Probabilistic Models for Supervised Dictionary Learning

Xiao-Chen Lian[1], Zhiwei Li[2,3], Changhu Wang[3], Bao-Liang Lu[1,2] and Lei Zhang[3]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China

[2]MOE-MS Key Lab for Intelligent Computing and Intelligent Systems, Shanghai Jiao Tong University, China

[3]Microsoft Research Asia

lianxiaochen@gmail.com, {zli,chw}@microsoft.com, bllu@sjtu.edu.cn, leizhang@microsoft.com

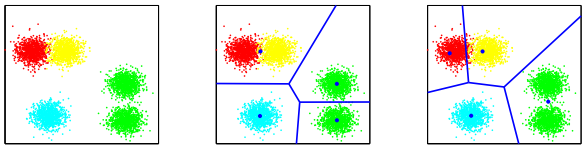
Abstract

Dictionary generation is a core technique of the bag-of-visual-words (BOV) models when applied to image categorization. Most of previous approaches generate dictionaries by unsupervised clustering techniques, e.g. k-means. However, the features obtained by such kind of dictionaries may not be optimal for image classification. In this paper, we propose a probabilistic model for supervised dictionary learning (SDLM) which seamlessly combines an unsupervised model (a Gaussian Mixture Model) and a supervised model (a logistic regression model) in a probabilistic framework. In the model, image category information directly affects the generation of a dictionary. A dictionary obtained by this approach is a trade-off between minimization of distortions of clusters and maximization of discriminative power of image-wise representations, i.e. histogram representations of images. We further extend the model to incorporate spatial information during the dictionary learning process in a spatial pyramid matching like manner. We extensively evaluated the two models on various benchmark dataset and obtained promising results.

1. Introduction

The bag-of-visual-words (BOV) model is an important component for a recently popular image categorization framework (i.e. local features, BOV models and SVM classifiers) [7, 15, 19, 22–24], which achieves state-of-the-art performances in PASCAL VOC challenges (*e.g.* [4, 5, 19]). A core technique of the BOV model is to generate a dictionary which is applied to quantize continuous local features to the so called discrete visual words.

Various dictionary generation approaches have been proposed in literatures, *e.g. k-means* [23], *mean-shift* [12] and manifold learning [10]. These approaches are designed to train a dictionary that contains sufficient information for representing the original features by minimizing the reconstruction error or expected distortion. Vocabularies are usually learned without taking category information into account. Consequently, the histogram representations of images over the learned dictionary may not be optimal for classification task. We illustrate the problem on a toy data shown in Figure 1(a). *k-means* groups the red and yellow clusters into one word (Figure 1(b)) and separate the two green clusters to two words because it only considers to minimizing the overall distortion. Histogram features obtained by this dictionary (blue dots in the figure) are not optimal for classification.



(a) toy data (b) k-means (c) our model

Figure 1. (a) The 2-D synthetic data. Colors of points indicate their categories. (b) K-means results and Voronoi boundary of the words. Red and yellow points are grouped onto the same word and as a result cannot be distinguished. (c) Partition generated by our model. Categories information are completely preserved at the price of distortion.

Therefore, the discriminative issue should be considered when constructing dictionaries, especially when the task is classification. In our perspective, a good dictionary for classification is: with *small distortion*, *discriminative*, and *compact in size*. Dictionaries with small distortion (or say, reconstruction error) are relatively robust against intra-class variations and noise [25]. "Discriminative" means that the histogram representations, which are obtained by applying a dictionary, of images belonging to different categories should be distinguishable. The size of dictionary is also an important factor for applications that need efficiency [8]. However, the three properties are contradict in some extent: Overemphasizing on discriminative ability may increase the size of a dictionary and weaken its generalization ability, and over-compressing to a dictionary will more

978-1-4244-6985-7/10/$26.00 ©2010 IEEE

or less lose the information as well as its discriminative power [8, 14, 25]. Thus a good dictionary learning method should find a balance between reconstruction, discrimination and compactness. As shown in Figure 1(c), although the dictionary obtained by our model has a bigger distortion than the dictionary obtained by *k-means*, it is more discriminative than the first one; and no dilemma exists if the dictionary size is larger than four.

Motivated from recent works on supervised dictionary learning [14, 18], and progress on supervised topic models [1, 24], we propose a probabilistic model for supervised dictionary learning, which is named as Supervised Dictionary Learning Model (SDLM) in this paper. SDLM essentially is a two-layer hidden variable model, which is composed of two logic parts: a generative part and a discriminative part. The first part describes the generation of an image which is assumed as a discrete distribution (i.e. a histogram) over words of a dictionary. The dictionary is formulated as a Gaussian Mixture Model (GMM) in the space of local descriptors. The second part requires that the histograms should be distinguishable in the perspective of a logistic regression loss function. Intuitively, the supervisory information containing in image labels will first be passed to histogram features of images by the logistic function, and then further be passed to affect parameters of the GMM, i.e. dictionary.

