Computing and Informatics, Vol. 30, 2011, 311–319

THE VARIANT OF LATENT DIRICHLET ALLOCATION FOR NATURAL SCENE CLASSIFICATION

Tang YINGJUN

School of Software and Communication Engineering
Jiangxi University of Finance and Economics
Nanchang, Jiangxi, China, 330032
e-mail: Yingjun.T@gmail.com, 07112047@bjtu.edu.cn

Manuscript received 21 January 2010; revised 30 June 2010
Communicated by Iveta Zolotová

Abstract. The paper proposes a novel model based on classic LDA (latent Dirichlet allocation), which is used to learn and recognize natural scene category. Uunlike previous work, the model performs variational Bayesian inference (VB) two times in order to get more precise prior Dirichlet parameters for each scene category. Although the scenes is represented in common topic simplex, the model has retained the diversities of each scene category based on the same topic simplex. Furthermore, two discriminations have been done to get good performance. We investigated the classification performance with classic 13 scenes image database and the experiments had demonstrated that our method can get better performance with less training time.

Keywords: LDA, CCLDA, topic, visual visterm, scene classification

1 INTRODUCTION

Scene classification groups images into semantically meaningful categories, which has been a feasible method to organize image dataset in a more efficient way. Natural scene is usually intended as the one of a semantically coherent and nameable human-scaled view of an real world environment [1]. Scene category, such as Coast, Bedroom, Forest, means a particular scene type, and greatly rests on particular co-occurrences of a large number of visual components (named visterm in our paper), which are connected with semantic information (named topic). Scene is generally

composed of several entities (car, house, door, tree, rocks, ...), which had been organized in often unpredictable layouts. Hence, the content of images from a specific scene category exhibits a large variability. Since natural scene classification becomes a difficult task and an open research field, many efforts are spent, and a huge amount of different approaches are present [2, 3, 4, 5].

In recent years, LDA [6] had been widely used to solve computer vision problems. For example, LDA was used to discover objects from a collection of images [7] and to classify images into different scene categories [5, 2]. In visual surveillance, LDA was employed to model atomic activities and interactions in a crowded and busy scene [10]. In these applications, LDA clustered low-level visual words (which were image patches, spatial and temporal interest points or moving pixels, namely visterm) into semantic topics (which may correspond to objects, parts of objects, human actions or atomic activities) utilizing their co-occurrence information. Since LDA assumed topics of document occurred in random way, Dirichlet distribution could be used as a prior on the parameters to a multinomial distribution of topics. The Dirichlet distribution had often been turned to Bayesian statistical inference, and was a convenient prior distribution to place over proportional data. Because LDA was originally applied in text classification, borrowing it to solve vision problems would bring some difficulties. First, users need to define the meaning of "documents" in vision problems, which often implies some assumptions on vision problems. Second, how to model to decide the parameters of Dirichlet are prior. Last, how to model to finish classification after getting the topic distribution.

In this paper, we will solve these three problems mentioned above, and the rest of the paper will be arranged as follows: We will propose our model using a new inference way in Section 2, and Section 3 will explain our decision method. A part of the figures from experiments will be shown and analyzed in Section 4. Finally, we will conclude our method.

2 OUR MODEL

We assume the images coming from the same scene category would share most of the topics, while most topics in images coming from different scene categories would be different. For example, Coast scene usually has beach, sand and sea, while Forest scene might have just trees and earth. In these two scenes, there are different topics, which are major components in their images. For LDA model, the hyper-parameter α denoted the prior distribution parameter of all topics in the image corpus, while the hyper-parameter β was interpreted as the prior observation count of visterms sampled from a topic before any visterm from the image corpus is observed. In order to get unique topic distribution for each category, we need unique prior distribution parameter for each scene category. Our work is different from previous works [5, 10], which destined the same topic constitution using the same prior parameter among all natural scene images. Further-

more, the common method to infer the hyper-parameters for LDA is the Variational Bayesian (VB) inference combined with the expectation-maximization (EM) algorithm. However, the previous approaches assumed all scenes had the same topic distribution in advance, which just roughly estimated the topic distribution for all scenes and could not represent the true topic distribution for each scene. Our work is to build the model to represent all image scenes in a more plausible and precise way and to solve the three difficulties for LDA mentioned in Section 1.

