Improved CNN Based on Super-Pixel Segmentation

Yadong Yang(✉) and Xiaofeng Wang

College of Information Engineering, Shanghai Maritime University,
Shanghai, China
yangyadong03@stu.shmtu.edu.cn, xfwang@shmtu.edu.cn

Abstract. Convolutional neural network has unique superiority in images processing, it can effectively extract features and reduce data dimensions by convolution and pooling. But it takes the "violent segmentation" method in the process of pooling. This method cannot guarantee the final selected pixel value can be a good representative of the partial features, neither average pooling nor max pooling. Therefore, three pooling methods based on super-pixel segmentation are proposed, called "super-pixel average pooling", "super-pixel max pooling" and "super-pixel smooth pooling". Firstly, the super-pixel segmentation is performed on the feature images, and then the value of the point which has the smoothest gradient in each super pixel is selected to represent the feature of the local area. Compared to the violent segmentation pooling operation, this method exhibits more stable characterization ability, it can retain image features perfectly while reduce the data dimensions. Experiments show that the improved convolutional neural network achieved better results than normal algorithm in the standard data sets.

Keywords: Convolutional neural network · Super-pixel segmentation · Super-pixel average pooling · Super-pixel max pooling · Super-pixel smooth pooling

1 Introduction

Convolutional neural network (CNN) is a deep neural network model with convolution structure. It can effectively reduce the number of weights and lower the complexity of the network. This model has some invariance for scaling and other forms of deformation. Each neuron only needs to perceive the local image area, The network obtain global information by combining these neurons that perceive different local regions at higher levels. In 1989, LeCun et al. proposed a CNN model "LeNet-5 [1] " for character recognition. LeNet-5 consists of convolutional layers, sub-sampling layers and fully connected layers. The system achieves good results in small-scale handwritten numeral recognition. In the ImageNet contest, "AlexNet [2] ", a new network architecture of CNN, designed by Krizhevsky et al. won the 2012 championship. AlexNet and OverFeat [3] are quite powerful CNN models trained on large natural image datasets. By improving the network performance, Girshick et al. proposed R-CNN [4] (Regions with CNN) can complete the target detection task

Z. Shi et al. (Eds.): ICIS 2017, IFIP AICT 510, pp. 305–310, 2017.
DOI: 10.1007/978-3-319-68121-4_33

effectively. He utilize spatial pyramid pooling (SPP) to deal with different size and aspect ratio of the input images, this new network called SPP-Net [6]. In recent years, more and more popular method to optimize CNN models is to develop deeper and more complex network structures, and then to train them with massive training data. VGG [5] (visual geometry group), a 19-layer depth network, mainly explores the importance of depth for the network. GoogLeNet [6] is the particular incarnation for ILSVRC14 submit by Szegedy et al., a 22 layers deep network, the quality of which is evaluated in classification and detection. In order to improve the performance of CNN, the researchers not only discuss the structure of CNN and its applications, but also improve the design of the network layers, the loss function, activation function, regular items and many other aspects of the existing network. They achieved a series of results, for example, Inception network [5], stochastic pooling, ReLU [2], Leakly ReLU [7], Cross entropy loss, Batch normalization [8], etc.

2 Related Work

2.1 Convolutional Neural Network

In a typical CNN model, the beginning layers are usually alternating between the convolution layer and the sub-sampling layer. The last few layers of the network near the output layer are usually fully connection networks. The four basic elements of CNN are convolutional layers, sub-sampling layers, all connected layers and back propagation (BP) algorithm.

The convolutional layers receive the input images or the feature maps from the previous layer, they carry out the convolution operation through the N convolution cores and activate operation using the activation function, then N new feature maps are generated and directed into the next sub-sampling layer. The function of convolution is defined by

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

where l denotes the convolution layer, j is the j-th channel, x_j^l is the output value from the j-th channel of convolution layer l, f(·) is the activation function, $\mathrm{M_j}$ represents the subset of feature maps, k_{ij}^l is the convolution kernel matrix, * is the convolution operation, b_j^l is offset of l.