Another limitation of the traditional bag-of-visual-words model is the ignorance to spatial relationships of local features. This simplification is reasonable in object classification task due to large variances of objects' poses and shapes. While as stated in [15], in the scene categorization task where images are considered in holistic, statistics of local features over subregions provide rich cues for semantics. This assumption has been proved to be reasonable even for object images in VOC challenges [5, 19]. We therefore embed spatial constraints into SDLM and obtain the Spatial-SDLM (S^2DLM) model. Extensive experiments on various benchmark datasets demonstrate the effectiveness of the proposed supervised dictionary learning models.

2. Related Work

Supervised dictionary learning have attracted much attention in recent years. Existing methods can be roughly divided into three categories.

Some approaches construct multiple dictionaries or category-specific dictionaries. Zhang *et al*. wrap dictionary construction inside a boosting procedure and learn multiple dictionaries with complementary discriminative power [27]. [21] learns a category-specific dictionary for each category. Yang *et al*. [26] unifies the dictionary generation with classifier learning. Compared with them, our method produces a universal dictionary for all categories which can be applied to any BOV-based image analysis approaches.

Another category of approaches compresses an initial dictionary by merging visual words. The merging process is guided by mutual information between visual words and categories [8], or trade-off between intra-class compactness and inter-class discrimination power [25]. The performance of such approaches is highly affected by the initial dictionary since only merging operation is considered in them. To ease this problem a large dictionary is required at the beginning to preserve as much discriminative abilities as possible.

The third category of approaches learning a dictionary via pursuing a descriptor-level discriminative ability, e.g. empirical information loss minimization method [14], randomized decision forests [20], and sparse coding-based approaches [9, 17, 18]. Most of these approaches are first motivated from coding of signals, where a sample (or say signal) is only analogous to a local descriptor in an image rather than a whole image which is composed of a collection of local descriptors. Actually, this requirement is over strong since local descriptors of different objects are often overlapped (i.e. a white patch may appear both in the sky and on a wall). Instead, our model only requires the image-wise representations should be distinguishable This relaxation is critical for dictionary learning: it lead a good trade-off between distortions and discriminative abilities in a learned dictionary.

Topic models were first proposed to simulate a generative process of a document which is represented by a bag of words [2]. They have been developed for supervised tasks recently [1]. With the popular of BOV models, supervised topic models have been widely applied to image classification and segmentation [7, 24].

3. Supervised Dictionary Learning Models

In this section, we introduce the SDLM and discuss how to learn the parameters. We also present the analysis on how the discriminative part of the model affects the dictionary construction by examining the update rules. In section 3.4 we extend SDLM to a so called Spatial-SDLM (S^2DLM).

3.1. SDLM

The motivation of our model is that a discriminative dictionary should make histogram representations of images over it discriminative with respect to image categories. Therefore, we integrate a dictionary learning module, a image quantization module and a discriminative ability verification module in a single probabilistic model. Let C be the number of categories, N be the number of local features in an image and M be the dimension of feature descriptors. Our model assumes that an image containing N local features $w_{1:N}$ arises from the following generative process:

1. Draw a discrete distribution (*i.e*. a histogram represen-

tation) $\theta \sim \text{Dir}(\alpha)$.

2. For each image descriptor w_n, $n \in \{1, 2, \ldots, N\}$:
   (a) Draw a word assignment $z_n \mid \theta \sim \text{Mult}(\theta)$.
   (b) Generate a descriptor $w_n \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$.
3. Draw class label $c \mid z_{1:N} \sim \text{softmax}(\bar{z}, \eta)$, where $\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n$ is the empirical word frequencies, and the probability of choosing class label c is subject to

$$p(c \mid \bar{z}, \eta_{1:C}) = \exp\left(\eta_c^T \bar{z}\right) / \sum_{l=1}^{C} \exp\left(\eta_l^T \bar{z}\right)$$

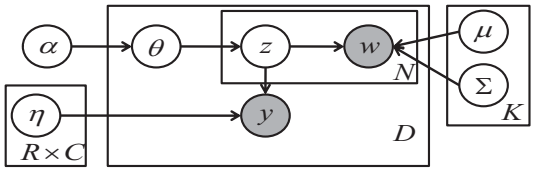
It is noted that, to facilitate the computation, we add the discriminative constraint on $\bar{z}$ rather on θ in Step 3, although the latter formulation is more straightforward. Figure 2(a) is a graphical illustration for SDLM. The parameters of our model are K word assignments $z_{1:K}$, K codeword parameters $\{\mu_k, \Sigma_k\}_{k=1}^{K}$, and C class coefficients $\eta_{1:C}$, where μ_k and Σ_k is the mean and covariance matrix for a M-dimensional multivariate Gaussian distribution and each η_c is a K-dimensional vector.

