Neurocomputing 119 (2013) 26–32

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Regularized Semi-Supervised Latent Dirichlet Allocation for visual concept learning

Liansheng Zhuang [a,*], Haoyuan Gao [a], Jiebo Luo [b], Zhouchen Lin [c]

[a] University of Science and Technology of China, Hefei 230027, PR China
[b] Department of Computer Science, University of Rochester, NY 14627, USA
[c] Key Laboratory of Machine Perception (MOE), Peking University, Beijing 100871, PR China

ARTICLE INFO

Available online 11 January 2013

Keywords:
Visual concept learning
Semi-supervised learning
Latent Dirichlet Allocation
Low rank graph

ABSTRACT

Topic model is a popular tool for visual concept learning. Most topic models are either unsupervised or fully supervised. In this paper, to take advantage of both limited labeled training images and rich unlabeled images, we propose a novel regularized Semi-Supervised Latent Dirichlet Allocation (r-SSLDA) for learning visual concept classifiers. Instead of introducing a new complex topic model, we attempt to find an efficient way to learn topic models in a semi-supervised way. Our r-SSLDA considers both semi-supervised properties and supervised topic model simultaneously in a regularization framework. Furthermore, to improve the performance of r-SSLDA, we introduce the low rank graph to the framework. Experiments on Caltech 101 and Caltech 256 have shown that r-SSLDA outperforms both unsupervised LDA and achieves competitive performance against fully supervised LDA with much fewer labeled images.

1. Introduction

Visual concept detection is a key problem in image retrieval. It aims at automatically mapping images into predefined semantic concepts (such as indoor, sunset, airplane, and face), so as to bridge the so-called semantic gap between low-level visual features and high-level semantic content of images. Although there have been many studies over the last decades [1–3], it is still a challenging problem within multimedia and computer vision communities. Recently, topic models have been introduced to solve this problem, and achieve impressive results [4–9]. In these applications, each image is treated as a document, and represented by a histogram of visual words. A visual word is equivalent to a text word, and often generated by clustering various local descriptors such as SIFT. Topic models cluster co-occurring visual words into topics, which are used to image classification.

Among current topic models, Latent Dirichlet Allocation (LDA) [10] is one of the most popular ones. Classic LDA is an unsupervised model without using any prior label information. The lack of useful supervised information usually leads to slow convergence and unsatisfactory performance. Moreover, only the visual words in the training images are modeled in classic LDA. During classification, class labels are simply treated as features extracted from the topic distribution [5]. Since class label is not part of the model, classic LDA is not well suited for classification problems, thus resulting in not so robust performance in visual concept detection.

To make LDA more effective for classification and prediction problem, Blei et al. introduced a supervised Latent Dirichlet Allocation (sLDA) model [11,7]. In the sLDA model, label parameter is a domain structure and topics are trained to best fit the corresponding variables or labels. Both visual words and class labels are modeled at the same time. Similarly, Wang et al. [6] proposed a Semi-Latent Dirichlet Allocation for human action recognition. Different from sLDA, Semi-LDA introduces supervised information into its model by associating image class labels with visual words. That is, Semi-LDA assumes that the topic of a visual word is observable and equal to the image class label. Fig. 1 shows the difference between classic LDA, sLDA and Semi-LDA. By modeling the class label, both sLDA and Semi-LDA outperforms classic LDA significantly for classification problems. Beside sLDA and Semi-LDA, Pang et al. [12] also proposed a supervised topic model called Travelogue Model, which can extract both local and global topics with each local topic corresponding to some semantics that characterize a few specific locations.

However, all these models (sLDA, Semi-LDA and Travelogue Model) improve the model performance in a fully supervised fashion, and therefore require all training images to be labeled. For a large dataset, any label information is labor intensive and expensive, making fully supervised topic models greatly restricted to only a few concepts. On the other hand, huge amounts of unlabeled images are available in the Internet and easy to obtain. These unlabeled images contain enough information to train visual

* Corresponding author.
E-mail address: lszhuang@ustc.edu.cn (L. Zhuang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.04.043

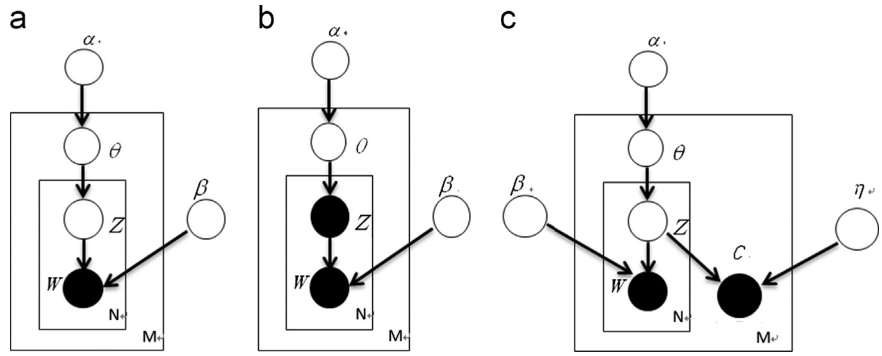

Fig. 1. Graph model representation of classic LDA (a), Semi-LDA (b) and full supervised LDA (c).

concept classifiers, and can help avoid overfitting. Therefore, learning visual concepts classifiers with a fully supervised topic model in a semi-supervised manner, which aims to utilize a large amount of unlabeled images, is a promising direction to explore.

