Neurocomputing 178 (2016) 36–45

Contents lists available at ScienceDirect

Neurocomputing

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

Deep Boosting: Joint feature selection and analysis dictionary learning in hierarchy

Zhanglin Peng [a], Ya Li [a], Zhaoquan Cai [b], Liang Lin [a,*]

[a] Sun Yat-sen University, Guangzhou, China
[b] Huizhou University, Huizhou, China

ARTICLE INFO

Article history:
Received 13 February 2015
Received in revised form
15 July 2015
Accepted 16 July 2015
Available online 6 November 2015

Keywords:
Representation Learning
Compositional boosting
Dictionary learning
Image Classification

ABSTRACT

This work investigates how the traditional image classification pipelines can be extended into a deep architecture, inspired by recent successes of deep neural networks. We propose a deep boosting framework based on layer-by-layer joint feature boosting and dictionary learning. In each layer, we construct a dictionary of filters by combining the filters from the lower layer, and iteratively optimize the image representation with a joint discriminative-generative formulation, i.e. minimization of empirical classification error plus regularization of analysis image generation over training images. For optimization, we perform two iterating steps: (i) to minimize the classification error, select the most discriminative features using the gentle adaboost algorithm; (ii) according to the feature selection, update the filters to minimize the regularization on analysis image representation using the gradient descent method. Once the optimization is converged, we learn the higher layer representation in the same way. Our model delivers several distinct advantages. First, our layer-wise optimization provides the potential to build very deep architectures. Second, the generated image representation is compact and meaningful by jointly considering image classification and generation. In several visual recognition tasks, our framework outperforms existing state-of-the-art approaches.

1. Introduction

Visual recognition is one of the most challenging domains in the field of computer vision and smart computing. Many complex image and video understanding systems employ visual recognition as the basic component for further analysis. Thus the design of robust visual recognition algorithm is becoming a fundamental engineering in computer vision literature and has been attracting many related researchers. Since the inadequate visual representation will greatly influence the performance of visual recognition system, almost all of the related methods are concentrated on developing the effective visual representation.

Traditional visual recognition systems always adopt the shallow model to construct the image/video representation. Among them, the *bag-of-visual-words* (BoW) model, which is the most successful one for visual content representation, has been widely adopted in many computer vision tasks, such as object recognition [1,2] and image classification [3,4]. The basic pipeline of BoW model consists of local feature extraction [5,6], feature encoding [7–9] and pooling operation. In order to improve the performance of BoW, two crucial schemes have been involved. First, the traditional BoW model discards the spatial information of local descriptors, which seriously limited the descriptive power of the feature representation. To overcome this problem, the Spatial Pyramid Matching method was proposed in [3] to capture geometrical relationships among local features. Second, dictionaries adopted to encode the local feature in traditional methods are learned in a unsupervised manner and can hardly capture the discriminative visual pattern for each category. This issue inspired a series of works [10–12] to train more discriminative dictionaries via supervised learning, which can be implemented by introducing the discriminative term into dictionary learning phase as the regularization according to various criteria.

As the research going, the deep models, which can be seen as a type of hierarchical representation [13–15] have played an significant role in computer vision and machine learning literature [16–18] in recent years. Generally, such hierarchical architecture represents different layer of vision primitives such as pixels, edges, object parts and so on [19]. The basic principles of such deep models are concentrated on two folds: (1) layerwise learning philosophy, whose goal is to learn single layer of the model individually and stack them to form the final architecture; (2) feature combination rules, which aim at utilizing the combination (linear

* Corresponding author.
E-mail address: linliang@ieee.org (L. Lin).

http://dx.doi.org/10.1016/j.neucom.2015.07.116
0925-2312/

or nonlinear) of low layer detected features to construct the high layer impressive features by introducing the activation function.

In this paper, the related exciting researches inspire us to explore how the traditional image classification pipelines, which include feature encoding, spatial pyramid representation and salient pattern extraction (*e.g.*, max spatial pooling operation), can be extended into a deep architecture. To this end, this paper proposes a novel deep boosting framework, which aims to construct the effective discriminative features for image classification task, jointly adopting feature boosting and dictionary learning. For each layer, followed the famous boosting principle [20], our proposed method sequentially selects the discriminative visual features to learn the strong classifier by minimizing empirical classification error. On the other hand, the analysis dictionary learning strategy is involved to make the selected features more suitable for the object category. A two-step learning process is investigated to iteratively optimize the objective function. In order to construct high-level discriminative representations, we composite the learned filters corresponding to selected features in the same layer, and feed the compositional results into next layer to build the higher-layer analysis dictionary. Another key to our approach is introducing the model compression strategy when constructing the analysis dictionary, that reduces the complexity of the feature space and shortens the model training time. The experiment shows that our method achieves excellent performance on general object recognition tasks. Fig. 1 illustrates the pipeline of our deep boosting method (applying two layers as the illustration). Compared with the traditional BoW based method [7], the analysis operation in our model (*i.e.*, convolution) is same as the encoding process that maps the image into the feature space. While the pooling stage is same as the traditional method to compute the histogram representation adopting spatial pyramid matching. Different from traditional models capturing the salient properties of visual patterns by max spatial pooling operation, we adopt the feature boosting to the discriminative features mining for image representation.