In Step 3, we use the same setup as the multi-class sLDA [24] for modeling the image labels, where multi-class logistic regression is applied on empirically estimated $\bar{z}$. However the meanings of the $\bar{z}$ are different. In [24] a pre-computed dictionary is required; given a topic z, a codeword index is picked from a multinomial distribution associated with that topic. Therefore, $\bar{z}$ is the empirical frequency of topics. In SDLM, we denote by z the codeword index to a pool of multivariate Gaussian distributions which are used to generate image descriptors in Step 2. Hence $\bar{z}$ is exactly the codeword frequency of the image. We evaluate the quality of a dictionary by checking its discriminative ability with respect to histogram representations of images over it rather than on topic frequencies. In terms of dictionary learning, modeling the image labels with respect to histograms over codewords is more straightforward than over topics.

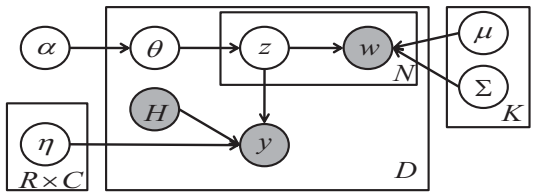
An important insight of Step 3 is that logistic parameter η_c plays the role as a codeword filter for category c. Positive/negative value of a particular component η_{ci} indicates that the model prefers the presence/absence of the i-th codeword in favor of enlarging inter-class difference. As will be explained later, SDLM utilizes this property to refine the codeword parameters.

3.2. Variational inference

Like the Latent Dirichlet Allocation (LDA) model [2], direct estimation of the posterior distribution of the latent variables given an image and its label is intractable. Thus we employ the variational inference algorithm [11] to approximate the posterior. Since the detailed derivation is not necessary for understanding our main ideas, we only present an outline in this paper.



(a) SDLM



(b) S^{2}DLM

Figure 2. Graphical model representations for (a) SDLM and (b) S^{2}DLM. The shaded circles stand for the observations and others are variables to be inferred.

To start we lower bound the log likelihood of a single image with Jensen's inequality:

$$\begin{aligned}
&\log p(y, w_{1:N} \mid \alpha, \eta_{1:C}, \mu_k, \Sigma_k) \\
&\geq \mathcal{L}(\gamma, \phi_{1:N}; \alpha, \eta_{1:C}, \mu_k, \Sigma_k) \\
&= \mathbb{E}_q\left[\log p(\theta \mid \alpha)\right] + \sum_{n=1}^{N} \mathbb{E}_q\left[\log p(z_n \mid \theta)\right] \\
&\quad + \sum_{n=1}^{N} \mathbb{E}_q[\log p(w_n \mid z_n, \mu_{1:K}, \Sigma_{1:K})] \\
&\quad + \mathbb{E}_q\left[\log p(y \mid \eta_{1:C}, z_{1:N})\right] + H(q)
\end{aligned} \tag{1}$$

where q is a variational distribution defined as

$$q(\theta, z_{1:N} \mid \gamma, \phi_{1:N}) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n) \tag{2}$$

where the Dirichlet parameter γ and the multinomial parameters $\phi_{1:N}$ are variational variables.

The coordinate ascent update equation of γ is the same as that in [2]:

$$\gamma = \alpha + \sum_{n=1}^{N} \phi_n. \tag{3}$$

To optimize $\mathcal{L}$ with respect to ϕ_n, we select terms which contain ϕ_n

$$\begin{aligned}
&\mathcal{L}_{\phi_n} \\
&= \sum_{i=1}^{K} \phi_{ni} \left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K} \gamma_j) + \log p(w_n | \mu_i, \Sigma_i) \right) \\
&+ \frac{\eta_y^T \phi_n}{N} - \mathbb{E}_q\left[\log\left(\sum_{l=1}^{C} e^{\eta_l^T \bar{z}}\right)\right] - \sum_{i=1}^{K} \phi_{ni} \log \phi_{ni}
\end{aligned} \tag{4}$$

The third term in equation 4 is not efficiently computable as its expectation is taken on variables in log-sum of exponentials. By lower bounding this term with Jensen's inequality again, we get

$$-\mathbb{E}_q\left[\log\left(\sum_{l=1}^{C} e^{\eta_l^T \bar{z}}\right)\right] \geq -\log\left(\sum_{l=1}^{C}\mathbb{E}_q\left[e^{\eta_l^T \bar{z}}\right]\right) = -\log\left(\sum_{l=1}^{C}\prod_{n=1}^{N}\left(\sum_{j=1}^{K}\phi_{nj}e^{\frac{\eta_{lj}}{N}}\right)\right) \quad (5)$$

