Image and Vision Computing 69 (2018) 33–43

Contents lists available at ScienceDirect

Image and Vision Computing

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

Designing a symmetric classifier for image annotation using multi-layer sparse coding[☆]

Amara Tariq[a,*], Hassan Foroosh[b]

[a] Department of Computer Science, Forman Christian College, Pakistan
[b] The Computational Imaging Lab., University of Central Florida, USA

ARTICLE INFO

Article history:
Received 28 June 2016
Received in revised form 14 July 2017
Accepted 10 November 2017
Available online 20 November 2017

Keywords:
Automatic image annotation
Sparse coding
Symmetric classifier response

ABSTRACT

Automatic annotation of images with descriptive words is a challenging problem with vast applications in the areas of image search and retrieval. This problem can be viewed as a label-assignment problem by a classifier dealing with a very large set of labels, i.e., the vocabulary set. We propose a novel annotation method that employs two layers of sparse coding and performs coarse-to-fine labeling. *Themes* extracted from the training data are treated as coarse labels. Each *theme* is a set of training images that share a common subject in their visual and textual contents. Our system extracts coarse labels for training and test images without requiring any prior knowledge. Vocabulary words are the fine labels to be associated with images. Most of the annotation methods achieve low recall due to the large number of available fine labels, i.e., vocabulary words. These systems also tend to achieve high precision for highly frequent words only. On the other hand, text mining literature discusses a general trend where relatively rare/moderately frequent words are more important for search retrieval process than the extremely frequent words. Our system not only outperforms various previously proposed annotation systems, but also achieves symmetric response in terms of precision and recall. Our system scores and maintains high precision for words with a wide range of frequencies. Such behavior is achieved by intelligently reducing the number of available fine labels or words for each image based on coarse labels assigned to it.

1. Introduction

Automatic annotation of images with accurate textual labels or words is a challenging yet important problem. Such systems have potentially vast applications in the areas of image search and retrieval where search engines have to retrieve appropriate images in response to textual queries of users. The image annotation process can be viewed as a label-assignment problem where a classifier has to choose appropriate labels or words for each image from a large variety of labels, i.e., the set of vocabulary words. Therefore, we use the terms 'word' and 'label' interchangeably in this paper.

Performance of image annotation systems is generally measured in terms of mean precision-per-word and mean recall-per-word scores. Precision(P) and recall(R) scores are dependent on *true positive*(TP), *false positive*(FP) and *false negative*(FN) values.

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN} \tag{1}$$

Let us assume that an annotation system is based on a random classifier. This classifier chooses a set of unique labels $\mathcal{Z}$ ($z = |\mathcal{Z}|$) for a given image, from the set of all available labels $\mathcal{W}$ ($M = |\mathcal{W}|$). In this case, $\mathcal{W}$ is the vocabulary set and $\mathcal{Z} \subset \mathcal{W}$. The number of all possible subsets of size z is given by

$$Z_{all} = \binom{M}{z} = \frac{M!}{z!(M-z)!} \tag{2}$$

If Z_l denotes the number of all possible subsets of size z containing a certain label/word w_l, then

$$Z_l = \binom{M-1}{z-1} = \frac{(M-1)!}{(z-1)!(M-z)!} \tag{3}$$

Let $p(w_l)$ denote the probability of assigning label/word w_l to the image while $\hat{p}(w_l)$ is the probability of not assigning this label/word to the image.

$$p(w_l) = \frac{Z_l}{Z_{all}} = \frac{(M-1)!}{(z-1)!(M-z)!} \times \frac{z!(M-z)!}{M!} \tag{4}$$

[☆] This paper has been recommended for acceptance by Sven Dickinson.
[*] Corresponding author.
E-mail address: amara_tariq@knights.ucf.edu (A.i. Tariq).

https://doi.org/10.1016/j.imavis.2017.11.002
0262-8856/© 2017 Elsevier B.V. All rights reserved.

$$p(w_l) = \frac{z}{M} \tag{5}$$

$$\hat{p}(w_l) = 1 - p(w_l) = \frac{M-z}{M} \tag{6}$$

If X is the fraction of all images that have label/word w_l as their true label, then probabilities of *true positive*, *false positive* and *false negative* are as follows:

$$p(TP) = \frac{zX}{M} \tag{7}$$

$$p(FP) = \frac{z(1-X)}{M} \tag{8}$$

$$p(FN) = \frac{X(M-z)}{M} \tag{9}$$

In general, $z << M$, i.e., only a few labels/words are assigned to each image whereas a large number of words are part of the vocabulary set. Using all these relation in Eq. (1), we gain the following insight into the precision and recall scores.

$$P \propto X, \quad \text{whereas} \quad R \propto \frac{1}{M} \tag{10}$$

For a simple annotation system based on a random classifier,

- Precision for a word is directly proportional to its frequency in the set (X: number of images with which the word is associated). As a result, highly frequent words or labels, i.e., words associated with a large number of images, tend to get better precision scores.
- Recall is adversely affected because of the availability of large number of vocabulary words or labels.

