Correlated PLSA for Image Clustering

Peng Li [1, 2], Jian Cheng [1, 2], Zechao Li [1, 2], and Hanqing Lu [1, 2]

[1] National Laboratory of Pattern Recognition,Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, P. R. China
[2] China-Singapore Institute of Digital Media, Singapore 119613, Singapore
{pli, jcheng, zcli, luhq}@nlpr.ia.ac.cn

Abstract. Probabilistic Latent Semantic Analysis (PLSA) has become a popular topic model for image clustering. However, the traditional PLSA method considers each image (document) independently, which would often be conflict with the real occasion. In this paper, we presents an improved PLSA model, named Correlated Probabilistic Latent Semantic Analysis (C-PLSA). Different from PLSA, the topics of the given image are modeled by the images that are related to it. In our method, each image is represented by bag-of-visual-words. With this representation, we calculate the cosine similarity between each pair of images to capture their correlations. Then we use our C-PLSA model to generate K latent topics and Expectation Maximization (EM) algorithm is utilized for parameter estimation. Based on the latent topics, image clustering is carried out according to the estimated conditional probabilities. Extensive experiments are conducted on the publicly available database. The comparison results show that our approach is superior to the traditional PLSA for image clustering.

Keywords: Correlated PLSA, topic model, image clustering

1 Introduction

Image clustering is the process of grouping similar images together. It is a basic problem in many applications such as image annotation, object recognition, image retrieval. Although it has been studied for many years, it is still a challenging problem in multimedia and computer vision communities.

There are many widely used clustering methods for image clustering, e.g K-means, Gaussian Mixture Model (GMM), etc. However, most clustering methods (eg. K-means) perform clustering intuitively by calculating the distance between the data points and the cluster centers, which often lead to poor clustering results. In recent years, topic models such as Latent Semantic Analysis (LSA) [3] and Probabilistic Latent Semantic Analysis (PLSA) [5] have become popular tools to handle the problem. For these topic models, the images are modeled by some latent topics, which are semantic middle level layers upon the low level features and more discriminative.

The topic models were originally proposed to handle text corpus problems. In [3], LSA used Singular Value Decomposition (SVD) of the word-document matrix to

identify a latent semantic space. However, LSA has a number of deficits due to its unsatisfactory statistical formulation [5]. In order to overcome this problem, Hofmann proposed a generative probabilistic model named Probabilistic Latent Semantic Analysis (PLSA) in [5]. PLSA models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of "topics". Due to the success of the topic models in text analysis, they have been introduced into the field of computer vision and multimedia to solve various problems [1, 8, 9, 10, 11, 12, 15]. PLSA is used for image classification in [1, 8]. Shah-hosseini A. et al. [12] did semantic image retrieval based on the PLSA model. Zhang R.F. et al. [15] used PLSA for hidden concept discovery by segmenting the images into regions. In [9], a latent space is constructed with the PLSA model and image annotation is done based on the latent space. Peng Y.X. et al. [10] constructed an audio vocabulary and proposed an audio PLSA model for semantic concept annotation.

However, there is a problem with the PLSA model. It doesn't consider the image correlations when estimating the parameters, which often leads to inaccurate latent topics. For example, in image clustering task, some similar images that should be in the same cluster always have different topic distributions, which leads to bad clustering results. Actually, there exists much latent semantic correlation among images or image regions. Therefore, it is natural that the correlations between images should be incorporated into the topic model in order to derive more accurate latent topics. Inspired by [4], we propose a Correlated Probabilistic Latent Semantic Analysis (C-PLSA) model in this paper. In our model, we introduce a correlation layer between the images and the latent topics to incorporate the image correlations. We apply the C-PLSA model to image clustering and the experiment results show that our model can get very promising performance.

The rest of this paper is organized as follows: in Section 2, we give a brief review of the PLSA model. Section 3 gives the detail of our C-PLSA model. Experiment results are presented in Section 4. Finally, we conclude our paper in Section 5.