Noting that the term in logarithm in equation 5 can be written as a linear function of ϕ_n, we obtain a computable lower bound for $\mathcal{L}_{\phi_n}$:

$$\mathcal{L}'_{\phi_n} = \sum_{i=1}^{K}\phi_{ni}t_{ni} + \frac{\eta_y^T\phi_n}{N} - \log\left(h^T\phi_n\right) - \sum_{i=1}^{K}\phi_{ni}\log\phi_{ni} \quad (6)$$

where

$$t_{ni} = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j) + \log p(w_n \mid \mu_i, \Sigma_i) \quad (7)$$

and $h = (h_1, \ldots, h_K)$ does not contain ϕ_n. We bound $\log\left(h^T\phi_n\right)$ using the inequality proposed in [3]

$$\log\left(h^T\phi_n\right) \leq \zeta h^T\phi_n - \log\zeta - 1, \quad (8)$$

where the equality holds if and only if $\zeta = (h^T\phi_n)^{-1}$. By treating ζ as a new variational parameter, we can update ϕ_n through a two-step iteration: in the first step ζ is set to $(h^T\phi_n^{old})^{-1}$ where ϕ_n^{old} is obtained in the last iteration; in the second step ϕ_{ni} is updated by

$$\phi_{ni} \propto \exp\left(\Psi(\gamma_i) + \log p(w_n|\mu_i, \Sigma_i) + \frac{\eta_{yi}}{N} - \zeta h_i\right) \quad (9)$$

3.3. Parameter estimation

In the variational E-step, we maximize the lower bound of posterior distribution for each image with respect to the variational parameters. In the M-step, we estimate the model parameters $\eta_{1:C}$ and $\{\mu_i, \Sigma_i\}_{i=1}^K$ by maximizing the log-likelihood of the corpus of images $\mathcal{D} = \{(w_{1:N_d}^d, y^d)\}_{d=1}^D$:

$$\mathcal{L}(\mathcal{D}) = \sum_{d=1}^{D}\log p(w_{1:N_d}^d, y^d \mid \alpha, \eta_{1:C}, \mu_{1:K}, \Sigma_{1:K}) \quad (10)$$

We simply fix α_i to $\frac{1}{K}$ for $i = 1, \ldots, K$ in practice. The codeword parameters are optimized as

$$\mu_i = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{ni}^d w_n^d}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{ni}^d}, \quad (11)$$

$$\Sigma_i = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{ni}^d (w_n^d-\mu_i)(w_n^d-\mu_i)^t}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{ni}^d} \quad (12)$$

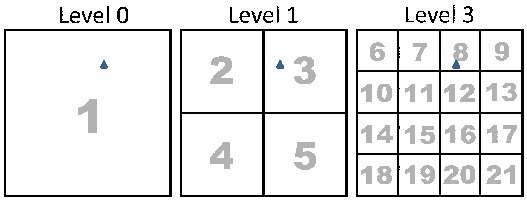


Figure 3. An example of a three-level pyramid.

As in [24], parameters of softmax function are optimized by minimizing the following function with conjugate gradient method

$$\begin{aligned}&\mathcal{L}(\mathcal{D})_{\eta_c}\\ &= \sum_{d=1}^{D}\mathbb{E}_q\left[\mathrm{I}(y^d = c)\eta_{y^d}^T\bar{z}^d - \mathbb{E}_q\left[\log\left(\sum_{l=1}^{C}\exp(\eta_l^T\bar{z}^d)\right)\right]\right]\\ &\geq \sum_{d=1}^{D}\left[\mathrm{I}(y^d = c)\eta_{y^d}^T\bar{\phi}^d - \log\left(\sum_{l=1}^{C}\prod_{n=1}^{N_d}(\sum_{i=1}^{K}\phi_{ni}^d\exp(\frac{\eta_{li}}{N_d}))\right)\right]\end{aligned} \quad (13)$$

Here we use the lower bound derived in equation 5 again.

3.4. S^{2}DLM

In [15] Lazebnik *et al*. propose a spatial pyramid matching (SPM) scheme to measure the similarity between images. Figure 3 shows an example pyramid structure with $L = 3$ levels and in total $R = 21$ regions. The R histograms of an image are computed for each region over the visual words falling in the corresponding regions and then are concatenated to form a $R \times K$-dimensional vector. This "long" vector is used to train Support Vector Machine (SVM) classifiers with the histogram intersection kernel [15]. The motivation is that the semantics of images are well captured by aggregating statistics of local features over fixed subregions. We demonstrate that this scheme can be easily embedded into SDLM. We term this extension as S^{2}DLM, in which beyond the expectation that the original histogram representations of images are distinguishable, we expect images are distinguishable in the perspective of SPM [15]. Consequently, the learned dictionary is optimal for this objective.