The above analysis explains the behavior of various previously proposed annotation system. Many previously proposed systems achieve much lower recall score than the precision score. The imbalance seems rooted in the large size of the available vocabulary set. Moreover, annotation systems achieve better precision scores for highly frequent words, i.e., words associated with a large number of training images. Research in the field of text search and retrieval concluded that the words that occur with moderate document-frequency, i.e., words occurring in a few documents of the documents' set, are more important in search and retrieval scenario than the words with high document-frequency (words occurring in almost all documents of the set) [1]. If a word appears in almost every document, it has little distinctive power to differentiate one document from the other in reference to a query. In case of image annotation, caption or annotations' set of each image is one document. Hence, X denotes the document-frequency of word w_l. In terms of information theory, words with high document-frequency are the *"expected"* events and words with moderate or low document-frequency are the *"surprising"* events. *Surprising* events have more information content than *expected* events. From this point onwards, we use the word 'frequency' to denote document-frequency in this manuscript.

In this paper, we propose a novel image annotation system that overcomes the above-mentioned shortcomings. Our system performs coarse-to-fine labeling of images. System identifies distinct *themes* present in the training data without requiring any additional prior knowledge and uses these *themes* as coarse labels for test images. Each *theme* is defined by a set of training images that share a common subject in their visual and textual contents. Vocabulary words are the fine labels which are then assigned to the test image in light of the coarse labels or *themes* assigned to it.

Sparse coding is the process of learning a sparse representation of an input signal in terms of coefficients of a set of basis vectors or predictor variables. Structured sparse coding assumes that there is inherent group structure among predictor variables and employs that structure while modeling input signal. Our system employs two layers of sparse coding that use training images as predictor variables. The first layer takes the group structure of the predictor variables into account, in terms of the *themes* available in the training data. This layer assigns coarse labels or *themes* to test images. The second layer deal with only images and labels of the *themes* relevant to the given test image. Previously proposed sparse coding based annotation systems require explicit specification of coarse labels [2,3]. Such labels are not explicitly available with all datasets. Our system increases its scope of application by identifying coarse labels without requiring any prior knowledge.

Mean precision per word (P) and mean recall per word (R) are used as evaluation measures for image annotation system. In general, one of these measures can be improved at the cost of the other. F-score incorporates the trade-off between precision and recall and is defined as harmonic mean of precision and recall.

$$F = \frac{2PR}{P+R} \tag{11}$$

Our system is designed to overcome the tendency of low recall for annotation systems without sacrificing precision. Hence, it outperforms other systems in terms of F-score. Our system also maintains its high precision score for words with wide range of frequencies. Thus our system is practically much more precise for high information content words, i.e., words with low-to-moderate frequency, than other systems. We performed thorough experimentation to prove the qualities of our system.

The rest of the paper is structured as follows. We discuss literature regarding image annotation and sparse coding in Section 2. The problem of image annotation is formally defined in Section 3. Section 4 describes the overall architecture of the proposed system. Results of experimental evaluation are presented in Section 5 followed by the concluding remarks in Section 6.

2. Related work

2.1. Image annotation

Automatic image annotation is the task of predicting individual words as labels for any given image. Image meta-data may be useful in predicting annotations for the image [4] but it may not always be readily available. Therefore, we avoid the use of such meta-data in our system. Relevance model from the domain of machine translation was adapted to treat visual features as a language that needs to be translated into words [5–8]. Relevance model is a generative modeling scheme. Relevance model and support vector machine, generative and discriminative models respectively, are combined in [9]. Performance of all of these systems tends to be highly imbalanced in terms of precision and recall, indicating that the systems are precise for only highly frequent words. In contrast, our system is designed to maintain its performance for a wide range of word-frequencies. Annotation systems like [10–14] identify 'nearest neighbors' of the test image from which to propagate the labels to the test image. Canonical correlation analysis has been used to improve the performance of such nearest-neighbor type systems [15].

The problem of image annotations has often been treated as multi-label problem while multiple labels are simultaneously employed in learning classifiers or dictionaries [16–18]. Our