Computer Vision and Image Understanding 132 (2015) 3–11

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Large margin learning of hierarchical semantic similarity for image classification ☆

Ju Yong Chang, Kyoung Mu Lee *

School of Electrical Engineering, ASRI, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-744, Republic of Korea

ARTICLE INFO

Article history:
Available online 10 December 2014

Keywords:
Image classification
Similarity learning
Semantic representation
Large-margin framework

ABSTRACT

In the present paper, a novel image classification method that uses the hierarchical structure of categories to produce more semantic prediction is presented. This implies that our algorithm may not yield a correct prediction, but the result is likely to be semantically close to the right category. Therefore, the proposed method is able to provide a more informative classification result. The main idea of our method is two-fold. First, it uses semantic representation, instead of low-level image features, enabling the construction of high-level constraints that exploit the relationship among semantic concepts in the category hierarchy. Second, from such constraints, an optimization problem is formulated to learn a semantic similarity function in a large-margin framework. This similarity function is then used to classify test images. Experimental results demonstrate that our method provides effective classification results for various real-image datasets.

1. Introduction

Recognizing categories of objects and scenes is one of the most critical problems in computer vision. Although continuous progress has been made in this field, there still remains a large gap between machine performance and human intelligence. Unlike machines, humans can categorize at least tens of thousands of objects and scenes without any difficulty [1]. Furthermore, they can build a hierarchy of categories by simply observing images, and exploit it to produce semantically more meaningful judgement. For example, someone may mistakenly classify a dog as a cat but hardly misclassify a dog as a car. This example shows that one can produce a more informative classification result by considering the similarity between two semantic concepts. In the current work, we focus on this issue and attempt to make the image classification algorithm more semantic and human-like.

To achieve this goal, image classification algorithm should be developed under a new performance evaluation criterion. This criterion can be formulated by utilizing the hierarchical loss, which reflects the hierarchy of various semantic concepts, and not the flat 0/1 loss. Similar to [2], the hierarchical loss can be defined based on WordNet [3], a lexical semantic network for modeling human psycholinguistic knowledge. Under the hierarchical loss-based criterion, misclassifying an image as a different but semantically close category incurs a smaller loss than misclassifying it as a semantically distant category. Therefore, image classification can be significantly more informative by learning the algorithm based on the hierarchical loss. As shown in Fig. 1, the use of the hierarchical loss can provide substantial benefit to the results of classification.

In the present paper, a new image classification method is presented, which utilizes the hierarchical loss function and produces semantically more meaningful results. The key idea of this approach is a novel combination of semantic representation and similarity function learning. Several recent works are available that explicitly estimate high-level semantic attributes for various applications, such as description of generic or unfamiliar images [4,5], zero-shot transfer learning [6], and intermediate features that can aid in visual recognition [5–8]. We adopt semantic representation for the latter purpose to produce low-dimensional semantic feature vectors. The low dimensionality of our feature vector helps in the subsequent similarity learning being performed very efficiently compared with the conventional similarity or distance function learning [9–11] that usually directly handles high-dimensional low-level feature vectors. Moreover, because our feature vector is semantic, enforcing constraints among semantic concepts for the minimization of hierarchical loss is easy. More specifically, learning the similarity function, which is the core learning problem of our approach, is formalized within *a large-margin framework* and is guaranteed to minimize empirical hierarchical loss.

☆ This paper has been recommended for acceptance by Chung-Sheng Li.

* Corresponding author. Fax: +82 2 878 1452.
E-mail addresses: jangbon@snu.ac.kr (J.Y. Chang), kyoungmu@snu.ac.kr (K.M. Lee).

