

Semi-Supervised Topic Modeling for Image Annotation

Yuanlong Shao, Yuan Zhou, Xiaofei He, Deng Cai, Hujun Bao
{shaoyuanlong, zhouyuan, xiaofeihe, dengcai, bao}@cad.zju.edu.cn
State Key Laboratory of CAD&CG, Zhejiang University
No. 38, Zheda Road, Hangzhou, Zhejiang, P.R.China

ABSTRACT

We propose a novel technique for semi-supervised image annotation which introduces a harmonic regularizer based on the graph Laplacian of the data into the probabilistic semantic model for learning latent topics of the images. By using a probabilistic semantic model, we connect visual features and textual annotations of images by their latent topics. Meanwhile, we incorporate the manifold assumption into the model to say that the probabilities of latent topics of images are drawn from a manifold, so that for images sharing similar visual features or the same annotations, their probability distribution of latent topics should also be similar. We create a nearest neighbor graph to model the manifold and propose a regularized EM algorithm to simultaneously learn a generative model and assign probability density of latent topics to images discriminatively. In this way, databases with very few labeled images can be annotated better than previous works.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*Indexing methods*

General Terms

Algorithms, Theory, Languages

Keywords

Automatic Image Annotation, Semantic Indexing, Laplacian Regularization, Semi-supervised Learning

1. INTRODUCTION

In recent years, as digital cameras become increasingly affordable and widespread, personal digital photos are growing exponentially and photo sharing through the Internet also becomes a common practice. To achieve the potential value of large image collections, users have to be able to

search and access images that are wanted effectively. Image annotation, the task of associating text to the semantic content of images, is a good way to reduce the semantic gap and can be used as an intermediate step to image retrieval. It enables users to retrieve images by text queries and often provides semantically better results than content-based image retrieval. In recent years, it is observed that image annotation has attracted more and more research interests.

The fundamental problem of image annotation is how to model the relationship among different modalities, including visual features and textual annotations, associated with the possibly existed latent topics of images, as well as the relationship among different images. Latent topic modeling has long been a promising approach for this problem [3], [1], [8], [9]. As is common, model based approaches have the benefit of better efficiency and stability, while it suffers mostly from probably insufficient modeling, *i.e.*, when the model does not fully describe the problem domain, the inferred quantities may not be accurate, *e.g.*, if data is not distributed in a Gaussian distribution, modeling it with a Gaussian will cause problems. For image annotation, it is always difficult for the probabilistic modeling to be sufficient due to the high variance of image contents.

In the contrast, traditional similarity-based approaches, *e.g.*, spectral clustering and manifold regularization, does not have to assume the specific probability structure of data, and requires only a similarity function defined on pairs of data instances. Recently, this approach has been shown to be successful in the semi-supervised context [2], [12]. When regarded as a regularization, it is also applicable to probabilistic models [6],[7],[4]. We introduce this approach into model based image annotation.

In practice, for images sharing some collective properties, it is reasonable to assume they have similar semantic concepts. In this paper, we incorporate the manifold assumption to say that the probabilities of latent topics of images reside on or close to a manifold, so that for images sharing similar visual features or the same annotations, they should have similar probabilities over different latent topics. We then fit the generative model with respect to the manifold structure which is modeled by a nearest neighbor graph. Using graph Laplacian, the manifold structure can be incorporated in the standard EM algorithm as a regularization term [6],[7],[4], and the semi-supervised annotation can be carried out in a consistent fashion. Our experimental results show that the latent topics learned in this way catches better the similarity relationship between images, which reflect some properties of discriminative learning.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'09, October 19–24, 2009, Beijing, China.

Copyright 2009 ACM 978-1-60558-608-3/09/10 ...\$10.00.

2. LAPLACIAN REGULARIZED IMAGE ANNOTATION MODEL

Inspired by [8], [9], [11], we use a pLSA-like model to explicitly represent the image level latent topics for modeling the connection between visual features and textual annotations of an image, as depicted in Fig. 1(a). We suppose that there are I images in the database. Among them, I_a images are labeled, *i.e.*, they are with textual annotations. Another I_u images are unlabeled. The index sets of the three are \mathbf{I} , \mathbf{I}_a and \mathbf{I}_u , respectively. The task is to find out the most probable textual annotations of the unlabeled images.

