J. Vis. Commun. Image R. 56 (2018) 275–286

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Can a machine have two systems for recognition, like human beings? [☆]

Jiwei Hu [a], Kin-Man Lam [b], Ping Lou [a], Quan Liu [a,*], Wupeng Deng [a]

[a] *School of Information Engineering, Wuhan University of Technology, China*
[b] *Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, China*

ARTICLE INFO

Article history:
Received 17 October 2017
Revised 27 July 2018
Accepted 11 September 2018
Available online 4 October 2018

Keywords:
Image annotation
Multi-labeling
Hierarchical tree structure
Feature-pool selection

ABSTRACT

Artificial Intelligence has attracted much of researchers' attention in recent years. A question we always ask is: "Can machines replace human beings to some extent?" This paper aims to explore the knowledge learning for an image-annotation framework, which is an easy task for humans but a tough task for machines. This paper's research is based on an assumption that machines have two systems of thinking, each of which handles the labels of images at different abstract levels. Based on this, a new hierarchical model for image annotation is introduced. We explore not only the relationships between the labels and the features used, but also the relationships between labels. More specifically, we divide labels into several hierarchies for efficient and accurate labeling, which are constructed using our Associative Memory Sharing method, proposed in this paper.

1. Introduction

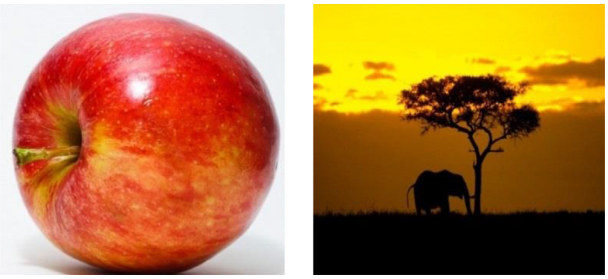
Many computer-vision applications, such as scene analysis and image segmentation, are ill-suited for traditional classification, in which each image can only be associated with a single class or label. However, in the real world, an image is usually associated with multiple labels, which are characterized by different regions of the image. Thus, image classification is naturally considered either as a multi-label learning or a multi-instance learning problem. Most of the recent work in multi-label classification task, such as scene recognition and multi-object recognition [1–3], has focused on the method of tagging a given image with multiple class labels. A serious problem with most of these existing approaches is that they do not exploit the correlations between the class labels.

For multi-label learning, a straightforward method of achieving the goal of correctly classifying the multiple labels of an image is to consider images with the same multiple labels as a new class, and to build a model for this new, multi-label class. However, the problem with this approach is that the samples belonging to the multi-label classes are usually too sparse to build usable models. To solve this problem, the multi-label samples are used more than once during training. Each sample is considered a positive example of each of the label classes it belongs to. This training method is called 'cross-training' [4]. Another approach [5] to multi-label learning is to perform image segmentation first. As an image is divided into a number of non-overlapping regions, and each region may be described by one label, this can roughly determine the maximum numbers of classes it can fit. Image segmentation is the process of dividing an image into different regions such that each region is nearly homogeneous, whereas the union of any two regions is not. It serves as a key task in image analysis and pattern recognition, and is a fundamental step toward low-level vision, which is significant for object recognition, image retrieval and other computer-vision-related applications [6–8]. However, segmentation itself is a difficult, imperfect task. Segmentation always results in the problem of complexity, and unsuccessful segmentation also degrades the performance of the image-annotation task. Nevertheless, a lot of research is still being devoted to achieving a good segmentation performance.

Human beings see image-annotation tasks as an easy problem. The related tags that we assign to an image can be classified into two categories. As shown in Fig. 1, one category includes those basic or obvious tags that we do not need to think about, e.g. apple, sky, dog, etc. The other includes the more complex or abstract tags that we need to think over, e.g. market, African, indoor, etc. The book titled "Thinking, Fast and Slow" [35] surmises that humans have two systems; one is used to solve the problems without requiring thinking, while the other requires some thought. Can a machine have two such systems, like human beings, for the image-annotation tasks? Motivated by this book, we wondered if a machine could have two such systems, like human beings? Therefore, in this paper, we propose a hierarchical framework to mimic the two systems for handling tags, i.e. with solid concepts and abstract concepts, respectively.

[☆] This paper has been recommended for acceptance by Dr Zicheng Liu.
[*] Corresponding author.
E-mail addresses: enkmlam@polyu.edu.hk (K.-M. Lam), quanliu@whut.edu.cn (Q. Liu).

https://doi.org/10.1016/j.jvcir.2018.09.008
1047-3203/© 2018 Elsevier Inc. All rights reserved.



(a) Apple (b) African

Fig. 1. (a) An image with a simple, solid tag, and (b) an image with a confusing, abstract tag.

