Neurocomputing xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Automatic image annotation by combining generative and discriminant models

Ping Ji*, Xianhe Gao, Xueyou Hu

Department of Electronics and Electric Engineering, Hefei University, Hefei 230601, China

ARTICLE INFO

Keywords:
Image annotation
Multimedia
Content Analysis
Discriminant model

ABSTRACT

Generative model based image annotation methods have achieved good annotation performance. However, due to the problem of "semantic gap", these methods always suffer from the images with similar visual features but different semantics. It seems promising to separate these images from semantic relevant ones by using discriminant models, since they have shown excellent generalization performance. Motivated to gain the benefits of both generative and discriminative approaches, in this paper, we propose a novel image annotation approach which combine the generative and discriminative models through local discriminant topics in the neighborhood of the unlabeled image. Singular Value Decomposition(SVD) is applied to group the images of the neighborhood into different topics according to their semantic labels. The semantic relevant images and the irrelevant ones are always assigned into different topics. By exploiting the discriminant information between different topics, Support Vector Machine(SVM) is applied to classify the unlabeled image into the relevant topic, from which the more accurate annotation will be obtained by reducing the bad influence of irrelevant images. The experiments on the ECCV 2002 [3] and NUS-WIDE [34] benchmark show that our method outperforms state-of-the-art annotation models.

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 [1–5,8]. However, due to the well known "semantic gap" problem, the performance of image auto-annotation still needs to be improved.

Generative models and discriminative techniques have been widely applied to address the problem of image annotation. By predicting the joint probability of the semantic keywords and visual features to annotate unlabeled images, the generative model based annotation methods have shown significant scalability in database size and number of concepts of interest and provide a natural ranking of keywords for each new image to annotate [8]. In their labeling process, the common semantic keywords shared in the training images with high visual generative probability usually have high ranking scores. So the performance of generative model based method is impaired by the training images having high visual similarities but different semantics [5] with unlabeled images. We denote such training images as false images in the following discussion. Since its excellent generalized performance, it seems promising to use the discriminative techniques such as SVM to differentiate such images from relevant ones. There is much interest in a combination of both generative and discriminative approaches. Some efforts have been done in related fields, such as visual object classification [6,9]. However, different from the classification task, in image annotation, each sample always has multiple correlated labels, which leads to be difficult to apply these methods to image annotation.

By exploiting the correlation between semantic labels, and establish the semantic topics in the "neighborhood" of the unlabeled image, we propose a novel image annotation method which achieves both the advantages of the generative and discriminative approaches. In our method, the discriminant information between local topics is exploited to alleviate the false images problem in generative models. Specifically, to label a new image, our method first generate 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, SVD is applied

* Corresponding author.
E-mail addresses: jiping@hfuu.edu.cn (P. Ji), gaoxh@hfuu.edu.cn (X. Gao), xueyouhu@hfuu.eddu.cn (X. Hu).

http://dx.doi.org/10.1016/j.neucom.2016.09.108
Received 10 March 2016; Received in revised form 12 August 2016; Accepted 17 September 2016
Available online xxxx
0925-2312/

to group the images of the "neighborhood" into different topics. The false training images and the relevant ones in the "neighborhood" are grouped into different topics according to their semantic labels. Regarding each topic as a class, we use SVM to classify the unlabeled image into relevant topic. With a small number of examples and a small number of classes in the neighborhood, SVM achieves better performance. Finally, the joint probability of the unlabeled images and semantic keywords is predicted from relevant topic where the bad influence of false images are reduced.

The rest of the paper is organized as follows. Section 2 introduces the related work in image annotation. Section 3 presents the proposed annotation method. We discuss the results of our experiments in Section 5. Section 6 concludes the paper.

2. Related work

In the past decades, many researches have been done to address the problem of image auto-annotation [1–9,18,19,35]. Specifically, many statistical learning models have been proposed to associate visual features with semantic concepts by using a training set of annotated images. In [2], Duygulu et al. generate the image visual words(blobs) vocabulary by clustering and discretizing the region features. Then, they utilized a machine translation model to collect the links between the words and image visual features, and use these links to annotate new image. Mori et al. used a Co-occurrence Model in which they looked at the co-occurrence of words with image regions created using a regular grid [11]. Latent semantic analysis (LSA) [20] and probabilistic latent semantic analysis (PLSA) [28] introduce latent variables to link image features with keywords. Florent et al. build a linked pair of PLSA models to attach more importance to textual features [28]. Barnard et al. introduced a hierarchical aspect model for image annotation in order to account for the fact that some words are more general than others [30]. Gaussian Mixture Model (GMM), Latent Dirichlet Allocator (LDA), and correspondence LDA, have also applied to the image annotation problem [32]. Jeon et al. introduced a cross-media relevance model(CMRM), in which the region features are represented by discrete blobs [15]. The CMRM modeling was subsequently improved through a continuous relevance model [7] and a multiple Bernoulli relevance model [1].

