A (very) brief survey of automatic image annotation

David A Torres
Department of Computer Science and Engineering, University of California at San Diego, 92093

Introduction

The problem of image annotation is to automatically label an image with keywords that describe it. Formally, given an image $I$ and a vocabulary of keywords $L$, find a set of key words $\{w\}$ in $L$ that describes $I$.

Automatic annotation comes from the problem of Content Based Image Retrieval (CBIR), which grew out of the need to mine the large collection of image repositories available today. CBIR is the problem of querying a database of images based on their content. One of the most natural ways for humans to interact with such a system is through text descriptions, thus giving importance to the task of associating keywords to images.

Since these associations are a pain to produce manually, researchers have looked at methods of automatically annotating images.

Supervised Methods

Recent attempts at image annotation focused on creating binary classifiers for each keyword in the library. The classifier outputs +1 if an image exhibits the class, and -1 otherwise.

Unsupervised Methods

Recent unsupervised methods have yielded the best results. These methods attempt to model the joint distribution of image features and words by adding a hidden state to the model. One of the simplest of these models [3] takes the form of Figure 3:

The model is built by segmenting an image then extracting feature vectors $x$. Given training data in the form of image-keywords pairs $\{(x_n,w_n),\ldots,(x_m,w_m)\}$, our task is to learn the density $p(x|\omega=\omega)$ with some density estimation technique such as direct estimation, or model averaging [1]. We make a classification using Bayesian decision theory, that is, a key word $\omega$ is annotated to an image $x$ if $P(x|\omega=1)p(\omega=1) \geq P(x|\omega=-1)p(\omega=-1)$

Consider this toy example: We find that training images containing grass contain mostly green color with probability $P(X=\text{Green}|W=\text{Grass})=0.9$, then we annotate the test image with “grass” because $P(\text{Green}|\text{Grass})=P(\text{Green})$ (i.e. $0.9 \geq 0.1$).

The problem with these supervised methods is that they do not scale easily. Estimating the density $P(X|\omega=\omega)$ requires going over all of the negative examples of a class which is usually a large number. Also, if images are not completely labeled, then the missing concepts are incorrectly assigned to the negative example class and so the model learns an incorrect distribution.

We can compute the joint probability of this model by

$P(x,\omega) = \sum_{\omega} P(x|\omega)p(\omega)$

(1)

Where $S$ is all the possible states of $L$. We learn the parameters of (1) using expectation maximization.

Another model, Supervised M-ary Labeling or Mix-Heir [1], eliminates the hidden variable altogether. This model makes the image features directly dependent on the keyword $W$

$P(x,w) = \sum_{\omega} P(x|\omega)p(\omega)P(\omega|w)$

(2)

Fig. 4. To generate a feature, first we pick a keyword $w$ from the variable $W$, then generate the feature according to $p(x|w)$.

Based on this model $P(w|x)$ or $P(x|w)$, By maximizing $w$ for a given $x$ we solve the annotation problem.

The advantages of these methods compensate for the supervised methods’ shortcomings. One of biggest advantages is that the training complexity is much simpler, since it is equivalent to learning only the positive example densities [1]. The complexity of the annotation process in M-ary Labeling is also superior to other methods. The next best unsupervised classifier (MBRM) [3] has annotation complexity $O(T)$, $T$ is the training size. M-ary complexity is $O(C)$, $C$ is the number of classes or keywords.

The future...

The Corel dataset has been established as the set to evaluate performance. It consists of a corpus of images which have been hand-labeled with four relevant keywords. Though the use of this standard is critical to the development of better annotation methods, real life applications will eventually need to deal with the absence of a good keyword library and the problem of how to acquire such a library.

A perfect example can be found in Google image search. How do you select the most descriptive keywords from the body of noisy hypertext surrounding the image on the web. Studies attacking this problem are emerging.

In one paper [4] the authors compute the mutual information $I(W;X)$ between a keyword and an image feature vector. Mutual information measures the mutual dependence between two random variables. It can be thought of intuitively as the information that is shared between two variables. Calculating $I(W;X)$ amounts to calculating the conditional entropy of $X$ given $W$, which again reduces to the problem of estimating a probability density function.

Another approach [5] mines a collection of music reviews for descriptive keywords. The authors create a simple measure of keyword salience and then train a kernalized least-squares classifier to predict the salience of future keywords.

Literature cited