USING LOCAL DISCRIMINANT TOPIC TO IMPROVE GENERATIVE MODEL BASED IMAGE ANNOTATION

Mei Wang[1], *Lan Lin*[2], *Xiangdong Zhou*[1]

[1]Department of Computing and Information Technology, Fudan University, Shanghai, China
[2]Department of Electronic Science and Technology, Tongji University, Shanghai, China

ABSTRACT

Statistical generative model based image annotation propagates the semantic labels of the training images to the unlabeled ones according to their visual generative probabilities. However, it suffers from the problem of "semantic gap", that is, sometimes visual similarity does not reflect semantic similarity. In order to alleviate this problem, we propose a novel image annotation approach which combines the advantages of the generative model and discriminative classification. Based on generative model, we exploit the local discriminants of the visual similar training images (neighborhood) of the unlabeled image. The semantic similar images in the neighborhood are grouped as topics by Singular Value Decomposition (SVD). The discriminative information between different topics is exploited to obtain the semantic relevant topic, which reduces the influence of the images with high visual similarity but irrelevant semantics. Thus, the joint probability of the semantic keyword and the unlabeled image estimated on the obtained relevant topic is more accurate. The experimental results on the ECCV2002 benchmark [1] show that our method outperforms state-of-the-art annotation models MBRM and ASVM-MIL.

Index Terms— automatic image annotation, generative model, discriminant classification

1. INTRODUCTION

Image semantic annotation – associating keywords or captions to the image, is the key step leading to the semantic keyword based image retrieval, which is considered to be convenient and easy for most ordinary users. The early annotation approaches rely on professionals or experts for annotation. It suffers from the problems of labor intensity and subjectivity. With the rapid growth of image archives, both the statistical generative model and the discriminant methods of machine learning have been applied to address the problem of image annotation [1, 2, 3, 4, 5, 6, 7]. However, due to the well known "semantic gap" problem, the performance of image auto-annotation still needs to be improved.

The basic idea shared by most previous work is that the visual features associated with the same keyword are coherent. Therefore, the images or image regions with similar visual features can be grouped together and associated with a certain set of keywords. However, due to the "semantic gap", such unsupervised labeling process is easily influenced by the images with high visual similarities but different semantics [7]. For example, in statistical generative model based image annotation [2], the common semantic keywords shared in the training images with high visual generative probability usually have high ranking scores. However, the high generative probability does not always reflect high semantic similarity. The training images which have high generative probability but irrelevant semantics with the unlabeled image will propagate the false labels to the unlabeled images. We denote such training images as false images in the following discussion. When each keyword is regarded as one semantic classes, discriminative approaches such as SVM have been applied in image annotation by exploiting the supervised discriminant information [6]. However, training SVM directly on the global training set suffers from the problem of timing-consuming and heavy imbalance of positive samples and negative ones [7, 8].

In this paper, we present a novel image annotation method which augments the traditional generative model via local discriminant topics. Instead of training the SVM globally, we use local classification strategy [8, 9] to differentiate the false images from the relevant ones by exploiting the local discriminations. In our method, to label a new image, we choose the "neighborhood" training image set of the new image which consists of the training samples with high visual generative probability. Because each training image will bring multiple semantically correlated labels, directly classifying the neighborhood images will result in a complex multi-class classification problem. Specifically, because an image with multiple annotations should belong to multiple class simultaneously, the overlaps among different semantic classes will heavily impair the classification power of the discriminative method. To deal with this problem, semantic similar images in the "neighborhood" of the labeling target are grouped into topics accord-

This work was partially supported by the Natural Science Foundation of China Grant No.60403018, the Natural Science Foundation of China Grant No.60773077, Fundamental Research 973 Program of China Grant No.2005CB321905.