As shown in figure 2(b), we add an observation H to the original model which represents the pyramid structure of an image. It has two attributes: pyramid level L and region number R. We assign R softmax parameters, $\{\eta_{c1}, \ldots, \eta_{cr}\}$ to each class c. Every visual word will fall in L regions, one on each level (*e.g*. the blue triangle in figure 3 falls in three regions: region 1, 3 and 8). We denote by $I(z_n; H)$ the set of indices of regions that z_n falls in. The generative process is nearly the same as that for SDLM, ex-

cept that the distribution over class labels is changed to

$$p(y \mid \bar{z}, \eta_{1:C}) = \exp\left(f(\{\eta_{yr}\}_{r=1}^{R}, \bar{z})\right) / \sum_{l=1}^{C} \exp\left(f(\{\eta_{lr}\}_{r=1}^{R}, \bar{z})\right) \quad (14)$$

where

$$f(\{\eta_{yr}\}_{r=1}^{R}, \bar{z}) = \frac{1}{N} \sum_{n=1}^{N} z_n^t \sum_{r \in I(z_n;H)} \eta_{yr} \quad (15)$$

Most of the inferring steps are similar with SDLM, while the update equation for ϕ_{ni} is modified as

$$\phi_{ni} \propto \exp\left(\Psi(\gamma_i) - \log p(w_n|\mu_i, \Sigma_i) + \sum_{r \in I_n} \frac{\eta_{yri}}{N} - \zeta h_i\right) \quad (16)$$

and the lower bound of equation 13 is changed to

$$\mathcal{L}(\mathcal{D})_{\eta_c} \geq \sum_{d=1}^{D} \left(\mathrm{I}(y^d = c) \eta_{y^d r}^t \sum_{r \in I(z_n;H)} \frac{\phi_n}{N_d} \right) - \sum_{d=1}^{D} \left(\log \left(\sum_{l=1}^{C} \prod_{n=1}^{N_d} \left(\sum_{i=1}^{K} \exp\left(\frac{\eta_{lri}}{N_d}\right) \sum_{r \in I(z_n;H)} \phi_{ni} \right) \right) \right) \quad (17)$$

3.5. Discussion

Mean vectors $\mu_{1:C}$ and covariance matrix $\Sigma_{1:C}$ are major parameters for a dictionary. They have apparently physical explanations from their update equations, *i.e.* Equation (11) and (12). The mean μ_i for the i-th codeword is obtained by taking weighted mean of all descriptors in all images, and the Σ_i has the same explanation. The weight $\phi_{ni}^d / \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{ni}^d$ for the n-th descriptor of the d-th image reflects the contribution of w_n^d when forming the i-th codeword. Therefore, the essence of SDLM lies on the equation 9 for updating ϕ_{ni}, which depends on the four terms in the exponential. The first term $\psi(\gamma_i)$ is proportional to ϕ_{ni}'s prior, *i.e.* the Dirichlet parameter α_i. The second term is the likelihood that a local descriptor w_n is generated by the i-th codeword. This term encourages ϕ_{ni} to decrease the reconstruction error. The last two terms are related to the logistic function which punish the classification error. As explained earlier, codewords with larger η_{ci} can be seen as critical features for histogram representations of the images belonging to category c. The effects of the two terms are refining the weights, ϕ_{ni}, of descriptors to updating the i-th codeword in the perspective of discriminative learning. We can treat these terms in ϕ_{ni}'s equation as forces which pull the codewords with their values as the strengths. Generative forces (the first two terms) pull the words in order to reduce the reconstruction error, while the discriminative forces (the last two terms) pull the words to make the histograms of images distinguishable. Therefore, the learned dictionary is a consequence of a trade-off between minimization of distortion and maximization of discriminative power.

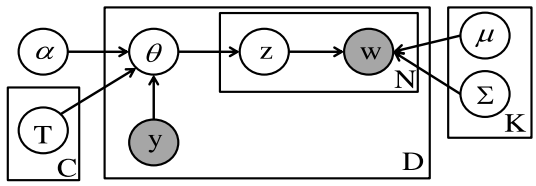


Figure 4. Graphical model representation for an alternative probabilistic dictionary learning model.

3.6. Other supervised dictionary learning models

Inspired by SDLM, we proposed two more supervised dictionary learning models. The first model is derived from the update rule for ϕ_{ni} in equation 9. We directly optimize a hybrid generative and discriminative energy function:

