Applied Soft Computing 67 (2018) 183–196

Contents lists available at ScienceDirect

Applied Soft Computing

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

Softly combining an ensemble of classifiers learned from a single convolutional neural network for scene categorization

Shuang Bai*, Huadong Tang

School of Electronic and Information Engineering, Beijing Jiaotong University, No. 3 Shang Yuan Cun, Hai Dian District, Beijing, China

ARTICLE INFO

Article history:
Received 29 August 2017
Received in revised form 14 January 2018
Accepted 4 March 2018
Available online 6 March 2018

Keywords:
Scene categorization
Convolutional neural networks
Ensemble learning
Soft combination
Genetic algorithm

ABSTRACT

In this paper we propose to train an ensemble of classifiers from a single convolutional neural network (CNN) and softly combine these classifiers for scene categorization. Specifically, we explore the hierarchical structure of a CNN to extract multiple types of features from images, and train a multi-class classifier corresponding to each type of features. To combine these classifiers effectively, a soft combination strategy is introduced. Considering the fact that different images may need to be discriminated by using different types of features, we train a set of auxiliary binary-class classifiers to estimate the quality of categorizing an image by using the corresponding multi-class classifiers, so that a dynamic weight can be assigned to each of the multi-class classifiers for combination. On the other hand, because features extracted from different layers of a CNN differ largely in their levels of abstraction, classifiers trained based on these features have quite different capabilities for scene categorization. To address this issue, in the soft combination strategy we adopt the genetic algorithm to learn another set of static weights for the multi-class classifiers for combination. The static weights are to adapt the multi-class classifiers to given datasets. Finally, to categorize an image, the multi-class classifiers are combined by using both dynamic and static weights. We conduct experiments on two challenging benchmark datasets, MIT-indoor scene 67 and SUN397. Experiment results show that the proposed method is effective for scene categorization and can give superior results to state-of-the-art approaches.

1. Introduction

The ability to categorize various scenes is important for humans to understand the environments [1]. Although humans are able to categorize scenes accurately and rapidly, it is a challenging task for computers. Aiming to make computers categorize scenes like humans, scene categorization has become one of the fundamental problems in the field of computer vision. With a wide range of applications such as image retrieval, autonomous vehicles, intelligent robotics and human–computer interaction, scene categorization is attracting more and more attention [2–4].

Since scene images are characterized by huge inter-class similarities and intra-class variabilities, scene categorization is more challenging than other image understanding tasks. The category of an image is not only determined by its contents, but also by the arrangements and interactions of the contents. For example, Fig. 1 shows sample images from scene categories: meeting room and classroom. It can be seen that images from both categories contain desks and chairs. Therefore, it is difficult to distinguish these two categories just by the objects they contain. The arrangements of the objects in the images are also essential.

Due to complexities of scene images, in order to perform scene categorization accurately, methods that can explore information in scene images effectively are needed. In the last decade, varieties of scene categorization approaches have been proposed. From methods that utilize low-level image features [5,6] to methods that are based on high-level object features [7–9], features of different levels of abstraction have been employed, and considerable progress has been made. However, features used in these methods are all hand-engineered. The design of hand-engineered features needs solid professional knowledge and large amounts of time and efforts. Furthermore, they are hard to be adapted to other tasks. As a result, further improvements of performances of hand-engineered features are limited.

On the other hand, recently remarkable successes have been made by convolutional neural networks (CNNs) in various visual tasks [10–15]. After being trained on large-scale image datasets, such as ImageNet [16,17] and Places [18], CNNs can generate deep features that are rich in semantic information. Such features

* Corresponding author.
E-mail address: shuangb@bjtu.edu.cn (S. Bai).

https://doi.org/10.1016/j.asoc.2018.03.007
1568-4946/© 2018 Elsevier B.V. All rights reserved.

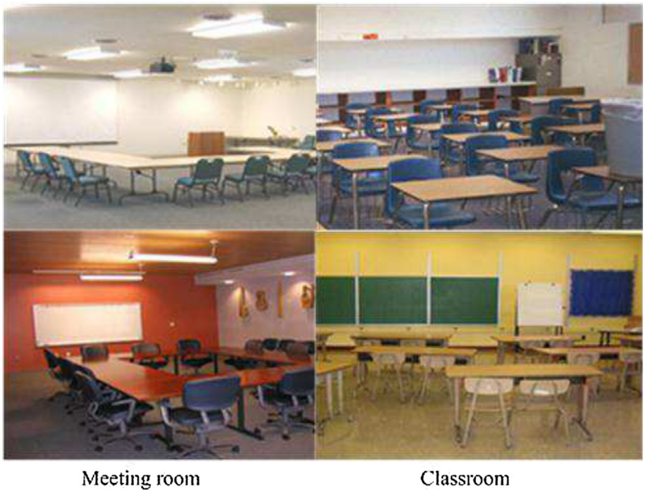

Fig. 1. Sample images from categories meeting room and classroom. Meeting room images are in the first column, while classroom images are in the second column.

are widely used for image classification and understanding tasks [19–21].