We suppose that each image is associated with a single hidden variable Z representing a specific latent topic. And there are K latent topics denoted as $\{z_1, \dots, z_K\}$. The probability of image i to have latent topic z_k is denoted as p_{ki} . Note that p_{ki} may not be equal to $P(z_k|\mathbf{F}_i, \mathbf{W}_i)$ when the probabilistic model is not sufficient on describing the problem. The words in the dictionary are denoted as $\{w_1, \dots, w_V\}$, where V is the size of the vocabulary. The visual features of images are vector quantized into U clusters denoted as $\{f_1, \dots, f_U\}$ to form a visual vocabulary [5], so that visual features of an image is regarded as visual words and processed in the same way as textual words.

For the i th image in the database, we use a U -nary feature vector \mathbf{F}_i to denote all the features it have, *i.e.*, the u th entry F_{iu} is the occurrence count of features of the i th image which are assigned to visual word f_u . Similarly, \mathbf{W}_i is the V -nary word vector denoting the occurrence counts of words of the i th image. For an unlabeled image $i \in \mathbf{I}_u$, \mathbf{W}_i is an all zero vector.

We assume that visual words and textual words are conditioned on latent topics independently. They are generated as follows: for each image, first, choose a topic Z_i for the i th image from the distribution Multinomial(α), then generate N_i visual features repeatedly from the distribution Multinomial(β_{Z_i}) to form the feature vector \mathbf{F}_i , where N_i is the sum of entries of \mathbf{F}_i . The word vector \mathbf{W}_i is generated in a similar way from the distribution Multinomial(ϕ_{Z_i}) with M_i words.

The log likelihood to maximize with standard EM is,

$$\mathbb{L} = \sum_{i=1}^I \log \sum_{k=1}^K P(z_k|\alpha) \prod_{u=1}^U P(f_u|z_k, \beta)^{F_{iu}} \prod_{v=1}^V P(w_v|z_k, \phi)^{W_{iv}} \quad (1)$$

In the standard M step, expectations are taken w.r.t. the posterior probability $P(z_k|\mathbf{F}_i, \mathbf{W}_i)$'s. According to a recent view of EM [10], it is equivalent to taking expectations w.r.t. the variational distributions $\{p_{ki}\}$, and maximizing the negative free energy as follows. In this way, the two-step EM algorithm becomes a single-objective optimization which makes it easy for us to add regularization,

$$\mathbb{F} = \sum_{i=1}^I \sum_{k=1}^K p_{ki} \log P(z_k|\alpha) \prod_{u=1}^U P(f_u|z_k, \beta)^{F_{iu}} \prod_{v=1}^V P(w_v|z_k, \phi)^{W_{iv}} - \sum_{i=1}^I \sum_{k=1}^K p_{ki} \log p_{ki} \quad (2)$$

A variant of this model named GM-Mixture is also proposed in [3], where the feature descriptors are modeled with a Gaussian component. However, the various restrictions imposed on the Gaussian component in [1] suggest that it has numerical problems when dealing with high dimensional

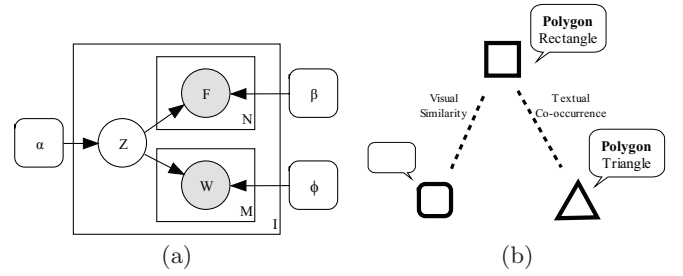


Figure 1: (a) Graphical model representation for the generative part; (b) Similarity graph for the discriminative part. Images as nodes are connected to each other according to visual similarity or co-occurrence of words. Some images have no textual annotations.

feature descriptors. Thus, we introduce the visual vocabulary approach to cope with this problem [5]. Additionally, in contrast to those in [3], [1], [11], our model assigns a single latent topic to each image. This is suitable for the discriminative part to put constraints on them based on image-level similarity.

