Pattern Recognition 71 (2017) 94–105

Contents lists available at ScienceDirect

Pattern Recognition

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

A deep convolutional neural network module that promotes competition of multiple-size filters

Zhibin Liao*, Gustavo Carneiro

Australian Centre for Robotic Vision University of Adelaide North Terrace, Adelaide, Australia

ARTICLE INFO

Article history:
Received 23 December 2016
Revised 3 May 2017
Accepted 25 May 2017
Available online 31 May 2017

Keywords:
Deep learning
Multi-size filter
Filter co-adaptation
Classification

ABSTRACT

We introduce a new deep convolutional neural network (ConvNet) module that promotes competition amongst a set of convolutional filters of multiple sizes. This new module is inspired by the inception module, where we replace the original collaborative pooling stage (consisting of a concatenation of the multiple size filter outputs) by a competitive pooling represented by a maxout activation unit. This extension has the following two objectives: 1) the selection of the maximum response amongst the multiple size filters prevents filter co-adaptation and allows the formation of multiple sub-networks within the same model, which has been shown to facilitate the training of complex learning problems; and 2) the maxout unit reduces the dimensionality of the outputs from the multiple size filters. We show that the use of our proposed module in typical deep ConvNets produces classification results that are competitive with the state-of-the-art results on the following benchmark datasets: MNIST, CIFAR-10, CIFAR-100, SVHN, and ImageNet ILSVRC 2012.

1. Introduction

The use of competitive activation units in deep convolutional neural networks (ConvNets) is generally understood as a way of building one network by the combination of multiple sub-networks, with each one being capable of solving a simpler task when compared to the complexity of the original problem involving the whole dataset [1]. Similar ideas have been explored in the past using multi-layer perceptron models [2], but there is a resurgence in the use of competitive activation units in deep ConvNets [1,3]. For instance, rectified linear unit (ReLU) [4] promotes a competition between the input sum (usually computed from the output of convolutional layers) and a fixed value of 0, while maxout [5] and local winner-take-all (LWTA) [3] explore an explicit competition amongst the input units. As shown by Srivastava et al. [1], these competitive activation units allow the formation of sub-networks that respond consistently to similar input patterns, which facilitates training [3–5] and generally produces superior classification results [1].

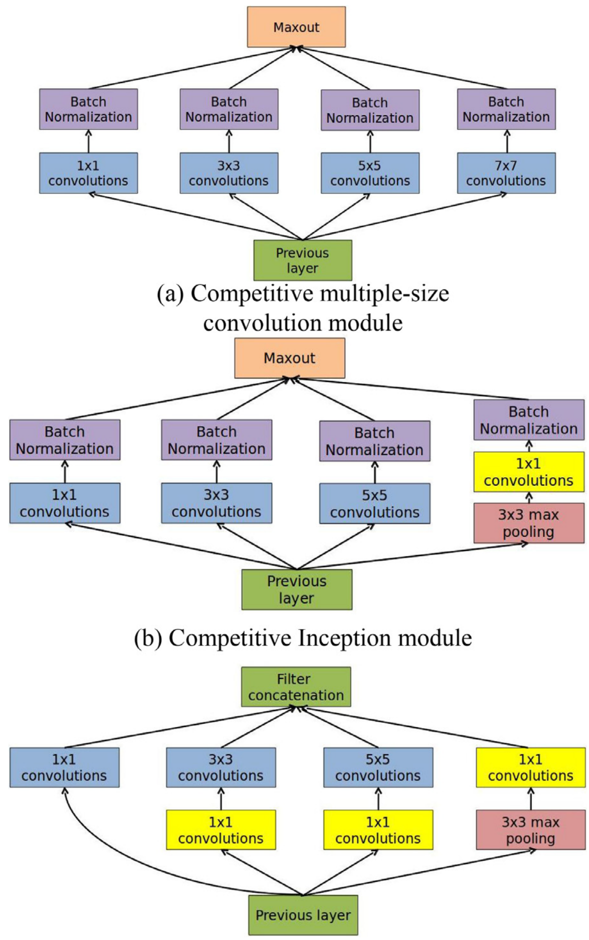
In this paper, we introduce a new module for deep ConvNets composed of several convolutional filters of multiple sizes that are joined by a maxout activation unit, which promotes competition amongst these filters. Our idea has been inspired by the recently proposed inception module [6], which currently produces state-of-the-art results on the ILSVRC 2014 classification and detection challenges [7]. The gist of our proposal is depicted in Fig. 1, where we have the data in the input layer filtered in parallel by a set of convolutional filters of multiple sizes [6,8,9]. Then the output of each filter of the convolutional layer passes through a batch normalisation unit (BNU) [10] that weights the importance of each filter size and also pre-conditions the model (note that the pre-conditioning ability of BNUs in ConvNets containing piece-wise linear activation units has been empirically shown in [11]). Finally, the multiple size filter outputs, weighted by BNU, are joined with a maxout unit [5] that reduces the dimensionality of the joint filter outputs and promotes competition amongst the multiple size filters. We empirically show that the competition amongst filters of multiple size prevents filter co-adaptation and allows the formation of multiple sub-networks. Furthermore, we show that the introduction of our proposal module in a typical deep ConvNet produces competitive results in the field for the benchmark datasets MNIST [12], CIFAR-10 [13], CIFAR-100 [13], street view house number (SVHN) [14], and ImageNet ILSVRC 2012 [7].

