Neural Networks 105 (2018) 340–345

Contents lists available at ScienceDirect

Neural Networks

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

A sparsity-based stochastic pooling mechanism for deep convolutional neural networks

Zhenhua Song [a], Yan Liu [a], Rong Song [a], Zhenguang Chen [b], Jianyong Yang [b], Chao Zhang [a,*], Qing Jiang [a]

[a] School of Engineering, Sun Yat-sen University, Guangzhou 510006, PR China
[b] The First Affiliated Hospital of Sun Yat-sen University, Sun Yat-sen University, Guangzhou 510080, PR China

ARTICLE INFO

Article history:
Received 21 March 2017
Received in revised form 2 May 2018
Accepted 23 May 2018
Available online 15 June 2018

Keywords:
Deep learning
Pooling mechanism
Degree of sparsity
Representative feature value
Recognition accuracy

ABSTRACT

A novel sparsity-based stochastic pooling which integrates the advantages of max-pooling, average-pooling and stochastic pooling is introduced. The proposed pooling is designed to balance the advantages and disadvantages of max-pooling and average-pooling by using the degree of sparsity of activations and a control function to obtain an optimized representative feature value ranging from average value to maximum value of a pooling region. The optimized representative feature value is employed for probability weights assignment of activations in normal distribution. The proposed pooling also adopts weighted random sampling with a reservoir for the sampling process to preserve the advantages of stochastic pooling. This proposed pooling is evaluated on several standard datasets in deep learning framework to compare with various classic pooling methods. Experimental results show that it has good performance on improving recognition accuracy. The influence of changes to the feature parameter on recognition accuracy is also investigated.

1. Introduction

The field of deep learning has advanced rapidly due to the ability to perform deep structure optimization (Hinton & Salakhutdinov, 2006) on computational models designed to perform a wide range of inference tasks. Convolutional neural networks (CNNs) are a type of multilayer-structured learning algorithm; they have gained wide attention from researchers, in improving deep network performance via reducing the number of parameters (weight sharing) using relative spatial relationships (LeCun et al., 1989; LeCun, Bottou, Bengio, & Haffner, 1998) at the earlier levels of processing. Pooling processes can be employed to classify perform dimensionality reduction, discouraging large CNNs models from over-fitting (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Being designed to maintain the properties of rotation, translation, scale invariance of features, pooling has been extensively harnessed in deep convolutional networks. Pooling has been shown to achieve better robustness to noise and clutter with high compactness for accurate recognition of mass images in automatic geometry molding of FEM and 3D printing (Xie, Tian, Wang, & Zhang, 2014).

Several popular pooling methods, such as max-pooling, average-pooling and stochastic pooling have been employed in practice. Each of these pooling methods has advantages and disadvantages due to the existence of inevitable quantization error in the pooling process. It is usually considered that average-pooling could reduce the variance increases, retaining background information (LeCun et al., 1990); while the max-pooling could reduce the deviation of the estimated average value caused by convolution layer parameter error while preserving the foreground texture details (Boureau, Bach, LeCun, & Ponce, 2010). These pooling methods adopt down-sampling operations, which reduce the feature dimensionality for the following computations (Sun, Song, Jiang, Pan, & Pang, 2017). Meanwhile, Yang, Yu, Gong, and Huang (2009) pointed out that the maximum value is better than the average value of a feature for classification to represent its activity over a region of interest. Boureau, Ponce, and LeCun (2010) compared the differences and features between average-pooling and max-pooling by theoretical analysis, and summarized that max-pooling and average-pooling have the potential drawbacks of losing background information and foreground texture information, respectively. Indeed, both average pooling and max-pooling have advantages and disadvantages in practice on performance, which highly depends on the specific application cases or datasets (Wang, Gao, Liu, & Meng, 2017). Thus, a principle/standard is to be established for automatically choosing the better one between

* Correspondence to: Sun Yat-sen University, No.132 Waihuan East Rd. of University Town, Guangzhou 510006, China.
E-mail address: zhchao9@mail.sysu.edu.cn (C. Zhang).

https://doi.org/10.1016/j.neunet.2018.05.015

these two pooling methods in the specific cases, which could promote the generalization ability of pooling.