The main contributions of this paper are three folds. (1) A novel deep boosting framework is proposed and it leverages the generative and discriminative feature representation. (2) It presents a novel formulation which jointly adopting feature boosting and analysis dictionary learning for image representation. (3) In the experiment on several standard benchmarks, it shows that the learned image representation well discovers the discriminative features and achieves the good performance on various object recognition tasks.

The rest of the paper is organized as follows. Section 2 presents a brief review of related work, followed by the overview of background technique details in Section 3. Then we introduce our deep boosting framework in Section 4. Section 5 gives the experimental results and comparisons. Section 6 concludes the paper.

2. Related work

In the past few decades, many works have been done to design different kinds of features to express the characteristics of the image for further visual tasks. These hand-craft features vary from global expressions [21] to the local representation [5]. Such designed features can be roughly divided into two types [22], the one is geometric features and the other is texture features. Geometric features which explicitly record the locations of edges are employed to describe the noticeable structures of local areas. Such features include Canny edge descriptor [23], Gabor-like primitives [24] and shape context descriptor [25,26]. In contrast, the texture features express the cluttered object appearance by histogram statistics. SIFT [5], HoG [6] and GIST [27] are delegates of such feature representation. Beyond such hand-craft feature descriptors, Bag-of-Feature (BoF) model seems to be the most classical image representation method in computer vision area. A lot of illuminating studies [4,3,7,8] were published to improve this traditional approach in different aspects. Among these extensions, a class of sparse coding based methods [7,8], which employ spatial pyramid matching kernel (SPM) proposed by Lazebnik et al., has achieved great success in image classification problem. However, despite we are developing more and more effective representation methods, the lack of high-level image expression still plagues us to build up the ideal vision system.

On the other hand, learning hierarchical models to simultaneously construct multiple levels of visual representation has been paid much attention recently. The proposed hierarchical image representation is partially motivated by recent developed deep learning approaches [13,14,28]. Different from previous hand-craft feature design method, deep model learns the feature representation from raw data and validly generates the high-level semantic representation. And such abstract semantic representations are expected to provide more intra-class variability. Recently, many vision tasks achieve significant improvement using the convolutional architectures [16–18]. A deep convolutional architecture consists of multiple stacked individual layers, followed by an empirical loss layer. Among all of these layers, the convolutional layer, the feature pooling layer and the full connection layer play major roles in abstract feature representation. The stochastic gradient descent algorithm is always applied to the parameters

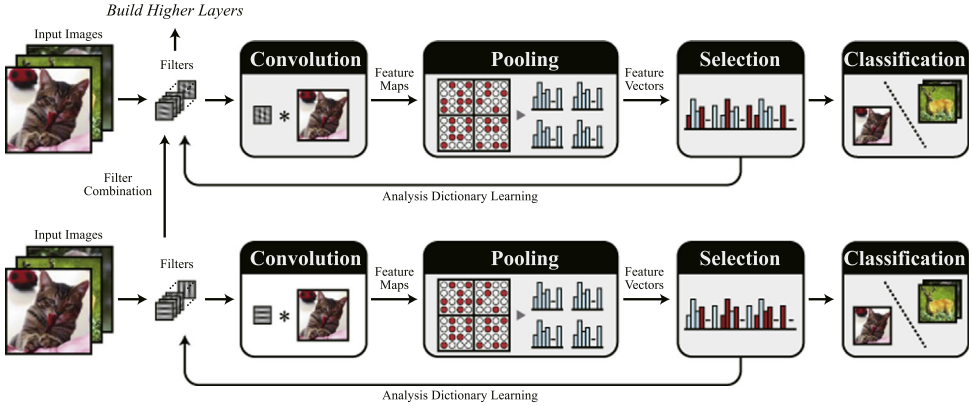

Fig. 1. A two-layer illustration of proposed deep boosting framework. The horizontal pipelines show the layer-wised image representation via joint feature boosting and analysis dictionary learning. When optimization in the single layer is done, the compositional filters are fed into the higher-layer to generate the novel analysis dictionary for further processing. Note that the feature set in the higher-layer only dependents on the training images and combined filters in the relevant layer.