Category Constrained Learning Model for Scene Classification

Yingjun TANG††(a), Student Member, De XU†(b), Member, Guanghua GU†††, Nonmember, and Shuoyan LIU†, Student Member

SUMMARY We present a novel model, named Category Constraint-Latent Dirichlet Allocation (CC-LDA), to learn and recognize natural scene category. Previous work had to resort to additional classifier after obtaining image topic representation. Our model puts the category information in topic inference, so every category is represented in a different topics simplex and topic size, which is consistent with human cognitive habit. The significant feature in our model is that it can do discrimination without combined additional classifier, during the same time of getting topic representation. We investigate the classification performance with variable scene category tasks. The experiments have demonstrated that our learning model can get better performance with less training data.

key words: scene classification, pLSA, LDA, CC-LDA

1. Introduction

Thousands of images are generated every day, which implies the necessity to classify, organize and access them using an easy, fast and efficient way. Scene classification, which is defined as classify images into semantic categories (e.g. coast, office, kitchen, bedroom and so on). Therefore, this has recently become an increasingly important challenge. It is also valuable in image retrieval from databases because an understanding of the scene content can be used for efficient and effective database organization and browsing [1]. Scene modeling for classification using a semantic intermediate representation was proposed recently in order to bridge the semantic gap between pixels and image understanding. Pioneering work [2]–[5] used it to match particular scene. But these intermediate representations usually are high dimension data, it is emergent to find a validate method to reduce their dimension and get top topic representations.

For scene classification task, probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) are common methods to infer the low dimension intermediate representation (topic representation), which had been originally used in text classification. pLSA presented by Hofmann [9], aims at identifying and distinguishing between different contexts of visual term (named visterm be-

Copyright © 2009 The Institute of Electronics, Information and Communication Engineers
2. Our Approach

In light of the flaws among pLSA, LDA and Li-model, our goal is to achieve a model that best represents the distribution of visterms in topic simplex by taking the changes of scene categories into account. Therefore we model the topics with the category information and produce an extended LDA model named Category Constraint-Latent Dirichlet Allocation (CC-LDA).

2.1 CC-LDA Model

In previous works [6]–[8], their models shared topics among all documents (images) and visterms no matter what category image belongs to. Contrasting with Li-model in Fig. 1 (b) and classic LDA model in Fig. 1 (c), the CC-LDA model’s graphic is displayed in Fig. 1 (a).

Similar to Li-model, our model need no supervision but a single category label to the training image and explicitly introduces a category variable for classification. However, Li-model need fixup intermediate themes (topics) while generating each patch of the scene and to share them in all categories. Topics in our model are nominal and will be varied with scene categories. Considering the variation of scene content, our model train category model separately. We model topics for each category and share them in its own image corpus than all categories. Furthermore, the size of topic-set is adaptively changed to its scene category. For own image corpus than all categories. Furthermore, the size varied with scene categories. Considering the variation of scene content, our model train category model separately.

Our model includes three levels: visterm-level, image level and corpus level. The parameters $\alpha$ and $\beta$ are corpus-level parameters, assumed to be sampled once in the training time with each category corpus. The variable $\theta$ is image-level variable, sampled once for each image. Finally, the variable $z$ and $v$ belong to visterm-level, and are sampled once for each visterm in each image. We define some terms for the CC-LDA model, contrasting explicitly the use of terminology in [7]:

A v is the basic unit of an image, defined to be a patch membership from codebook [8]. The node shaded in Fig. 1 (a) indicates that it is an observed.

An image is a sequence of N patches denoted by $W = (v_1, v_2, \ldots, v_N)$. A corpus is a collection of images denoted by $D = (W_1, W_2, \ldots, W_M)$, it also a collection of same scene category during the training time.

$$p(W|\alpha, \beta, c)$$

Equation (1) is the marginal distribution of a image, in which the corpus-level parameter $\alpha$ is a $k$-dimensional vector, defining relative strength of topics in the scene category. The $k$ is the size of topic-set variable $z$. The visterms probabilities are parameterized by a $k \times v$ matrix $\beta$, where $\beta_{ij} = p(v_j = 1|z_i = 1)$. After estimated $k$, $\alpha$ and $\beta$ in training process, they are treated as fixed quantities. A $k$-dimensional Dirichlet random variable $\alpha$ can take values in the $(k-1)$-simplex, and has the Dirichlet probability distribution with $\alpha$ on this simplex. The variable $z$ can be obtained by variable $\theta$ decided by scene category $c$.