2.1 Model Description

In order to explain our model clearly, its graphic model is shown in Figure 1 a). Our model is inspired by the topic representation method of LDA and CCLDA. The LDA model shown in Figure 1 c) used the same hyper-parameters to infer the topic set for all categories. According to Figure 1, our model is similar with CCLDA model [2] in Figure 1 b), since they used different set of topics for each category. However, each category model of CCLDA is inferred with its own set of topics independently. By contrast, the category model of our model is based on common set of topics shared in all categories.

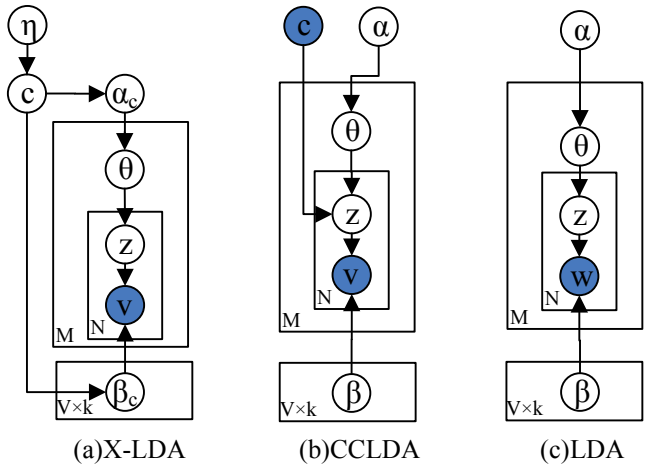

Fig. 1. Graphs of LDA model and its extension

In order to solve the first problem of LDA applied in vision field, the terms of our model are defined as follows. Our model includes three levels: visterm-level, image level and corpus level. The hyper-parameters α and β are corpus-level parameters (also named category model parameters), assumed to be sampled once in the training process and decided by the category label c. The variable θ sampled once for each imag, is image-level variable and denotes topic distributions. Finally,

the variable z and the visterm v belong to the visterm level, and are sampled once for each visterm in each image. The variable v denotes the visterm, which is the basic unit of an image, defined to be a patch membership from Codebook. The shaded node indicates that it is an observed variable. An image is a sequence of N visterms denoted by $W = (v_1, v_2, \ldots, v_N)$. Namely, an image is presented by help of BoV (Bag of Visterms). A corpus is a collection of images denoted $D = (W_1, W_2, \ldots, W_M)$, it means a collection of the same scene category during the training time, while it means a collection of all images irrespective of its category in test time.

In our model, it is easy to obtain the marginal probabilistic distribution of image, given hyper-parameters α and β of specific scene (cf. Equation (1).

$$p(W|\alpha, \beta, c) = p(c|\eta) \int p(\theta|\alpha_c) \times (\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(v_n|z_n, \beta_c)) d\theta \tag{1}$$

In Equation (1) the category parameter α_c is K-dimensional vector, defining relative length of topics in the corpus. The visterms' probabilities are parameterized by category parameter β_c, which is a $K \times N$ matrix and $\beta_{ij} = p(v_j = 1 \mid z_i = 1)$. The parameter c denotes the scene category, decides the category parameters, and belongs to uniform distribution. In our model, θ belongs to Dirichlet distribution with the parameter α_c, while z_n belongs to Multinomial distribution with the parameter θ.