The sub-sampling layers perform sub-sampling operations on the input feature maps, which can effectively reduce the computational complexity and extract more representative local features. The function of sub-sampling is defined by

$$x_j^l = f(\beta_j^l down(x_j^{l-1}) + b_j^l) \tag{2}$$

where β_j^l denotes is the sub-sampling weight coefficient, down(·) is sub-sampling function.

The fully connected layers consist of an input layer, several hidden layers and an output layer. The upper layer's feature maps is spliced into one-dimensional vector as input data, and the final outputs is obtained by weighting and activating.

$$x^l = f(\varpi^l x^{l-1} + b^l) \tag{3}$$

where ϖ^l is the neural network weight.

The BP algorithm is a common neural network method in supervised learning. For the convolutional neural network, the main job is to optimization the convolution kernel parameters, the sub-sampling layers weights, the network weights of the full connection layer and the bias parameters of all layers. The essence of the BP algorithm is to allow us to calculate each network layer's effective errors and to derive the learning rules of network parameters, and then drive the actual network outputs are closer to the target values.

2.2 Super-Pixel Segmentation

Super pixels are divided a pixel-level image into district-level image or sets of pixels by image segmentation. The goal of super pixel segmentation is to change or simplify the representation of an image into something that is more convenient and significant to analyze. The super pixel segmentation method has been under intensive study, now there are many super pixel segmentation algorithms.

Super pixels are obtained by using NCut [9] and SLIC [10] algorithm has a high compactness. Using SLIC and Watershed algorithm for ultra-pixel segmentation is very productive. But if you put more emphasis on edge accuracy and regional merging, you can choose Marker-based Watershed and Meanshift algorithm.

3 Improved CNN

CNN have been substantiated to provide a powerful approach for image processing. In this work, we focus on the pooling part of the CNN.

The common CNN pooling methods include average pooling, max pooling and multi-scale mixing pooling. These methods have achieved some good results, they calculate a value representing the local region feature in the pre-divided local area of feature maps. However, those violent segmentation methods have their own unreasonable place, including ignore some local features or neutralize some salient features. This paper presents three pooling methods based on super pixel segmentation. Firstly, take super-pixel segmentation action on the feature maps, and the pixel feature values in each local region (i.e., super pixels) we got are similar. Then three kinds of pooling methods are come up with super pixel segmentation.

(1) Calculate the average value of pixels within each super pixel called super pixel average pooling.
(2) Get the max value of pixels within each super pixel, that is, super pixel max pooling.
(3) And super pixel smooth pooling trying to take the value of the point which has the smoothest gradient in each super pixel.

This paper only considers the super pixel level, does not deal with regional merger after image segmentation. Based on SLIC has the low time complexity and high compactness, we choose to use SLIC algorithm for super pixel segmentation (Fig. 1).

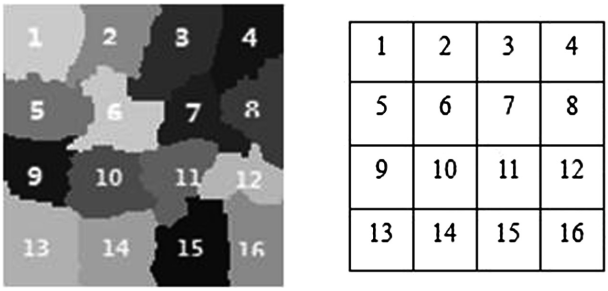

Fig. 1. The super-pixel pooling schematic diagram

These segmentation methods can not only extract the necessary features, but also the local stability of the images has no destruction at all.

4 Results

The structure of CNN we used in experiment for MNIST has 7 layers. The network is composed of 2 convolutional layers, 2 pooling layers and 3 fully connected layers. The experiment for CIFAR-10 use AlexNet. While the convolutional layers are followed by rectified linear operator (ReLU) layers and the pooling layers take the value of local regions with two-pixel strides, the dropout layers having a 0.5 dropout ratio. The last fully connected layer employ softmax function as a multi-class activation function.

