7.1 Class Specific Topic Discovery in css-LDA

We start with a set of experiments that provide insight on the topics discovered by css-LDA. Fig. 9 presents a visualization of the topic-conditional distributions Λ^y_z (marked #1 to #10) discovered for classes "Bocce" (S8, left) and "Highway" (N13, right), using 10 topics per class. Also shown is the class conditional distribution Λ^{flat}_y (marked flat model) of the flat model. The visualization was produced by a 2D embedding of the word simplex, using nonmetric multidimensional scaling [12] from the matrix of KL divergences between topic- and class-conditional distributions. Note how, for both classes, the flat model is very close to the average of the topic-conditional distributions. This shows that, on average, topics discovered by css-LDA represent the class conditional distribution of the flat model. In fact, the KL divergence between the average of the topic conditional distributions of css-LDA and the class conditional distribution of the flat model is very close to zero (0.013 ± 0.019 for N13, 0.014 ± 0.008 for S8). Also shown in the figure, for some sample topics, are the two images closest to the topic conditional distribution. Note that the topics discovered by css-LDA capture the visual diversity of each class. For example, "Bocce" topics #9, #7, #8, and #1 capture the diversity of environments on which sport can be played: indoors, sunny-outdoor, overcast-outdoor, and beach. These variations are averaged out by the flat model, where each class is, in effect, modeled by a single topic.

7.2 Classification Results

We have previously established that topic-supervision yields better classification accuracy than LDA models learned without topic-supervision. However, topic supervision is not enough to outperform the flat model. This can be seen more clearly in Fig. 10, which reports the performance of ts-sLDA (which has the best performance amongst the topic-supervised models) and the flat model for N13, N8, and S8. In all cases, the performance of the latter is very close to that of the former. This is in stark contrast to css-LDA, which has a clearly better performance than the two, across datasets and codebook sizes. Since css-LDA is an extension of the flat model, this gain can be attributed to its topic-discovery mechanism. We have also implemented the MFM model and verified that, as discussed in Section 6.3, it collapses to a flat model. In all experiments, the parameters of the collapsed MFM (learned using EM) were identical to those of the flat model (learned by maximum likelihood estimation). Hence, the performance of MFM is identical to that of the flat model.

Fig. 11a presents the performance of css-LDA as a function of the $\boldsymbol{\alpha}$ parameter on S8 (for different codebook sizes and 60 topics). It is clear that css-LDA is not very sensitive to this parameter. Fig. 11b shows the performance of css-LDA as a function of the smoothing hyperparameter of the topic-distributions. Again, the model is not very sensitive to this parameter. Table 1 summarizes the best

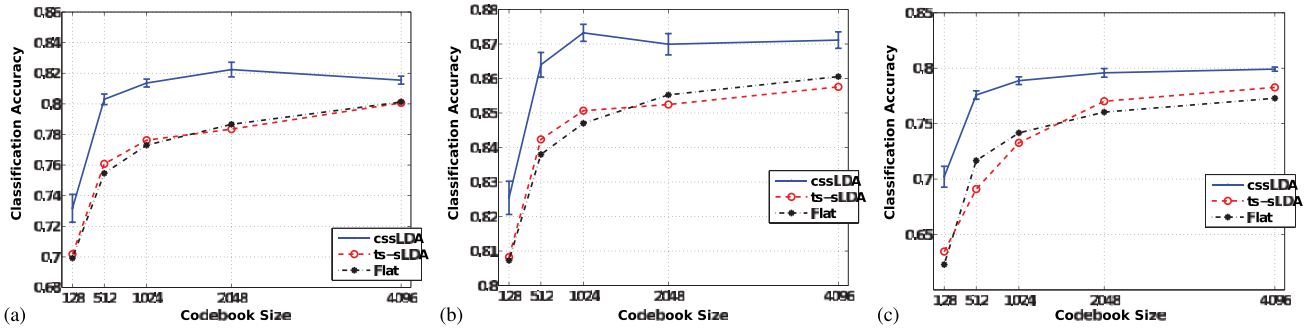


Fig. 10. Classification accuracy of css-LDA, ts-sLDA, and the flat model, as a function of codebook size on (a) N13, (b) N8, and (c) S8. The reported css-LDA performance is the best across number of topics, while for ts-sLDA the number of topics is equal to the number of classes.