2.2 Inference and Parameters Estimation

The key inferential problem that we need to solve is that of computing the posterior distribution of the hidden variables given in Equation (2).

$$p(\theta, z|W, \alpha, \beta, c) = p(c|\eta) \frac{p(\theta, z, W|\alpha_c, \beta_c)}{p(\mathrm{W}|\alpha_c, \beta_\mathrm{c})} \tag{2}$$

Unfortunately, this distribution is intractable to compute in general. Although the posterior distribution is intractable for exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximation, variational approximation, and Markov chain Monte Carlo. In this section we borrow a simple convexity-based Variational Bayesian algorithm [6] to approximate the hyper-parameters. Furthermore, double inferences will be done in order to get more precise prior parameters.

We consider to getting a tractable family of lower bounds as a surrogate for the posterior distribution, which has been defined as a family of distributions.

$$q(\theta, z|\gamma, \phi) = q(\theta, \gamma) \prod_{z} q(z|\phi_z) \tag{3}$$

In Equation [3] the variational parameters γ and ϕ are set via an optimization procedure by minimizing the Kullback-Leibler (KL) divergence between the varia-

tional distribution $q(\theta, z \mid \gamma, \phi)$ and the true posterior $p(\theta, z \mid W, \alpha, \beta, c)$. So we have:

$$\log p(W|\alpha, \beta, c) = L(\gamma, \phi; \alpha, \beta) + D(q(\theta, z \mid \gamma, \phi) \mid p(\theta, z \mid W, \alpha, \beta, c) \approx L(\gamma, \phi; \alpha_c, \beta_c) \tag{4}$$

In our paper, we adopt EM algorithm to get the unknown parameters α_c, β_c, γ and ϕ, which also is the different part from [6, 2]. First, we train the classic LDA model with all category images to get the approximately Dirichlet prior parameters α_c and β, which can be interpreted as a prior observation count for each topic and the prior observation count for visterms of each topic in all training images. Second, EM algorithm has been done to infer category model parameters using each category training images again. We apply LDA's prior parameters to initialize category hyper-parameters of EM algorithm in Equation (5), which helps to get more precise prior parameters, and solves the second problem of LDA used in vision field.

$$\beta_c = \beta_{LDA}, \alpha_c = \alpha_{LDA} - H(\alpha_{LDA})^{-1} g(\alpha_{LDA}) \tag{5}$$

The third step is the second EM algorithm including E-step and M-step, the E-step is used to compute γ and ϕ for each image of specific scene category in E-step.

$$\phi_{ni} = \beta_{iv_n} \exp\{E_q[\log(\theta_i)|\gamma]\}, \gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{6}$$

In Equation (6), the exponential is a Digamma function. The M-step of EM maximizes the likelihood and computes the new hyper-parameters iteratively displayed in Equation (7).

$$\beta_{ij}^c = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} v_{dn}^j, \alpha_{new}^c = \alpha_{old} - H(\alpha_{old})^{-1} g(\alpha_{old}) \tag{7}$$

After the iterative computation in EM algorithm, we obtain the new model composed by each category model with different hyper-parameters α_c and β_c.

3 CLASSIFYING IMAGES

For recognizable task, our method includes three steps. The first one is to computer the likelihood value $L(\gamma, \phi, \alpha_c, \beta_c)$ for our model. During the computation, the hyper-parameters will not be alternated and used to compute γ and ϕ alternatively. So several likelihoods have been computed with category parameters, in which the number is equal to the number of scene category. The scene category can be judged with the category parameters in Equation (8) by the ML (Maximum Likelihood) computed.

$$i = \arg\max_j \{L(\gamma, \phi; \alpha_j, \beta_j), j \in 1 \ldots C\} \tag{8}$$

The topic constitution for our model is represented in common topic simplex, which is shared among all kinds of images. On the contrary, CCLDA [2] was based

on topic simplex category, in which each simplex category is unique. As a result, we assume the precise decision can be obtained based on the result of our model and CCLDA model. Consequently, second step is used to get the other decision as in [2], in order to refer to the decision of CCLDA. At last, what needs to be decided is to judge the correct answer by comparing the result of two steps. Since they all applied ML to get scene category, it can be reasonable that the right answer c should be far away with the mean of the values of likelihood computed by category model in Equation (9).

