Correspondence with Category Latent Dirichlet Allocation for Image Annotation

Xiaoxu Li, Xiaojie Wang, Chunxiao Wu, Haipeng Liu, Peng Lu
Center for Intelligence Science and Technology
School of Computer Science, Beijing University of Posts and Telecommunications
Beijing, China
xiaoxulibupt@gmail.com

***Abstract*—We present correspondence with category Latent Dirichlet Allocation (corr-c-LDA), a novel probabilistic topic model for the task of image and video annotation. The heart of our annotation model lies in introducing the class label information and assuming the dependence relationships between class label and image feature, as well as class label and annotation words. Instead of modeling the image and annotation words in the formulation of correspondence LDA, our model models the image with class label and annotation words, and tries to avail category information to promote image annotation. We demonstrate the power of our model on 2 standard datasets: a 1791-image subset of UIUC-dataset and a 2400-image LabelMe dataset. The proposed association model shows improved performance over several existing models as measured by F-measure.**

Keywords-image annotation; probabilistic model; maximum likelihood estimation; variational inference

I. INTRODUCTION

Image annotation has been an important task in computer vision, and the goal of which is to annotate an image with several keywords which can describe the content of the image. It would allow one to further exploit the fast indexing and retrieval architecture of Web image search engines for improved image search. So that the problem of annotating images with relevant text keywords has immense practical meaning.

At present, much work has been done on image annotation, and can be broadly summarized into two groups. The first group of methods places annotation words a higher level over images features, and treats the image annotation a classification problem [5]. This kind of methods focuses on modeling the class-conditional density which is the distribution over image features conditioned on an annotation word. When the number of annotation words is less, the method is feasible. However, when the number of annotation words is enormous, the method will be powerless. The second group of methods places annotation words and image feature at the same level, [1], [4], [6]. Using the latent variable framwork, this kind of methods applies itself to construct a joint probability distribution for learning the inner pattern between image feature and annotation words. [1] proposes the LDA-based classical method corr-LDA, which assumes that image and annotation words share the same latent topic variable, and that annotation words generate from subsets of empirical image topics, so as to achieve the correspondence between textual modal and image modal.

In this work, we build on several previous work in probabilistic topic models, corr-LDA in [1] and the classification model proposed by [3]. [3] is the LDA-based seed work for image classification. In the work, each category is identified with its own Dirichlet prior, which is optimized to distinguish between each other. Meanwhile, each class has its own Dirichlet prior, see Figure 1(a). We note the fact that category information can provide certain evidence, or valuable information for image annotation. Once the category of the images was ascertained, it is equivalent to reduce some uncertainty of annotation. It motivates us to construct a new model which models the relation between the image with class label and annotation words. We embed corr-LDA into the model proposed by [3]. Our model is in spirit similar to [3]. We derive the parameters estimation procedure based variatinal EM framework, and give an approximate method for annotating a new image with class label.

The remaining sections of the paper are organized as follows. We introduce basic notation and terminology, and describe our model in Section 2. In Section 3, we show variational inference and parameter estimation for our model. In Section 4, we study the performance of our model on annotation for two real-world image datasets. Finally, we present our conclusions in section 5.

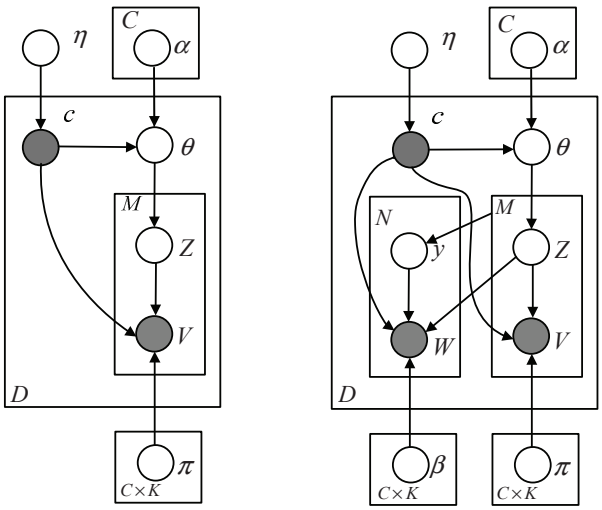
II. MODELING IMAGES, LABELS AND ANNOTATIONS

