

SUPERVISED TOPIC MODEL FOR AUTOMATIC IMAGE ANNOTATION

*D. Putthividhya*¹, *H. T. Attias*^{2, 3}, *S. S. Nagarajan*³

¹Institute for Neural Computation, UCSD
9500 Gilman Drive,
La Jolla, CA 92093

²Golden Metallic, Inc.
P.O. Box 475608
San Francisco, CA 94147

³Dept. of Radiology, UCSF
513 Parnassus Ave,
San Francisco, CA 94143

ABSTRACT

This paper presents a new probabilistic model for the task of image annotation. Our model, which we call sLDA-bin, extends supervised Latent Dirichlet Allocation (sLDA) model to handle a multi-variate binary response variable of the annotation data. Unlike correspondence LDA (cLDA), the association model in sLDA allows each caption word to be associated with more than 1 image region and is thus more appropriate for annotation words that globally describe the scene. By modeling the response variable as a multi-variate Bernoulli, we introduce a tight convex variational bound for the logistic function and derive an efficient variational inference algorithm based on mean-field approximation. Our model compares favorably with cLDA on an image annotation task, as demonstrated by a superior caption prediction probability.

Index Terms— Statistical Topic Models, Probabilistic Graphical Models, Automatic Image Annotation, Image Retrieval, Multimedia Signal Processing.

1. INTRODUCTION

The problem of image and video retrieval has been at the forefront of computer vision research over the past decades. Nonetheless, with recent unprecedented volumes of images and videos available online, there is a growing demand for an efficient algorithm to search and navigate through large-scale collections. Current state-of-the-art image search engines rely heavily on the use of annotated text or captions to identify and retrieve images. While such an approach allows for high-level semantic querying, the caption information, which is vital to the success of the text-based search technology, is often manually obtained—a process that cannot scale with the continually growing size of today’s multimedia corpus. There is therefore an immediate need to automate this annotation process. With its potential impact on a wide array of applications involving digital media archives, considerable amount of attention in recent years has been given to the design and development of automated tools to annotate images and videos.

Given an image with no captions, the task of an annotation algorithm is to predict the missing captions by learning patterns of association between image and text. Previous work in

this area can be broadly categorized into 2 groups. In the first line of work, the problem of image annotation is cast as a supervised learning problem where annotation words are treated as concept classes [1, 2]. For each word in the vocabulary, the class conditional density is learned from all the images tagged with that word. During annotation, the posterior over class labels is computed, and the concepts with highest probability are then used as predicted captions. In practice, this line of approach suffers a scalability issue and can only handle a small annotation vocabulary as the class-conditional density must be learned for each word.

Another set of techniques treat annotation and image data on a more equal footing by modeling the joint statistical correlation between the 2 data types. Using a latent variable framework, these models learn the joint probability distribution of text and image features by assuming that for each document there is a small set of hidden factors that govern the association between the image features and the corresponding caption words, see [3, 4, 5]. Several variations of the same association model are presented with the different forms of probability distribution assumed for caption words (multinomial vs. bernoulli) and different image modeling (non-parametric density estimation vs. mixture of Gaussians).

In this paper, we propose an annotation model that builds on several previous work in probabilistic topic models, namely Latent Dirichlet Allocation (LDA), correspondence LDA (cLDA), and supervised LDA (sLDA). More specifically, our model, which we call sLDA-bin, extends sLDA to account for multi-dimensional binary response variables suitable for the annotation data. We model the binary response using a multi-variate Bernoulli distribution, and in process introduce the logistic link function which makes the computation for inference and parameter learning intractable. We derive an efficient variational inference algorithm based on mean-field approximation and adopt a tight convex variational bound for the logistic function, similar to [6]. Using a subset of the COREL dataset with 5000 images, we demonstrate the power of our model on an image annotation task. Experimental results show that sLDA-bin performs competitively with cLDA as measured by caption prediction probabilities.

2. PROPOSED MODEL

2.1. Data Representation

We borrow a tool from statistical text document analysis and represent an image as a bag of words. In such a representation, word ordering is ignored and an image is simply reduced to a vector of word count. We adopt the following notation. Each image is a collection of N patches and is denoted as $\mathbf{R} = \{r_1, r_2, \dots, r_N\}$ where r_n is a unit-basis vector of size T_r with exactly one non-zero entry representing the membership to only 1 codeword in the dictionary of T_r visual words. For caption data, we simply record the presence/absence of each caption word in an image, similar to [3]. We denote this as a $T_t \times 1$ binary vector \mathbf{w} , where the entry w_i takes a value 1 if word i is present in an image and 0 otherwise. A collection of D image-caption pairs is denoted as $\{\mathbf{R}_d, \mathbf{w}_d\}$, $d \in \{1, 2, \dots, D\}$.