classification accuracy achieved by all methods considered in this work, plus SVM-LDA [4], [5], [18], on all datasets. Note that css-LDA outperforms all existing generative models for image classification and a discriminative classifier, SVM-LDA, on all five datasets, with a classification accuracy of 76.62 percent on N15, 81.03 percent on N13, 87.97 percent on N8, 80.37 percent on S8, and 46.04 percent on C50. On average, it has a relative gain of 3.0 percent over the flat model, 3.5 percent over ts-sLDA, 4.9 percent over ts-cLDA, 6.7 percent over ts-LDA, 8.5 percent over sLDA, 17.2 percent over cLDA, and 4.0 percent over SVM-LDA.

7.3 Time Complexity

Fig. 12 compares the time required for training and testing of the different models on N13, using a 1,024-word codebook. All experiments were conducted on a 2× Intel Xeon E5504 Quad-core 2.00 GHz processor, with average image size of 270×250 pixels. The figure does not account for the time required to compute the BoW representation—800 (20) milliseconds per image to compute SIFT (DCT) features, 15 minutes to learn a 1,024 codeword codebook (see Appendix IV-B, which is available in the online supplemental material), and 1.5 seconds per image to compute word histograms—which is common to all methods. Both training and testing of the flat model are significantly faster than those of the LDA models. Among the latter, the decoupled parameter learning of topic supervised models (ts-cLDA/ts-sLDA) enables an order of magnitude learning speed-up over the nonsupervised versions (cLDA/sLDA). The learning time of css-LDA is quite close to that of cLDA, and about one order of magnitude smaller than sLDA. Furthermore, since the topic distributions of css-LDA can be learned in parallel for the different classes, this time can be trivially decreased by a factor of K (e.g., 13× for this dataset). This makes css-LDA one of the fastest LDA models to learn. This type of computational efficiency is not possible for the other LDA models, where topics are shared between all classes. During testing, css-LDA is an order of magnitude

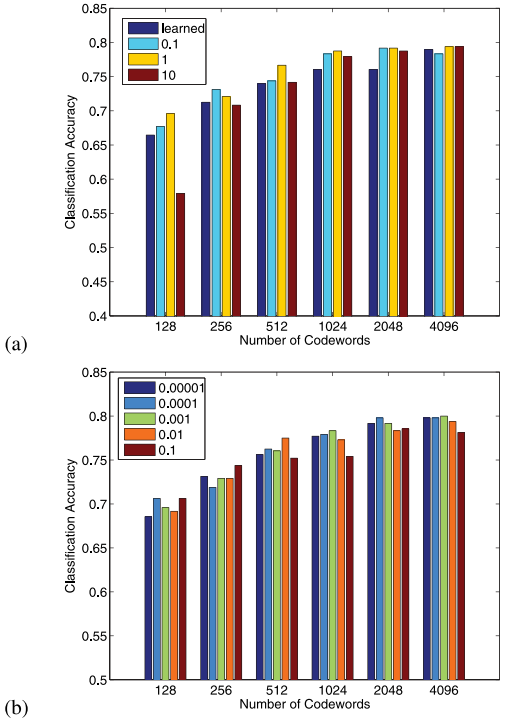


Fig. 11. Classification accuracy of css-LDA (with 60 topics) on S8 as a function of (a) α_k and (b) hyperparameter of topic-distributions.