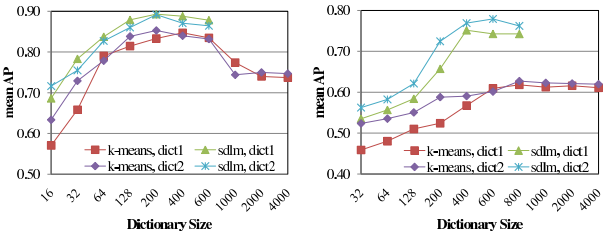
$$\min_{\eta,\mu} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{K} \phi_{ni}^d \left\|\mu_i - w_n^d\right\|^2 + C_1 \mathcal{L}(y^d, \phi^d | \eta_{1:C}) + C_2 \sum_{c=1}^{C} {\eta_c}^T \eta_c$$
$$s.t. \qquad \textstyle\sum_{i=1}^{K} \phi_{ni}^d = 1, \qquad \phi_{ni}^d \geq 0$$

where ϕ_n^d is the soft assignment for w_n^d over codewords, $\phi^d = \frac{1}{N_d} \sum_{n=1}^{N_d} \phi_n^d$ is the histogram representation of the d-th image, and C_1 and C_2 are coefficients to balance the three terms. The second term $\mathcal{L}(y^d, \phi^d | \eta_{1:C})$ can be any loss function to punish classification error, *e.g.* the Hinge loss or the logistic loss function. This formulation is similar as those in [9, 17, 18]. However, their approaches adopt a classification error term which is based on a patch-wise representation in loss function.

The second model is a probabilistic model in which the topic mixture (θ) of an image is generated from a category-specific prior distribution, that is, $\theta \sim \mathrm{Dir}(T_y \alpha)$. Similar with the linear transformation T^y in DiscLDA [13], the matrix T_y selects a group of codewords for the associated category. For example, suppose $C = 2$, the two matrices could be $T_1 = \mathrm{diag}(\mathbf{1}_{n_1 \times n_1}, \mathbf{0}_{n_2 \times n_2}, \mathbf{1}_{n_0 \times n_0})$, $T_2 = \mathrm{diag}(\mathbf{0}_{n_1 \times n_1}, \mathbf{1}_{n_2 \times n_2}, \mathbf{1}_{n_0 \times n_0})$. By doing this, each category will exclusively possess some codewords and meanwhile share some with other categories. Figure 4 shows the graphical representation of the model.

4. Experiments

In this section, we compare performance of the proposed models with state-of-the-art dictionary generation approaches on various benchmark dataset.



(a) (b)

Figure 5. (a) Performance on caltech10 dataset for different approaches. (b) Performance on VOC 2006 dataset for different approaches

4.1. Performance of SDLM

To evaluate the performances of SDLM, we choose two benchmark dataset: a subset of Caltech101 [6] and VOC 2006 [5]. We select the biggest 10 categories from Caltech101 dataset, which is termed as caltech10 in following experiments. Images in caltech10 are typical object images: objects usually locate at the centers of images and their backgrounds are clean. Images in VOC 2006 are more like real images: objects may appear at any positions in images, with occlusions and in various sizes, and their backgrounds are complex.

We extract SIFT descriptors every six pixels [16]. The support of each descriptor is a 16×16 patch. *k-means* is one of the state-of-the-art dictionary generation approaches [5], which is chosen as a baseline algorithm. Classifiers are trained by SVM with χ^2 kernel. 40 images from each category are utilized to train dictionaries and classifiers, and the rest is used for test on caltech10 dataset. We train two dictionaries under two different settings: use all local features and only use local features in objects' rectangles. The two dictionaries are termed as *dict1* and *dict2* respectively. The binary classification performance for each object class is quantitatively measured by mean average precision (mAP).

With the increasing of dictionary size, performances of all approaches increase first but finally drop. Figure 5(a) and Figure 5(b) show the mean average precision (mean AP) obtained by all approaches. When the dictionary size is too big, the dictionary is likely to be overfitting on training set, or say, may have low generalization capabilities.

Table 1 and Table 2 list best results of all approaches under different dictionary learning settings. On both of the two dataset, SDLM significantly outperform *k-means*. On the caltech10 dataset, the best performance of SDLM is achieved when the dictionary size is around 200, while the *k-means* obtains its best performance when the dictionary size is around 400. On the VOC 2006 dataset, we can get a similar observation. However, the sizes of dictionaries achieving the best results on VOC 2006 are bigger than on caltech10. This results indicate that 1) The dictionary learned by SDLM is more discriminative than *k-means*, 2) the dictionary learned by SDLM is likely to be more compact, and 3) The dictionary size may be bigger on a complex dataset than a simple one. It is interesting to note the performance of the *dict1* is a little better than the *dict2* on caltech10 dataset. By checking the images in caltech10, we find lots of artifacts on background of images, which may lead the improvement of performance of *dict1*.

Table 3. Performance on the fifteen scene dataset of different approaches under different dictionary sizes.

	32	64	128	256	400	1000
k-means [14]	0.5950	0.6580	0.7040	0.7330	0.7671	0.7532
info. loss [14]	0.6390	0.6800	0.7160	0.7470		
SDLM	0.6407	0.7112	0.7449	0.7734	0.7850	

As we have discussed in Section 3.1, η_c has a functionality to select the important words for a category. We plot the top 40 most significant words for each category in Figure 6. Most of the significant words are on objects. This property may have potential applications in object localization. At least, supervised learned dictionaries are able to help determine the candidate windows of objects.