2 The PLSA Model

The PLSA model was originally developed for topic discovery in a text corpus, where each document is represented by its word frequency. The core of PLSA model is to map high dimensional word distribution vector of a document to a lower dimensional topic vector. Therefore, PLSA introduces a latent topic variable $z_k \in \{z_1, \ldots, z_K\}$ between the document $d_i \in \{d_1, \ldots, d_N\}$ and the word $w_j \in \{w_1, \ldots, w_M\}$. Then the PLSA model is given by the following generative scheme:

1. select a document d_i with probability $P(d_i)$,
2. pick a latent topic z_k with probability $P(z_k \mid d_i)$,
3. generate a word w_j with probability $P(w_j \mid z_k)$.

As a result one obtain an observation pair (d_i, w_j) while the latent topic variable z_k is discarded. This generative model can be expressed by the following probabilistic model:

$$P(w_j, d_i) = P(d_i)P(w_j \mid d_i), \tag{1}$$

$$P(w_j \mid d_i) = \sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \cdot \tag{2}$$

The model is graphically in Fig. 1.

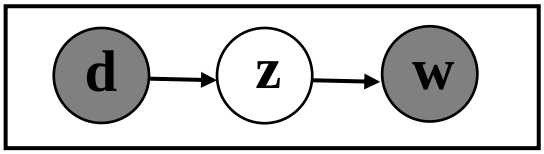

Fig. 1. The PLSA model

We learn the unobservable probability distribution $P(z_k \mid d_i)$ and $P(w_j \mid z_k)$ from the complete dataset using expectation maximization (EM) algorithm [2]. The log-likelihood of the complete dataset is:

$$\begin{aligned} L &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log P(d_i, w_j) \\ &\propto \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \end{aligned} \tag{3}$$

where $n(d_i, w_j)$ is the number of occurrences of word w_j in document d_i. The E-step is given by

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k)P(z_k \mid d_i)}{\sum_{l=1}^{K} P(w_j \mid z_l)P(z_l \mid d_i)}, \tag{4}$$

and M-step is given by

$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)P(z_k \mid d_i, w_j)}, \tag{5}$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{j=1}^{M} n(d_i, w_j)}. \tag{6}$$

Iteratively perform E-step and M-step until the probability values are stable.

3 Our Correlated PLSA Model

3.1 Overview

Although the PLSA model was originally developed for topic discovery in a text corpus, it has been introduced into multimedia field due to its success in recent years, for example, image annotation, object recognition, etc. When applied to images, each image represents a single document and the words can be replaced by visual words, image regions, etc. However, there is a problem with the PLSA model: it doesn't consider the image correlations when estimating the parameters. In order to derive more accurate latent topics, we propose an improved Correlated PLSA (C-PLSA) model to address the correlations between the images in the dataset.

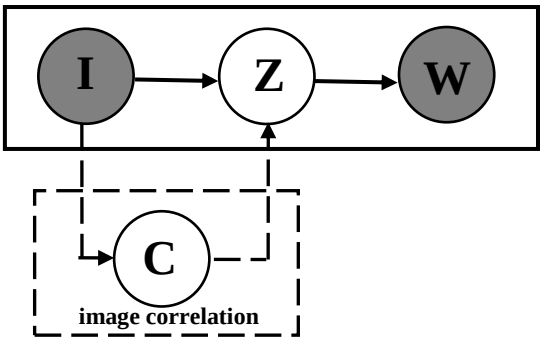

Fig. 2. The C-PLSA model

We depict an overview of our C-PLSA model in Fig.2. Given the image $I_i \in \{I_1,\ldots,I_N\}$, the visual word $W_j \in \{W_1,\ldots,W_M\}$ and the latent topic $Z_k \in \{Z_1,\ldots,Z_K\}$, we adopt the same generative scheme as that of PLSA, which is shown in the solid box in Fig.2:

1. select an image I_i with probability $P(I_i)$,
2. pick a latent topic Z_k with probability $P(Z_k \mid I_i)$,
3. generate a visual word W_j with probability $P(W_j \mid Z_k)$.