To be specifically, we assume that the probabilities of latent topics of images should be similar if they share common annotations or similar visual features, *i.e.*, p_{ki} for the i th image and p_{kj} for the j th image with $k \in \{1, \dots, K\}$ should have similar values if

$$|\{\mathbf{W}_i\} \cap \{\mathbf{W}_j\}| \geq T_w \text{ or } |\{\mathbf{F}_i\} \cap \{\mathbf{F}_j\}| \geq T_f \quad (3)$$

here $\{\cdot\}$ denotes the set of feature elements or word elements. $|\cdot|$ is the cardinality of the set. T_w and T_f are threshold numbers. When condition in Eq.(3) is satisfied, we create an edge $\langle i, j \rangle$. All these edges together with all the images as nodes form a similarity graph denoted by \mathbb{S} , which is illustrated in Fig. 1(b).

This assumption is usually referred to as *manifold assumption* [2], in which we assume that the probabilities are reside on or close to a manifold and vary smoothly on the manifold [6],[7],[4]. We aim to fit the probabilistic model with respect to this manifold structure. However, in reality the actual manifold structure is unknown. Thus we constructed a nearest neighbor graph \mathbb{S} to model the geometric structure. We can define the adjacency matrix S on the graph according to the similarity of the visual features and the annotations of the images. There are commonly three ways to define a non-zero element of S with an edge $\langle i, j \rangle \in \mathbb{S}$ [6]:

1) **0-1 Weighting:**

$$S_{ij} = 1 \quad (4)$$

2) **Heat Kernel Weighting:**

$$S_{ij} = e^{-\|\mathbf{F}_i - \mathbf{F}_j\|^2 / \sigma_f - \|\mathbf{W}_i - \mathbf{W}_j\|^2 / \sigma_w} \quad (5)$$

where $\|\cdot\|$ is the norm operator, σ_f and σ_w are the heat kernel parameters.

3) **Dot-Product Weighting:**

$$S_{ij} = \mu_f \mathbf{F}_i^T \mathbf{F}_j + \mu_w \mathbf{W}_i^T \mathbf{W}_j \quad (6)$$

where μ_f and μ_w are relative weights of different modalities.

We define D as a diagonal matrix where $D_{ii} = \sum_j S_{ij}$ and $L = D - S$. L is called graph Laplacian. Also define vector \mathbf{p}_k as $[p_{k1}, \dots, p_{kI}]^T$. We introduce the Laplacian

regularization term as:

$$\mathbb{R} = \frac{1}{2} \sum_{k=1}^K \sum_{(i,j) \in \mathcal{S}} S_{ij} (p_{ki} - p_{kj})^2 = \sum_{k=1}^K \mathbf{p}_k^T L \mathbf{p}_k \quad (7)$$

The regularizer with our choice of S_{ij} incurs a heavy penalty if images associated by S_{ij} have very different probability distributions over latent topics. Therefore, minimizing \mathbb{R} is an attempt to ensure that the probabilities of latent topics of the images are similar if they share common annotations or similar visual features. By minimizing \mathbb{R} , we get the probability p_{ki} 's that are sufficiently smooth on the intrinsic manifold.

We incorporate the Laplacian regularizer into the standard objective function in Eq.(1) and get a newly defined objective as:

$$\mathbb{Q} = \mathbb{F} - \lambda \mathbb{R} \quad (8)$$

where λ is the regularization parameter, $\lambda \geq 0$.

It would be important to note that our method is closely related to the NetPLSA algorithm [7, 4]. However, unlike NetPLSA which regularizes the expected complete likelihood only in M step, we add the regularization term to form a globally consistent objective function.

3. PARAMETER ESTIMATION BY REGULARIZED EM ALGORITHM

The new objective function in Eq.(8) can be carried out in an EM like procedure. We denote the parameters of the model as Θ , such that $\Theta = \{\alpha, \beta, \phi\}$ and \mathbb{Q} is a function of Θ and $\{\mathbf{p}_k\}$. We denote the current parameters as $\Theta^{(t)}$, the parameters obtained in the previous EM iteration as $\Theta^{(t-1)}$. It is easy to see that M step remains unchanged since \mathbb{R} does not depend on Θ . However, E step is a bit difficult to solve. We adopt the optimization scheme discussed in [7, 4]. To simplify the optimization, our algorithm is also divided into ‘‘E step’’ and ‘‘M step’’, but the meanings are different.