2.2. Supervised Latent Dirichlet Allocation with Multi-variate Binary Response Variables (sLDA-bin)

Supervised LDA builds on the basic LDA model which uses hidden variables, loosely termed topics, to cluster words and model word co-occurrences. Given a collection of documents (in a bag-of-words form), LDA decomposes the distribution of word counts from each document into contributions from K topics. A document under LDA is modeled as a proportion of topics (mixture of topics), while each topic, in turn, is a multinomial distribution over terms. When the number of topics K is much smaller than the size of vocabulary, LDA can be seen as performing dimensionality reduction by learning a small set of projection directions (topics) that account for most of the correlations in the data. The low-dimensional subspaces (topics) uncovered by LDA often reveal semantic structures useful for visualization, browsing, and navigation through large repositories of text collections.

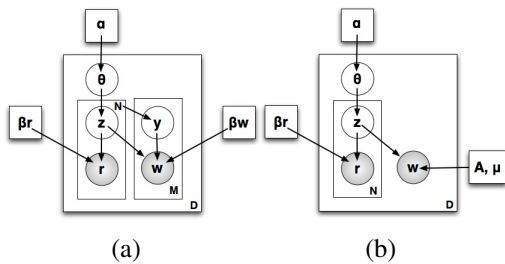


Fig. 1. (a) Correspondence LDA (cLDA). (b) Supervised Latent Dirichlet Allocation with a multi-variate binary response variable (sLDA-bin).

When working with labeled documents or documents paired with their associated response values, where the goal is to predict the response values given the document, it might be useful to incorporate these response variables into learning the low-dimensional mapping of documents. This is precisely the motivation behind the supervised LDA (sLDA) model proposed in [7]. Instead of inferring hidden topics that

best explain correlations between words in documents, sLDA finds latent topics that are best predictive of the response variables. More specifically, sLDA adds to LDA a linear regression module that allows a real-valued response variable to be linearly predicted from the empirical topic proportion $\bar{\mathbf{z}} = \frac{1}{N} \sum_{n=1}^N z_n$. Instead of regressing over the mean topic proportion θ , the formulation in [7] makes the response variable directly dependent on the topics that actually occur in the document. Such a setup prevents the learning of topics that are used entirely to explain either the response variables or the words in the documents, as these topics will not be of use in predicting the response variables. In [5], sLDA was extended to handle response variables that are of categorical type (e.g. class labels) using a softmax function.

Motivated by such a success of sLDA in a prediction task, here we adopt the association model of sLDA for the task of annotation. In such a setting, documents are images while caption data is the response variable that we want to predict. The goal is to learn image topics that are predictive of the caption words. In order to handle the multi-variate binary response variables of the annotation data, we model the distribution of the response variables as a multi-variate Bernoulli and use the logistic link function to define its probability. By adopting the same empirical topic frequency covariates, the probability of tagging the current image with caption word w_i is given by $p(w_i | \mathbf{Z}, \mathbf{A}, \mu) = \sigma(\mathbf{a}_i^\top \bar{\mathbf{z}} + \mu_i)^{w_i} \sigma(-\mathbf{a}_i^\top \bar{\mathbf{z}} - \mu_i)^{1-w_i}$, where $w_i \in \{0, 1\}$, $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic function, and $\{\mathbf{a}_i, \mu_i\}$ are regression coefficients for word i . To generate an image-caption pair with N image patches, we follow the generative process of sLDA-bin as described below and illustrated in the graphical model of Fig. 1(b):

- Draw a topic proportion $\theta | \alpha \sim \text{Dir}(\alpha)$
- For each image patch r_n , $n \in \{1, 2, \dots, N\}$
 1. Draw topic assignment $z_n | \theta \sim \text{Mult}(\theta)$
 2. Draw word $r_n = t | z_n = k \sim \text{Mult}(\beta_{kt}^r)$
- Given the empirical topic proportion $\bar{\mathbf{z}} = \frac{1}{N} \sum_{n=1}^N z_n$, for each caption word $i \in \{1, 2, \dots, T_t\}$, draw a Bernoulli r.v. $w_i \sim p(w_i)$ where $p(w_i)$ is given by:

$$p(w_i | \mathbf{Z}, \mathbf{A}, \mu) = \sigma(\mathbf{a}_i^\top \bar{\mathbf{z}} + \mu_i)^{w_i} \sigma(-\mathbf{a}_i^\top \bar{\mathbf{z}} - \mu_i)^{1-w_i}.$$