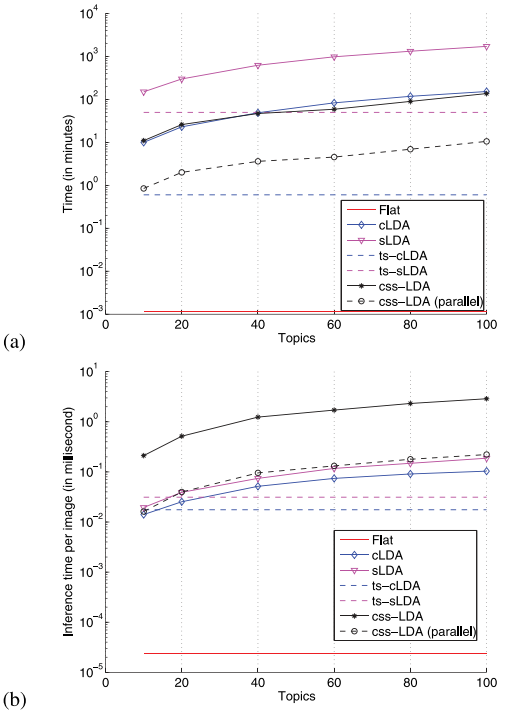


Fig. 12. Time complexity of various models on N13. (a) Learning time (in minutes). (b) Inference time per image (in milliseconds).

slower than the other LDA models. Again, the computations can be parallelized, in which case test time is equivalent to those of cLDA and sLDA. In this case, the average inference time of css-LDA was about 0.1 milliseconds per image for 100 topics.

8 Conclusion

In this paper, we proposed two novel families of LDA models, viz. ts-LDA and css-LDA, for image classification. We have argued that current supervised extensions of LDA are driven by generic image regularities, rather than the semantic regularities of interest for classification, and thus not suitable for image classification. These arguments were shown to hold by 1) a theoretical analysis of the algorithms used to learn these models, which revealed a vanishingly small impact of class supervision on topic discovery for large training samples, and 2) experiments where unsupervised topic discovery was shown to achieve performance at par with these models. To address this limitation, we proposed ts-LDA models where, instead of automated topic *discovery*, topics are *specified* and equal to the classes of interest for classification. In particular, we introduced topic-supervised versions of LDA, cLDA, and sLDA viz. ts-LDA, ts-cLDA, and ts-sLDA, respectively. In all cases, topic supervised models outperformed the corresponding models without topic-supervision.

It was then noted that even the topic supervised models fail to outperform the much simpler flat model. This was explained by the fact that the bulk of the modeling power of LDA rests on a dimensionality reduction from a high-dimensional word simplex to a low-dimensional topic simplex, and not on the modeling of class distributions on the latter. Due to the weak influence of class labels in cLDA and sLDA topic discovery, the resulting topic simplex has limited discrimination ability. While this problem is solved by the topic-supervised extensions, whose topic-conditional distributions were shown equivalent to the class-conditional distributions of the flat model, the price is a loss of ability to discover unanticipated structure of interest for image classification. Since the topic simplex has to be aligned with the class labels, there is little room to outperform the flat model.

It was shown that this limitation is shared by any model that constrains the class-conditional distributions to lie on a common topic simplex. This was addressed by the introduction of css-LDA, where class supervision occurs at the level of visual features. This induces a topic-simplex per class, freeing topic distributions to populate the word simplex in the most discriminant manner. Thus, the model combines ability to discriminate between classes with capacity to richly model intraclass structure. In particular, css-LDA was shown to extend both the flat model and LDA, combining the labeling strength of the former with the flexibility of topic discovery of the latter. Its effectiveness was tested through an extensive experimental comparison to the previous LDA-based image classification approaches, on several benchmark datasets. In this evaluation, css-LDA consistently outperformed all previous models.

References

[1] C. Bishop, *Pattern Recognition and Machine Learning*, vol. 4. Springer, 2006.

[2] D. Blei and M. Jordan, "Pattern Recognition and Machine Learning," *Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 127-134, 2003.

[3] D. Blei and J. McAuliffe, "Supervised Topic Models," *Advances in Neural Information Processing Systems*, vol. 20, pp. 121-128, 2008.

[4] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *The J. Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[5] A. Bosch, A. Zisserman, and X. Muoz, "Scene Classification Using a Hybrid Generative/Discriminative Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712-727, Apr. 2008.

[6] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning Mid-Level Features for Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2559-2566, 2010.