Since these methods did not explicitly treat semantic labels as image classes, Yang et al. posed above methods as unsupervised learning approaches and point out that their performance are strongly influenced by the quality of unsupervised learning and suffer from the semantic gap [5]. For example, in discrete feature models [2,20], regions with different semantic concepts but similar appearance may be clustered together. In the same way, the performance of continuous relevant model may be damaged by high generative probability training images but different semantics. Cross Media Relevance Models (CMRM) [29], Continuous Relevance Model (CRM) [26], and Multiple Bernoulli Relevance Model (MBRM) [27] assume different, nonparametric density representations of the joint word-image space. In particular, MBRM achieves a robust annotation performance using simple image and text representations: a mixture density model of image appearance that relies on regions extracted from a regular grid, thus avoiding potentially noisy segmentation, and the ability to naturally incorporate complex word annotations using multiple Bernoulli models. However, the complexity of the kernel density representations can be an obstacle for large-scale application. Alternative approaches based on graph representation of joint queries [25], and cross-language LSI [24], offer means for linking the word-image occurrences, but they still do not perform as well as the non-parametric models. Li et al. develop a real-time ALIPR image annotation system that uses multi-resolution 2D Hidden Markov Models to model concepts based on a training set [21]. Gao et al. propose a hierarchical classification approach that explores the tree structure of ontology to accomplish image annotation. Some methods accomplish image annotation by finding the neighbors of each to-be-labeled in a large set of labeled dataset [22]. Wang et al. collect 2.4 million high-quality web images with abundant surrounding texts, and a new image is tagged by mining common phrases from the surrounding texts of its visually similar images [23]. They further extend the search-based annotation method to a 2-billion web image dataset [17]. Torralba et al. collect about 80 million tiny images of 32×32 pixels and confirm that, with simple nearest neighbor methods, the recognition performance improves when the image database expands [16]. Deng et al. on the other hand, structuralize 3.2 million web images and observe improved recognition performance assisted by image ontology [33].

Classification techniques have been applied to image annotation task by imposing strict semantic constraints on the training data by viewing each keyword as an independent class, and creating a different classification model for each keyword [2,4,7]. Model-based methods [7] and SVM-based approaches [2] are both applied in image annotation. Gao et al. introduce a multi-resolution grid-based annotation framework for image content representation and a hierarchical boosting algorithm to address the problem in image annotation using classification technique [30]. Shi et al. proposed a novel Bayesian learning framework of hierarchical multinomial mixture models of concepts for automatic image annotation [4]. Due to the diversity and richness of object classes and image concepts, the concept models may contain hundreds of parameters in high-dimensional feature space and thus large-scale labeled samples are needed for model-based methods [2]. The SVM based methods need to learn the separate hyper plane for each class, bring high computation cost. Furthermore, they often suffer from the imbalanced training images. Namely, the negative samples is much more than positive ones [11]. In [8], Carneiro et al. formulated the image annotation as supervised multi-class problem and learn a distribution model for each class, while their method did not exploit the discriminative information between different classes.

However, the images in those above mentioned approaches, are represented by either local or global features. The performances of annotation are not stable, i.e., for certain concepts, local feature are more suitable while global features are better for others. A unified learning framework for annotation that combines the global and local features is presented in [35]. In [19], they claimed that the multiple instance learning can be cast to a single-instance learning problem and can thus be solved by traditional supervised learning methods. However, the approaches for feature mapping usually overlook the discriminative ability and the noises of the generated features. They proposed an multiple instance learning method with discriminative feature mapping and feature selection. Another work deserved to be mentioned is [18]. The motivation of our work is similar to them, i.e., to annotate images by leveraging on the advantages of both generative and discriminative models. The performance of our proposed method is comparable to theirs while ours is with much few computational costs.

The philosophy of our work is partly similar to that local classification techniques such as discriminant adaptive nearest neighbor classification [10], SVM-KNN [6]. However, these methods are applied to classification, where each sample corresponds to only one class label. Their work are not driven to adapt the image annotation problem, where each sample is related to multiple correlated labels.

3. Motivation

Generative model based image annotation approach such as Relevance Model [1] 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: