Image and Vision Computing 51 (2016) 22–35

Contents lists available at ScienceDirect

Image and Vision Computing

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

A discriminative and sparse topic model for image classification and annotation☆,☆☆

Liu Yang [a,b], Liping Jing [a,*], Michael K. Ng [c], Jian Yu [a]

[a] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
[b] College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China
[c] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

ARTICLE INFO

Article history:
Received 14 June 2014
Received in revised form 22 August 2015
Accepted 16 March 2016
Available online 5 April 2016

Keywords:
Graphical model
Discriminative topic
Sparsity
Image classification
Image annotation

ABSTRACT

Image classification is to assign a category of an image and image annotation is to describe individual components of an image by using some annotation terms. These two learning tasks are strongly related. The main contribution of this paper is to propose a new discriminative and sparse topic model (DSTM) for image classification and annotation by combining visual, annotation and label information from a set of training images. The essential features of DSTM different from existing approaches are that (i) the label information is enforced in the generation of both visual words and annotation terms such that each generative latent topic corresponds to a category; (ii) the zero-mean Laplace distribution is employed to give a sparse representation of images in visual words and annotation terms such that relevant words and terms are associated with latent topics. Experimental results demonstrate that the proposed method provides the discrimination ability in classification and annotation, and its performance is better than the other testing methods (sLDA-ann, abc-corr-LDA, SupDocNADE, SAGE and MedSTC) for LabelMe, UIUC, NUS-WIDE and PascalVOC07 images.

1. Introduction

Image classification and annotation are important problems in computer vision. Image classification is to assign a category of an image, and the purpose is to describe an image globally. Traditional supervised learning methods are widely used for image classification, including support vector machine [9,33], random forest [7,26], and probabilistic graphic model [6,14,23,34,29]. Image annotation is to describe individual components of an image by using some annotation terms. Matrix factorization [36], multi-label learning [39,34] and generative techniques [2,4,15,16,28] have been developed.

Image classification and image annotation are strongly related. For examples, an image labeled as a coast scene is likely to be annotated with "boat" and "beach", but is unlikely to be annotated with "door" and "window". An image annotated with "door" and "window" is likely to be labeled as a street scene, but is unlikely to be labeled as a coast scene. Label information can be used to enhance the performance of image annotation, and annotation information can be employed to increase classification accuracy. Therefore, it would be an interesting and challenging research problem how to combine both label and annotation information to design efficient and effective methods for image classification and annotation simultaneously.

Recently, graph model [8,11], neural network model [38], supervised non-negative matrix factorization model [21], regression model [19], and generative graphical model [32,24] are adopted here. Among them, the graph model and the neural network model have higher complexities for the process of building graph or network. The ideas of other models are to identify latent topical bases and represent images in the topical space for later image classification and annotation. However, the learned topics are not discriminative enough to identify the image category because they are related to all visual and annotation terms. An ideal situation is that each topic has its own relevant visual words and annotation terms as shown in Fig. 1.

In this paper, we propose a discriminative and sparse topic model (DSTM) to generate latent topics such that relevant visual words and annotation terms can be identified and irrelevant words and terms can be ignored. The essential features of DSTM include:

(i) The label information is enforced in the generation of visual words and annotation terms, which guarantees that each latent

☆ This paper has been recommended for acceptance by Y. Aloimonos, PhD.

☆☆ The work of L. Yang, L. Jing, and J. Yu was supported in part by the National Natural Science Foundation of China under Grant 61105056, under Grant 61370129, and Grant 61375062, in part by the Fundamental Research Funds through the Central Universities under Grant 2014JBM029, and in part by the CCF-Tencent Open Research Fund. The work of M. K. Ng was supported in part by the Research Grants Council, Hong Kong, through the General Research Fund, Hong Kong Baptist University (HKBU), Hong Kong, under Grant 202013 and Grant 12302715, and in part by HKBU under Grant FRG2/14-15/087.

* Corresponding author.
E-mail addresses: yangliubjtu@bjtu.edu.cn (L. Yang), lpjing@bjtu.edu.cn (L. Jing), mng@math.hkbu.edu.hk (M.K. Ng), jianyu@bjtu.edu.cn (J. Yu).