Although much work on semi-supervised learning (SSL) algorithms has been developed, few considered combining semi-supervised properties with topic models to solve the visual concept learning problem. In [8], Zhuang et al. proposed a method called Semi-supervised pLSA (Ss-pLSA) for image classification. By introducing category label information into the EM algorithm during training, they can train classifiers with pLSA in a semi-supervised fashion. Although supervised information effectively speeds up the convergence to achieve desire results, Ss-pLSA does not encode class labels into its model, and seems to be a loosely coupled way of simple label propagation in conjunction with a unsupervised pLSA model. Different from [8], [13–15] carried out semi-supervised topic models in a more consistent fashion by incorporating the manifold assumption into the topic model. They assumed that the probabilities of latent topics of images resided on or close to a manifold, and incorporated the manifold structure into the standard EM algorithm as a regularization term. Since the underlying manifold was unknown, they simply used a nearest neighbor graph to approximate it. However, a nearest neighbor graph is mainly based on pairwise Euclidean distances, and thus is very sensitive to data noise. Since only taking local pairwise relationship into account, a nearest neighbor graph cannot well capture the global geometric structure of the manifold, thus having poor performance. Moreover, all these methods use only class label information to help model learning, while not modeling the class label in their models. As the above analysis, this will decrease the performance of visual concept classifiers.

In this paper, we propose a novel semi-supervised topic model called regularized Semi-Supervised Latent Dirichlet Allocation (r-SSLDA) for visual concept learning. Inspired by Wang et al. [16], instead of attempting to introduce a new Bayesian statistical model, we try to find a simple and an efficient semi-supervised way to learn visual concept classifier with topic models. Unlike the loosely coupled solution in [8], we consider both semi-supervised properties and topic models simultaneously in a regularization framework. By minimizing the cost function of the regularization framework, we provide a direct solution to the semi-supervised topic model problem. Different from current semi-supervised topic models [8,13–15], our r-SSLDA encodes class labels into its framework by adopting a supervised LDA model to learn the visual concept classifiers. Meanwhile, instead of using a nearest neighbor graph, r-SSLDA uses the low rank graph (LR-graph) [17] to approximate the manifold. Compared with existing popular graphs (k NN-graph [18], ℓ_1-graph [19,20], LLE-graph [21,22]), LR-graph uses both the global property and local property of the graph, and thus is better at capturing the global structure of all data. Experimental results showed that r-SSLDA significantly outperformed classic unsupervised LDA and achieved competitive performance compared with fully supervised LDA with fewer labeled images.

The rest of this paper is organized as follows: In Section 2, we give the detail of low rank graph construction. Then, we introduce the regularized Semi-supervised LDA framework in Section 3. Experiments and result analysis follow in Session 4. Section 5 is our conclusions.

2. Low rank graph construction

Let $X=[x_1,x_2,\ldots,x_n]\in\mathbb{R}^{d\times n}$ be a set of data points drawn from a manifold. Each column of X is a data point in $\mathbb{R}^d$. Since the manifold is unknown, we construct a graph from these data points to approximate it. Let $\mathcal{G}=(V,E)$ be a graph, where $V=\{v_1,\ldots,v_n\}$ is the set of graph vertices (node v_i corresponds to data point x_i), and E is the set of graph edges and associated with a weight matrix $W\in\mathbb{R}^{n\times n}$. For any two neighboring nodes v_i and v_j, $W_{ij}>0$ if they are connected with an edge $E_{ij}\in E$, otherwise $W_{ij}=0$. Fixing the nodes set V, the goal of graph construction is to learn the edge weights matrix W.

To construct a low rank graph, we assume that (1) Data points are drawn from a union of low rank and independent subspaces,[1] and each data point can be represented as a linear combination of few other points and (2) A fraction of the data vectors are corrupted by noise or contaminated by outliers, or to be more precise, the data contains *sparse* and *properly* bounded errors. These assumptions are the same to [23]. The independence assumption is mild, because this is usually true especially when the subspaces are low-rank. For clean data, we have

$$\min_Z \quad rank(Z)+\beta\|Z\|_0, \quad \text{s.t.} \quad X=XZ, \quad Diag(Z)=0, \tag{1}$$

where $Z=[z_1,z_2,\ldots,z_n]$ is the coefficient matrix with each z_i being the *reconstruction coefficient* of point x_i. $\beta>0$ is a parameter to trade off between low rank and sparsity.