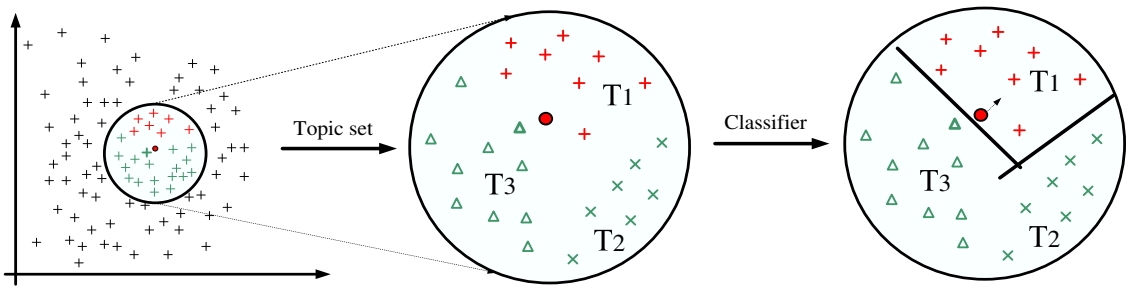

Fig. 1. The basic idea of our method: The left side shows the K-HGPN region bounded by the circle of the unlabeled image (red solid dot). We set $K = 25$ in this example. The relevant images (red cross) and the irrelevant ones (green cross) with the new image are spread in this region. According to their semantic annotation, the images in K-HGPN compose of three local topics as shown in the middle. The semantic relevant training images and the false images are assigned into different topics. The right side illustrates that taking each topic as a class and considering the class distribution, the underlying relevant topic will obtain largest posterior probability. So the new image is classified to the relevant topic.

ing to their semantic labels. The false training images and the relevant ones in the "neighborhood" are grouped into different topics. Regarding each topic as a class, the discriminant information between different topics is exploited to obtain the semantic relevant topic, where the bad influence of the false images is reduced. The joint probability of the semantic keyword and the unlabeled image estimated on the obtained relevant topic is more accurate.

The rest of the paper is organized as follows. Section 2 introduces our motivation. Section 3 presents local discriminant topic based annotation method. We discuss the experiment results in Section 4. Section 5 concludes this paper.

2. MOTIVATION

Generative model based image annotation approach such as Relevance Model [2] has shown significant performance improvements. First consider a standard Relevance Model procedure: For a given labeled image set L, let $|L|$ denote the size of L. Each annotated image J_i in the collection can be described using a set of image regions and annotation words. Given a new image I, the ranking score for a word w to be an annotation keyword for I is calculated as follows:

$$P(w|I) \propto P(w,I) = \sum_{i=1}^{|L|} P(w,I|J_i)P(J_i) \quad (1)$$
$$= \sum_{i=1}^{|L|} P(I|J_i)\,P(w|J_i)\,P(J_i),$$

where $P(J)$ is assumed to be uniformly distributed, $P(I|J)$ is the probability density of image I generated from J. From Eqn.1, we could observe that the labels of I is dominated by the labels contained in the training images which have highest generative probability $P(I|J)$. However, if image J has high generative probability but different semantics with I, the labels of J are easily propagated to I, leading to a false annotation. To avoid this problem, we exploit the supervised learning technique by using classification method. The basic idea of our algorithm is illustrated in Fig. 1.

The left side of Fig. 1 shows a new image I (red solid dot) and its K Highest Generative Probability Neighborhood(HGPN) region bounded by the circle. Due to the "semantic gap", some semantically irrelevant training samples (green cross) are contained in this region. The image with the highest probability of "generating" I, is irrelevant with I. According to Eqn.1, its false labels are easily propagated to I. According to their semantic annotation, in this example, images in K-HGPN are grouped into three semantic topics as shown in the middle figure. The semantic relevant training images and the false images are assigned into different topics. From the right side we observe that, taking each topic as a class, the underlying relevant topic and I follow the same class distribution. So we use classifier to assign image I to the relevant topic, which reduces the risks of false labeling due to the false images.

3. THE PROPOSED METHOD

The main steps of our method is listed at below: Given a new image I,

- Establish the K Highest Generative Probability Neighborhood(HGPN) of I.
- Generate local topics $\{T_1, T_2, \ldots, T_D\}$ based on K-HGPN, each of which consists of their semantically correlated images in K-HGPN.
- Obtain the relevant topic T_r of I by using classification technique.
- Calculate the joint probability based on the relevant topic T_r.