http://dx.doi.org/10.1016/j.imavis.2016.03.005
0262-8856/

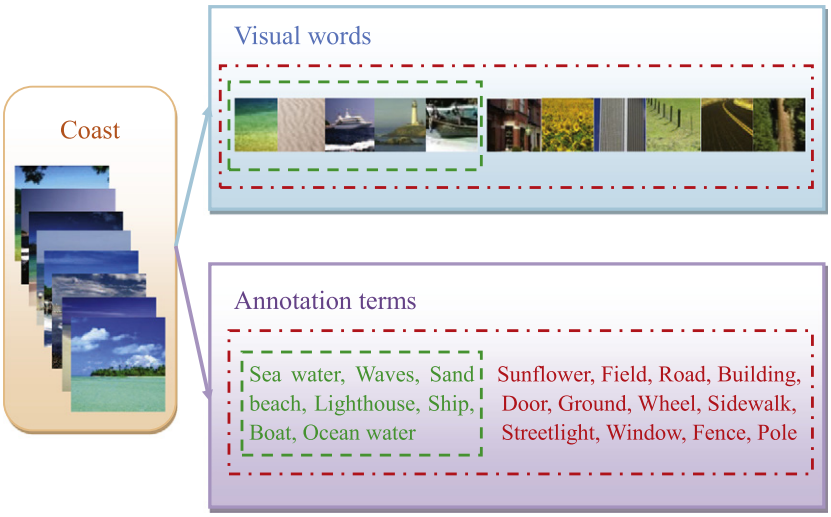

Fig. 1. Illustration of an identified latent topic on visual words or annotation terms related with "coast". The ideal topic marked by green box contains few visual words/annotation terms related with one category. The traditional topic marked by red box often contains many irrelevant visual words/annotation terms.

topic consists of visual words or annotation terms which closely corresponding to a category, i.e., the learned topics are more discriminative.

(ii) The zero-mean Laplace distribution is added to a topic generative process, which makes each topic contain a few visual words and annotation terms, and then an image can be represented sparsely by the latent topics.

(iii) The sparse image representation in the identified topic space is helpful to learn a training model and then to improve classification and annotation performance. A series of experiments on LabelMe, UIUC, NUS-WIDE and PascalVOC07 images are conducted to demonstrate the performance of DSTM is better than the other testing methods (sLDA-ann, abc-corr-LDA, SupDocNADE, SAGE and MedSTC).

The rest of this paper is organized as follows. In Section 2, we describe related work. In Section 3, we introduce DSTM model and its parameter estimation algorithm. In Section 4, experimental results are presented to show the DSTM is effective. Finally, some concluding remarks are given in Section 5.

2. Related work

Probabilistic graphical model has become popular in the data mining community due to its solid theoretical foundation and promising performance. Latent Dirichlet allocation (LDA) [5] is a hierarchical Bayesian model that projects a data point into a latent low dimensional space spanned by a set of automatically learned topical bases. Each topic is a multinomial distribution over the original features. A wide variety of its extensions has been proposed in different contexts for different modeling purposes [35,18,29,34]. In the area of image processing, the final understanding performance benefits from the graphical model based on its ability to integrate multi-resource. For instance, Li and Fei-Fei [23] proposed an integrative model to classify the types of the events and object components of images.

Traditional methods handle image classification and annotation as two dependent tasks. Wang et al. [34] proposed the max-margin latent Dirichlet allocation model (MMLDA) for image classification and annotation. However, they also conduct these two tasks separately. Wang et al. [32] firstly proposed a joint probabilistic graphical model (sLDA-ann) to simultaneously classify and annotate images. sLDA-ann assumes that image annotation and classification share the same latent topic space as shown in Fig. 2(a), where each topic is a distribution over a vocabulary containing all image visual words or annotation terms. As shown in Fig. 3(a), a topic learned by sLDA-ann from LabelMe training data [30] contains terms "sand beach", "sea water", "door", "window", and "sky-scraper" with high probabilities. In this case, it is hard to identify whether this topic is related to "coast", "tall building" or "street" category, i.e., the learned topics have no ability to distinguish the category. Li et al. [24] presented an annotation-by-class corr-LDA (abc-corr-LDA) model by enforcing the label information in the annotation term generative process, as shown in Fig. 2(b). This model narrows the scope of annotation terms for the training images, and then improves the annotation performance. However, abc-corr-LDA ignores the explicit relation between categories and visual words. This will result in that each topic is represented by many visual words distributed over all the dictionary for a testing image as shown in Fig. 4(a) and (b) (this image comes from the LabelMe testing dataset [30]). From these figures, we can find that the topics are not sparse and it is hard to identify the proper category for the testing image.

Zheng et al. [38] proposed the supervised document neural autoregressive distribution estimator (SupDocNADE) model to simultaneously deal with the image classification and annotation tasks. It obtains the hidden topic features by using the neural network on the mixed representation with all visual and annotation words, and learns the connection between hidden layer and the class label. SupDocNADE uses both visual and annotation words to learn the topics, but it does not consider the sparsity of topic and make use of the label information during topic learning process.

In data mining and statistical community, sparse learning becomes a significant research direction to learn parsimonious models. For identifying latent topic, sparsity is a good choice to shrink the vocabulary range of each topic. The existing sparsity-favoring models include L_1 norm, Normal-Gamma, Laplace, spike-and-slab distributions, Student's t [27,31,25]. Among them, only the Laplace distribution is log-concave, which leads to a posterior whose log density is a concave function, then it has a single local maximum [31]. Thus, the Laplace distribution is often used in Bayesian models where sparsity is desired such as the sparse additive generative model (SAGE) [12]. Recently, Zhu and Xing [40] presented the maximum margin supervised sparse topical coding model (MedSTC) which adopted L_1 norm strategy to learn sparse topics via a constrained optimization process, later Zhang et al. [37] extended it to the online version. These two models only consider single task (image classification), i.e., the learned topics are related to only one resource (image visual information), which cannot be used to simultaneously classify and annotate images.