Problem (1) is difficult to solve due to the discrete nature of the rank function and the ℓ_0 norm. Fortunately, as suggested by matrix completion methods [24–26], the following convex optimization can provide a good surrogate for problem (1):

$$\min_Z \quad \|Z\|_*+\beta\|Z\|_1, \text{ s.t.} \quad X=XZ, \quad Diag(Z)=0. \tag{2}$$

here $\|\bullet\|_*$ denotes the nuclear norm [27] of a matrix and $\|\bullet\|_1$ is the ℓ_1-norm of a matrix. In real applications, observations are often noisy, or even grossly corrupted, and may be missing. For small Gaussian noise, a reasonable strategy is simply to relax the equality constraint

[1] The subspaces are independent if and only if $\sum_{i=1}^k \mathcal{S}_i=\oplus_{i=1}^k \mathcal{S}_i$, where $\oplus$ is the direct sum.

in problem 2, similar to [28]. If a fraction of the data vectors are grossly corrupted, a more reasonable optimization model is

$$\min_{Z,E} \quad \|Z\|_* + \beta\|Z\|_1 + \lambda\|E\|_{2,1}, \text{ s.t. } \quad X = XZ + E, \quad Diag(Z) = 0, \tag{3}$$

where $\|E\|_{2,1} = \sum_{j=1}^{n}\sqrt{\sum_{i=1}^{n}([E]_{ij})^2}$ is called the $\ell_{2,1}$-norm, which encourages the columns of E to be zero. The underlying assumption here is that the corruptions are "sample-specific", i.e., some data vectors are corrupted and the others are clean. The parameter $\lambda > 0$ is used to balance the effects of the three terms.

To solve problem (3), we first convert it to the following equivalent problem:

$$\min_{Z,E,W} \quad \|Z\|_* + \lambda\|E\|_{2,1} + \beta\|W\|_1,$$
$$\text{s.t.} \quad X = XZ + E, \quad W = Z, \; Diag(W) = 0. \tag{4}$$

Problem (4) can be solved by minimizing the following augmented Lagrange multiplier (ALM) function:

$$\begin{aligned} L(Z,W,E,Y_1,Y_2,\mu) = {} & \|Z\|_* + \lambda\|E\|_{2,1} + \beta\|W\|_1 \\ & + \langle Y_1, X - XZ - E\rangle + \langle Y_2, W - Z\rangle \\ & + \frac{\mu}{2}(\|X - XZ - E\|_F^2 + \|W - Z\|_F^2) \end{aligned} \tag{5}$$

where Y_1 and Y_2 are Lagrange multipliers and μ is the penalty parameter which is positive. If we drop the terms independent of Z, a linearization of L w.r.t. Z at Z_k is

$$\begin{aligned} \tilde{L}(Z,W,E,Y_1,Y_2,\mu) = {} & \|Z\|_* + \mu\left\langle Z - Z_k, X^T\left(-X + XZ_k + E - \frac{Y_1}{\mu}\right)\right. \\ & \left. - W + Z_k - \frac{Y_2}{\mu}\right\rangle + \frac{\mu\eta}{2}\|Z - Z_k\|_F^2 \end{aligned} \tag{6}$$

where $\eta = 1 + \sigma_{\max}^2(A)$ and $\sigma_{\max}(A)$ is the largest singular value of A. We can minimize over function $\tilde{L}$ to update Z, and minimize over function L to update W and E [29,23]. The complete algorithm is outlined in Algorithm 1.

Algorithm 1. Solving problem 4 by inexact ALM.

Input: data matrix X, parameter $\lambda > 0$
Initialize: $Z = W = 0$, $E=0$, $Y_1 = Y_2 = 0$, $\mu = 0.1$, $\rho = 1.1$, $\varepsilon_1 = 10^{-8}$, $\varepsilon_2 = 10^{-1}$

1: **while** not converged **do**

2: Update the variable Z by

$$Z_{k+1} = \underset{Z}{\operatorname{argmin}}\, \tilde{L}(Z, W_k, E_k, Y_{1,k}, Y_{2,k}, \mu_k t)$$
$$= \Theta_{(\eta\mu_k)^{-1}}\left(Z_k + \frac{X^T\left(X - XZ_k - E_k + \frac{Y_{1,k}}{\mu_k}\right) + W_k - Z_k + \frac{Y_{2,k}}{\mu_k}}{\eta}\right)$$

where Θ is the singular value shrinkage operator [30].

3: Update the variable E by

$$E_{k+1} = \underset{E}{\operatorname{argmin}}\, L(Z_{k+1}, W_k, E, Y_{1,k}, Y_{2,k}, \mu_k)$$
$$= \Omega_{\lambda\mu_k^{-1}}\left(X - XZ_{k+1} + \frac{Y_{1,k}}{\mu_k}\right)$$

where Ω is the $l_{2,1}$ minimization operator [23].

4: Update the variable W by

$$W_{k+1} = \underset{Wi,i=0}{\operatorname{argmin}}\, L(Z_{k+1}, W, E_{k+1}, Y_{1,k}, Y_{2,k}, \mu_k)$$
$$= D\left(S_{\lambda\mu_k^{-1}}\left(X - XZ_{k+1} + \frac{Y_{2,k}}{\mu_k}\right)\right)$$

where S is the shrinkage operator and $D(X)$ is an operator that sets the diagonal zeros.

5: Update the multiplier using the newly updated variables:

$$Y_{1,k+1} = Y_{1,k} + \mu_k(X - XZ_{k+1} - E_{k+1})$$
$$Y_{2,k+1} = Y_{2,k} + \mu_k(W_{k+1} - Z_{k+1}).$$

6: Update the parameter μ by

$$\mu_{k+1} = \begin{cases} \rho\mu_k & \text{if } \frac{\mu_k(\|Z_{k+1} - Z_k\|_F + \|E_{k+1} - E_k\|_F + \|W_{k+1} - W_k\|_F)}{\|X\|_F} < \varepsilon_2, \\ \mu_k & \text{otherwise.} \end{cases}$$

7: Check the convergence conditions:

$$\frac{\|X - XZ_{k+1} - E_{k+1}\|_F + \|W_{k+1} - Z_{k+1}\|_F}{\|X\|_F} < \varepsilon_1.$$

8: Update $k : k \leftarrow k+1$.

9:**end while**
Output: an optimal solution (Z^*,E^*).

After solving problem (3), we can obtain the reconstruction coefficient matrix $Z^* = (z_1^*, \ldots, z_n^*)$ of data X. This coefficient matrix naturally reveals the relationship among samples: the reconstruction coefficients z_i^* reflect a closeness relationship between point x_i and the other samples, and the magnitude of the corresponding coefficients naturally weighs the closeness of the relationship. The graph weight matrix $W \in \mathbb{R}^{n\times n}$ is defined as

$$W^* = \frac{|Z| + |Z^*|}{2} \tag{7}$$

In practice, Z^* is often dense due to noise. To make it sparse, we often zeroize those elements with small absolute values in W^*.

3. Framework of regularized semi-supervised LDA

Given an image set $X = \{x_1, \ldots, x_l, x_{l+1}, \ldots, x_n\} \subset R^d$ and a label set $C = \{1, \ldots, c\} \subset R$, the first l images $X^L = \{x_1, \ldots, x_l\}$ are labeled and the others $X^U = \{x_{l+1}, \ldots, x_n\}$ are unlabeled. Let $y = (y_1, y_2, \ldots, y_n)^T$ be the label vector of all images. For labeled image $x_i \in X^L$, y_i is set to one of the elements in C. For unlabeled images $x_i \in X^U$, y_i can be any limited value beyond C. To simplify our discussion, this paper only considers binary classification with $C = \{1, -1\}$. In this case, y_i is set to 1 for positive labeled images, -1 for negative labeled images. For unlabeled images, we set y_i to be 0. The goal of regularized semi-supervised LDA is to learn an efficient binary classifier from X and y. The basic idea behind r-SSLDA is to use labeled images to predict the unlabeled images, and train final classifiers with all training images and their labels.

3.1. Low rank graph based label propagation

In essence, the goal of label propagation is to estimate a function f on a graph. It is based on two basic assumptions: *local consistency assumption* and *manifold assumption*. The former says that nearby points are likely to have the same label, whereas the latter says that points lying in the same manifold are likely to have the same label. Based on these two assumptions, we first build a low rank graph [17] to approximate the underlying manifold, and then propagate existing labels to unlabeled images along the graph.

Let F denote the set of $n \times 1$ vectors. A vector $f \in F$ corresponds to a classification function defined on X. $\forall f \in F$ assigns a real value f_i to each image x_i, where f_i is the i-th element of f. The label of an unlabeled image $x_u \in X^U$ is determined by the sign of f_u. To find the optimal vector f^* to classify X, we design a cost function $Q(f)$ as follows:

$$f^* = \arg\min_f Q(f) = \arg\min_f (Q_{smoothness} + \mu Q_{fitting}^L) \tag{8}$$

The first term $Q_{smoothness}$ is the smoothness cost, meaning that a good classification function should not change too much between nearby sample points. That is, images that are close nearby in the feature space (thus similar) tend to have the same labels. Similar to the standard SSL algorithm [31], we define the smoothness cost function as follows:

$$Q_{smoothness} = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}\right)^2 \tag{9}$$

where W_{ij} represents the similarity between two images x_i and x_j, d_i is the sum of the i-th row of W. In our framework, we obtain W by constructing the low rank graph from observed samples.

The second term $Q^L_{fitting}$ means that a good classification function should not change too much from the initial label assignment. So we define the fitting cost as follows:

$$Q^L_{fitting} = \sum_{x_i \in X^L} (f_i - y_i)^2 \tag{10}$$

where X^L means a set of labeled images. Note that $Q^L_{fitting}$ is only used on the labeled images. For unlabeled images, y_i is indefinite. The regularization parameter μ controls the trade-off between constrains, and is empirically set to 1/9 in our experiments.

Thus, the cost function in our semi-supervised topic model is defined as

$$Q(f) = \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}\left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}}\right)^2 + \mu \sum_{x_i \in X^L} (f_i - y_i)^2 \tag{11}$$

To minimize Eq. (11) with respect to f, we assume that the affinity matrix W is symmetric and irreducible. Let D be a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of W. Therefore, we rewrite the cost function as

$$Q(f) = f^T(I - D^{-1/2}WD^{-1/2})f + \mu(f-y)^T I^L (f-y) \tag{12}$$

where I^L is a diagonal matrix. I_{jj} is set to 1 if $y_j = 1$, otherwise 0.