A. Data representation

Suppose that there are D images with class label and annotations in datasets $\boldsymbol{D}$. We adopt a 'bag-of-word' representation for both images and annotations text. For images, we construct the image codebook first. The length of codebook is denoted as V_s (The more details see section 4). Each image is reduced as a collection of M image words, denoted as $V = \{v_1, v_2, ..., v_M\}$, in which an image word v_m is a unit-basis vector of size V_s. For annotation text, all different annotation words construct the textual vocabulary, is denoted as V_t.

This research was supported by the Major Research Plan of the National Nature Science Foundation of China (90920006).

978-1-61284-774-0/11/$26.00 ©2011 IEEE



(a) (b)

Figure 1. Example of a figure caption.

The annotation of each image is denoted as $W=\{w_1,w_2,...,w_N\}$, in which each annotation word w_n is a unit-basis vector of size V_t. The class label c is a unit-basis vector of size C. So, the dataset $\boldsymbol{D}$ consisting of D image-class-annotation triples is represented as $\boldsymbol{D}=\{(V_d,c_d,W_d)|d\in\{1,2,....,D\}\}$.

B. Correspondence with category Latent Dirichlet Allocation

In this section, we will introduce our model which we call correspondence with category Latent Dirichlet Allocation (corr-c-LDA). The generative process of our model for an image-class-annotation triple (V,c,W) is given as follows:

1) Draw a class label $c\sim\text{Multi}(\eta)$.

2) Draw a topic proportions $\theta\sim\text{Diriclet}(\alpha_c)$

3) For each image word $\boldsymbol{v}_m$, $m\in\{1,2,...,M\}$:

a) Draw topic assignment $z_m|\theta\sim\text{Multi}(\theta)$.

b) Draw image word $v_m|z_m\sim\text{Multi}(\pi_{c,z_m})$

4) For each annotation word w_n, $n\in\{1,2,...,N\}$:

a) Draw a topic identifier of a region $y_n\sim\text{unif}\{1,2,...,N\}$.

b) Draw annotation word $w_n|z_{y_n}\sim\text{Multi}(\beta_{c,z_{y_n}})$.

The graphical model representation of our model is depicted in Figure 1(b). Our model specifies a joint distribution over latent variables and observation variables. Let $\Omega=\{\alpha,\eta,\pi,\beta\}$, $\text{H}=\{\theta,z,y\}$ and $\text{E}=\{V,c,W\}$, then

$$p(\text{E},\text{H}|\Omega)=p(c|\eta)p(\theta|c,\alpha)\prod_{m=1}^{M}p(z_m|\theta)p(v_m|z_m,\pi)\prod_{n=1}^{N}P(y_n|M)P(w_n|y_n,z,\beta,c) \quad (1)$$

From the step 1 and step 2, we need first choose a class label from the multinomial distribution Multi(η), then draw a proportion of topic from the c th Dirichlet distribution which is specified by α_c. The topic refers to a multinomial distribution over image vocabulary and annotation vocabulary. From the step 3, we choose a topic π_{c,z_m} from the image topics of the c th class, and draw an image word v_m from the topic. Repeating step 3 N times, the image parts in a triple can be obtained. From the step 4, we choose an identifier of the images topics which have been drawn, and draw an image word v_m from the topic with the identifier. Repeat step 4 M times, the annotation parts in a triple can be obtained. From the generative process, we can see that in our model, each class has its own annotation topics and images topics, and all the classes in the corpus share the only image or annotation vocabulary. So that the model can learn the differences of annotation words in each class, and allow some similar annotation words among classes at the same time.

III. VARIATIONAL INFERENCE VIA PARAMETER ESTIMATION

We treat $\Omega=\{\alpha,\eta,\pi,\beta\}$ as unknown constants to be estimated, rather than random variables, and carry out approximate maximum-likelihood estimation.

