Image classification method based on local receptive field extended S-Mobilenet model

Hongmei Liu, Chuanwu Tan*

Hunan Railway Professional Technology College, Zhuzhou, 412001, China

*corresponding author e-mail: 229522351@hnrpc.edu.cn

Abstract. Aiming at the problem that the lightweight deep neural network mobilenet will reduce the classification accuracy, the hollow convolution kernel is introduced into a convolution layer of mobilenet model, and a S-Mobilenet model based on local receptive field expansion is proposed. The model is divided into three structures according to the location of the hole convolution kernel. Without increasing the number of parameters, it can expand the local receptive field of the convolution kernel and improve the classification accuracy. The experiment was carried out on caltech-101 data set, caltech-256 data set and animal classification database of Tubingen University. The results show that S-Mobilenet model can obtain better classification accuracy than Mobilenet, which can be improved by 2% at most.

Keywords: Image classification; Deep neural network; MobileNet; Cavity convolution; S-MobileNet.

1. Introduction

With the continuous progress of science and technology and the continuous development of Internet technology, the number of images produced every day has grown exponentially. Image classification, as the basis of computer vision tasks such as image recognition and image segmentation, has been widely studied. Traditional image classification methods focus on feature extraction and classifier selection. Most of the features extracted by this method are artificially designed, and the effect will have some deviation. In recent years, deep learning has made great achievements in the field of image classification and has become the main research content of image classification. However, the image classification method based on deep learning has the disadvantage of memory intensive and computation intensive, so it cannot be better applied to mobile devices with low memory. Therefore, the network model needs to be compressed and accelerated. In this paper, an improved model structure based on mobilenet lightweight network is proposed. By improving mobilenet lightweight network, parameters can be reduced or classification accuracy can be improved. The convolutional neural network model has strong assumptions about natural images, namely statistical smoothness and local correlation. Convolution operation can effectively reduce the learning complexity of the network model, and make the network connection and weight parameters less, which is easier to train than the fully connected network of the same scale. The image classification technology based on Convolutional Neural Networks (CNN) has achieved good results in many fields of Internet application and will have a good application prospect.

In recent years, great progress has been made in the research of compression and acceleration of deep neural networks in the field of deep learning. These methods can be roughly divided into the following categories: low rank factorization, parameter pruning and sharing, knowledge distillation and special compression networks. The low order factorization technique uses matrix/tensor factorization to estimate the information parameters of deep convolutional neural networks. Parametric pruning and sharing method is used to study the redundancy of model parameters and try to eliminate redundant and unimportant parameters. Learning a distillation model, the useful information extracted from a complex network with better performance but more redundant parameters is transferred to a smaller and simpler network, so that the simple network has similar performance to the complex network. Several special lightweight networks, such as Heatmap, MobileNets and ShuffleNet, use special network models to achieve less parameters and computation than the original network, while losing less or no loss of classification accuracy.

2. Basic theory of convolutional neural networks

2.1 Convolutional neural network

Convolutional neural network mimics biological neural network and adopts the core weight sharing network structure, which enables it to adjust the network model magnitude by changing the depth and width of the network. Convolutional neural network, as a machine learning model for automatic feature extraction, is the earliest deep learning network model applied in various fields. Convolutional neural network reduces the number of weights and the complexity of network model through local acceptance domain and weight sharing, and avoids the over-fitting problem caused by too many parameters in general deep learning models. At the same time, convolutional neural network can directly take the image as the input of the network, avoiding the complex feature extraction process in the traditional image classification method, and has a high degree of invariance to the image translation, scale, etc.

The general convolutional neural network can contain five network layers: input layer, convolution layer, down-sampling layer, full-connection layer and output layer, as shown in Figure 1. The neurons in each layer are made up of three dimensions: width, height and depth. Width and height refer to the size of neurons, and depth refers to the number of channels or output feature maps in the previous layer. Convolutional neural network can not only control the scale of the whole network well, but also realize the robustness of the network to the deformation of the recognition image such as displacement, scaling and deformation through local connection, weight sharing and pooling operations in the structure. Two-dimensional images can be directly as the convolution of the input data and pool layer covering the structure of the feature extraction, and to extract the feature dimension reduction will continue convolution layer all output characteristics, arranged in a one-dimensional vector diagram, characteristic vector, as the input connection, and then after a few full connection layer and output layer connection, the final output of the final classification result. The network adjusts weight parameters by back propagation algorithms that minimize residuals.