[7] W. Buntine, "Operations for Learning with Graphical Models," *J. Artificial Intelligence Research*, vol. 2, pp. 159-225, 1994.

[8] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394-410, Mar. 2007.

[9] G. Csurka, C. Dance, L. Fan, and C. Bray, "Visual Categorization with Bags of Keypoints," *Proc. European Conf. Computer Vision Workshop Statistical Learning in Computer Vision*, vol. 1, pp. 1-22, 2004.

[10] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 524-531, 2005.

[11] T. Hofmann, "Probabilistic Latent Semantic Indexing," *Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 50-57, 1999.

[12] J. Kruskal, "Nonmetric Multidimensional Scaling: A Numerical Method," *Psychometrika*, vol. 29, no. 2, pp. 115-129, 1964.

[13] S. Lacoste-Julien, F. Sha, and M. Jordan, "DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification," *Proc. Advances in Neural Information Processing Systems Conf.*, vol. 21, 2008.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2169-2178, 2006.

[15] T. Minka, "Estimating a Dirichlet Distribution," vol. 1, p. 3, http://research.microsoft.com/minka/papers/dirichlet/, 2000.

[16] D. Putthividhya, H. Attias, and S. Nagarajan, "Supervised Topic Model for Automatic Image Annotation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 1894-1897, 2010.

[17] D. Putthividhya, H. Attias, and S. Nagarajan, "Topic Regression Multi-Modal Latent Dirichlet Allocation for Image Annotation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3408-3415, 2010.

[18] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A Thousand Words in a Scene," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1575-1589, Sept. 2007.

[19] D. Ramage, D. Hall, R. Nallapati, and C. Manning, "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora," *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 248-256, 2009.

[20] J. Rennie, "Improving Multi-Class Text Classification with Naive Bayes," PhD thesis, Massachusetts Inst. of Technology, 2001.

[21] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[22] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *Proc. Ninth IEEE Int'l Conf. Computer Vision*, pp. 1470-1477, 2003.

[23] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," *Handbook of Latent Semantic Analysis*, vol. 427, no. 7, pp. 424-440, Psychology Press, 2007.

[24] M. Vasconcelos, N. Vasconcelos, and G. Carneiro, "Weakly Supervised Top-Down Image Segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1001-1006, 2006.

[25] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous Image Classification and Annotation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1903-1910, 2009.

[26] Y. Wang, P. Sabzmeydani, and G. Mori, "Semi-Latent Dirichlet Allocation: A Hierarchical Model for Human Action Recognition," *Proc. Second Conf. Human Motion: Understanding, Modeling, Capture, and Animation*, pp. 240-254, 2007.
[27] J. Winn, A. Criminisi, and T. Minka, "Object Categorization by Learned Universal Visual Dictionary," *Proc. 10th IEEE Int'l Conf. Computer Vision*, pp. 1800-1807, 2005.
[28] J. Zhu, A. Ahmed, and E. Xing, "MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification," *Proc. 26th Ann. Int'l Conf. Machine Learning*, pp. 1257-1264, 2009.

Nikhil Rasiwasia received the BTech degree in electrical engineering from the Indian Institute of Technology Kanpur in 2005 and the MS and PhD degrees from the University of California, San Diego, in 2007 and 2011, respectively. He is currently a scientist at Yahoo Labs! Bangalore. His research interests include the areas of computer vision and machine learning. He was recognized as an "emerging leader in multimedia" in 2008 by IBM T.J. Watson Research. He was also awarded the Best Student Paper Award at ACM Multimedia 2010. He is a member of the IEEE.

Nuno Vasconcelos received the Licenciatura degree in electrical engineering and computer science from the Universidade do Porto, Portugal, and the MS and PhD degrees from the Massachusetts Institute of Technology. He is a professor in the Electrical and Computer Engineering Department at the University of California, San Diego, where he heads the Statistical Visual Computing Laboratory. He has received the US National Science Foundation CAREER Award, a Hellman Fellowship, and has authored more than 150 peer-reviewed publications. He is a senior member of the IEEE.