Our model is closely related to correspondence LDA (cLDA) [4] which also extends the basic LDA model for an image/video annotation task. The main difference, however, lies in how image features are associated with their captions. With the goal of image region annotation, each caption word in [4] is restricted to be associated with 1 particular image region. As seen from the graphical model of cLDA in Fig.1(a), each caption word is generated by first selecting an image region it associates with; now with the hidden topic of that region, we generate a word. In practice, however, some annotation words do globally describe the scene as a whole, using such a restrictive association model could prove to be very inaccurate. By regressing over the empirical topic proportion, our formulation allows each caption word to be influenced by the topics from all image regions as well as by a particular image region

depending on the corresponding regression coefficients. Our association model is thus more general and accurately reflects the process of how true annotation is generated.

2.3. Variational Inference

To infer the posterior over hidden variables, we begin with the expression of the log-likelihood for an image-caption pair:

$$\log p(\mathbf{w}, \mathbf{R}|\Psi) \geq \int q(\mathbf{Z}, \theta) \log p(\mathbf{w}, \mathbf{R}, \mathbf{Z}, \theta|\Psi) d\mathbf{Z}d\theta - \int q(\mathbf{Z}, \theta) \log q(\mathbf{Z}, \theta) d\mathbf{Z}d\theta. \quad (1)$$

where Ψ denotes the model parameters $\{\beta^r, \mathbf{A}, \mu\}$. Equality in (1) holds when the posterior over the hidden variables $q(\mathbf{Z}, \theta)$ equals the true posterior $p(\mathbf{Z}, \theta|\mathbf{w}, \mathbf{R})$. Like in LDA, computing the exact joint posterior is computationally intractable as the hidden variables $\{\mathbf{Z}, \theta\}$ become highly dependent when conditioned on the observed document $\{\mathbf{w}, \mathbf{R}\}$. We employ a mean-field approximation to approximate the joint posterior distribution with a variational posterior in a factorized form: $p(\mathbf{Z}, \theta|\mathbf{w}, \mathbf{R}) \approx \prod_n q(z_n)q(\theta)$. The problem now becomes one of finding, within such family of factorized distributions, the variational posterior that maximizes the lower bound of the data log-likelihood. Taking the expectation w.r.t the variational posterior, the lower bound in RHS of (1) can be expressed as:

$$\mathcal{F} = \sum_n E[\log p(r_n|z_n, \beta_r)] + E[\log \theta] + E[\log p(\theta|\alpha)] + \sum_i E[\log p(w_i|\mathbf{Z}, \mathbf{A}, \mu)] - \sum_n E[\log q(z_n)] - E[\log q(\theta)]. \quad (2)$$

The logistic function complicates computing the expectation in the term $E[\log p(w_i|\mathbf{Z}, \mathbf{A}, \mu)]$. We make use of convex duality and represents the logistic function as a point-wise supremum of a square function of $\mathbf{a}_i^\top \bar{\mathbf{z}} + \mu_i$ (same derivation as in [6]). More specifically, we have:

$$\log p(w_i|\mathbf{Z}, \mathbf{A}, \mu) \geq \frac{2w_i - 1}{2} (\mathbf{a}_i^\top \bar{\mathbf{z}} + \mu_i) - \log(2 \cosh(\frac{\xi_i}{2})) - \lambda(\xi_i) (\mathbf{a}_i^\top \bar{\mathbf{z}} \bar{\mathbf{z}}^\top \mathbf{a}_i + 2\mu_i \mathbf{a}_i^\top \bar{\mathbf{z}} + \mu_i^2 - \xi_i^2), \quad (3)$$

where we introduce variational parameters ξ_i which correspond the point of contact where the lower bound touches the logistic function. $\lambda(\xi_i)$ here denotes $\frac{\tanh(0.5\xi_i)}{4\xi_i}$.

