Neurocomputing 76 (2012) 28–35

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A hierarchical latent topic model based on sparse coding

Wenjun Zhu *, Liqing Zhang, Qianwei Bian

MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China

ARTICLE INFO

Available online 3 September 2011

Keywords:
Latent topic model
Sparse coding
Laplace distribution

ABSTRACT

We propose a novel hierarchical latent topic model based on sparse coding in this paper. Unlike the other topic models applied in the computer vision field, the words in our model are not discrete but continuous. They are generated by sparse coding and represented with n-dimensional vectors in $\mathbf{R}^n$. In sparse coding, only a small set of components of each word is active, so we assume the probability distribution over these continuous words is Laplace and the parameters of the Laplace distribution depend on topics, which are the latent variables in this model. The relationship among word, topic, document and corpus in our model is similar to Latent Dirichlet Allocation (LDA). Thereby this model is a generalization of the traditional LDA by introducing the concept—continuous words. We use an EM algorithm to estimate the parameters in our model. And the proposed method is applied to some significant computer vision problems such as natural scene categorization and object classification. The experimental results show the method is a valuable direction to generalize topic models.

1. Introduction

Topic models such as probabilistic Latent Sematic Analysis (pLSA) [1] and Latent Dirichlet Allocation (LDA) [2] were originally used in text understanding and information retrieval. In recent years, they also have been widely used to solve computer vision problems. In [3,4], topic model was used to learn and recognize natural scene categories. Refs. [5–9] employed topic model to discover objects from a collection of images. In [10], the topic model was also applied to human action classification.

To borrow the algorithms from text analysis, the researchers of above-mentioned applications construct visual words from images, which are the basic elements in topic models. These words are discrete variables and represented by unit-basis vectors that have a single component equal to one and all other components equal to zero (e.g. (0,…,0,1,0,…,0)). For example, a collection of patches are firstly sampled from training images and resized into the same scale in [3]. Then a codebook is learned by performing k-means algorithm. Codewords are defined as the centers of the learned clusters. Each image patch is assigned to the nearest codeword. In the other applications, although other local descriptors [11] instead of image patches are used to construct visual words, these words are still discrete and represented by unit-basis vectors. Although this way is easy and efficient, the difference between the original patch and the assigned codeword is discarded and the discarded information may be helpful to the classification task. For example, when using above-mentioned codewords to encode image patches, a patch which is just the superposition of several codewords will be assigned only to the nearest one and some useful information is ignored. Fig. 1(a)–(e) illustrates this situation. (a) is an image patch. (b), (c) and (d) are three codewords. As expressed in (e), (a) should be assigned to (b) which is the nearest codeword to (a). The representation of (a) with respect to this three codewords is (1,0,0) now, so the information about (b) and (c) is ignored. A better way to represent an image patch is using the coordinates with respect to a set of basis images, which will avoid the above limitation. The image patch can be absolutely reconstructed by the linear combination of the basis images with the coordinates. Fig. 1(f) illustrates how to code an image patch with respect to three basis images. The coordinates is (1,1,1) in this example. We can see the representation of a word is not a unit-basis vector but a general vector in this way. We call this kind of words "continuous words" to differ from the traditional discrete word. Can we construct topic model based on continuous words like the traditional topic models based on discrete words?

In this paper, we deal with the above problem by proposing a three-level hierarchical Bayesian model similar to LDA. The essential difference between the traditional LDA and our model lies in the basic elements—words. In traditional LDA, a word is discrete and represented by a unit-bases vector. But in our model, it is continuous and represented with a general n-dimensional vector. We also define a reasonable probability distribution over these words and present an EM algorithm to estimate the parameters of the model.

Several matrix factorization algorithms such as PCA, ICA [12] and NMF [13] can be used to generate the continuous visual words. And we generate them by sparse coding in our model. The sparse code is

* Corresponding author.
E-mail address: donghaiwilson@sjtu.org (W. Zhu).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.11.038

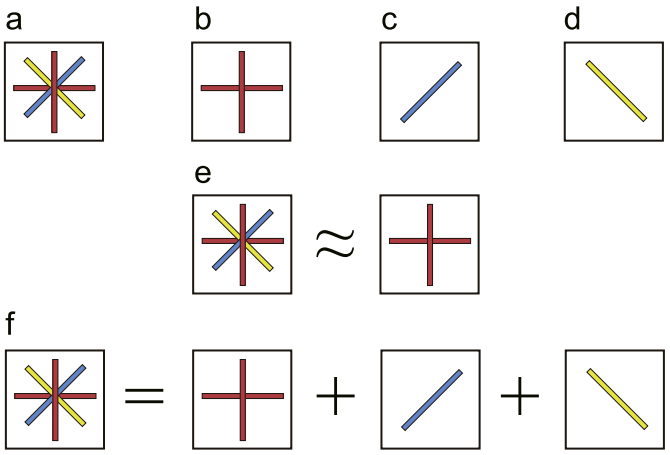

Fig. 1. (a) An image patch. (b)–(d) Three codewords or bases. (e) Representing an image patch with the nearest codeword. (f) Representing an image patch with the combination of three bases.