E Step:

While it is possible to solve the new estimation of p_{ki} 's in the E step using Newton-Raphson algorithm, the updating formula is too complex to be efficiently computed, we optimize the two terms in Eq.(8) separately in the hope of finding a solution which is better than the last iteration. For the \mathbb{F} part, the updating formula is the same as standard EM [10],

$$p_{ki} \propto P(z_k | \alpha) \prod_{u=1}^U P(f_u | z_k, \beta)^{F_{iu}} \prod_{v=1}^V P(w_v | z_k, \phi)^{W_{iv}} \quad (9)$$

The symbol of proportionate ‘‘ \propto ’’ in the above equation means a normalization step over all k 's for each i after computing Eq.(9), so that the results are valid probability distributions.

After carrying out this standard E step, we then use these p_{ki} 's as the initial values and continue to use the Newton-Raphson method to update the p_{ki} 's until \mathbb{R} converged, in the hope that the evaluation of the objective function after M step will give better results than the last iteration.

By taking the first and second derivatives of \mathbb{R} with respect to every p_{ki} , we get the following updating scheme:

$$p_{ki}^{(t)} = (1 - \gamma) p_{ki}^{(t-1)} + \gamma \frac{\sum_j S_{ij} \cdot p_{kj}^{(t-1)}}{\sum_j S_{ij}} \quad (10)$$

Algorithm 1 Laplacian Regularized EM

Input:

Visual features of all images $\mathbf{F}_i, i \in \mathbf{I}$.
Words of annotated images $\mathbf{W}_i, i \in \mathbf{I}_a$.

Output:

Parameters $\Theta = \{\alpha, \beta, \phi\}$,
Probabilities of latent topics p_{ki} 's for $i \in \mathbf{I}$.

Set Values of Constants $\varepsilon, \lambda, \gamma, T_w, T_f, \sigma_w, \sigma_f, \mu_w, \mu_f$.

Initialize $\Theta^{(0)}$ and $\{\mathbf{p}_k\}^{(0)}$ randomly.

Construct S using any of Eq.(4)(5)(6).

$t \leftarrow 0$

repeat

$t \leftarrow t + 1$.

E Step:

Compute probabilities $\{\mathbf{p}_k\}^{(t)}$ using Eq.(9).

Smooth $\{\mathbf{p}_k\}^{(t)}$ using Eq.(10) until \mathbb{R} converged.

M Step:

Compute new estimation of parameters $\Theta^{(t)}$ by Eq.(11).

Evaluate $\mathbb{Q}(\Theta^{(t)}, \{\mathbf{p}_k\}^{(t)})$ using Eq.(8)(1)(7).

until $\mathbb{Q}(\Theta^{(t)}, \{\mathbf{p}_k\}^{(t)}) - \mathbb{Q}(\Theta^{(t-1)}, \{\mathbf{p}_k\}^{(t-1)}) \leq \varepsilon$

where $0 \leq \gamma \leq 1$ is the step parameter. It is easy to show that the graph Laplacian is positive semi-definite, so each updating step will decrease \mathbb{R} . Moreover, by properly choosing the threshold parameters T_f and T_w , the weight matrix S is highly sparse, so the updating of p_{ki} 's are very efficient.

M Step:

M step is the same as in standard EM algorithm,

$$\begin{aligned} \alpha_k &\propto \sum_{i=1}^I p_{ki}, \quad k \in \{1, \dots, K\}; \\ \beta_{ku} &\propto \sum_{i=1}^I F_{iu} \cdot p_{ki}, \quad (k, u) \in \{1, \dots, K\} \times \{1, \dots, U\}; \\ \phi_{kv} &\propto \sum_{i=1}^I W_{iv} \cdot p_{ki}, \quad (k, v) \in \{1, \dots, K\} \times \{1, \dots, V\}; \end{aligned} \quad (11)$$

Normalization steps are required here over k 's, u 's and v 's for the three equations respectively.

Annotation:

The annotation of unlabeled images is done by computing the marginal probability of words given the probability of latent topics computed at the training stage,

$$P(w_v) = \sum_{k=1}^K p_{ki} \cdot P(w_v | z_k, \phi), \quad i \in \mathbf{I}_u \quad (12)$$

The most probable textual annotations are obtained by choosing the words with highest probabilities.