Differentiating $Q(f)$ with respect to f, we have

$$\left.\frac{dQ}{df}\right|_{f=f^*} = 2 \times [(I - D^{-1/2}WD^{-1/2})f^* + \mu I^L (f^* - y)] = 0 \tag{13}$$

With simple deduction, we obtain

$$f^* = (I - \alpha S - \beta A^L)^{-1} \beta I^L y \tag{14}$$

where $S = D^{-1/2}WD^{-1/2}$, $A^L = I - I^L$, $\alpha = 1/(1+\mu)$ and $\beta = \mu/(1+\mu)$. When the number of data is large, we can replace it with an iteration process

$$f(t+1) = (\alpha S + \beta A^L)f(t) + \beta I^L y \tag{15}$$

When the iterative process converges, we obtain the modified classification score vector f^*. To achieve a good precision, we first use supervised LDA to train an initial classifier from labeled images, and estimate the labels of all unlabeled images. That is, we first provide a good estimation f^0 based on initial labeled images, and then refine it under above regularization framework.

3.2. Supervised Latent Dirichlet Allocation

To improve the performance of image classification, r-SSLDA adopts supervised LDA [11,7] as its learning model, which simultaneously models both visual words and class label. The idea behind this model is that images and class label are related, and we can leverage that relationship by finding a latent space predictive of both. These latent topics will best predict the categories for unlabeled images.

Each image is represented as a bag of visual words $w_{1:N}$. The category c is a discrete class label. We fix the number of topics K and let C denote the number of class labels. The parameters of sLDA are a set of K image topics $\pi_{1:K}$, and a set of C class coefficients $\eta_{1:C}$. Each coefficient η_c is a K-vector of real values. Each "topic" is a distribution over a visual words vocabulary. An image and its class label is given by the following generative process:

1. Draw topic proportions $\theta \sim \mathrm{Dir}(\alpha)$.
2. For each visual word w_n, $n \in \{1,2,\ldots,N\}$:
   (a) Draw topic assignment $z_n | \theta \sim \mathrm{Mult}(\theta)$.
   (b) Draw visual word $w_n | z_n \sim \mathrm{Mult}(\pi_{z_n})$.
3. Draw class label $c | z_{1:N} \sim \mathrm{softmax}(\overline{z}, \eta)$, where $\overline{z} = (1/N)\sum_{n=1}^{N} z_n$ is the empirical topic frequencies and the softmax function provides the following distribution, $p(c|\overline{z},\eta) = \exp(\eta_c^T \overline{z}) / \sum_{l=1}^{C} \exp(\eta_l^T \overline{z})$

Note here that different from [11], the class label variable is assumed drawn from a generalized linear model with input given by the empirical distribution of topics that generated the visual words. In essence, above the sLDA model just simplify the model in [7] by ignoring annotations. So, we can use variational EM algorithm to infer the model, which is similar to [7].

4. Experiments

4.1. Data preparation

The datasets used in this paper were Caltech 101 and Caltech 256, two popular image datasets in the literature of image classification. Compared with Caltech 101, Caltech 256 is more challenging because of containing more complex clutters. In our experiments, only 10 categories were selected, and 200 images were randomly selected from each category, 100 images for training and 100 images for test. Specifically, we chose five categories (leopard, motorbike, watch, airplane and face) from Caltech 101 and five categories (bathtub, billiard, binocular, gorilla and grape) from Caltech 256. We selected these categories only because these categories contain enough images (over 200 images). Sample images are shown as Fig. 2.

From these images, we extracted key points and their SIFT descriptors, and then used k-means algorithm to quantize these SIFT descriptors into visual words [32,33]. In the end, we generated 300 visual words to form our visual codebook, and represented each image by the popular "bag of visual words" model.

4.2. Regularized Semi-Supervised LDA vs. fully supervised LDA

To validate the performance of our r-SSLDA, we conducted image classification experiments on Caltech 101 and Caltech 256,[2] and compared r-SSLDA with classic unsupervised LDA and fully supervised LDA. In our experiments, we converted the multi-class classification problem into a set of binary classification problem, and trained binary classifiers for all categories. For any category, there were totally 200 images to train its binary classifier, 100 from its training images and 100 from the rest categories. For r-SSLDA, only 20% of the training images were randomly selected and labeled. That is, we randomly labeled 40 images out of 200 images training, 20 images from the given category as positive samples and 20 images from the rest categories as negative samples. For fully supervised LDA (sLDA), we considered two cases, *sLDA-40* and *sLDA-200*. In the former