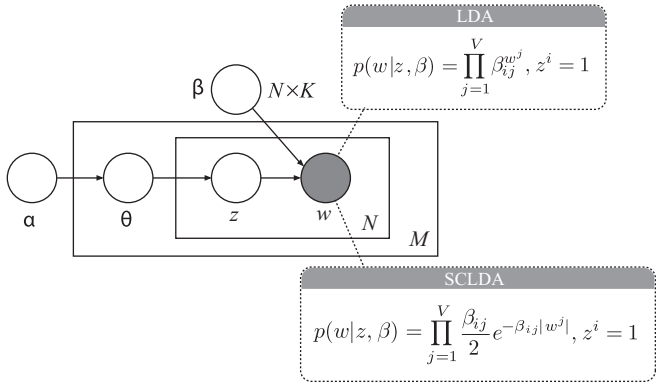

Fig. 2. Graphical model representation of LDA (SCLDA). w represents word, z represents topic, θ are parameters of multinomial variable z, α and β are hyperparameters. α are parameters of Dirichlet distribution and every column of β are parameters of a multinomial distribution (Laplace distribution). The difference of LDA and SCLDA lies in the conditional probability distribution over w i.e. $p(w|z,\beta)$.

a kind of neural code in which each item is encoded by the strong activation of a relatively small set of neurons. It is also an efficient encoding method for visual signals. The advantages of sparse coding have been discussed in these papers [14–16] from both physiological and information processing viewpoints.

Fig. 2 shows the dependence relationship of variables in our model. It seems the same as the traditional LDA. The critical difference is the probability distribution over the word w. Our model inherits the strength of LDA and improves it by introducing continuous words, which are more natural representation of images. Because our continuous words are generated by sparse coding, we formally call the model Sparse Coding Latent Dirichlet Allocation, or SCLDA for short.

The rest of the paper is organized as follows: In Section 2 we introduce SCLDA, a novel hierarchical latent topic model based on sparse coding including model construction and parameter estimation algorithm. Section 3 illustrates the experimental results of SCLDA on scene categorization and object classification. Finally, we summarize our work and discuss future directions in Section 4.

2. SCLDA

In this section, we propose our algorithm—SCLDA. We briefly introduce sparse coding at first. Then we explain some notations and terminologies used in our model. Next the generative process of our model is described in detail. Finally, we present the variational inference and parameter estimation in our EM algorithm.

2.1. Sparse coding

Sparse coding generally refers to a representation where a small number of neurons are active, with the majority of the neurons inactive or showing low activity. It has been proposed as a guiding principle in neural representations of sensory input, particularly in the visual system.

In this paper, we use an efficient sparse coding algorithm proposed in [17]. Let $X\in\mathbf{R}^{k\times m}$ be the input matrix (each column is an input vector), $B\in\mathbf{R}^{k\times n}$ be the basis matrix (each column is a base), and $S\in\mathbf{R}^{n\times m}$ be the coefficient matrix (each column is a coefficient vector). Sparse coding is formulated into an optimization problem:

$$\min_{(B,S)}\frac{1}{2\sigma^2}\|X-BS\|_F^2+\lambda\sum_{i,j}\phi(S_{ij}) \tag{1}$$

subject to $\sum_i B_{ij}^2\le c$ $(\forall j=1,2,\ldots,n)$, where $\phi(\cdot)$ is a sparsity function. The prior distribution over each coefficient S_{ij} is given by $p(S_{ij})\propto\exp(-\lambda\phi(S_{ij}))$. In our implementation, $\phi(\cdot)$ is L_1-norm, so the prior distribution over S_{ij} is Laplace with the location parameter is 0 and the scale parameter is $1/\lambda$ i.e.

$$p(S_{ij})=\frac{\lambda}{2}e^{-\lambda|S_{ij}|}. \tag{2}$$

Fig. 3 illustrates 120 sparse coding basis learned in a natural scene categorization experiment. This experiment will be detailedly introduced in Section 3. Some example images are showed in Fig. 4. We extracted totally 13 000 12×12 patches from these images for training and obtained these 120 basis. Most of them appear to represent simple orientations and illumination patterns, similar to the ones that the early human visual system responds to.

2.2. Notation and terminology

Fig. 5 is the flow chart of how to form a "corpus" from a collection of images. At first, patches are sampled from images. Then we use a sparse coding algorithm [17] to learn a basis. Each

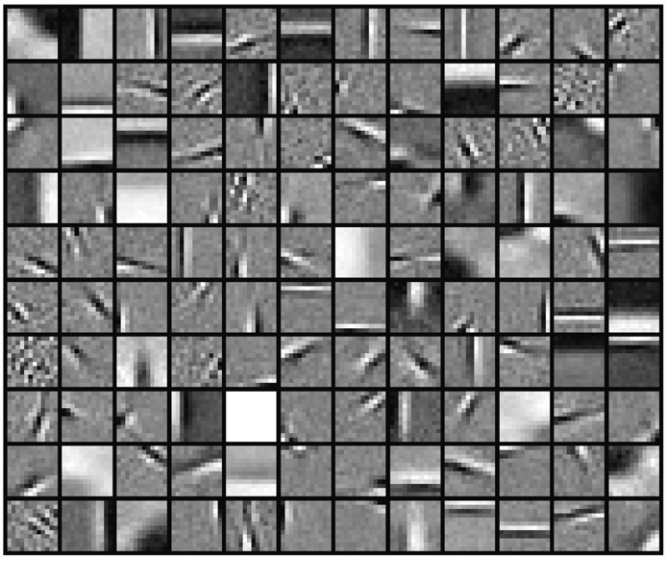

Fig. 3. One hundred and twenty bases learned on 12×12 image patches by sparse coding. Most of them appear to represent simple orientations and illumination patterns, similar to the ones that the early human visual system responds to.