4. EXPERIMENTS

We use the Corel image dataset from Barnard et. al. [1] to evaluate the performance of our algorithm. This dataset uses a subset of 80 Corel CDs and is comprised of roughly 14000 images. For each image, there are no more than 10 blob features and 1 to 5 annotations. These images are divided into 10 overlapping subsets and each set is further

divided into a big set (75%, about 5200 images) and a small set (25%, about 1800 images). The vocabulary size of each subset is around 150.

There are different measures to evaluate the annotation performance of an algorithm, such as hit rate [11], complete length [1], accuracy [9] and normalized score [1]. Among them, complete length and accuracy are two measures that are complementary and representative, we use them as our performance measurements.

- **Complete Length:** the average minimum length of top returned annotations that contains the ground truth.
- **Accuracy:** the average value of R/M , where M is both the number of predications and the number of ground truth annotations, R is the number of correctly predicated annotations.

In our experiments, we use the small set of each subset as labeled images and try to annotate the large set of each subset. We cluster all visual features of the training images into 600 clusters ($U = 600$) and the number of latent topics K varies from 10 to 100. The regularization parameter λ is tuned to 2.11 and the Newton step parameter γ is set to 0.1. T_w is set to infinite since annotated images are very few. T_f is set to 3 to ensure sparsity. We compared our results with pLSA-Words introduced in [9], which is the state-of-the-art result for using pLSA to do image annotation. From Fig. 2, we see that our method out performs pLSA-Words overwhelmingly. For both methods, the annotation performance gets better with increasing number of latent topics. With more detailed inspection on the annotation results we can see that pLSA-Words tend to prefer annotations that are empirically dominating, while our method achieves rather perfect results even in many tough images, as shown in Fig. 3.

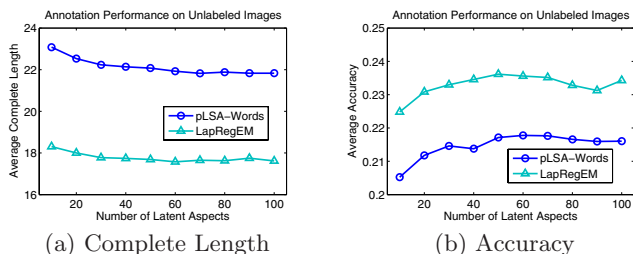


Figure 2: Annotation performance.

Acknowledgment

This work was supported by National Key Basic Research Foundation of China under Grant 2009CB320801 and National Natural Science Foundation of China under Grant 60875044.

5. REFERENCES

- [1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
- [2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

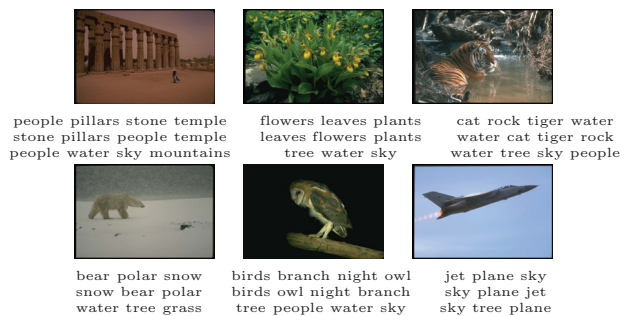


Figure 3: Annotation examples with top returned words. First line: ground truth annotations; Second line: our Laplacian regularized image annotation model; Third line: pLSA-Words model.

- [3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (ACM SIGIR)*, pages 127–134, 2003.
- [4] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *Proc. ACM Conf. on Information and knowledge management (CIKM'08)*, pages 911–920, 2008.
- [5] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
- [6] F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, pages 524–531, 2005.
- [7] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *Proc. ACM Int. Conf. on World Wide Web (WWW'08)*, pages 101–110, 2008.
- [8] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia (SIGMM'03)*, pages 275–278, 2003.
- [9] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. In *Proc. ACM Int. Conf. on Multimedia (SIGMM'04)*, pages 348–351, 2004.
- [10] R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. 1999.
- [11] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV'05)*, pages 846–851, 2005.
- [12] X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. Int. Conf. Machine Learning (ICML'05)*, 2005.