[2] Available athttp://www.vision.ethz.ch/projects/categorization/

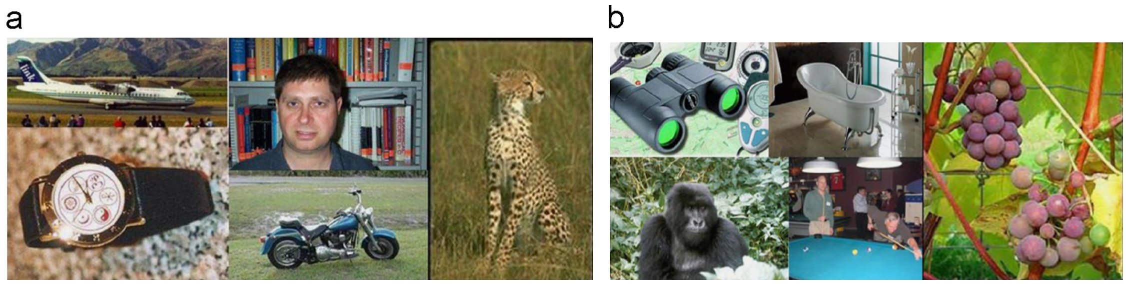

Fig. 2. Sample images in our experiments: (a) images from Caltech 101, including airplane, face, leopard, watch, and motorbike; (b) images from Caltech 256, including bathtub, billiard, binocular, gorilla, and grape.

case (*sLDA-40*), sLDA only used the 40 labeled images to train a binary classifier, which had less training images than r-SSLDA (totally 200 training images). In most real applications, this case often happens because labeled images are hard to obtain. In the latter case (*sLDA-200*), sLDA labeled all the 200 training images, and had the same number of training images to r-SSLDA. This case is often restricted to small amounts of categories, and is very unfair for r-SSLDA. For each classifier, we performed binary classification on 200 test images (100 from the corresponding category and 100 from the other categories). In all experiments, the topic number was set to 30. To keep authority, all experiments ran eight times and averaged all results. The final results are shown in Fig. 3.

From Fig. 3, we can see that

- r-SSLDA significantly outperformed classic LDA for all 10 classes. This suggests that supervised information is important to improve the classifier performance.
- When having identical labeled images, our r-SSLDA also outperformed sLDA (see *sLDA-40*) in all 10 categories. This indicates that unlabeled images provide enough information to boost the classifier performance.
- At last, even compared with *sLDA-200*, our r-SSLDA only incurred little loss on the classification rate while significantly reducing the required labeled data. In practice, labeled images are often very costly to obtain, while unlabeled images are readily available from the Internet. This makes our r-SSLDA more suitable and advantageous for real applications.

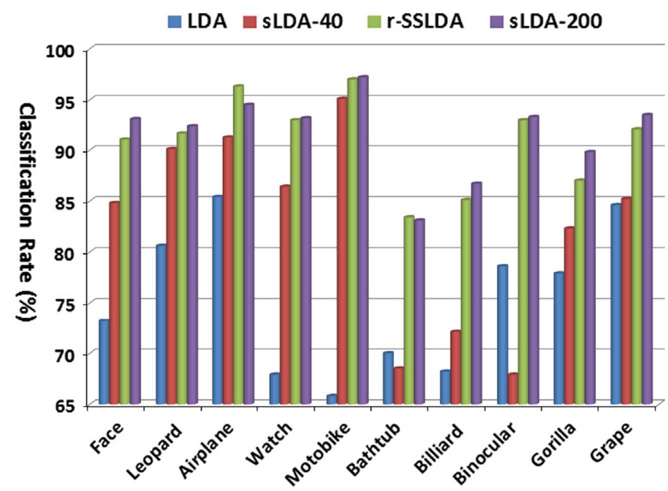

Fig. 3. Recognition results of unsupervised LDA, r-SSLDA and fully supervised LDA on the ten categories. The percentage of labeled images for r-SSLDA is 20%. The topic number was 30 for all the methods.

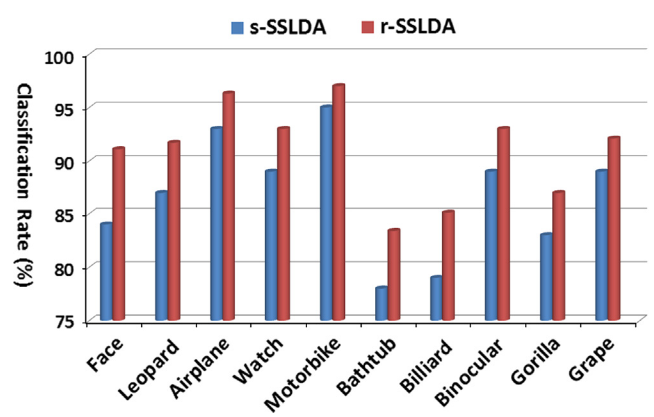

Fig. 4. Performance comparison between r-SSLDA and s-SSLDA across all the ten categories. 20% of the training images were labeled. The topic number was 30.

4.3. Regularized semi-supervised LDA vs. Simple semi-supervised LDA

There are many strategies to learn a topic model in a semi-supervised way. One of the simple strategies is to implement topic models twice. First, we use labeled images to train an initial classifier with sLDA. Then, we use the initial classifier to predict the label of unlabeled training data. After obtaining all the labels for all the training images, we use sLDA to train the visual concept classifier. We call this strategy simple Semi-supervised LDA (s-SSLDA). s-SSLDA is vulnerable to prediction errors because of data noise and model bias. To reduce the prediction errors, our r-SSLDA refines the predictions using a regularization framework that simultaneously considers smoothness and consistency. To validate the efficiency of our regularization framework, we compared r-SSLDA with s-SSLDA in all 10 categories. Fig. 4 show the performance comparison between r-SSLDA and s-SSLDA, when the percentage of labeled training images was 20%. As we can see, our r-SSLDA outperformed s-SSLDA in all cases. This suggests that the regularization framework is more efficient than the simple semi-supervised strategy for combining supervised and unsupervised information.