2.2 Bayesian Learning and Inference

The Eq. (1) is intractable due to the coupling between $\alpha$ and $\beta$ in the summation over topics. Borrowing from the approach used by Blei [7], we consider to get a tractable family of lower bounds by simplifying the original graphical model in which some of the edges and nodes are removed. Then we define a family of distributions:

$$q(\theta, z|\gamma, \phi) = q(\theta, \gamma) \prod_z q(z|\phi)$$

The optimizing values of parameters $\gamma$ and $\phi$ can be found by minimizing the Kullback-Leibler divergence between the variational distribution $q$ and the true posterior $p$, and they can be achieved by computing the derivatives of the KL divergence and setting them equal to zero.

$$\log p(W|\alpha, \beta, c)$$

$$= L(\gamma, \varphi; \alpha, \beta) + \min D(q(\theta, z|\gamma, \varphi)||p(\theta, z|W, \alpha, \beta, c))$$

$$\approx \max c L(\gamma, \varphi; \alpha', \beta')$$

Therefore, the funtion can be replaced approximately by log likelihood $\max L$ in Eq. (3). For each $L(\gamma, \varphi; \alpha', \beta')$, it can be computed with $\alpha$ and $\beta$ of each category model (referring $L(\gamma, \varphi; \alpha, \beta)$ listed in the appendix A.1 of [7]). Consequently, we need to compute $L$ for C times in order to getting the maximum of L.

In our paper, we adopt EM algorithm to iteratively compute the unknown parameters $\alpha$ and $\beta$ for each scene category model, and get image-level variables $\gamma$ and $\phi$. Then, we can obtain the CC-LDA model composed by C category models (C is the number of category). The $\gamma$ is

![Fig. 1](a) CC-LDA model, (b) Li-model [6], (c) LDA model [7].
image-specific, we view the Dirichlet parameter $\gamma$ as representation of image in the topic simplex. For recognizable task, we compute the optimizing values of the variation parameters $\gamma$ and $L$ for each image with EM algorithm, and each $L$ is computed for $C$ times (because of $C$ category models). Since the maximum $L$ have been achieved with corresponding category model, the process of computing $\gamma$ can be used to decide its category label. Namely, finding the maximum $L$ also means decide scene category model that used to decide its category label. Consequently, we can recognize the category label $c$ while iteratively computing the $\gamma_m$ and $L$ than seek other additional classifier do, which is our significant feature.

$$c = \arg \max_L \{L(\gamma, \varphi; \alpha', \beta')\}, c \in 1 \ldots C$$  \hspace{1cm} (4)

By contrast, Li-model recognized the scene category with ML (Maximum Likelihood) after getting the image intermediate representation $\gamma$, and [1] use K-NN to do decision.

3. Experimental Setup and Result

Our dataset come from [6], that contains 13 category of natural scenes: highway, inside of cities, tall buildings, streets, suburb residence, forest, coast, mountain, open country, bedroom, kitchen, living room and office. In order to compare with the approach of [8], we choose 8 categories including bedroom, suburb residence, coast, forest, mountain and tall building and choose 4 categories including bedroom, forest, mountain and highway. Each scene category is split randomly into two separate set of images. Our experiment has implemented in Matlab 7 by computer with 1.6 GHz processor.

Compared to the global features, local patches are more robust to occlusions and spatial variations. Our choice was motivated by the findings in the [10], where Scale Invariant Feature Transform (SIFT) was found to work best. Furthermore, our main task is to evaluate our learning model, so we just choose grey SIFT to represent image than paying more attention to the selection of feature. Since our model doesn’t share the topic-set among all categories, the differences in the same category will be relatively small and there has no need to give k with a bigger value as done in [7] (which give k with 40 to obtain lower perplexity). In our experiment, we give k with 10 can get better performance. As for special case, some scene such as bedroom, living-room and kitchen, have more changes in content, so the topic-set size will adaptively be adjusted with training performance.

Figure 2 is a summary of our approach. The routine with black dashed is the procedure of training, while blue solid lines mean recognition procedure. In training time, codebook is learnt from patches drawn from the training set, represented by grey SIFT descriptor. Although codebook size is pay a certain role in the classification, we must settle for a solution that trades off efficiency with completeness. So we give 500 visterms to the codebook. We represent an image as Bag of Visterms (histogram of quantized local visual features). Then, we use these representations of each category corpus to train CC-LDA and obtain the category parameters $\alpha$ and $\beta$. In recognition (testing) procedure, we compute $\gamma$ with all category parameters and get the category label by choosing the maximum likelihood.