4.2. Comparisons with other supervised approaches

Due to difficulties on re-implementation, we have to compare our approaches with previous supervised dictionary learning approaches on the same dataset, same features and same experimental configurations as that utilized in their papers. Two approaches are selected: an information loss minimization-based approach [14], and a word merging-based approach [8].

In the first experiment, we compare the performance of our approaches with the approaches proposed in [14] on the fifteen scene dataset [15]. This dataset consists of fifteen kinds of scene images, *e.g.* highway, kitchen and street. The results are shown in Table 3. Our approach outperforms the baseline approaches. With the increasing of sizes of dictionaries, the differences between performance of the *information loss* approach [14] and *k-means* are becoming smaller and smaller. However, our approach has reached the upper bound (observed in experiments) of the baseline approaches even when the size of dictionary is very small (*e.g.* 256).

In the second experiment, we compare the performance of our approaches with the approaches proposed in [8] on the Graz-02 dataset. This dataset consists of three categories of images, *i.e.* cars, people and bicycles. Actually, their experiments target on object localization rather than image categorization. They use the trained classifiers to classify each pixel, and then evaluate how many pixels on objects are correctly classified. We evaluate our approaches under the same settings. The results when the sizes of dictionaries are 200 are reported in Table 4. It is noted that all dictionaries are trained with local features on whole images rather than in rectangles of objects. The performance of SDLM is significantly better than the baseline approaches.

Table 1. The best performance on caltech10 dataset for different approaches and dictionary settings. Results are reported with dictionary size 400 and 200 for *k-means* and SDLM.

	airplanes	motorbikes	faces	watch	leopards	bonsai	car side	ketch	chandelier	hawksbill	mAP
k-means,dict1(400)	0.8105	0.8958	**0.9823**	0.7625	0.9157	0.8784	0.7333	0.9153	0.8276	0.7500	0.8471
SDLM, dict1(200)	0.8803	0.9314	0.9772	**0.9750**	0.9157	0.8649	0.8000	0.9153	**0.9310**	**0.7375**	**0.8928**
k-means,dict2(400)	0.8304	0.8658	0.9742	0.7450	0.9215	**0.8872**	0.7125	**0.9217**	0.8270	0.7141	0.8399
SDLM, dict2(200)	**0.8981**	**0.9431**	0.9655	0.9511	**0.9231**	0.8573	**0.8239**	0.9064	0.9253	0.7218	0.8916

Table 2. The best performance on VOC 2006 dataset for different approaches and dictionary settings. Results are reported with dictionary size 800 and 400 for *k-means* and SDLM.

	bicycle	bus	car	cat	cow	dog	horse	motor	person	sheep	mAP
k-means,dict1(800)	0.7439	0.6226	0.6694	0.7217	0.5061	0.688	0.5241	0.5286	0.5852	0.5911	0.6181
SDLM, dict1(400)	0.8198	0.8538	0.8122	0.7537	0.7581	0.7234	0.6322	**0.7946**	0.6307	0.7376	0.7516
k-means,dict2(800)	0.7682	0.6433	0.6875	0.7318	0.539	0.6592	0.5142	0.5309	0.5863	0.6102	0.6271
SDLM, dict2(400)	**0.8242**	**0.8652**	**0.8232**	**0.8103**	**0.7893**	**0.7407**	**0.6645**	0.7739	**0.6486**	**0.7507**	**0.7691**

Table 4. A comparison of the pixel precision-recall equal error rates on Graz-02 dataset. Dictionary size is 200.

	cars	people	bicycles
AIB200-KNN [8]	0.5090	0.4970	0.6380
AIB200-SVM [8]	0.4010	0.5070	0.5990
SDLM	0.5531	0.5485	0.6628

Table 5. Performance of the six approaches on the fifteen scene dataset. Dictionary size is 200.

	L = 0	L = 0+1	L = 0+1+2
k-means+SVM [15]	0.7220		
SDLM+SVM	0.7687		
S^2DLM+SVM	0.7845		
k-means+SPM+SVM [15]	0.7220	0.7900	0.8110
SDLM+SPM+SVM	0.7687	0.8156	0.8227
S^2DLM+SPM+SVM	0.7845	0.8189	0.8276

4.3. Performance of S^2DLM

Spatial information has been observed very useful for image categorization [4, 5, 19]. We compare the proposed models with a state-of-the-art approach, *i.e*. spatial pyramid matching (SPM) kernel [15], on the fifteen scene dataset under different configurations. Six approaches are evaluated: *k-means*+SVM (obtain a dictionary by *k-means* and train SVM classifiers with histogram intersection kernel, SDLM+SVM (learn a dictionary with SDLM and train SVM classifiers with histogram intersection kernel), S^2DLM+SVM(learn a dictionary with S^2DLM), *k-means*+SPM+SVM(train SVM classifiers with SPM kernel), SDLM+SPM+SVM(learn a dictionary with SDLM and train SVM classifiers with SPM kernel), and S^2DLM+SPM+SVM. Three level of SPM schemes are evaluated, *i.e*. 1, 2×2 and 4×4. The results when the sizes of dictionaries are 200 are shown in Table 5.