4.4. Influence of graph construction methods

Label propagation is one of the key components in r-SSLDA. In this paper, we introduced the low rank graph (LR-graph) [17] into our r-SSLDA framework. To verify the advantages of low rank graph, we compared it with other popular graphs (k NN-graph [18], ℓ_1-graph [19,20], LLE-graph [21,22]) under the framework of r-SSLDA. That is, we constructed different graphs to predict unlabeled images, and then trained different binary classifiers for each categories using r-SSLDA. To achieve the best performance, parameters of different graphs were set manually. More specifically, we set the number k of nearest neighborhoods to 3 for the k NN-graph and LLE-graph. For the LR-graph, we set $\lambda = 2$ and $\beta = 0.3$. Other experiment settings were

Table 1
Recognition results of the r-SSLDA framework using different graphs on the ten categories. The percentage of labeled images for r-SSLDA is 20%. The topic number was 30 for all the experiments.

Data set	k NN-graph	ℓ_1-graph	LLE-graph	LR-graph
Face	89.0	86.6	84.0	**91.1**
Leopard	87.5	84.1	88.8	**91.7**
Airplane	95.3	94.8	95.5	**96.3**
Watch	91.2	88.8	92.0	**93.0**
Motobike	96.0	96.2	95.5	**97.0**
Bathtub	81.2	80.8	81.7	**83.4**
Billiard	80.5	75.7	82.5	**85.1**
Binocular	90.7	89.2	91.5	**93.0**
Gorilla	84.4	85.3	83.3	**87.0**
Grape	91.3	91.4	90.0	**92.1**

similar to Section 4.2. The final results are shown in Table 1. As we can see, the LR-graph significantly outperformed other popular graphs for SSL. This suggests that the LR-graph is more informative and discriminative than other graphs for SSL problems. Maybe it is because the LR-graph can capture the global structure of all samples, and is more robust to noises and outliers than other popular graphs.

5. Conclusion

In this work, we developed a novel regularized Semi-Supervised Latent Dirichlet Allocation (r-SSLDA) for visual concept learning. r-SSLDA considered both semi-supervised properties and topic models simultaneously in the regularization framework. Also, we introduced the low rank graph into the framework to improve the performance. Experiments on Caltech 101 and Caltech 256 showed that our r-SSLDA could effectively utilize both labeled images and unlabeled images and achieved competitive performance against fully supervised LDA (sLDA), while drastically reducing the requirement of labeled training images. However, current experiments are only limited to small-scale datasets. Extending r-SSLDA to large-scale datasets is an important direction in the future.

Acknowledgments

We would like to thank Dr. Yi Ma (Microsoft Research Asia) for his helpful conversations about sparse representation and low rank representation. We also thank anonymous reviewers for their constructive comments. This work is partially supported by the National Science Foundation of China (No. 60933013, No. 61103134), the National Science and Technology Major Project (No. 2010ZX03004-003), the Fundamental Research Funds for the Central Universities (WK210023002, WK2101020003), and the Science Foundation for Outstanding Young Talent of Anhui Province (BJ2101020001).