Fig. 4. Examples of scenes. Each column is a class of scenes and there are total 13 classes.

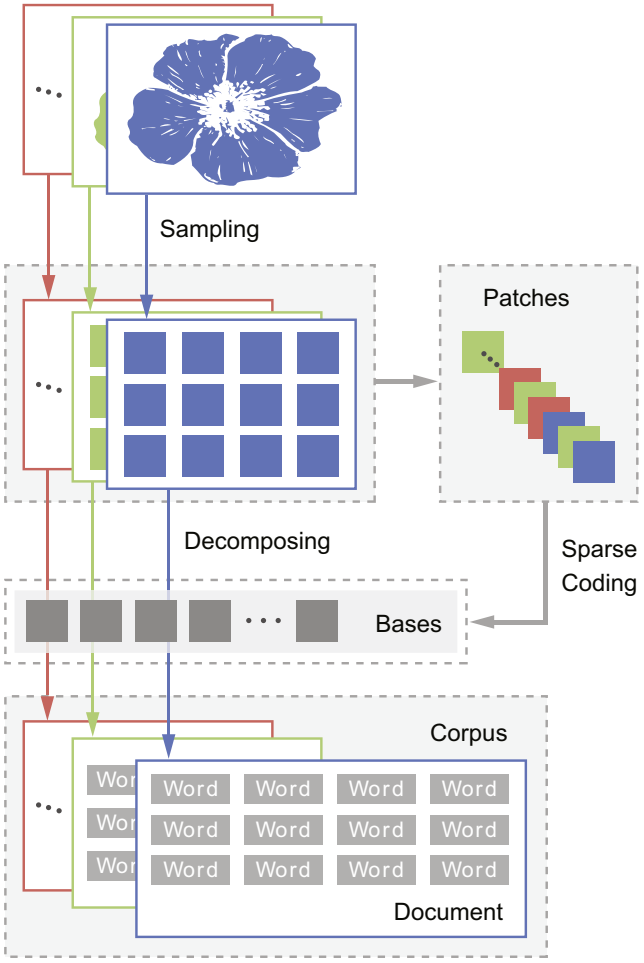

Fig. 5. Flow chart of forming a corpus.

patch is represented by the coefficients with respect to the basis. These coefficients are "words" in our model. An image, which is a "document" in our model, is a bag of words. And a "corpus" is a collection of these "documents" (images).

Formally, we define the following terms:

- A word is the basic element, defined to be sparse coding coefficients of a image patch. It is a general n-dimensional vector and denoted by w.
- A document is a sequence of N words denoted by $\mathbf{w}=(w_1,w_2,\ldots,w_N)$, where w_n is the nth word in the sequence.
- A corpus is a collection of M documents denoted by $\mathcal{D}=\{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_M\}$.

2.3. Generative process

The conditional probability distribution over w i.e. $p(w|z,\beta)$ is multinomial in the traditional LDA. It is given by

$$p(w|z,\beta)=\prod_{j=1}^{V}\beta_{ij}^{w^j},\quad z^i=1, \tag{3}$$

w denotes a word, z denotes a topic. They are both unit-basis vectors, w^j is the jth component of w and z^i is the ith component of z. β is a matrix and each row of it is the parameters of a multinomial distribution. However, in our model, w is a general n-dimensional vector. The probability distribution over w cannot be multinomial now. In the sparse coding algorithm, we use L_1 penalty function, it means we have assumed that w^j $(j=1,2,\ldots,V)$ are independent and identically distributed Laplace random variables [17]. Here w^j is the same with S_{ij} in (2). The probability density is given by

$$p(w^j)=\frac{\lambda}{2}e^{-\lambda|w^j|},\quad j=1,2,\ldots,V. \tag{4}$$

And the joint density function of w is given by

$$p(w)=\prod_{j=1}^{V}p(w^j)=\prod_{j=1}^{V}\frac{\lambda}{2}e^{-\lambda|w^j|}. \tag{5}$$

To accord with this, we also assume the probability distribution of w^j conditioned on z is independent Laplace with the location parameter is 0. But they are not identical now. The scale parameters are related to the value of the topic variable z i.e.

$$p(w|z,\beta)=\prod_{j=1}^{V}\frac{\beta_{ij}}{2}e^{-\beta_{ij}|w^j|},\quad z^i=1. \tag{6}$$

In (6), we replace the uniform scale parameter $1/\lambda$ in (5) with $1/\beta_{ij}$. Because we use L_1 regularization in our sparse coding algorithm, the absolute value of w^j is taken in (6). So the data likelihood is the same if w^j is taken the opposite sign. It means we only care the intensity of neuron response in object recognition.