In addition to works mentioned above, convolutional neural networks can be used to solve various other tasks. For example, in [22], the authors propose a rice disease identification method, where 10 common rice diseases are identified by utilizing deep convolutional neural networks with an accuracy much higher than traditional methods. Lu et al. utilize convolutional neural networks to solve the biomimetic robot path planning problem and accomplish superior results to conventional methods [23].

CNNs are constructed by layers of neural network units, each of which applies certain operation to its inputs and feeds the outputs to the layer that follows it. Typically, CNNs are composed of convolution, pooling, non-linearity and full-connection layers [11], of which convolution and full-connection layers contain learnable weights. After training, these layers can be used to extract features from images. CNNs use hierarchical structures to compute representations for input images in a layer by layer manner [24].

Inspired by achievements made by CNNs in image classification [11], researchers begin to adopt CNNs for scene categorization widely. Donahue et al. use deep CNNs trained on the dataset ImageNet for scene categorization directly [25]. Gong et al. [26] use the VLAD method [27] to pool features extracted with CNNs from multi-scale local image patches. Zhou el at train CNNs over a scene-centric image database to perform scene categorization [18]. Herranz et al. design multi-scale CNN architectures for categorizing scene images [28]. Although remarkable progress is made by methods that use CNNs for scene categorization, there are few approaches on exploring features extracted from multiple layers of CNNs for scene categorization.

Because both low-level [5,6] and high-level features [7–9] are useful for scene categorization, and have been applied to tackle this task with certain achievements, it is intuitive to explore features of multiple levels of abstraction to take full advantage of the information that can be extracted from images for effective scene categorization. Considering that CNNs can extract features which are of different levels of abstraction from images, we propose to train an ensemble of classifiers based on features from different layers of CNNs for scene categorization.

The theoretical basis of the proposed method is as follows. Diversity is crucial for obtaining accurate ensembles [29–31]. Therefore, different individual classifiers in an ensemble are encouraged to represent different subspaces of the classification problem for specializing [32,33]. In feature subset based methods, diversity is realized by manipulating input feature sets for creating ensemble members [34], so that each classifier is given a different projection of the training set [35]. In these methods, the diversity of the ensemble is obtained by diversifying the feature subsets.

On the other hand, deep neural networks can generate progressively more abstract features through the hierarchical architectures, where more abstract representations at higher layers can be constructed by less abstract ones in lower layers [36,37]. Particularly, in CNNs, this abstraction is explicitly built via a pooling mechanism [24]. Furthermore, work on understanding features generated by CNNs has demonstrated that low layers in CNNs can capture low-level visual patterns such as corners and edges, while high layers can capture high-level features such as objects and their parts [38–40]. Previous research shows that features generated at higher layers of CNNs are more semantically meaningful and more abstract.

From properties of CNNs, it can be seen that outputs of each layer of a CNN are different from each other, capturing visual concepts of different levels of abstraction. Therefore, based on the diversity theorem of ensemble classifiers, it is reasonable to train an ensemble of classifiers from outputs of different layers of a CNN for scene categorization.

Since CNNs mainly depend on layers that contain learnable weights to extract features from images, corresponding to each convolution and full-connection layer, we train a multi-class classifier. Consequently, an ensemble of multi-class classifiers can be obtained from a single CNN model. However, because features from different layers of CNNs vary largely in their levels of abstraction, the obtained classifiers differ remarkably in their abilities to distinguish scene categories.

In order to obtain satisfactory results, proper weights are supposed to be assigned to the ensemble of classifiers for combination. To address this issue, we propose a soft combination strategy. On one hand, this strategy takes variations in scene image contents into consideration. Given an input image, it estimates the matching degree between this image and each classifier in the ensemble, and dynamically computes a weight for each classifier for combination. On the other hand, this strategy takes global properties of the classifiers into consideration and adapts the classifiers in the ensemble to given datasets by computing another set of static weights for them.

Specifically, to compute dynamic weights, we train a set of auxiliary classifiers based on features extracted with CNNs and associate each auxiliary classifier with a classifier in the ensemble, which is trained with the same type of features as the auxiliary classifier. The auxiliary classifiers are trained to generate dynamic weights for classifiers in the ensemble to reflect the quality of categorizing an image by using the corresponding classifier. To compute static weights, we adopt the genetic algorithm [41,42] to make the combination of the ensemble of classifiers to give optimum performances over validation sets of given datasets.

Given an input image, we first use the multi-class classifiers to perform classification based on the corresponding features independently, then we use the auxiliary classifiers to compute a set of dynamic weights for the classifiers. After that, the outputs of the multi-class classifiers are combined with both the dynamic weights and static weights to predicate the category of the input image.

Our main contributions in this paper can be summarized as follows:

1. We introduce an architecture to train an ensemble of classifiers from a single CNN.
2. We propose to train a set of auxiliary classifiers to predicate dynamic weights for an ensemble of classifiers for combination,