3.1. Generate Local Topic set

For K-HGPN of I, we use SVD to get a set of local topics. Specifically, the images and the keywords appear in K-HGPN are summarized in a co-occurrence matrix A, where each entry in the matrix denotes how often the keyword occurrs in the image. We apply the standard SVD to the co-occurrence matrix A, and then obtain:

$$A = U\Sigma V^T, \tag{2}$$

where a diagonal matrix Σ contains the singular values of A, the corresponding singular vectors form the columns of two orthogonal matrices U and V, sorted by decreasing singular values. Each singular vector corresponds to a coordinate dimension. We choose the first D singular vectors in U and analyze the images in the corresponding coordinate dimension to form the topic set. If the largest singular value is smaller than a predefined threshold e, means the semantics in HGPN are too diverse to construct the topic set. In such situation, we set $D = d = 1, T_d = \{J_j | J_j \in$K-HGPN(I)$\}$; else, the most correlated images in dimension $d \in D$ compose the dth local topic, which is given by:

$$T_d = \{J_j | j \in Top_s(\, Abs(U_{jd})\,)\}, \tag{3}$$

where U_{jd} is the coordinate value of image J_j in dimension d. $Abs(.)$ means the absolute value. $Top_s(.)$ means the largest s elements.

3.2. Obtain Relevant Topic

Denote the local topic set of unlabeled image I as $\{T_1, \ldots, T_D\}$. Assuming each topic as a class, our goal is to find out which topic is relevant with I. Therefore, if $D > 1$, our problem becomes a supervised classification problem. In this paper, SVM is applied to classify the new image to its relevant topic.

We line up all images in all topics together and re-index them as $\{x_1, x_2, \ldots, x_m\}$, where $m = \sum_i |T_i|$. The corresponding labels are denoted by $\{y_1, y_2, \ldots, y_m\}$. If image x_i belongs to topic T_k, we have $y_i = k$. For training data from the ith and the jth classes, we solve the following two-class classification problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C(\sum_t \xi_t) \\ \text{subject to:} \quad & \mathbf{w}^T\phi(x_t) + b \le 1 - \xi_t, \text{if } x_t \text{ in the } i\text{th class}, \\ & \mathbf{w}^T\phi(x_t) + b \ge 1 - \xi_t, \text{if } x_t \text{ in the } j\text{th class}, \\ & \xi_t \ge 0 \end{aligned} \tag{4}$$

The decision function is:

$$f(x) = sign(\mathbf{w}^T\phi(x) + b) \tag{5}$$

if $f(x) > 0$, $y = i$, else $y = j$. ϕ is the mapping function, and $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, where K is the positive definite matrix, namely kernel matrix. In this paper, we use predefined kernel matrix. Assume the generative probability is a kind of distance measure. We calculate the pairwise generative probability of images in all local topics, and then kernelize the distance and transform it in a straightforward way to the kernel for the SVM according to [8].

We train classifiers for each pair of the classes. In classification, the image is designated to be in a class r with maximum number of $y = r$. Then, we obtain the relevant topic of I: T_r. If $D = 1$, we set $T_r = T_D$.

3.3. Annotation

After obtaining the relevant topic $T_r = \{J_{T_{r1}}, J_{T_{r2}}, \ldots, J_{T_{rs}}\}$, we estimate the joint probability $P(w, I)$ based on the local relevant topic T_R:

$$\begin{aligned} P(w, I) &= \sum_{i=1}^{|T_r|} P(w, I | J_{T_{ri}}) P(J_{T_{ri}}) \\ &= \sum_{i=1}^{|T_r|} P(I | J_{T_{ri}})\, P(w | J_{T_{ri}})\, P(J_{T_{ri}}) \end{aligned} \tag{6}$$

where $P(J)$ is assumed to be uniformly distributed, $P(I|J)$ and $P(w|J)$ are the probability of drawing the image I or keyword w from the model of J, being estimated same as [2].

4. EXPERIMENTS

The experiment was conducted on the Corel data set provided in [1], which consists of 5000 images from 50 Corel Stock Photo CDs. Each CD includes 100 images on the same topic, and each image is labeled with 1-5 annotation words. Similar to the previous studies for image annotation, we use recall, precision and F_1 to measure the quality of the algorithm. Given query word w, if there are $|W_G|$ human annotated images with label w in the test set, $|W_M|$ model annotated images with this label, where $|W_C|$ are correct. The recall γ and precision ρ are defined as: $\gamma = \frac{|W_C|}{|W_G|}, \rho = \frac{|W_C|}{|W_M|}, F_1 = \frac{2*\gamma*\rho}{\gamma+\rho}$.