A. Variational inference(E-step)

The posterior distribution of the latent variables conditioned on a triple image-class-annotation, $p(H|V,c,W)$ is intractable to compute. We adopt variational inference method [9] to approximate the posterior. For convenience, let $\Lambda=\{\gamma,\varphi,\lambda\}$ represent variational parameters, $\Omega=\{\alpha,\eta,\pi,\beta\}$ represent model parameters. For a triple, we maximize the lower bound of log probability, which has the form:

$$L(\Lambda,\Omega)=E_q[\log P(E,H|\Omega)]-E_q[\log q(H|\Lambda)], \quad (2)$$

where q is variational distribution over the latent variables which is defined as factorized distribution: $q(H|\Lambda)=q(\theta|\gamma)\prod_{m=1}^{M}q(z_m|\varphi_m)\prod_{n=1}^{N}q(y_n|\lambda_n)$, where γ is K -dimensional Dirichlet parameter, φ_m is a K -dimensional multinomial parameter and λ_n is a M -dimensional multinomial parameter.

For the update of the variational parameters γ, the procedure is the same to [2]:

$$\gamma_i=\alpha_i+\sum_{m=1}^{M}\varphi_{mi} \quad (3)$$

For the update of the parameters φ_m (λ_{nm}), we pick up the terms including φ_m (λ_{nm}) from L. Maximizing the terms under the constraint $\sum_{i=1}^{K}\phi_{mi}=1$ ($\sum_{m=1}^{M}\lambda_{nm}=1$) leads to

$$\varphi_{mi} \propto \pi_{civ_m} \exp\left(\sum_{n=1}^{N}\sum_{l=1}^{C}\sum_{j=1}^{V_t} c^l \lambda_{nm} w_n^j \log \beta_{lij} + \psi(\gamma_i)\right) \quad (4)$$

$$\lambda_{nm} \propto \exp\left(\sum_{l=1}^{C}\sum_{i=1}^{K}\sum_{j=1}^{V_t} c^l \varphi_{mi} w_n^j \log \beta_{lij}\right) \quad (5)$$

We use coordinate ascent, repeatedly optimizing with respect to each parameter while holding the others fixed. It naturally leads to an inference algorithm. Equations 3, 4 and 5 invoke repeatedly until the lower bound Equation 2 converges.

B. *Parameter estimation(M-step)*

After E-step, we have obtained the approximate posterior of each document via variational inference. In M-step, we maximize the lower bound on the log probability of a collection $\boldsymbol{D}$ given by summing Equation 2 over the documents, namely maximize $L(\boldsymbol{D}) = \sum_{d=1}^{D} L(\Lambda_d; \Omega)$ w.r.t model parameters $\Omega = \{\alpha, \eta, \pi, \beta\}$. We isolate the terms including π_{lij} (β_{lij}) and maximizing the terms under the constraint $\sum_{j=1}^{V_s} \pi_{lij} = 1$ ($\sum_{j=1}^{V_t} \beta_{lij} = 1$), then

$$\pi_{lij} \propto \sum_{d=1}^{D}\sum_{m=1}^{M_d} c^l \varphi_{dmi} v_{dm}^j \quad (6)$$

$$\beta_{lij} \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{m=1}^{M_d} c^l w_{dn}^j \varphi_{dmi} \lambda_{dnm} \quad (7)$$

For the optimization of α_c, we adopt the optimization method [2]. In the paper, we don't optimize η. The variational EM algorithm alternates between the two steps until the bound $L(\boldsymbol{D})$ converges.

C. *Image annotation*

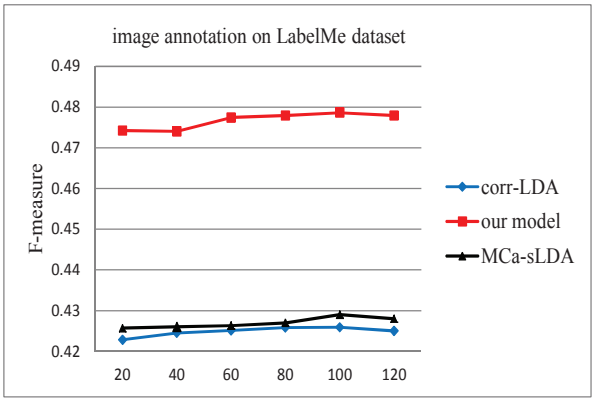
In the previous sections, we have introduced our model and provided the scheme solving model parameters. In this section, we will introduce the procedures predicting annotations using our model. We use LDA [2] inference step for solving the $q(\theta, z)$ first, and it is equivalent to remove the terms including λ or β from the Variational distribution and Equation 4. Given class label and image, we compute the probability appearing every word in the annotation vocabulary, and the words with high probability will be chosen. In particular, the formulation is:

$$p(w|v,c) \approx \sum_{n=1}^{N}\sum_{z_n} p(w|c, z_n, \beta) q(z_n|\phi). \quad (8)$$

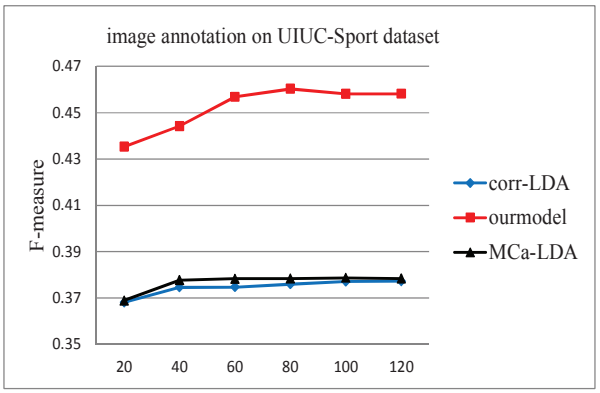
IV. EXPERIMENT

A. *Datasets and preprocessing*

We test our model on the subset of LabelMe dataset from [11], 8 class UIUC-Sport data from [7]. The LabelMe data contains 8 classes: 'street,' 'tall building,' and so on.



(a)



(b)

Figure 2. Comparisons of F-measure over all classes based on 5 random train/test subsets. The horizontal axis shows the number of topics, and red color is our model. (a). F-measure on the LabelMe dataset. (b). F-measure on the UIUC-Sport dataset.

We select randomly 300 images in each class, the total number of images is 2400. The UIUC-Sport data contains 8 classes: 'badminton,' 'polo,' and so on. The number of images in each class varies from 137 (bocce) to 329 (croquet). The total number of images is 1791.

Taking LabelMe data as an example, the steps of the preprocessing are:

1) Choosing the key points. Choose key points applying grid sampling technique (the grid size is 5 × 5), and extract a 16 × 16 patch at each key point. Represent the patches using 128-dimensional SIFT [8] region descriptor.

2) Constructing the codebook. Cluster these descriptors using the k-means algorithm [12]. Construct the codebook of images using all the cluster centers, in which each center is a codeword. We report on a codebook of 240 codewords. Construct the annotation codebook using all different annotations words.

3) Coding the image and annotation words. Code the images and annotations using the corresponding codeword. Finally, remove the annotation terms that occurred less than 3 times, then split each class to create the training and testing sets evenly.

Ground truth (class: polo)
Mallet, athlete, horse, ground, tree, fence.
Corr-LDA
Sky, tree, audience, *badminton racket*, *shutter*.
Mca-LDA
sky, tree, *wicket*, *trees*, *mallet*.
Our model
Mallet, athlete , horse, ground, tree.

Ground truth(class: badminton)
Athlete, badminton racket, ground, audience, billboard shutter chair wall flag
Corr-LDA
sky, athlete, *road*, *trees*, shutter.
Mca-LDA
sky, athlete, *mallet*, *fence*, *road*, *tree*.
Our model
badminton racket, shutter, athlete, battledore, floor.

Ground truth (class: rowing)
athlete, sailing boat, ball, woods, sky, water.
Corr-LDA
sky, *athlete*, *oar*, *rowboat*, *road*, *shutter*.
Mca-LDA
athlete, sky, *tree*, *oar*, *rowboat*.
Our model
athlete, sailing boat, water, trees, ball.

Ground truth (class: sailing)
referee, athlete, pleasure ground, fence, Lawn, tree, mallet, ball, plant.
Corr-LDA
sky, athlete, *shutter*, *road*, *trees*,
Mca-LDA
athlete, sky, ball, *water*, *battledore*.
Our model
athlete, *water*, pleasure ground, ball, fence.

Figure 3. Example results from UIUC-Sport dataset. The italic words indicate unrelated annnotions.

For UIUC-Sport data, the preprocessing step is similar to the above procedure. The only difference is that we extract 2800 32 × 32-size patches uniformly for each image in the dataset. Note that all testing is on unannotated images with class label.

B. Image annotation

In order to assess the annotation quality of our model, we adopt F-measure on our model and two state-of-the-art annotation models, corr-LDA (03) [1] and multi-classsLDA with annotation (MCa-sLDA) [4]. MCa-sLDA (09) reports a better annotation performance comparing with corr-LDA. We first annotate each image in the test set with 5 words by using Equation 8. And then we compute top-(N = 5) F-measure. As seen in Figure 2, our model performs better than the other two approaches. Compared with them, our model improves annotation by about 5% on LabelMe datasets and about 8% UIUC-Sport dataset.

The reason of the improvement is that our model annotates the images according to the category. Once the category of the image is ascertained, the scope of annotation will be narrowed, and the probability of generating irrelevant annotation words will be reduced.

Figure 3 shows the example results of our model, corr-LDA and MCa-sLDA on the two datasets. From Figure 3, we can find that the first image in Figure 3 is labeled as 'polo'. However, corr-LDA annotates the image using the 'badminton racket' and 'shutter', and MCa-sLDA annotates the image using 'wicket'. These words don't obviously belong to the class 'polo'. Other images in the figure present similar situation. Our model just improves the annotation aiming at the problem. The results show further that our model is a better annotation tool.

V. CONCLUSION

In this work, we propose a new probabilistic topic model corr-c-LDA for the task of image and video annotation, and the model models the relation between the image with label and annotation words. The main contribution of this work is proposing corr-c-LDA inspired by [3], and deriving the parameters estimation of the model and the procedure of a new image. Experimental results on image annotation show that the association model of corr-c-LDA has better performance than corr-LDA and MCa-sLDA in predicting annotation. In the future work, we plan on extending corr-c-LDA to predict free-form texts.

ACKNOWLEDGMENT

This research was supported by the Major Research Plan of the National Nature Science Foundation of China (90920006).

REFERENCE

[1] D. M. Blei and M. I. Jordan. Modeling annotated data. In ACM SIGIR, 2003.

[2] D. M.Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. JMLR, 3:993C 1022, 2003.

[3] L. Fei-fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In CVPR, 2005.

[4] C. Wang, D. M. Blei, and L. Fei-fei. Simultaneous image classification and annotation. In CVPR, 2009.

[5] D. Putthividhya, H. T. Attias, S. S. Nagarajan Supervised Topic model for automic Iamge Annotaion. In ICASSP, 2010.

[6] Duangmanee Putthividhya, Hagai T. Attias, Srikantan S.Nagarajan. Topic Regression Multi-Modal Latent Dirichlet Allocation for Image Annotation. In CVPR, 2010.

[7] L.-J. Li and L. Fei-Fei. What, where and who? Classifying event by scene and object recognition. In ICCV, 2007.

[8] D. Lowe. Object recognition from local scale-invariant features. In ICCV, 1999.

[9] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. Machine Learning, 37(2):183C233, 1999.

[10] J. Nocedal and S. J. Wright. Numerical Optimization. Springer, 2006.

[11] B. C. Russell, A. B. Torralba, K. P. Murphy, and W. T. Free-man. LabelMe: A database and web-based tool for image annotation. IJCV, 77(1-3):157C173, 2008.

[12] T. Kadir and M. Brady. Saliency, scale and image description. IJCV, 45(2):83C105,2001.