The following is the generative process of SCLDA:

1. For a document (image) d, a multinomial parameter θ_d over K topics is sampled from Dirichlet prior $\theta_d\sim Dir(\alpha)$.
2. For a word i in document d, a topic label z_{di} is sampled from multinomial distribution $z_{di}\sim Mult(\theta_d)$.
3. The value w_{di}^j, which is the jth component of word i in document d, is sampled from the Laplace distribution of topic z_{di}, $w_{di}^j\sim Laplace(\beta_{z_{di}j})$.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is

given by

$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta), \tag{7}$$

where θ is a K-dimensional Dirichlet random variable and the probability density is given by

$$p(\theta|\alpha)=\frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_K^{\alpha_K-1}, \tag{8}$$

z_n is a multinomial random variable and the probability mass function is given by

$$p(z_n|\theta)=\prod_{i=1}^{K}\theta_i^{z_n^i} \tag{9}$$

and $p(w|z,\beta)$ is given by (6). Then the marginal distribution of a document is given by

$$p(\mathbf{w}|\alpha,\beta)=\int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta, \tag{10}$$

and the probability of a corpus is given by

$$p(\mathcal{D}|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d. \tag{11}$$

2.4. Parameter estimation

The optimal parameters α and β are estimated by maximize the log likelihood term $\log p(\mathcal{D}|\alpha,\beta)$. But it is intractable due to the coupling between θ and β in the summation. Approximation inference or sampling methods can be used to solve such optimization problems. Because the conjugate condition is satisfied in our model, we choose variational approximation inference [18] to solve it.

We introduce variational parameters γ and ϕ. The variational distribution is given by

$$q(\theta,\mathbf{z}|\gamma,\phi)=q(\theta|\gamma)\prod_{n=1}^{N}q(z_n|\phi_n). \tag{12}$$

The log likelihood of a document can be expressed as

$$\log p(\mathbf{w}|\alpha,\beta)=\mathcal{L}(\gamma,\phi;\alpha,\beta)+\mathrm{D}(q(\theta,\mathbf{z}|\gamma,\phi)\|p(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta)), \tag{13}$$

where

$$\mathcal{L}(\gamma,\phi;\alpha,\beta)=E_q[\log p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)-\log q(\theta,\mathbf{z})] \tag{14}$$

and $\mathrm{D}(q\|p)$ is the KL divergence between q and p. For simplification, we use q, p and $\mathcal{L}$ to denote $q(\theta,\mathbf{z}|\gamma,\phi)$, $p(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta)$ and $\mathcal{L}(\gamma,\phi;\alpha,\beta)$. $\mathcal{L}$ is the lower bound of $\log p(\mathbf{w}|\alpha,\beta)$. Then we use an EM algorithm to maximize the lower bound instead of the log likelihood.

In the E-step, the hyper-parameters α and β are treated as known constants. Now, maximizing $\mathcal{L}$ is equivalent to minimizing $\mathrm{D}(q\|p)$:

$$(\gamma^*,\phi^*)=\arg\min_{(\gamma,\phi)}\mathrm{D}(q\|p). \tag{15}$$

The update rules of γ and ϕ are

$$\phi_{ni}\propto\exp\left(\Psi(\gamma_i)+\sum_{j=1}^{V}(\log\beta_{ij}-\beta_{ij}|w_n^j|)\right), \tag{16}$$

$$\gamma_i=\alpha_i+\sum_{n=1}^{N}\phi_{ni}. \tag{17}$$

In the M-step, γ and ϕ are known and fixed. Now, $\mathcal{L}$ is a function of hyper-parameters α and β. We wish to find parameters α and β that maximize the log likelihood of a corpus:

$$(\alpha^*,\beta^*)=\arg\max_{(\alpha,\beta)}\sum_{d=1}^{M}\log p(\mathbf{w}_d|\alpha,\beta), \tag{18}$$

$$(\alpha^*,\beta^*)\approx\arg\max_{(\alpha,\beta)}\sum_{d=1}^{M}\mathcal{L}_d, \tag{19}$$

where $\mathcal{L}_d$ is the lower bound on the log likelihood of the dth document. The update rule of β can be written out analytically:

$$\beta_{ij}=\frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}}{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}|w_{dn}^j|}. \tag{20}$$

The update for Dirichlet parameter α can be implemented by an efficient Newton–Raphson method [2]:

$$\alpha_{new}=\alpha_{old}-H(\alpha_{old})^{-1}g(\alpha_{old}), \tag{21}$$

where $H(\alpha)$ and $g(\alpha)$ are the Hessian matrix and gradient, respectively, at the point α. At the end of this subsection, we summarize our iterative variational EM algorithm:

1. (E-step) For each document, find the optimizing values of the variational parameters γ^* and ϕ^* to minimize $\mathrm{D}(q\|p)$.
2. (M-step) Maximize the lower bound on the log likelihood of the corpus ($\sum_{d=1}^{M}\mathcal{L}_d$) with respect to the hyper-parameters α and β.

The two steps are repeated until the lower bound converges.

3. Experiments

In this section, we apply the proposed algorithm SCLDA to solve some interesting application problems in computer vision field. There are two major ways to apply SCLDA—generative framework and discriminative framework. In generative framework, we treat each category of images as a "corpus" and learn model parameters for each one, then the decision of an unknown image is made by comparing the likelihood of each category. In discriminative framework, SCLDA is regarded as a dimension reduction algorithm, each image is projected to a fixed set of real-valued features—the posterior Dirichlet parameters $\gamma^*(\mathbf{w})$ associated with the image, then a support vector machine (SVM) is trained on these features.