In addition, we introduce a new correlation layer in our model such that the topic distributions of the given image can be updated by that of the images similar to it. The image correlations are parameterized by the image correlation matrix C as is shown in the dashed box in Fig.2. As we have incorporated the image correlations into the PLSA model, the related images can have similar topic distributions and we will get more accurate latent topics than the PLSA model.

3.2 Bag-of-visual-words Representation and Image Correlations

When we use the C-PLSA model, the bag-of-words image representation has to be generated first. Here the generation of bag-of-visual-words consists three major steps. First, Difference of Gaussian (DoG) filter is applied on the images to detect a set of key points and scales respectively. Then, we compute the Scale Invariant Feature Transform (SIFT) [7] over the local region defined by the key point and scale. Finally, we perform vector quantization on SIFT region descriptors to construct the visual vocabulary by exploiting the hierarchal k-means clustering methods. Then we can get the word-image matrix (see Fig.3). Each row in the matrix represents an image and $n(I_i, W_j)$ specifies the number of times the visual word W_j occurred in image I_i.

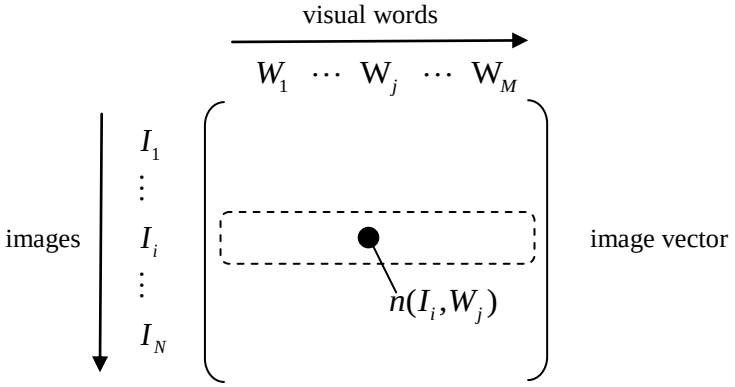

Fig. 3. word-image matrix

With the bag-of-visual-words vector representation introduced above, we compute the image correlation matrix C by cosine similarity. For each pair of images in the dataset, we first compute their cosine similarity as follows:

$$Sim_{ih} = \frac{\vec{I}_i \cdot \vec{I}_h}{|\vec{I}_i| \cdot |\vec{I}_h|}, \tag{7}$$

where $\vec{I}_i$ is the i-th image and represented by the i-th row in the word-image matrix. Then we can get a similarity matrix S where $S_{ih} = Sim_{ih}$. After we get the similarity matrix S, we only keep Q nearest neighbors for each image. In other words, only the top Q values in each row of S are kept and the others are set to zero. At last, we get the image correlation matrix C by normalizing the matrix S such that its row add up to 1.

$$C_{ih} = \frac{S_{ih}}{\sum_{h=1}^{N} S_{ih}} \tag{8}$$

Therefore, each element of C can be considered as the conditional probability $P(I_h \mid I_i)$ and the topic distribution of a given image can be updated by the topic distributions of the images that are related to the given image as follows:

$$P(Z_k \mid I_i) = \sum_{h=1}^{N} P(Z_k \mid I_h) P(I_h \mid I_i) \cdot \tag{9}$$

Since we have introduced a correlation layer between the images and the latent topics in our C-PLSA model, we can derive more accurate topic distributions than the traditional PLSA model.

3.3 Parameter Estimating

Following the maximum likelihood principle, we estimate the parameters $P(Z_k \mid I_i)$ and $P(W_j \mid Z_k)$ by maximizing the log-likelihood function:

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} n(I_i, W_j) \log \sum_{k=1}^{K} P(W_j \mid Z_k) P(Z_k \mid I_i), \tag{10}$$

and EM algorithm can be used to estimate the parameters. In order to incorporate the image correlations, we renew the probability $P(Z_k \mid I_i)$ by equation (9) at each end run of the M-step, thus resulting in a variation of EM algorithm through the following expectation (E-step) and maximization (M-step) solution.

The E-step is given by

$$P(Z_k \mid I_i, W_j) = \frac{P(W_j \mid Z_k) \overline{P(Z_k \mid I_i)}}{\sum_{l=1}^{K} P(W_j \mid Z_l) \overline{P(Z_l \mid I_i)}}, \tag{11}$$

and the M-step is given by

$$P(W_j \mid Z_k) = \frac{\sum_{i=1}^{N} n(I_i, W_j) P(Z_k \mid I_i, W_j)}{\sum_{i=1}^{N} \sum_{j=1}^{M} n(I_i, W_j) P(Z_k \mid I_i, W_j)}, \tag{12}$$

$$P(Z_k \mid I_i) = \frac{\sum_{j=1}^{M} n(I_i, W_j) P(Z_k \mid I_i, W_j)}{\sum_{j=1}^{M} n(I_i, W_j)}, \tag{13}$$

$$\overline{P(Z_k \mid I_i)} = \sum_{h=1}^{N} P(Z_k \mid I_h) P(I_h \mid I_i) \cdot \tag{14}$$

Iteratively perform E-step and M-step until the probability values are stable.

4 Experimental Evaluations

In this section, we evaluate our C-PLSA model by comparing it with the traditional PLSA and K-means on the Caltech-101 Object Categories [13]. Some category names and randomly selected sample images are shown in Fig.4. Each object category contains about 40 to 800 images and a unique label has been assigned to each image to indicate which category it belongs to, which serves as the ground truth in the performance studies. We first compute the word-image matrix by extracting SIFT features as described in Section 3.2. The dimension of the bag-of-visual-words is set to 1000 in the experiment. Then all the clustering methods are performed on the word-image matrix to generate K clusters.

For the topic models, we run the EM algorithm multiple times with random starting points to improve the local maximum of the EM estimates. To make comparison fair, we use the same starting points for PLSA and C-PLSA. The maximum iteration times is set to 150. After representing all the images in terms of latent topic space, each image can be assigned to the most probable latent topic according to the topic distributions $P(Z_k \mid I_i)$. As respect to K-means, we implement the algorithm on the word-image matrix by computing Euclidean distance between image vectors and the randomly initialized cluster centers until the cluster centers are not changed.

The clustering result is evaluated by comparing the obtained cluster label of each image with that provided by the dataset. The accuracy (AC) [14] is used to measure the clustering performance. Given an image I_i, let r_i and s_i be the obtained cluster label and the label provided by the dataset respectively. The AC is defined as follows:

$$AC = \frac{\sum_{i=1}^{N} \delta(s_i, map(r_i))}{N} \tag{15}$$

where N is the total number of images and $\delta(x, y)$ is the delta function that equals to 1 if $x = y$ and zero otherwise, and $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the dataset. The best mapping function can be found by using Kuhn-Munkres algorithm [5].

Fig. 4. Some sample images from the Caltech-101 Object Categories

The evaluations are conducted for different number of clusters K ranging from 2 to 10. At each run of the test, the images from a selected number K of categories are

mixed and provided to the clustering methods. For each given cluster number K, 10 test runs are conducted on different randomly chosen categories, and the final performance scores are obtained by averaging the sores over the 10 test runs.

We first test our C-PLSA model at different numbers of Q and the comparison results are given in Fig.5. As we expected, the accuracy decreases when Q is increasing, because more noise will be introduced into the image correlation matrix. We test different numbers of Q in our experiment and get the best results when Q is around 5.

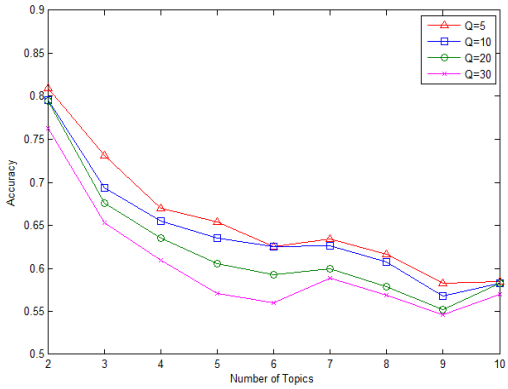

Fig. 5. Accuracy of C-PLSA at different numbers of Q

The accuracy comparisons between C-PLSA (Q=5) and other methods are reported in Fig.6, which shows that our C-PLSA model outperforms PLSA and traditional K-means in terms of accuracy. It is also in line with our expectation: the correlation information do offer help in deriving more accurate latent topics.