In order to exploit the correlations between the class labels, we introduce a method called Associative Memory Sharing (AMS), which classifies image labels into different levels of a hierarchy according to their level of abstraction, for the purpose of constructing a tree structure in the learning framework. In other words, the labels or graphs of labels are linked to each other through the tree structure. In pursuit of the ultimate goal of building an intelligent image-annotation system, it is also necessary to incorporate human knowledge into our proposed framework. In the training part, we will use human knowledge interactively to help the system to choose representative images for each label class.

The remainder of this paper is organized as follows. In Section 2, a brief introduction to related works will be given. We present our proposed method in detail in Section 3. The experiment set-up and results, and a conclusion, are given in Sections 4 and 5, respectively.

2. Related work

In this section, we will give a brief overview of the different models for solving multi-label learning problems. We will also discuss feature extraction and image representation, which play an important role in image-annotation frameworks.

2.1. Literature review

In recent years, various learning methods have been proposed for automatic image annotation. These methods have in common that they all rely on a set of labeled pictures to learn models, which can then predict the labels for unlabeled data. The literature can be grouped based on three models: generative models, discriminative models, and nearest-neighbor (NN)-based models. Most generative models [9,10] construct a joint distribution over image contents and the associated keywords while finding a mapping between the image features and the annotation keywords. These generative models aim to learn a single model for all the vocabulary terms, which yields a better modeling in terms of dependencies. Some methods treat the task of image annotation as several binary classification problems. This means that the joint distribution of the unobserved variables and the observed variables is not needed. In this situation, discriminative models [11–13] can generally yield a superior performance. Discriminative models learn a separate classifier for each single label, and use the classifier to judge whether the test image belongs to a particular label or not. Although the training process is complicated and time-consuming, this approach can, with a smart design, achieve more promising performances than the generative models. The third model – one of the oldest, simplest, and most effective methods for pattern classification – is the *k*NN-based model [14], which is accurate, especially with an increasing amount of training data. Recently, a NN-based keyword-transfer approach was proposed in [15]. In this method, the labels are transferred from neighbors to a given image after a simple distance calculation. The nearest neighbors are determined using Joint Equal Contribution (JEC) only, which finds the average distance obtained from the differences in image features. The method was extended in [16] to filter out most of the irrelevant labels, with a promising result obtained.

Although the learning stage plays an important role in an image-classification system, the features employed also affect the performance of the whole framework. In [17], a graph structure was proposed to describe the relationship between the features. In this approach, a pair-wise graph is constructed, with each vertex representing a single image that may be labeled or unlabeled. Two similar images are connected by an edge, and the edge weight is calculated as an image-to-image distance. In [18], a new graph-based model was proposed for recognition based on a semi-supervised framework, which can predict both the predefined labels and undefined labels. The concept of a simple graph was extended in [19] to a hypergraph, whose main argument is that the simple graph cannot completely represent the relations between images. Actually, a hypergraph can contribute to a better representation of the relations between images by considering not only the local grouping information, but also the similarities between the hyperedges that involve more than two images. The idea of a hypergraph was used in [20] to determine a suitable feature space for each class. It is a simple and efficient method for finding a good representative image patch for each label class, which can greatly enhance efficiency in the learning stage.

With the ongoing development of consumer electronics equipment, image databases are becoming larger and larger, with a growing number of labels. In [21], millions of photos have been captured as informative reports, and utilized for computer-vision tasks, such as situation recognition. In their work, a visual analytics system was built to understand the information that could be collected from their photo report streams. To learn about thousands of objects from millions of images, a model with a large learning capacity and considerable efficiency is needed. Deep Convolution Neural Networks (CNNs) [22–24], which have achieved great success on single-label image classification in recent years, constitute this model. Because of their strong capability for learning representative features, CNN models yield breakthrough performance on many other computer-vision tasks, which have attracted attention in the research of image understanding recently. Although many techniques that have been proposed in the last decade can give a reasonable performance, a large number of potential labels causes problems, in terms of decreasing their accuracy and efficiency. More and more researchers are now exploring the relationship between labels; many contributions, which represent a landmark in the research of image annotation. have been made regarding this.

2.2. A structured framework for image understanding

Recent progress on image annotation mainly focused on exploring semantic relations between different labels. Such relations can be modeled by graphical models [25,26] or recurrent neural networks (RNNS) [23]. Despite the great improvements achieved by exploiting semantic relations, existing approaches cannot capture the spatial relations of labels, for the reason that their spatial locations are not annotated for training. To address the problem of a large number of labels required for an image multi-labeling framework, contextual modeling has become a recent focus. For example, in object-class recognition, the presence of one class may suppress (or promote) the presence of another class that is negatively (or positively) correlated, e.g. [27,28]. In [24], the object-detection task is achieved by modeling the object co-occurrences