Due to the Dirichlet-multinomial conjugacy, the posterior $q(\theta)$ takes the form of a Dirichlet and we use $\tilde{\alpha}$ to denote its parameters. To simplify the notation, we write $q(z_n = k)$ as ϕ_{nk} . By making use of the following expectations $E[\bar{\mathbf{z}}] = \frac{1}{N} \sum_n \phi_n$ and $E[\bar{\mathbf{z}}\bar{\mathbf{z}}^\top] = \frac{1}{N^2} (\sum_n \text{diag}(\phi_n) + \sum_n \phi_n \sum_{m \neq n} \phi_m^\top)$, we can now differentiate \mathcal{F} w.r.t to the variational posterior parameters and obtain the update rules:

$$\log \phi_n = \log \beta_{r,t}^n + E[\log \theta] + \sum_{i=1}^{T_t} \left[\frac{2w_i - 1 - 4\lambda(\xi_i)\mu_i}{2N} \mathbf{a}_i - \frac{\lambda(\xi_i)}{N^2} \left(\text{diag}(\mathbf{a}_i \mathbf{a}_i^\top) + 2\mathbf{a}_i \mathbf{a}_i^\top \sum_{m \neq n} \phi_m \right) \right] \quad (4)$$

$$\tilde{\alpha}_k = \sum_n \phi_{nk} + \alpha_k \quad (5)$$

$$\xi_i^2 = \mathbf{a}_i^\top E[\bar{\mathbf{z}}\bar{\mathbf{z}}^\top] \mathbf{a}_i + 2\mu_i \mathbf{a}_i^\top E[\bar{\mathbf{z}}] + \mu_i^2 \quad (6)$$

In annotation, given a test image without caption \mathbf{R} , the task is to infer the most likely caption words. For this task, we run variational inference on \mathbf{R} until convergence and use the inferred ϕ_n and $q(\theta)$ to approximate the conditional probability $p(\mathbf{w}|\mathbf{R})$ as follows:

$$p(\mathbf{w}|\mathbf{R}) \approx \int p(\mathbf{w}|\mathbf{Z}, \mathbf{A}, \mu) q(\mathbf{Z}|\mathbf{R}) d\mathbf{Z} \approx \prod_{i=1}^{T_t} \sigma(\mathbf{a}_i^\top E[\bar{\mathbf{z}}] + \mu_i)^{w_i} \sigma(-\mathbf{a}_i^\top E[\bar{\mathbf{z}}] - \mu_i)^{1-w_i} \quad (7)$$

where $E[\bar{\mathbf{z}}] = \frac{1}{N} \sum_n \phi_n$, with ϕ_n inferred from each test image by withholding the caption.

2.4. Parameter Estimation

To update the model parameters $\Psi = \{\beta^r, \mathbf{A}, \mu\}$, we maximize the lower bound of the log-likelihood in (2) w.r.t. Ψ and obtain the following closed-form updates:

$$\beta_{kt} = \frac{\sum_{d,n} \phi_{nk}^d 1(r_n^d = t)}{\sum_{t,d,n} \phi_{nk}^d 1(w_n^d = t)} \\ \mathbf{a}_i = \left(2 \sum_d \lambda(\xi_i^d) E[\bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top] \right)^{-1} \left(\sum_d (w_i^d - \frac{1}{2} - 2\lambda(\xi_i^d)\mu_i) E[\bar{\mathbf{z}}_d] \right) \\ \mu_i = \frac{\sum_d (w_i^d - \frac{1}{2} - 2\lambda(\xi_i^d)\mathbf{a}_i^\top E[\bar{\mathbf{z}}_d])}{\sum_d 2\lambda(\xi_i^d)}$$

3. EXPERIMENTAL RESULTS

We test our model on the 5,000 image subset of the COREL dataset obtained from [2]. This subset contains 50 classes of images, with 100 images per class. Each image in the collection is reduced to size 117×181 (or 181×117). 4,500 images are used in training (90 images from each class), and 500 for testing (10 images per class). Each image is treated as a collection of 20×20 patches obtained by sliding a window every 15 pixels, resulting in 77 patches per image.

While there exists a great variety of image features in the literature to choose from, the SIFT feature descriptors have been shown to be discriminative in numerous classification and recognition tasks. In this work, we follow [8] and use the 128-dimensional SIFT patch descriptor computed on 20×20 gray-scale patches. In addition, we use the 36-dimensional robust color descriptors proposed in [9] which have been designed to complement the SIFT-descriptors extracted from the gray-scale patches. To learn a dictionary of visual words, we run a k-means algorithm on a collection of 164-dimensional features and obtain a set of T_r visual words.