3.1. Natural scene categorization

Natural scene categorization is very useful in our everyday life. Humans can categorize all kinds of complex scenes very quickly [19]. There are litter or no attention in this rapid procedure [20]. A number of recent studies use intermediate representations to improve performance and these intermediate features are manual annotated [21,22]. Our approach also takes the advantage of intermediate features, but avoids manual annotation.

The natural scene data-set used in this experiment is the same as [3], which contains 13 categories. They are suburb, coast, forest, highway, inside city, mountain, open country, street, tall building, office, bedroom, kitchen and living room. Some example images are illustrated in Fig. 4.

In each category, we randomly selected 100 samples for training and the rest for testing. To learn basis by sparse coding, we randomly extracted 20 patches per image from half training samples. The total number of the patches are $20\times50\times13=13\,000$. The patch size is

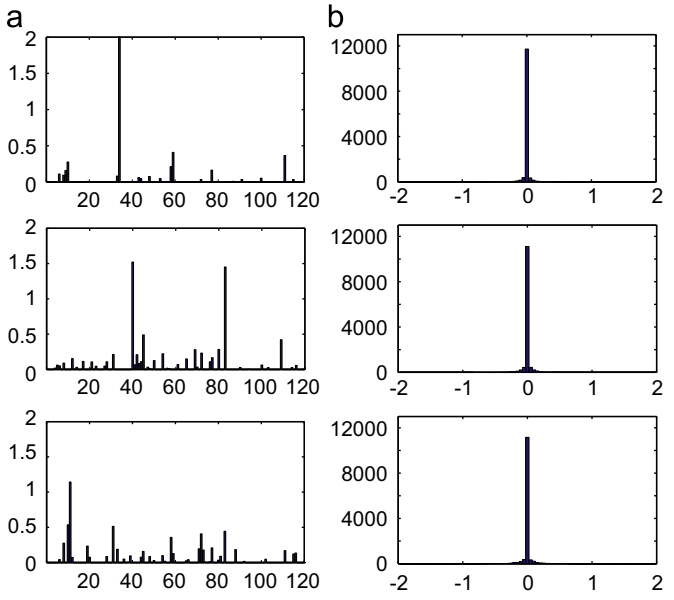

Fig. 6. (a) Examples of sparse coding coefficient vectors. We display absolute value of each component. Only a small set of components is active. (b) Histograms of three components of the coefficient vector. The distributions are similar to the Laplace.

12×12 in our experiments. Fig. 3 illustrated 120 bases learned by sparse coding algorithm.

Some examples of sparse coding coefficient vectors are shown in Fig. 6(a). We can see only a small set of components is active in this three vectors. The average sparseness of all continuous words is 0.87. The sparseness measure used here is given by

$$\text{sparseness}(\mathbf{x}) = \frac{\sqrt{n} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1}, \tag{22}$$

where n is the dimensionality of $\mathbf{x}$. This function is proposed in [23], which evaluates to unity if and only if $\mathbf{x}$ contains only a single non-zero component, and takes a value of zero if and only if all components are equal (up to signs). In Fig. 6(b), the distributions of three components of the coefficient vector are illustrated using histograms. We can see the distributions are similar to the Laplace.

After the sparse coding basis was prepared, we generated a corpus for each category by the flow described in Fig. 5. When a document was formed, 12×12 patches are evenly sampled every six pixels in each image.

In this experiment, we solved the classification problem by generative approach. The proposed algorithm was run to learn model parameters for each corpus. When asked to categorize one test image, the decision is made to the category label which gives the highest likelihood probability. The number of topics in this experiment is 55 and the number of bases is 120. We use a confusion table in Fig. 7 to demonstrate the performance of our approach. The rows denote true label and the columns denote estimated label. The average precision of 13 categories is 66.8%.

Table 1 lists the performance of three topic models. Spatial-LTM [8] and SCLDA are both improved from the traditional LDA [3]. Spatial-LTM tries to add spatial coherent information such as colors or texture features to LDA. And our SCLDA focuses on sufficiently using the information of the local image patches. We can see our method works best in this three algorithms. In this paper, we show the dedication of continuous words through generalizing the traditional LDA. In fact, continuous words can be introduced into not only LDA but also almost all topic models (including above mentioned Spatial-LTM). It means the performance of this experiment may be further improved by using continuous words in the other topic models such as Spatial-LTM.

	1	2	3	4	5	6	7	8	9	10	11	12	13
suburb 1	96	0	0	3	0	0	0	0	0	0	1	0	0
coast 2	0	86	1	0	0	1	0	0	0	1	0	8	3
forest 3	3	0	74	0	0	10	0	1	0	0	0	13	0
livingroom 4	3	0	0	32	2	1	6	4	1	19	33	0	0
insidecity 5	2	0	0	5	67	0	3	13	2	4	0	2	0
mountain 6	2	6	9	1	0	58	0	0	0	0	4	15	4
kitchen 7	0	0	0	15	9	0	37	2	0	22	15	0	0
street 8	1	0	1	3	4	1	0	83	2	0	2	1	5
tallbuilding 9	0	0	2	12	11	1	1	11	52	4	7	0	0
office 10	0	0	0	4	0	0	2	0	0	91	3	0	0
bedroom 11	0	0	0	21	3	1	9	3	0	14	50	0	1
opencountry 12	2	19	5	0	0	4	0	0	0	0	1	65	4
highway 13	2	9	1	1	1	0	0	4	1	1	2	3	77

Fig. 7. Confusion table of classifying 13 categories of natural scene. All the numbers stand for the percentage number. The average precision is 66.8%.