http://dx.doi.org/10.1016/j.cviu.2014.11.006
1077-3142/

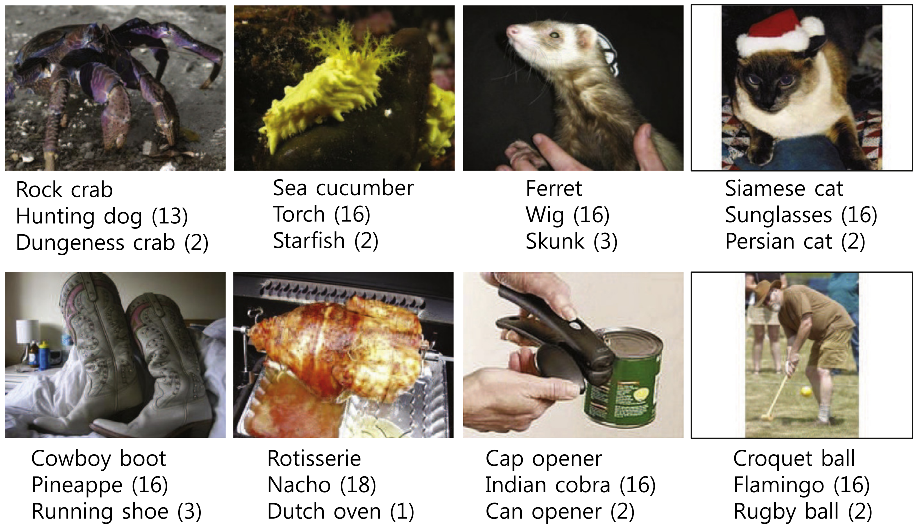

Fig. 1. Image classification results by exploiting class hierarchy versus those without considering hierarchy. The first label is the ground truth category. The second and third labels are the results of flat prediction (one-versus-all SVM) and hierarchical prediction (our method), respectively. The numbers indicate the hierarchical loss.

The contributions of the present work include the following:

- A novel large-margin formulation for semantic similarity learning is proposed. This learning problem can be viewed as an instance of a semidefinite program (SDP) [12], and an efficient optimization algorithm is developed.
- A thorough experimental study is conducted for comparing the performances of several algorithms for hierarchical image classification [13–15]. For this purpose, the Caltech [16,17] and ImageNet [18] datasets are used.
- The proposed method is shown to achieve a state-of-the-art classification result under the hierarchical-loss criterion. Furthermore, a noticeable gain in the conventional measures, such as accuracy and precision, can also be obtained.

The rest of the current paper is organized as follows. In Section 2, some related works are discussed. Section 3 presents the framework of the proposed method, followed by the presentation of the experimental results in Section 4. Finally, Section 5 concludes this paper.

2. Related works

The document categorization method proposed by Cai and Hofmann [13] treats the category structure above *flat* and considers the relationships among categories, which are commonly expressed in concept hierarchy or taxonomies. The task of exploiting these pre-determined taxonomies as additional information for classification fits well into the popular **structured learning framework** [19,20]. This idea enables us to use not just the flat 0/1 loss but also the more general loss function. In [13], the hierarchical loss function between two categories is defined and then minimized to provide the hierarchical classifier based on the structural support vector machines (S-SVMs). The experimental results of [13] on the document categorization show that S-SVM outperforms flat support vector machines (SVMs) in terms of hierarchical loss.

The main drawback of the structured learning approach is its high computational complexity, which renders the approach inherently slow even for the ordinary case of several dozen categories. To address this problem, Binder et al. [14] proposed an efficient alternative to the structured approach by decomposing the problem into several local tasks. The idea is to learn a binary SVM for each node in the taxonomy tree instead of solving the whole problem at once using the structured learning approach. At the final stage of their approach, **the ensemble of local binary SVMs** from all nodes is appropriately assembled by reflecting the taxonomy. They applied their classification method to real-world image data, such as Caltech256 [17] and VOC2006 [21], and reported that their local approach performed at par with the structured approach in terms of hierarchical loss while being considerably faster in training. However, the number of nodes in the taxonomy tree rapidly increases as the number of categories increases, which makes the method not free from computational burden.

To move from the flat classification to settings that utilize the category hierarchy information, Weinberger and Chapelle [15] proposed a very different approach from structured learning approaches. Instead of learning a classifier, their method solves a regression problem where images are mapped into a latent semantic space. This semantic space is learned in a supervised manner and underlies the category taxonomy, which is the reason why this method is referred to as **taxonomy embedding** (abbreviated to *taxem*). It first performs ridge regression to embed input features into a low-dimensional semantic space and then learns the distance function in the semantic space based on the large-margin framework. This two-step approach inspired our current work. Nevertheless, our method significantly differs in its essential features, such as the methods used in obtaining the semantic space and comparing two semantic vectors. Moreover, our approach is extensively tested on image data, whereas taxem was only applied to document categorization.

In [22], **hierarchical similarity** among semantic representation of images for large-scale image retrieval was proposed. Their approach incorporates prior hierarchy information and achieves significant improvements over state-of-the-art image retrieval methods. It performs learning to recognize the semantic attributes of images and then computes a similarity score by using a predefined comparison function. This comparison function is determined only based on a known hierarchical structure and does not utilize the training data, causing the hierarchical similarity method to produce suboptimal classification results compared with our method where the similarity function is determined by a large-margin learning framework with the training data. Further, they applied additional probabilistic calibration [23] to the semantic attributes after SVM, which is not necessary for our method.