The average true rate of performance of our model is 70.5%, using100 images to learn and 200 images to recognition for 13 categories task. The performance is quoted from the average of the confusion table similar to [6]. Our performance is much better than [6], they just tested 100 image per category. From the Fig. 3, the best performance comes from forest scene and the worst comes from kitchen scene, because forest scene has more monotone texture and kitchen scene has more variation in style.

We can further explain this situation in Fig. 4 that lists classification examples for some categories with high performance or low performance. The first 3 columns on the left show examples of correctly recognized images, and the last column shows examples of incorrectly recognized images. That can be seen home interior scenes are apt to confuse, since they have many same elements.

We have compared classification performance among four models with variable testing images in Table 1. The CC-LDA and pLSA model share same codebook, whose size is 500. LDA and Li-model use the setup in [6], [7], which use 40 topics to represent all scene categories and 174 visterms. Since SVM can get better performance in high dimension representation [1], we adopt SVM to do classifi-
Fig. 4 Classification examples.

Table 1 Performance comparison under 13 categories.

| # of test image | CC-LDA | Theme1 | LDA | pLSA |
|-----------------|--------|--------|-----|------|
| 100             | 52.92% | 52.5%  | 25.38% | 47.38% |
| 200             | 70.5%  | unknown | 49.88% |

Table 2 Performance comparison between CC-LDA and pLSA.

| Model  | # of test image | # of category |
|--------|-----------------|----------------|
| CC-LDA | 100             | 13             | 52.92% | 66% | 85.5% |
|        | 200             | 13             | 70.5% | 81.94% | 92.75% |
| Li-model [6] | 100             | 8              | 52.5% | unknown | 76.0% |
|        | 200             | 8              | 47.38% | 59.25% | 77.75% |
| pLSA   | 100             | 4              | 49.88% | 61.63% | 78.50% |
|        | 200             | 4              | 49.88% | 61.63% | 78.50% |

cation for pLSA and LDA. That can be seen the average performance of our model is higher than other models, the situation becomes more evident with more testing images to recognize.

We also investigate performance among three models with different category number tasks, using 100 images per category to train. From the Table 2, our performance is better than pLSA and Li-model, getting true rate of 92.75% in four category task.

4. Conclusion

We propose a new method to model natural scene categories by putting the category information in topic inference. Therefore, each scene category can be represented in different topic simplex with the perceptive way people have. Compared with previous work [6],[8], we have demonstrated that our learning model can get better performance with fewer training images, and obtain more improvement with more testing images. The dominant feature in our model is that it can do scene classification, without being dependent on additional classifier. In this paper we pay more attention to validating our learning model than feature descriptors, so we will introduce other feature descriptors to represent patches and to improve our approach’s performance in future work.

Acknowledgement

This work was supported by National Nature Science Foundation of China (60803072), and Youth Science Foundation project of Jiangxi Education Department, China (GJJ09563).

The authors would like to thank Jonathan Huang for providing part LDA code.

References

[1] A. Bosch and A. Zisserman, “Scene classification using a hybrid generative/discriminative approach,” IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.4, pp.712–727, April 2008.
[2] P. Quelhas, “A thousand words in a scene,” IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.9, pp.1575–1589, 2007.
[3] S. Lazebnik and C. Schmid, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” Proc. CVPR, vol.2, pp.2169–2178, 2006.
[4] D. Gokalp and S. Aksoy, “Scene classification using bag-of-regions representations,” Proc. CVPR, 2007.
[5] E. Nowak, “Sampling strategies for bag of features image classification,” ECCV, Part IV, LNCS 3954, pp.490–503, 2006.
[6] F.-F. Li and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” Proc. CVPR, vol.2, pp.524–531, 2005.
[7] D. Blei and Y. Andrew, “Latent Dirichlet allocation,” J. Machine Learning Research, vol.3, pp.993–1020, 2003.
[8] A. Bosch and A. Zisserman, “Scene classification via pLSA,” ECCV, Part IV, LNCS 3954, pp.517–530, 2006.
[9] T. Hofmann, “Probabilistic latent semantic analysis,” Uncertainty in Artificial Intelligence (UAI), 1999.
[10] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.10, pp.1615–1630, 2005.