Table 1
Performance comparison of three topic models.

Method	LDA	Spatial-LTM	SCLDA
Precision	65.2	66.4	66.8

Table 2
Classes used in our multi-class object classification experiment.

airplanes	bonsai	cannon	car_side
chair	cup	dalmatian	dolphin
dragonfly	elephant	ewer	Faces
hedgehog	kangaroo	ketch	laptop
leopards	Motorbikes	panda	snoopy

3.2. Multi-class object classification

In this experiment, we apply SCLDA to categorize 20 classes of objects. Different with the scene categorization experiment, we use SCLDA as a dimension reduction method this time. We only generate one corpus for all classes, then train a model for it. For each training and testing image, we use the posterior Dirichlet parameters $\gamma^*(\mathbf{w})$ to reduce it to a set of real-valued features. At last, we input these features and labels into a SVM [24] to finish the classification task.

The data-set used in this experiment is Caltech101. We select 20 classes from it, and most of these classes contain more than 60 images. We use 30 images each class for training and the rest for testing. Table 2 lists the classes used in this experiment. As a preprocessing procedure, we normalized all images to 140 pixels in height (width is re-scaled accordingly so that the image aspect ratio was preserved) and converted to gray values. Fifteen training images in each class were used to learn sparse coding basis. The number of patches extracted from every image was 40. So totally $20 \times 15 \times 40 = 12\,000$ patches were used in sparse coding algorithm. To form a document, we evenly sampled 12×12 patches every six pixel from an image. Then we generated a corpus for all classes including totally $20 \times 30 = 600$ documents and a model was trained for it by SCLDA. Next all documents were

reduced to a set of features by the procedure mentioned at the beginning of this subsection. The number of features equals to the number of topics used in SCLDA. Finally, the classification task was finished by SVM.

Fig. 8 illustrates four classes of objects. Every point denotes a sample. Every sample image is firstly reduced to 80 features by SCLDA. Then the 80-dimensional vector are projected to a three-dimensional vector by PCA. It seems these four classes of points can be easily categorized by a linear classifier.

Fig. 9 is the confusion table of the classification result. We use brightness instead of number to denote the precision. White means 100% and black means 0%. The average performance is 75.1% with 80 topics.

Fig. 10 illustrates the performance vs. the number of training examples and topics. In Fig. 10(a), the number of topics is fixed on 50 and the performance ascends as the increase of training examples. In Fig. 10(b), the number of training examples is fixed on 30 and the performance reaches the maximum when the number of topic is 80.

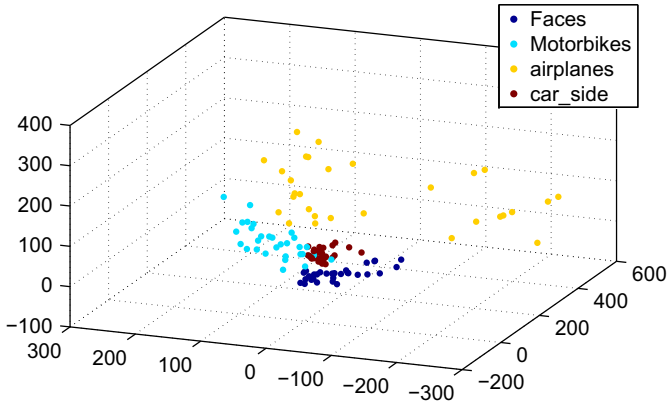

Fig. 8. Image samples of four classes are reduced to 80 dimensional features by SCLDA. The coordinates of every point denotes the first three principal components of 80 features.

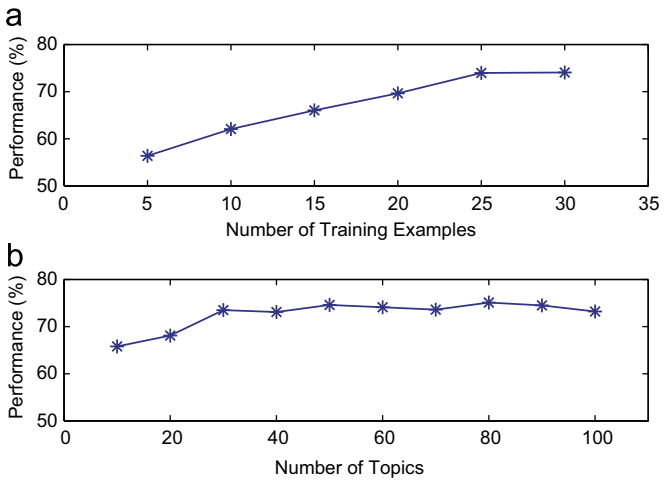

Fig. 10. (a) Number of training examples vs. performance (50 topics). (b) Number of topics vs. performance (30 training examples).