$$\begin{aligned} L_{step1} &= \max\{L(\gamma,\phi;\alpha_i,\beta_i) - Mean(L), i \in 1\dots C\}, \\ L_{step2} &= \max\{L_{CCLDA}(\gamma,\phi;\alpha_j,\beta_j) - Mean(L_{CCLDA}), j \in 1\dots C\}, \\ c &= \arg\max(L_{step1}, L_{step2}) \end{aligned} \tag{9}$$

4 EXPERIMENT ANALYSIS

Our dataset comes from [5], which contains 13 natural scene categories. Each scene category is split randomly into two separate sets of images. Our experiment was implemented in Matlab 7 by computer with 1.6 GHz processor. During the training process, 100 images of each category have been randomly picked out to estimate the hyper-parameters, and another 100 images of each category have been randomly chosen in the test process. In Figure 2, the performance of our method is shown. The color meter marks the right recognized count in each category.

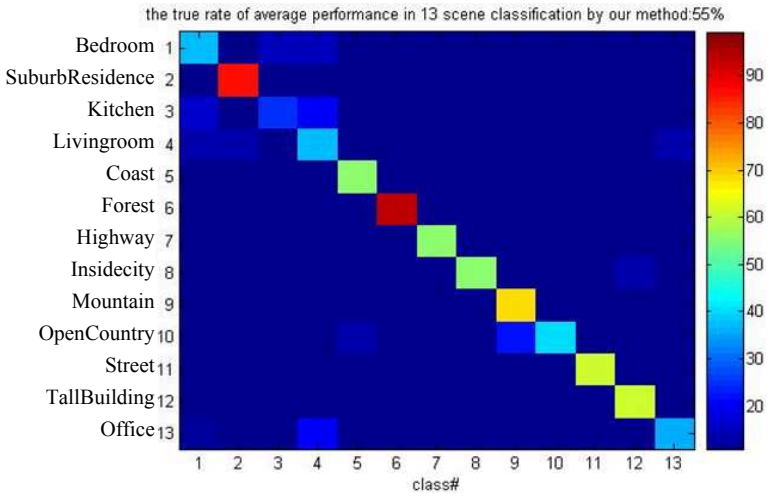

Fig. 2. The average performance of our method

Although our method needs to do double inferences and discrimination, the performance has proved it is adequate. The average correct rate by our mode in Step 1

is 53.3 %, which is higher than the CCLDA in Table 1. After double discrimination, our method gets the 55 % in Figure 2, which is the highest correct rate. It is obtained by synthetically considering the results of steps 1 and 2 in Equation (9). Our method not only keeps the high correct rate for some scene category in two steps, such as forest, suburb residence, highway, etc., but also improves performance in some indoor scenes, such as kitchen, living-room and office, which are badly classified by first two steps and previous methods.

Compared to the global features, local patches are more robust to occlusions and spatial variations, and SIFT (Scale Invariant Feature Transform) has been proved robust in many situations, so our experiment uses sparse grey SIFT to represent local feature and cluster to visterm of Codebook. Image is initially represented by BoV, which is high dimension. The dimension is close to the size of Codebook, so the codebook size plays an important role in scene classification. If the size is small then the correct rate is relatively low but the computation is fast. On the contrary, higher correct rate can be obtained with larger Codebook size. Therefore, we must settle for a solution that trades off efficiency with completeness, and we set the size of Codebook to 520. By the same token, the size of topic simplex also plays an important role in the scene classification. Although [5, 6] claimed 40 was the best size for simplex topic, our method presents the variation of correct rate in step 1 with different topic size in Figure 3. According to the curve of Figure 3 and the best topic size of [2], our double discriminator chooses 45 topics for Step 1 and 10 topics for Step2 separately. We assume the best performance can be obtained by the by the best setting for the above two steps.