References

[1] J. Tang, S. Yan, R. Hong, G. Qi, T. Chua, Inferring semantic concepts from community-contributed images and noisy tags, in: ACM Multimedia (ACM MM), 2009, pp. 223–232.
[2] J. Tang, H. Li, G. Qi, T. Chua, Image annotation by graph-based inference with integrated multiple/single instance representations, IEEE Trans. Multimedia 12 (2) (2010) 131–141.
[3] J. Tang, X. Hua, M. Wang, Z. Gu, X. Wu, Correlative linear neighborhood propagation for video annotation, IEEE Trans. Syst. Man Cybern.: Part B 39 (2) (2009) 409–416.
[4] Y. Chen, J. Wang, Image categorization by learning and reasoning with regions, J. Mach. Learn. Res. 5 (2004) 913–939.
[5] R. Fergus, F.-F. Li, P. Perona, A. Zisserman, Learning object categories from google's image search, in: IEEE International Conference on Computer Vision, vol. 2, 2005, pp. 1816–1823.
[6] Y. Wang, G. Mori, Human action recognition by semi-latent topic models, IEEE Trans. Pattern Anal. Mach. Intell. (Special Issue on Probabilistic Graphical Models in Computer Vision) 31 (10) (2009) 1762–1774.
[7] C. Wang, D. Blei, F.-F. Li, Simultaneous image classification and annotation, in: Proceedings of IEEE Computer Society Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 1903–1910.
[8] L. Zhuang, L. She, Y. Jiang, K. Tang, N. Yu, Image classification via semi-supervised plsa, in: Proceedings of the Fifth International Conference on Image and Graphics, Xi'an, China, 2009, pp. 205–208.
[9] Y. Pang, X. Lu, Y. Yuan, X. Li, Travelogue enriching and scenic spot overview based on textual and visual topic model, Int. J. Pattern Recognition Artif. Intell. 25 (3) (2011) 373–390.
[10] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[11] D. Blei, J. McAuliffe, Supervised topic models, Adv. Neural Inf. Process. Syst. 21 (2007) 34.
[12] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, L. Zhang, Summarizing tourist destinations by mining user-generated travelogues and photos, Comput. Vis. Image Understanding 115 (3) (2011) 352–363.
[13] Y. Shao, Y. Zhou, X. He, D. Cai, H. Bao, Semi-supervised topic modeling for image annotation, in: Proceedings of the 17th International ACM Conference on Multimedia, Beijing, China, 2009, pp. 533–536.
[14] Q. Mei, D. Cai, D. Zhang, C. Zhai, Topic modeling with network regularization, in: Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 2008, pp. 101–110.
[15] Y. Lu, C. Zhai, Opinion integration through semi-supervised topic modeling, in: Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 2008, pp. 121–130.
[16] C. Wang, L. Zhang, H.-J. Zhang, Graph-based multiple-instance learning for object-based image retrieval, in: Proceedings of the First ACM International Conference on Multimedia Information Retrieval, Vancouver, Canada, 2008, pp. 156–163.
[17] L. Zhuang, H. Gao, J. Huang, N. Yu, Semi-supervised classification via low rank graph, in: The Sixth International Conference on Image and Graphics (ICIG 2011), 2011.
[18] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[19] B. Cheng, J. Yang, S. Yan, Y. Fu, T. Huang, Learning with l1-graph for image analysis, IEEE Trans. Image Process. 19 (4) (2010) 858–866.
[20] S. Yan, H. Wang, Semi-supervised learning by sparse representation, in: SIAM International Conference on Data Mining, SDM, 2009, pp. 792–801.
[21] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323.
[22] J. Wang, F. Wang, C. Zhang, H. Shen, L. Quan, Linear neighborhood propagation and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1600–1615.
[23] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: Proceedings of the 26th International Conference on Machine Learning, Haifa, Israel, Citeseer, 2010.
[24] E. Candès, B. Recht, Exact matrix completion via convex optimization, Found. Comput. Math. 9 (6) (2009) 717–772.
[25] E. Candès, X. Li, Y. Ma, J. Wright, Robust Principal Component Analysis, preprint.
[26] R. Keshavan, A. Montanari, S. Oh, Matrix completion from noisy entries, J. Mach. Learn. Res. 2 (2010) 2057–2078.
[27] M. Fazel, Matrix Rank Minimization with Applications, Ph.D. Thesis, Stanford University, 2002.
[28] E. Candès, Y. Plan, Matrix completion with noise, Proc. IEEE 98 (6) (2010) 925–936.
[29] Z. Lin, M. Chen, L. Wu, Y. Ma, The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices, http://arxiv.org/abs/1009.5055.
[30] J. Cai, E.J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2008) 1956–1982.
[31] D. Zhou, O. Bousquet, N.T. Lal, et al., Learning with local and global consistency, in: Advances in Neural Information Processing Systems, vol. 16, 2004, pp. 321–328.
[32] D. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the International Conference on Computer Vision, vol. 2, Kerkyra, Corfu, Greece, 1999, pp. 1150–11573.
[33] T. Kadir, M. Brady, Saliency, scale and image description, Int. J. Comput. Vis. 45 (2001) 83–105.

Liansheng Zhuang received the B.Sc. degree and Ph.D. degree from University of Science and Technology of China (USTC), in 2001 and 2006, respectively. He is now a Lecturer in the School of Information Science and Techonlogy, USTC. His current research interests include computer vision, image & video retrieval, and machine learning. He is a member of the IEEE and ACM.

Haoyuan Gao received the B.Sc. degree from University of Science and Technology of China (USTC) in 2009. He is currently working toward the Master degree from USTC. His research interests include computer vision, image & video retrieval, and machine learning.

Jiebo Luo received the Ph.D. degree from the University of Rochester in 1995. He is a Professor in CS Department, University of Rochester since Fall 2011. Before that he was a Senior Principal Scientist leading research and advanced development at Kodak Research Laboratories, Rochester, New York. His research spans image processing, computer vision, machine learning, data mining, medical imaging, and ubiquitous computing. He has authored more than 150 technical papers and holds 50 US patents. He has been involved in numerous technical conferences, including serving as the program co-chair of ACM Multimedia 2010 and IEEE CVPR 2012. He is the Editor-in-Chief of the Journal of Multimedia, and has served on the editorial boards of the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, Pattern Recognition, Machine Vision and Applications, and Journal of Electronic Imaging. He is a Fellow of the SPIE, IEEE, and IAPR.

Zhouchen Lin received the Ph.D. degree in Applied Mathematics from Peking University in 2000. He is currently a Full Professor in Peking University. He is also now a Guest Professor to Beijing Jiaotong University, Southeast University and Shanghai Jiaotong University. He is also a Guest Researcher to Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer vision, computer graphics, image processing, pattern recognition, and machine learning.