2. Literature review

One of the main reasons behind the outstanding performance of deep ConvNets is attributed to the use of competitive activation units in the form of piece-wise linear functions [1,15], such as ReLU [4], maxout [5] and LWTA [3] (see Fig. 2). In general, these

* Corresponding author.
E-mail addresses: zhibin.liao@adelaide.edu.au (Z. Liao), gustavo.carneiro@adelaide.edu.au (G. Carneiro).

http://dx.doi.org/10.1016/j.patcog.2017.05.024
0031-3203/



(a) Competitive multiple-size convolution module

(b) Competitive Inception module

(c) Original inception module

Fig. 1. The proposed deep ConvNet modules are depicted in (a) and (b), where (a) only contains multiple size convolutional filters within each module, while (b) contains the max-pooling path, which resembles the original inception module [6] depicted in (c) for comparison.

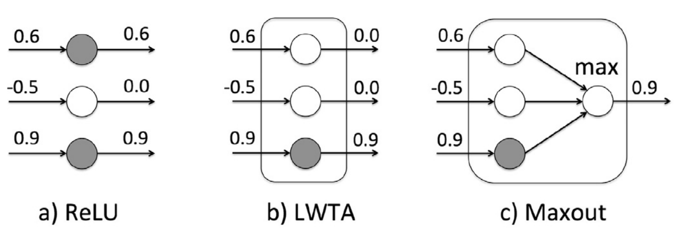


Fig. 2. Competitive activation units, where the grey nodes are the active ones, from which errors flow during backpropagation. ReLU [4] (a) is active when the input is bigger than 0, LWTA [3] (b) activates only the node that has the maximum value (setting to zero the other ones), and maxout [5] (c) has only one output containing the maximum value from the input. This figure was adapted from Fig. 1 of [1].

activation functions enable the formation of sub-networks that respond consistently to similar input patterns [1], dividing the input data points (and more generally the training space) into regions [15], where classifiers and regressors can be learned more effectively given that the sub-problems in each of these regions are simpler than the original problem involving the whole training set. In addition, the joint training of the sub-networks present in such deep ConvNets represents a useful regularisation method [3–5]. In practice, ReLU, maxout and LWTA allows the division of the input space in exponentially many regions as a function of the number of layers and the number of input nodes to each competitive activation unit, so this means that maxout and LWTA can estimate exponentially complex functions more effectively than ReLU because of the larger number of sub-networks that are jointly trained. An important aspect of deep ConvNets with competitive activation units is the fact that the use of batch normalisation units (BNU) helps not only with respect to the convergence rate [10], but also with the pre-conditioning of the model by promoting an even distribution of the input data points, which results in the maximisation of the number of the regions (and respective sub-networks) produced by the piece-wise linear activation functions [11]. Furthermore, training ConvNets with competitive activation units [1,11] usually involves the use of dropout [16] that consists of a regularisation method that prevents filter co-adaptation [16], which is a particularly important issue in such models, because filter co-adaptation can lead to a severe reduction in the number of the sub-networks that can be formed during training.

Another aspect of the current research on deep ConvNets is the idea of making the network deeper, which has been shown to improve classification results [17]. However, one of the main ideas being studied in the field is how to increase the depth of a ConvNet without necessarily increasing the complexity of the model parameter space [6,18]. For the Szegedy et al.'s model [6], this is achieved with the use of 1×1 convolutional filters [19] that are placed before each local filter present in the inception module in order to reduce the input dimensionality of the filter. In Simonyan et al.'s approach [18], the idea is to use a large number of layers with convolutional filters of small size (e.g., 3×3). In this work, we restrict the complexity of the deep ConvNet with the use of maxout activation units, which selects only one of the input nodes, as shown in Fig. 2.

Finally, the use of multiple size filters in deep ConvNets is another important implementation that is increasingly being explored by several researchers [6,8,9]. Essentially, multiple size filtering follows a neuroscience model [20] that suggests that the input image data should be processed by filters of different sizes (which can lead to filters of different scales) and then pooled together, so that the deeper processing stages can become more robust to scale changes [6]. We explore this idea in our proposal, as depicted in Fig. 1, but we hypothesise (and show supporting evidence) that the multiple size of the filters prevents their co-adaptation during training, leading to better generalisation. We also hypothesise and show evidence that what is driving this better generalisation is the fact that the multiple sizes of the filters promote the learning of features that are more different from each other within competitive units when compared to the single-size filters.

3. Methodology

Assume that an image is represented by $\mathbf{x} : \Omega \to \mathbb{R}$, where Ω denotes the image lattice, and that an image patch of size $(2k-1) \times (2k-1)$ (for $k \in \{1, 2, \ldots, K\}$) centred at position $i \in \Omega$ is represented by $\mathbf{x}_{i \pm (k-1)}$. The models being proposed in this paper follow the structure of the NIN model [19], and is in general defined as follows:

$$f(\mathbf{x}, \theta_f) = f_{out} \circ f_L \circ \ldots \circ f_2 \circ f_1(\mathbf{x}), \quad (1)$$

where $\circ$ denotes the composition operator, θ_f represents all the ConvNet parameters (i.e., weights and biases), $f_{out}(.)$ denotes an averaging pooling unit followed by a softmax activation function [19], and the network has blocks represented by $l \in \{1, \ldots, L\}$, with each block containing a composition of N_l modules with $f_l(\mathbf{x}) = f_l^{(N_l)} \circ \ldots \circ f_l^{(2)} \circ f_l^{(1)}(\mathbf{x})$. Each module $f_l^{(n)}(.)$ at a particular position $i \in$