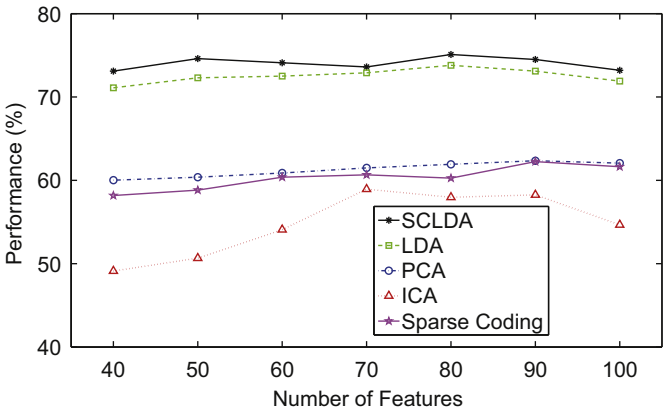

Fig. 11. Performance comparison of four dimension reduction algorithms.

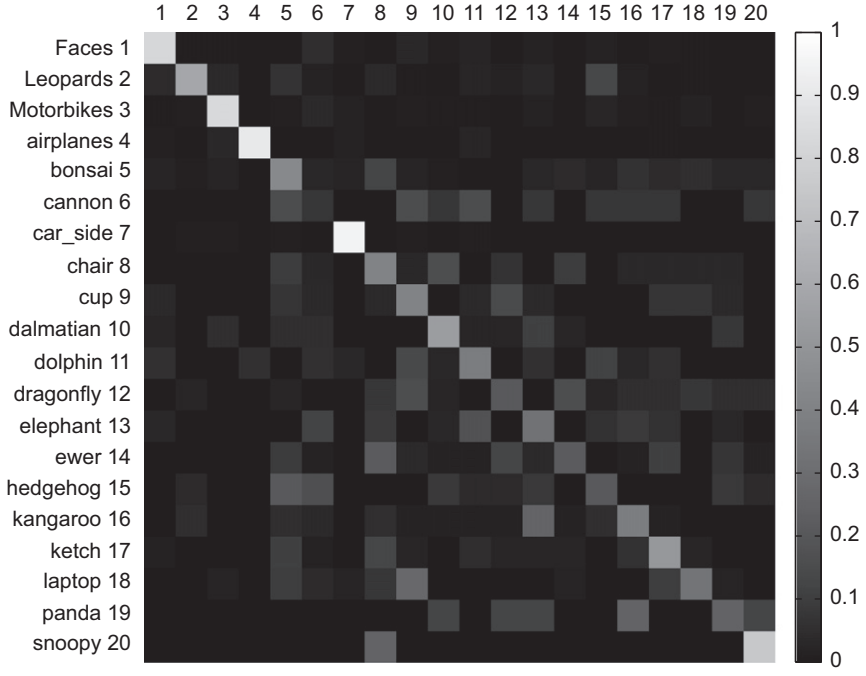

Fig. 9. Confusion table of 20 classes of objects. Brightness indicates precision. The average precision is 75.1%.

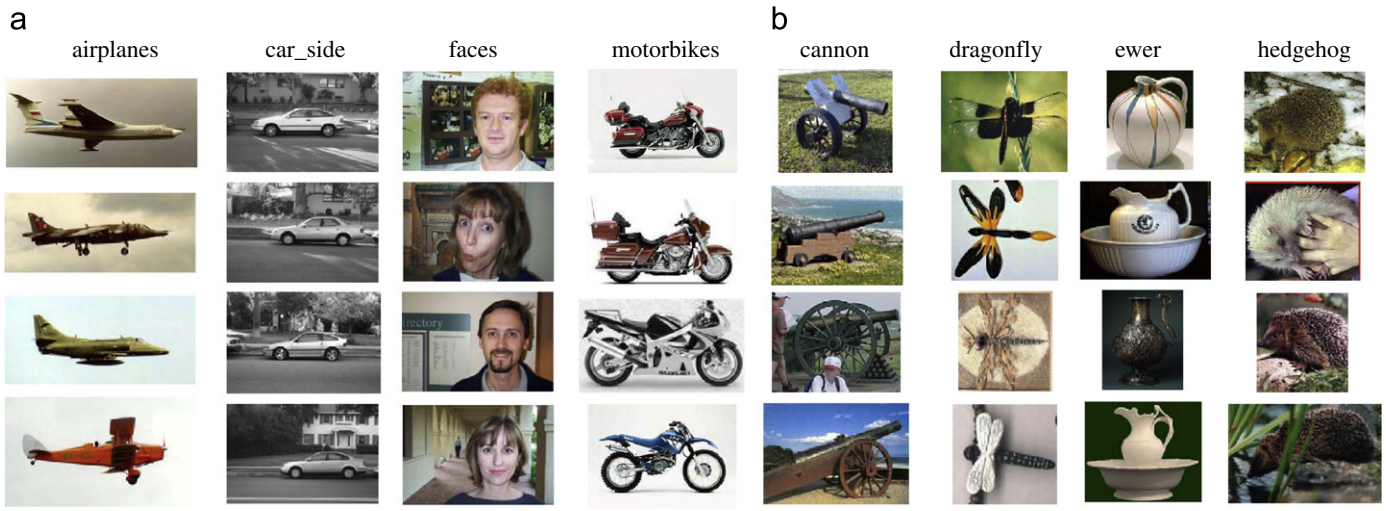

Fig. 12. (a) Classes on which our approach works best. (b) Classes on which our approach is least effective.