Because of the MNIST data have relatively simple characteristics. Here divide the MNIST test set into six parts randomly, each part is transformed into different scales or angles, and then normalize the images in 28*28 pixels to get a new test set. The training set remains the same. Test on data sets MNIST, New MNIST (Fig. 2).

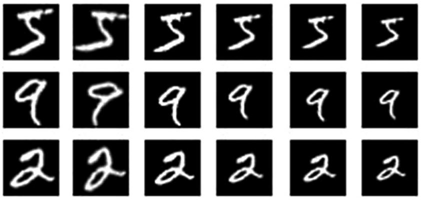

Fig. 2. Part of the new MNIST test images

In the following tables: “max”, “avg”, “sps”, “sp-max”, “sp-avg” each representing max pooling CNN method, average pooling CNN method, super-pixel smooth pooling CNN method, super-pixel max pooling CNN method and super-pixel average pooling CNN method. The results are shown as follows (Tables 1, 2, and 3).

Table 1. Test results of the five methods on the MNIST dataset.

Methods	Training error (%)	Test error (%)
max	0.01	0.8
avg	0.06	0.9
sps	0.04	0.7
sp+max	0.05	0.8
sp+avg	0.06	0.8

Table 2. Test results of the five methods on the new MNIST dataset

Methods	Training error (%)	Test error (%)
max	0.04	10.7
avg	0.1	9.8
sps	0.06	4.7
sp+max	0.06	8.9
sp+avg	0.08	6.6

Table 3. Test results of the five methods on the CIFAR-10 dataset

Methods	Training error (%)	Test error (%)
max	2.67	29.5
avg	12.50	28.8
sps	5.65	24.2
sp+max	6.54	26.4
sp+avg	8.26	25.1

Based on the above results it can be seen that: the average pooling method is better than the max pooling methods, because of the local regions average values are more representative. Super-pixel pooling method is better than the standard pooling method, which is due to better image segmentation. The super pixel smooth pooling method gets the best results, because this method can find the most representative values.

5 Conclusions

A neoteric and universal approach is proposed for improving CNN performance through using the super-pixel pooling to train the network. This approach makes the models have more stable characterization and better generalization, and it can be used for different CNN network structures. Extensive experiments on several standard data sets for the image classification prove that using the super-pixel pooling in the training process can significantly enhance performance of CNN models, in comparison with the same model trained without employing this method.

The super-pixel segmentation technique may be further used for the convolution operation. By introducing the fuzzy segmentation method, adjust the number of extra pixels and the number of pixels per super-pixel contains after segmentation. This method is suitable for convolution of large-size images, and convolution of image sets containing different sizes of images. Moreover, we can use the improved CNN to resolve many practical problems such as pedestrian detection, ship classification, text recognition or medical images processing.

References

1. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, vol. 25, pp. 1097–105. Curran Associates, Inc., Lake Tahoe (2012)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, pp. 580–587. IEEE (2014)
4. He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
5. Szegedy, C., Liu, W., Jia, Y.Q., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, pp. 1–9. IEEE (2015)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. http://arxiv.org/abs/1409.1556. Accessed 16 May 2016
7. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectier nonlinearities improve neural network acoustic models. In: Proceedings of ICML Workshop on Deep Learning for Audio, Speech, and Language Processing, Atlanta, USA. IMLS (2013)
8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, France, pp. 448–456. IMLS (2015)
9. Shi, J., Malik, J.: Motion segmentation and tracking using normalized cuts. In: Sixth International Conference on Computer Vision, pp. 1154–1160 (1998)
10. Jung, E.S., Ranka, S., Sahni, S.: Bandwidth allocation for iterative data-dependent e-science applications. In: EEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 233–242 (2010)