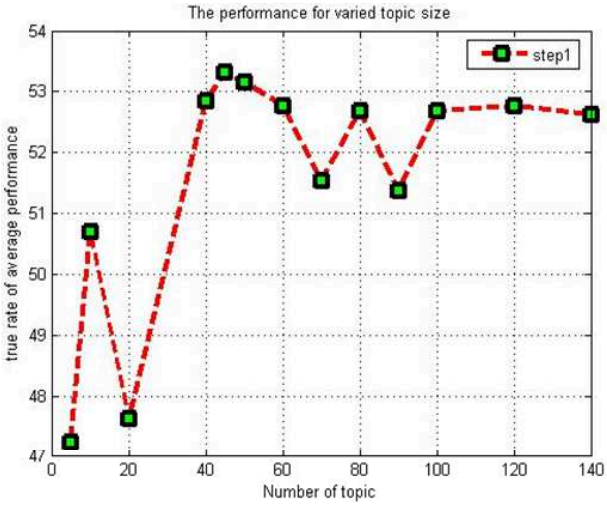

Fig. 3. The curves of performance varied with topic size

In order to investigate the performance of our method, the average correct rate with different models for 13 categories task are listed in Table 1. Each model uses its specific setting to get the best performance. It is obvious that our model can get the best performance (55 %) among these models.

Model	Our Model	Theme 1	LDA	CCLDA
Topic size	45	40	40	10
Performance	55.0 %	52.2 %[5]	37.0 %	49.46 %

Table 1. The performance comparison among models

5 CONCLUSION

The paper has proposed a new model to learn and recognize natural scene category. The model has inferred more precise prior of topic distribution by doing double inference; it produces different topic prior for each scene category based on common topic simplex, and can represent each image in the most correlational category topic distribution, which is also similar to human cognition. In order to get higher performance, two steps for discrimination have been done. Furthermore, since steps 1 and 2 are mutually independent, they can be executed side by side. Therefore, we will take it as future work and implement our method by concurrent computation.

Acknowledgement

This work was supported by the Fundamental Research Funds for the Central Universities (2009JBM024) and China Postdoctoral Science Foundation (201003044).

REFERENCES

[1] Henderson, J.: Introduction to Real-World Scene Perception. Visual Cognition, Vol. 12, 2005, No. 3, pp.849–851.

[2] Tang, Y.—Xu, D.: Category Constrained Learning Model for Scene Classification. IEICE Trans. Inf. & Sys., Vol. E92-D, No. 2, 2009.

[3] Bosch, A.—Zisserman, A.: Scene Classification Using a Hybrid Generative/Discriminative Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, 2008, No. 4, pp. 712–727.

[4] Lazebnik, S.—Schmid, C.—Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. IEEE Conference on Computer Vision and Pattern Recognition, New York, USA 2006, Vol. 2, pp. 2169–2178.

[5] Li, F.-F.—Perona, F.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA 2005, Vol. 2, pp. 524–531.

[6] Blei, D.—Andrew, Y.: Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, Vol. 3, pp. 993–1020.

[7] Yang, W.—Mori, G.: Human Action Recognition by Semi-Latent Topic Models. IEEE PAMI, Vol. 31, 2009, No. 10, pp. 1762–1774.

[8] Philbin, J.—Sivic, J.: Geometric LDA: A Generative Model for Particular Object Discovery. BMVC 2008.

[9] Wang, X.—Ma, X.—Grimson, E. L.: Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, 2009, No. 3, pp. 539–555.

[10] Elango, P. K.: Clustering Images Using the Latent Dirichlet Allocation Model. University of Wisconsin, 2005.

Tang Yingjun is the lecturer of Jiangxi University of Finance and Economics. She received Master and Ph.D. degrees, both in computer science. She is also (co-)author of numerous scientific papers and gives lectures at Jiangxi University of Finance and Economics. Her research field includes Image scene Classification, Image process and related topics.