We compare the performance of our approach with other dimension reduction algorithms in Fig. 11. The other four methods are PCA, ICA, sparse coding and the traditional LDA. We can see two topic models (SCLDA and LDA) are significantly better than three matrix factorization methods (PCA, ICA and sparse coding), also our method clearly improves the performance of the traditional LDA. This improvement is the dedication of introducing continuous words.

Fig. 12 shows some of the "easiest" and "hardest" classes for our approach. Fig. 12(a) illustrates four classes on which our approach works best, and Fig. 12(b) illustrates four classes on which our approach is least effective.

4. Summary and conclusion

In this paper, we introduce continuous words to traditional LDA and propose a novel hierarchical latent topic model. Our purpose is to generalize the powerful topic models and apply them to computer vision problems. The experimental results show our work is really a valuable direction to generalize topic models in computer vision field.

The continuous words used in our model are sparse coding coefficients of image patches. Of course other matrix factorization algorithms such as PCA, ICA, NMF, etc. can also be employed to construct continuous words. And, in this paper, we assume the probability distribution over the continuous words conditioned on the topic is Laplace. In the scene categorization and object classification applications, this assumption works well. What distribution should be chosen depends on the application. A direction of our future research is how to determine the probability distribution of continuous words automatically.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (Grant No. 90920014) and the NSFC-JSPS International Cooperation Program (Grant No. 61111140019).

References

[1] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (1–2) (2001) 177–196.
[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[3] F.F. Li, P. Perona, A Bayesian hierarchical model for learning natural scene categories, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2005, pp. 524–531.
[4] P. Quelhas, F. Monay, J.M. Odobez, D. Gatica-perez, T. Tuytelaars, L.V. Gool, Modeling scenes with local descriptors and latent aspects, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2005, pp. 883–890.
[5] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, W.T. Freeman, Discovering objects and their location in images, Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 1, 2005, pp. 370–377.
[6] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from google's image search, Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 2, 2005, pp. 1816–1823.
[7] B.C. Russell, W.T. Freeman, A.A. Efros, J. Sivic, A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1605–1614.
[8] L. Cao, L. Fei-Fei, Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2007, pp. 1–8.
[9] X. Wang, E. Grimson, Spatial Latent Dirichlet Allocation, The Neural Information Processing Systems (NIPS), vol. 20, 2007.
[10] J. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Comput. Vision 79 (2008) 299–318.
[11] D.G. Lowe, Object recognition from local scale-invariant features, Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 2, 1999, pp. 1150–1157.
[12] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley-Interscience, 2001.
[13] D.D. Lee, S.H. Seung, Algorithms for non-negative matrix factorization, in: The Neural Information Processing Systems (NIPS), 2000, pp. 556–562.
[14] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (6583) (1996) 607–609.
[15] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by v1? Vision Res. 37 (23) (1997) 3311–3325.
[16] M.S. Lewicki, B.A. Olshausen, Probabilistic framework for the adaptation and comparison of image codes, J. Opt. Soc. Am. A: Opt. Image Sci. Vision 16 (1999) 1587–1601.
[17] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, in: The Neural Information Processing Systems (NIPS), 2007, pp. 801–808.
[18] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 2006.
[19] S. Thorpe, D. Fize, C. Marlot, Speed of processing in the human visual system, Nature 381 (6582) (1996) 520–522.
[20] F.F. Li, R. Vanrullen, C. Koch, P. Perona, Rapid natural scene categorization in the near absence of attention, Proc. Natl. Acad. Sci. USA 99 (2002) 9596–9601.
[21] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vision 42 (3) (2001) 145–175.
[22] A. Vailaya, M.A.T. Figueiredo, A.K. Jain, H.J. Zhang, Image classification for content-based indexing, IEEE Trans. Image Process. 10 (1) (2002) 117–130.
[23] P.O. Hoyer, Non-negative matrix factorization with sparseness constraints, J. Mach. Learn. Res. 5 (2004) 1457–1469.
[24] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, Software available at 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉, 2001.

Wenjun Zhu was born in Changzhou, China, in 1978. He received his B.S. degree in Computer Science from Nanjing University, Nanjing, China, in 2001. He is currently a Ph.D. student in the Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China. His current research interests include statistical learning and inference, perception and cognition computing model, computational theory for cortical networks and computer vision.

Liqing Zhang received the Ph.D. degree from Zhongshan University, Guangzhou, China, in 1988. He was promoted to full professor position in 1995 at South China University of Technology. He worked as a research scientist in RIKEN Brain Science Institute, Japan from 1997 to 2002. He is now a Professor with Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests cover computational theory for cortical networks, brain signal processing and brain–computer interface, perception and cognition computing model, statistical learning and inference. He has published more than 160 papers in international journals and conferences.

Qianwei Bian was born in Yuncheng, China, in 1981. He received his B.S. and M.S. from Shenyang University of Technology, Shenyang, China, in 2006. He is currently a Ph.D. candidate in the Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China. His current research interests include 3D model retrieval, human–computer interaction, computer graphics, CAD and computer vision.