There are four adjustable parameters: K, the number of high generative probability training samples in the neighborhood HGPN of the new image for applying SVD and SVM; D and e, decide the number of local topics generated by SVD; s, the number of the images contained in each local topic. We used 500 images as the validation set to estimate optimal values for these parameters. The experimental results reported in the following section are obtained under the setting: $K = 25, D = 2, e = 4.0, s = 12$. We use LibSVM [10] to implement our classifiers.

We compare our method with MBRM [2] and ASVM-MIL [7], due to their good performance and representations. The comparison result is listed in Table 1. It is clear that the average recall and precision of our method LDT(Local Discriminant Topic) are better than MBRM. The F_1 measure increases 15 percent. In order to further evaluate the effectiveness of our method, we also compare our method with ASVM-MIL [7]. ASVM-MIL poses annotation as the problem of multiple-instance learning and proposes asymmetrical

					
MBRM	Hawaii field tree water grass	clouds sky shore grass water	people sky grass tree water	grass people sky tree water	city mountain tree temple water
1-HGPN	Hawaii field	grass shore water	grass tree water	grass sun rocks fox autumn	city mountain tree temple
LDT	frost snow water grass tree	waves coast water grass storm	cafe building sky people water	deer water forest tree sky	forest tiger cat bengal
Ground Truth	frost ice glass window	waves coast water grass	cafe building water shore	deer water river white-tailed	forest tiger cat bengal

Fig. 2. Comparison of the annotation results of five sample images provided by MBRM and LDT. The third line lists the annotation of 1-HGPN, namely the training images with the highest probability of generating the sample images.

	Avg.Recall	Avg.Precision	F_1 measure
	Results on all 263 keywords		
MBRM	16.1%	19.0%	0.174
LDT	**19.9%**	**20.1%**	**0.20**
	Results on 70 mostly used keywords		
ASVM-MIL	**39.7**%	31.2%	0.349
LDT	39.2%	**35.8**%	**0.374**

Table 1. The effectiveness of our method compared with MBRM and ASVM-MIL

support vector machine to address it. In order to be comparable, we report experimental results based on the 70 frequent keywords, which are the same as the test environment [7] used in ASVM-MIL based on standard Corel data set. The results provided in Table 1 demonstrate that the average recall and F_1 measure of our method LDA is higher than ASVM-MIL significantly.

Finally, Fig. 2 shows five sample images and their annotations provided by MBRM and LDT, we also give the annotations of the highest generative probability training images for the corresponding sample images(1-HGPN). It is obvious that the annotation of MBRM is strongly influenced by the images which have high generative probability. If these images have the irrelevant semantics with the new images, they tend to counteract the contribution of the relevant ones, and easily lead to a wrong annotation. In our method LDA, we can remove the bad influence of such images effectively, so the annotation results are more precise.

5. CONCLUSIONS

In this paper, we proposed a novel image auto-annotation method by using local discriminant topic to remove the bad influence made by the false images. The experimental results indicate that our method is effective. In the future, we will design more effective methods to decide the parameters in our method and extend our algorithm to other probability estimation approaches.

6. REFERENCES

[1] P. Duygulu, K. Barnard, J.F.G. de Freitas, and D.A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *ECCV*, 2002.

[2] S.L. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," *CVPR*, 2004.

[3] R. Shi, T.S. Chua, C.H. lee, and S. Gao, "Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers," *CIVR*, 2006.

[4] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," *CVPR*, 2005.

[5] X.D. Zhou, L. Chan, J.Y. Ye, Q. Zhang, and B.L. Shi, "Automatic image semantic annotation based on image-keyword document model," *CIVR*, 2005.

[6] Y.L. Gao, J.P. Fan, X.Y. Xue, and R. Jain, "Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers," *ACM Multimedia*, 2006.

[7] C.B. Yang and M. Dong, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," *CVPR*, 2006.

[8] H. Zhang, A.C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," *CVPR*, 2006.

[9] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *PAMI*, 1996.

[10] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*.