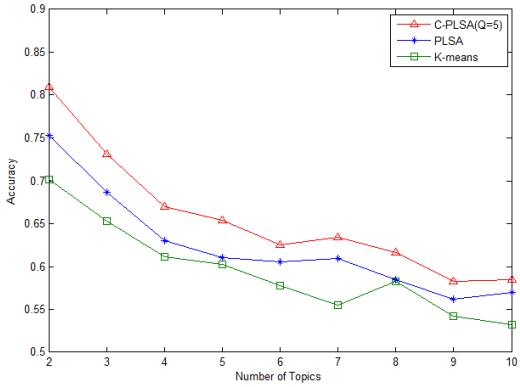

Fig. 6. Accuracy comparison between C-PLSA and other methods

5 Conclusions and Future Work

In this paper, we have presented a novel approach for topic modeling named Correlated Probabilistic Latent Semantic Analysis (C-PLSA). The C-PLSA model introduces a correlation layer between the images and the latent topics, which incorporates the image correlations for topic modeling. As a result, our model can generate more accurate latent topics and have more discriminative power than the traditional PLSA model. The experiment results also show that the image correlations do offer help in the process of topic modeling.

Several questions remain to be investigated in our future work:

1. We consider the image correlations in topic modeling and develop our model based on PLSA. The idea of exploiting image correlations can also be naturally incorporated into other clustering methods, eg., K-means.
2. We compute the image correlations only by bag-of-visual-words features in the paper. More visual features can be combined to get more accurate image correlations.
3. Visual features can't reflect the semantic information of images correctly in many cases. It is very interesting to explore other ways to capture image correlations. For example, web information such as image tags and hyperlink information may be a good way to construct the semantic correlations for web images.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (Grant No. 60975010, 60873185, 60833006).

References

1. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Proc. of the European Conference on Computer Vision (2006)
2. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from in complete data via the EM algorithm. Journal of the Royal Statistical Society, B 39, 1–38 (1977)
3. Deerwester, S., Dumais, G. W., Furnas, S. T., Landauer, T. K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407 (1990)
4. Guo, Z., Zhu, S.H., Chi, Y., Zhang, Z.F., Gong, Y.H.: A latent topic model for linked documents. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 720–721. NY: ACM Press, (2009)
5. Hoffmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1), 177-196 (2001)
6. Lovász, L., Plummer, M.D.: Matching Theory, North-Holland (1986)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal Computer Vision 60(2), 91-110. (2004)

8. Lu, Z.W., Peng, Y.X., Horace, H.S.Ip.: Image categorization via robust pLSA. Pattern Recognition Letters 31(1), 36-43. (2010)
9. Monay, F., Gatica-Perez, D.: PLSA-based image auto-annotation: constraining the latent space. In: Proceedings of ACM international conference on multimedia, pp. 348–351. (2004)
10. Peng, Y.X., Lu, Z.W., Xiao, J.G.: Semantic concept annotation based on audio PLSA model. In: Proceedings of ACM International Conference on Multimedia, pp. 841-844. (2009)
11. Rainer, L., Stefan, R., Eva, H.: Multilayer pLSA for multimodal image retrieval. In: Proceedings of the ACM International Conference on Image and Video Retrieval. (2009)
12. Shah-hosseini, A., Knapp, G.: Semantic image retrieval based on probabilistic latent semantic analysis. In: Proceedings of ACM International Conference on Multimedia, pp. 452-455. (2004)
13. The Caltech-101 Object Categories, http://www.vision.caltech.edu/feifeili/Datasets.htm
14. Xu, W., Liu, X., Gong, Y.H.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267–273. NY: ACM Press, (2003)
15. Zhang, R.F., Zhang, Z.F.: Effect image retrieval based on hidden concept discovery in image database. IEEE Transactions on Image Processing 16(2), 562-572. (2007)