For this aim, Yu, Wang, Chen, and Wei (2014) proposed a mixed pooling method that consisted in randomly choosing between max-pooling or average-pooling to generate the output. This mechanism is realized by adding together the maximum and average values which are multiplied by their own coefficients. One of the coefficients is either 0 or 1 randomly, and another is equal to the corresponding opposite value of the previous one (0 and 1 are the opposite values). This mechanism is praiseworthy on improving the overall performance of pooling results, except that it fails to reflect the advantages of both pooling methods at the same time because it can only adopt and reflect either max-pooling or average-pooling in each pooling process. Lee, Gallagher, and Tu (2015) made an improvement on this mixed pooling by replacing previous random coefficient with a real number ranging from 0 to 1, namely, the mixing proportion; consequently, the weights of maximum and average values are assigned by this real number. The features of both max-pooling and average-pooling could be reflected in each pooling process by this mixing proportion mechanism, although the randomness of the sampling process is sacrificed. Later on, stochastic pooling emerged to give the probability weights of the elements in feature map according to their numerical values, and to randomly take sample in accordance with the probability weights as well. Zeiler and Fergus (2013) proposed a classical stochastic pooling method by randomly picking the activations in pooling region on the basis of their activities. It has the advantages of being hyper-parameter free and the ability of combining with other regularization approaches, such as dropout and data augmentation. This stochastic pooling method presents smaller training and testing errors than those of max-pooling and average-pooling. Meanwhile, it is also reported that the pooling performance could also be improved using the method of Dropout (Iosifidis, Tefas, & Pitas, 2015). But the performance of classic Dropout is highly depended on the experience of position selection for random deleting, which makes it an experience-dependent method and limits its generalization ability (Cao, Li, & Zhang, 2015; Srivastava et al., 2014). Wu and Gu (2015) pointed out that the random sampling process of stochastic pooling for activation obeys multinomial distributions, which is same as that of max-pooling dropout. But in the case of specific retaining probabilities, the max-pooling dropout could perform better than stochastic pooling. It reveals that max-pooling dropout and stochastic pooling have their own advantages with respect to sampling. Therefore, if a novel pooling mechanism is designed to integrate the advantages of max-pooling, average-pooling and stochastic pooling together, it would be expected to not only improve the diversity of pooling results by taking a balance between highlighting foreground textures and preserving background information, but also promote the performance of recognition accuracy.

In this research, a novel sparsity-based stochastic pooling has been proposed to integrate advantages of max-pooling, average-pooling and stochastic pooling on taking a balance to highlight foreground and preserve background information at same time and improving randomness of sampling. The pooling mechanism was introduced by using an optimized representative feature value, which could automatically select to perform the features of max-pooling or average pooling primarily in specific application cases or databases for promoting the generalization ability of pooling since it has been defined by using the degree of sparsity and a special control function to generate a value ranging from average value to maximum value of a pooling region. And the probability weights of activations are assigned according to the distance between the feature value and value of each activation based on normal distribution, which could evaluate the contributions of all activations in the feature pooling region. A method of weighted random sampling (WRS) has been employed for this sampling operation to promote the performance of pooling by improving randomness of sampling. This proposed pooling was evaluated in terms of recognition accuracy within several classic datasets and its experimental test error compared with other classic pooling methods. The influence of changes to feature parameters on recognition accuracy is also discussed.

2. Pooling mechanism

2.1. Optimized representative feature value

A feature value is always employed for the pooling region, such as maximum value or average value, to be a benchmark for weight assignment and probability distribution of activations in pooling region. The weight of the maximum value of activations could be defined as 1 in max-pooling; in contrast, the weight of each activation is the same in average-pooling by this point of view. In rank-based stochastic pooling, no such feature value exists. Activations are arranged in descending order and given probability weights by exponential ranking (Michalewicz, 1994). This method could improve the performance of pooling by avoiding the mistake that of offering equal or highly imbalanced importance to each region since image features are highly spatially non-stationary (Shi, Ye, & Wu, 2016). Meanwhile, the authors also mentioned that an inevitable degeneration of rank-based stochastic pooling into max-pooling and loss of background information would occur if the maximum activation is much greater than the sum of others (probabilities of others are ignored by that of maximum activation).

To remedy above mentioned disadvantages and improve the pooling algorithm, an optimized representative feature value R is proposed to replace these common feature values (maximum and average values) and seek a reasonable balance between max-pooling and average-pooling to highlight foreground texture details while preserving enough background character information. The feature value R is defined by Eq. (1) (and shown in Fig. 1).

$$\frac{R - Avg}{Max - Avg} = F_p(\alpha) \tag{1}$$

where, Max and Avg are the maximum and average values of activations in pooling region, respectively. $F_p(\alpha)$ is the control function for optimizing this feature value, as shown in Eq. (2).

$$F_p(\alpha) = \begin{cases} 2^{p-1}\alpha^p, \left(0 \le \alpha \le \dfrac{1}{2}\right) \\ 1 - 2^{p-1}(1-\alpha)^p, \left(\dfrac{1}{2} \le \alpha \le 1\right). \end{cases} \tag{2}$$

Here, p is a positive integer as feature parameter for setting the curved shape of function $F_p(\alpha)$, and α is the degree of sparsity of convolved features in a pooling region as many researches showed that the performance of pooling methods are highly affected by the sparsity of the pooling region (Boureau, Ponce et al., 2010). For example, taking the maximum value works better than average value in a sparse region. Thus, a representative feature value designed based on the sparsity of activations in a pooling region is more reasonable.

There are three main advantages of using Eq. (2) to define R. First, if $p = +\infty$ (its value is set to be 100 in real case, which is a number large enough to meet the computing requirement), the value of R tends to be either maximum value or average value of activations in pooling region (Fig. 1). This pooling will degenerate into max-pooling or average-pooling to contain the features and functions of these two classic pooling methods; if $p = 1$, the value of R will be linearly distributed between maximum and average values, which simplifies it for high computational efficiency in