Image Annotation and Retrieval Using Dual Classifier Formulation

Pranav Kumar Joshi¹, Riddhi V. Shah²

1. Department of Electronics and Communication, Engineering Parul Institute of Engineering and Technology, Limda, Vadodara
2. Parul Institute of Engineering and Technology, Limda, Vadodara

Abstract: Automatic image annotation is a difficult and highly relevant machine learning task. Recent advances have significantly improved state-of-the-art image retrieval accuracy with algorithms based on nearest neighbor classification in a carefully learned metric space. But this comes at the price of increased computational complexity during training and testing. We propose FastTag, a novel algorithm that achieves comparable results with a simple learning mapping that incorporates a regularized joint convex loss function. The loss function can be efficiently optimized using blocked coordinate descent, which allows us to incorporate a large number of image descriptors cheaply. On several standard real-world benchmark data sets, we demonstrate that FastTag matches the current state-of-the-art in taggimg equality, yet reduces the training and testing time by several orders of magnitude and has lower asymptotic complexity.

Key Words: Image Annotation, Asymptotic complexity, Automatic Image Annotation (AIA), Translation Model (TM), continuous-space relevance model (CRM)

I. INTRODUCTION

With the widespread use of images available from various multimedia devices, effective technologies for organizing, searching and browsing these images are urgently required by common users. Ideally, these images should be indexed by a semantic description so that traditional information retrieval techniques may be adopted for precise image search. However, it is impossible to manually annotate many images, automatic image annotation (AIA) might be a promising solution.

The goal of AIA is to automatically assign relevant keywords to an image that can well describe the content. It Figure 1 illustrates a typical system of automatic image annotation. Given an image collection and a dictionary of keywords, a computer algorithm assigns keywords to each image automatically.

In recent years, many approaches to automatic image annotation have been proposed. Early work by Duygu et al. proposed a translation model (TM) to treat AIA as a process of translation from a set of word tokens, obtained by clustering image regions, to a set of keywords. Jeon et al. proposed the cross-media relevance model (CMRM), which extends the expectation over words in an predefined lexicon is also proposed. In addition, Carneiro et al. proposed a supervised multi-class labeling (SML), which utilizes the principle of minimum probability of error and treats annotation as a multi-class classification problem. As latent aspect models, probabilistic latent semantic analysis (PLSA), latent semantic analysis (LSA) and layered pictorial structures (LPS) have been successfully applied in automatic image annotation. In Feng et al., extended the PLSA model by adding some user information based on the visual words. Subsequently, Monay and Gatica-Perez have proposed the classical PLSA-Words and PLSA-Features models.

II. RELATED WORK

In this section, we review some of the popular methods for automatic image annotation. The first group of methods are based on parametric topic models. Monay and Gatica-Perez (2004) extend the probabilistic latent semantic analysis model, and Barnard et al. (2003) extend the latent Dirichlet allocation model to multimodal data. Each annotated image is modeled as a mixture of topics over visual and textual features. The mixture proportions are shared between feature modes, but the topic distribution is distinct. The second group of methods (Jeon et al., 2003; Lavenko et al., 2003; Feng et al., 2004) models the joint distribution of the image features and the tags with mixture models. The third group of methods trains discriminative models, such as SVM (Cusano et al., 2003), ranking SVM (Grangier & Bengio, 2008) and boosting (Hertz et al., 2004), to predict tags from image features. While these methods achieve promising annotation results, their complex training processes limit the number of descriptors that can be incorporated. Recently proposed models such as the Joint Equal Contribution model of Makadia et al. (2008) and the TagProp model of Guillaumin et al. (2009) rely on local nearest neighborhoods and work surprisingly well despite their simplicity. TagProp is the current state-of-the-art in automatic image annotation.
The method for image annotation is successful can be attributed to three elements:
1. It incorporates a large number of different visual descriptors;
2. It can be trained effectively on images within incomplete tag sets;
3. It treats rare tags specially. Although TagProp achieves superior performance on several benchmark datasets, the $O(n^2)$ training and $O(n)$ test complexity hinders its applicability to large-scale datasets (where $n$ is the number of examples in the training set). In this work, we introduce a new model that incorporates these three elements for successful annotation much more cheaply. Most existing models assume that a complete list of relevant tags for each image is available at training time. However, in practice, this is either impractical or impossible for a large training set. It is much easier to tag an image with a few of the most prominent visual features than to obtain the complete list from a tag dictionary. To alleviate the need for complete labeling, several existing approaches (Fergus et al., 2009, Schrö et al., 2007, Socher & Fei-Fei, 2010) resort to semi-supervised approaches to leverage unlabeled or weakly labeled data from the web. We adopt the same assumption of sparse training tags and incorporate partial supervision in our work.

III. DUO CLASSIFIER FORMULATION
In this section we introduce a new model for automatic image annotation from incomplete user tags. It jointly learns two classifiers on two sources, i.e., image and text, to agree upon the list of tags predicted for each image. It leads to an optimization problem which is jointly convex and has closed form solutions in each iteration of the optimization.

IV. CO REGULARIZED LEARNING
As we are only provided with an incomplete set of tags, we create an additional auxiliary problem and obtain two sub-tasks:
1. Training an image classifier $x_i^T W x_i$ that predicts the complete tag set from image features, and
2. Training a mapping $y_i^T B y_i$ to enrich the existing sparse tag vector $y_i$ by estimating which tags are likely to co-occur with those already in $y_i$. We train both classifiers simultaneously and force their output to agree by minimizing:

$$\frac{1}{n} \sum_{i=1}^{n} \|B y_i - W x_i\|^2.$$

Here, $B y_i$ is the enriched tag set for the $i$-th training image, and each row of $W$ contains the weights of a linear classifier that tries to predict the corresponding (enriched) tag base on image features. The loss function as currently written has a trivial solution at $B=0=W$, suggesting that the current formulation is underconstrained. We next describe additional regularizations on $B$ that guide the solution.

V. EXPERIMENTAL RESULTS
The outcome of the DualClassifier formulation method is annotation of objects in image images.

(a) Original Image
(b) Annotating Image
The quality of the resultant image based on how well they train without losing any properties.

VI. CONCLUSION AND FUTURE WORK

We present an image tagging method, FastTag, that performs on par with current state-of-the-art algorithms, at a fraction of the computation cost. We cast a supervised multi-label classification problem as an unlabeled multi-view learning problem. We define two classifiers, one for each view of the data, and coerce them into agreement via co-regularization in a joint loss function. We trade off complexity in the classifiers with non-linear mapping of the features and demonstrate that such a choice pays off. FastTag is computationally efficient during training and testing yet maintains tagging accuracy. It can effectively deal with sparsely tagged training data and rare tags that are often obstacles in such large-scale learning problems.

In future work, we propose a low complexity and fast algorithm for image annotation using a dual-classifier formulation. It is also applicable for video tagging.

REFERENCES


[6] Lei Wu Member, Rong Jin, Anil K. Jain “Tag Completion for Image Retrieval” Fellow IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, JANUARY 2011