3.1. Caption Prediction Probability

We compare the performance of sLDA-bin with cLDA in [4]. To measure the caption quality of the models, we compute the caption prediction probability as defined as follows:

$$\text{score} = \sum_d \log p(\mathbf{w}_d|\mathbf{R}_d), \quad (8)$$

where $p(\mathbf{w}_d|\mathbf{R}_d)$ is approximated using the formula given in (7) and the summation is performed over all 500 test images.

	cLDA	sLDA-bin
$T_r = 128$	$-7.8829 \pm .0071e3$	$-7.1362 \pm .0366e3$
$T_r = 256$	$-7.8327 \pm .0094e3$	$-7.1268 \pm .0346e3$
$T_r = 512$	$-7.7930 \pm .0233e3$	$-7.0381 \pm .0324e3$

Table 1. Comparison of caption prediction probabilities for $K = 30$, as the number of visual words T_r increases.

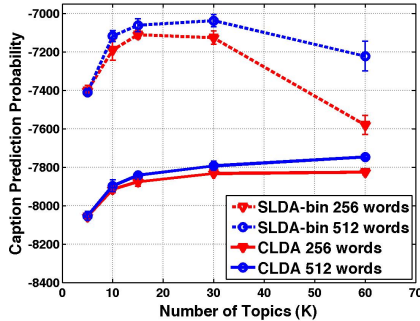


Fig. 2. Comparison of Caption Prediction Probability between cLDA and sLDA-bin as a function of number of topics.

As seen in the plot in Fig 2, sLDA-bin gives superior prediction probabilities compared to cLDA for all values of K . We attribute the good predictive capabilities to the direct association model between image topics and the corresponding captions. By eliminating the hidden variables denoting the topics for captions, image topics can be used directly to predict the captions without integrating over the posterior probability of the caption topics (the step necessary in cLDA). As we increase the number of visual words in the dictionary of code words, we also obtain better predictive probabilities as seen in Fig. 2 and Table 1.

Examples of predicted captions generated by sLDA-bin and cLDA are shown comparatively in Fig. 3. In most scenarios, captions predicted by our model contain specific words that are semantically related to the content of the image and the true captions, while cLDA has preference for more general words (e.g. sky, water, tree) in its choice of captions.

4. CONCLUSION

In this work, we propose a new image/video annotation model that extends the basic supervised LDA model in [7] to accommodate binary response variables of the annotation data. The main contribution of this work is the derivation of sLDA approximate inference algorithm that uses a convex dual representation of the logistic link function to simplify the computation. Experimental results on image annotation show that the association model of sLDA-bin is better suited than correspondence LDA in predicting annotation as seen in the superior annotation quality. In the future work, we plan on extending sLDA-bin to predict free-form texts.

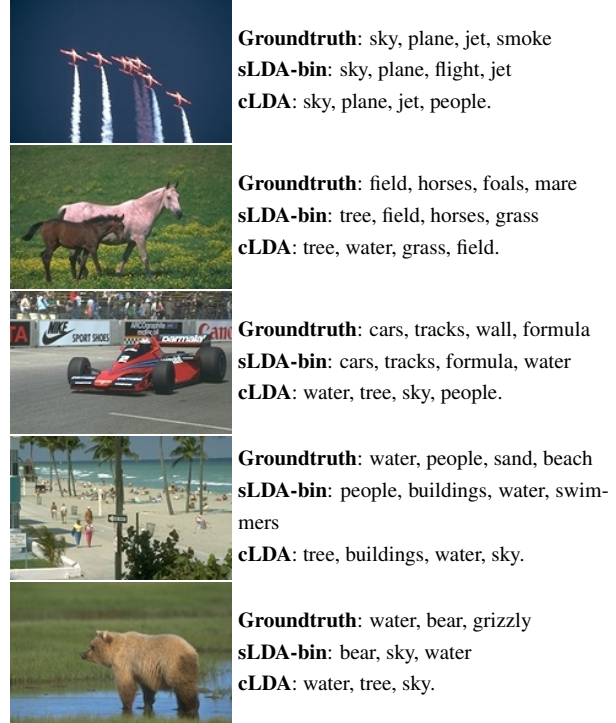


Fig. 3. Examples of predicted annotation.

5. REFERENCES

- [1] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, 2003.
- [2] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [3] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Advances in Neural Information Processing Systems (NIPS)*, 2003.
- [4] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *ACM SIGIR*, 2003.
- [5] C. Wang, D. M. Blei, and L. Fei-fei, "Simultaneous image classification and annotation," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [6] A. I. Schein, L. K. Saul, and L. H. Ungar, "Generalized linear model for principal component analysis of binary data," in *AISTATS*, 2003.
- [7] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Neural Information Processing Systems (NIPS)*, 2007.
- [8] L. Fei-fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [9] J. van de Weijer and C. Schmid, "Coloring local feature extraction," in *ECCV*, 2006.