All dictionaries when combined with SPM kernels obtain significant better performance. Although SPM kernel looks a little rigorous in terms of utilizing spatial constraints, it has been observed effective on both scene and object image categorization [5, 15]. Among approaches which do not been trained with the SPM kernel, S^2DLM+SVM obtains a much better result than the other two approaches. Its performance (0.7845) is comparable with the approach which is trained with up to two level pyramid matching kernel (0.7900). By utilizing the dictionaries learned by SDLM and S^2DLM, and SPM kernels together, we get performance which outperforms the state-of-the-art results on this dataset [15]. These observations indicate 1) the S^2DLM has effectively incorporated spatial information when generating a dictionary, and 2) the SPM kernel is so strong to complement the weakness of features, and 3) the performance of dictionaries learned by S^2DLM can be further improved by applying SPM kernel. These results demonstrate the effectiveness of the learned dictionaries again.

5. Conclusion

We have proposed two probabilistic models for supervised dictionary learning. The first model, SDLM, seeks a balance between minimization of distortions of clusters and maximization of discriminative power of image-wise representations, *i.e*. histograms of images. The balance guarantees that the learned dictionary is more discriminative and generalizable than most of existing approaches. The second model, S^2DLM, which incorporates spatial constraints in a spatial pyramid matching like manner, is able to further improve the discriminative capabilities of a learned dictionary.

Our experiments demonstrate that the proposed models are able to outperform both state-of-the-art unsupervised and supervised dictionary learning approaches. Especially, only with a small number of codewords, our models achieve the upper-bound performances of large unsupervised dictionaries. The proposed models meet the three criterions of a good dictionary in the perspective of image classification.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Grant No. 60773090 and

Figure 6. Visualization of the top 40 most discriminative words in *dict1* for each category on caltech10 dataset. Dictionary size is 200.

Grant No. 90820018), the National Basic Research Program of China (Grant No. 2009CB320901), and the National High-Tech Research Program of China (Grant No. 2008AA02Z315).

References

[1] D. Blei and J. McAuliffe. Supervised topic models. *NIPS*, 2008. 2

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 2003. 2, 3

[3] G. Bouchard. Efficient bounds for the softmax function, applications to inference in hybrid models. In *Proc. NIPS*, 2007. 4

[4] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. Overview and results of the pascal visual object class classification challenge (voc2009). 2009. 1, 7

[5] M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The 2006 pascal visual object classes challenge (voc2006) results. 2006. 1, 2, 6, 7

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 2007. 6

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, 2005. 1, 2

[8] B. Fulkerson, A. Vedaldi, and S. Soatto. Localizing objects with smart dictionaries. In *Proc. ECCV*, 2008. 1, 2, 6, 7

[9] K. Huang and S. Aviyente. Sparse representation for signal classification. *NIPS*, 2007. 2, 5

[10] Y.-G. Jiang and C.-W. Ngo. Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Comput. Vis. Image Underst.*, 2009. 1

[11] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine learning*, 1999. 3

[12] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. ICCV*, 2005. 1

[13] S. Lacoste-Julien, F. Sha, and M. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. *NIPS*, 2008. 5

[14] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *TPAMI*, 2009. 2, 6

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006. 1, 2, 4, 6, 7

[16] D. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, 1999. 6

[17] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Proc. CVPR*, 2008. 2, 5

[18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *NIPS*, 2008. 2, 5

[19] M. Marszale, C. Schmid, H. Harzallah, and J. van de Weijer. Learning representations for visual object class recognition. In *Visual Recognition Challange workshop*, 2007. 1, 2, 7

[20] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. *NIPS*, 2007. 2

[21] F. Perronnin. Universal and adapted vocabularies for generic visual categorization. *TPAMI*, 2008. 2

[22] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *Proc. ICCV*, 2005. 1

[23] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003. 1

[24] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *Proc. CVPR*, 2009. 1, 2, 3, 4

[25] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proc. ICCV*, 2005. 1, 2

[26] L. Yang, R. Jin, R. Sukthankar, and F. Jurie. Unifying discriminative visual codebook generation with classifier training for object category recognition. In *Proc. CVPR*, 2008. 2

[27] W. Zhang, A. Surve, X. Fern, and T. Dietterich. Learning non-redundant codebooks for classifying complex objects. In *Proc. ICML*, 2009. 2