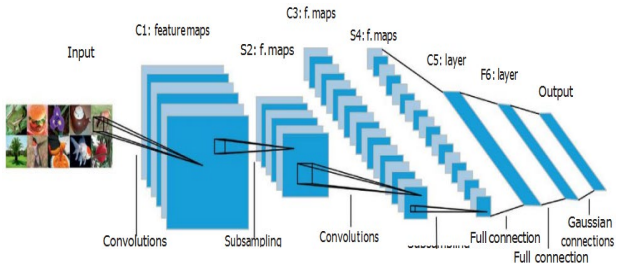


Figure 1. Convolutional neural network architecture

The convolutional layer extracts different features of images through convolution operation. This layer contains several sets of learnable parameters (convolution kernel), which is the core of the whole convolutional network. The convolution of the current layer performs convolution operation on the input feature image to extract the local features and obtain the feature image. Then the nonlinear feature mapping image can be obtained through nonlinear operation of the activation function. The local receptive field and weight sharing ideas of convolutional neural networks are embodied in the convolutional layer, and most of the computation in the network is also concentrated in the convolutional layer.

The down-sampling layer, also known as the pooling layer, uses a specific value as the output for a certain size area through the down-sampling operation. In short, the lower sampling layer "compresses" into a new image by removing the minor part of the feature graph output by the convolution layer of the previous layer, and extracts the main features in the feature graph. After the pooling layer, compared with the original image, the pooling feature map is greatly compressed, which reduces the network scale and further avoids the over-fitting problem. The pooling layer is different from the convolution layer in that it does not generate new training parameters. There are many forms of pooling, such as maximum pooling, average pooling, spatial pyramid pooling, norm

pooling and logarithmic probability pooling, etc. The most commonly used pooling is maximum pooling.

2.2 MobileNet network

MobileNet network is a lightweight CNN network focusing on mobile terminals or embedded devices. Compared with traditional neural network, it greatly reduces model parameters and computation on the premise of slightly reduced accuracy, which can be proved in the calculation of the following formula. The basic unit of MobileNet is deep Wisese parable convolution. As shown in Figure 2, this operation can be divided into two operations-depth wise convolution and Point wise convolution. In fact, pointwiseconvolution is common convolution, because the traditional convolution kernel is 3×3.However, 1×1convolution is adopted in pointwiseconvolution. Compared with the standard convolutional network, N DK×DK×M standard convolutional kernels (FIG. 3 (a)) are replaced by M DK×DK×1 deep convolutional kernel (FIG. 3 (b)) and N 1×1×M point convolutional kernels (FIG. 3 (c)), so the core depth separable convolution of MobileNet network can greatly help reduce the computation. To make it a much lighter model.

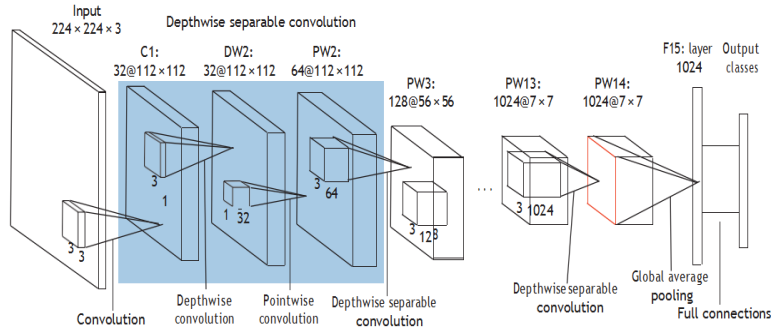


Figure 2. Architecture of MobileNet

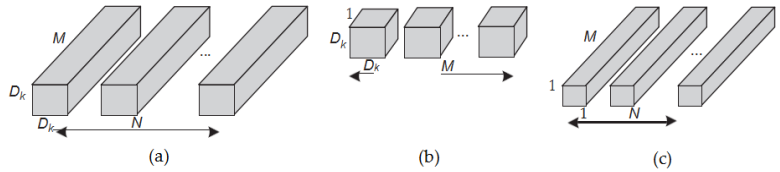


Figure 3. Standard convolution kernel and depth separable convolution kernel

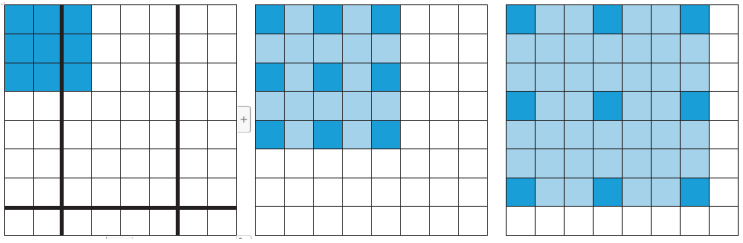
3. D - MobileNet model

In the standard of the convolution kernels of lightweight MobileNets network model has been used in small size 3 x3 convolution kernels, and while a small size of the convolution kernels can use fewer parameters and the amount of calculation, but in the previous layer closest to the input figure is higher, the resolution of the small size of the local receptive field of convolution kernels is too small, although can extract the characteristics of the tiny, But the resolution is too high to capture good features. If the convolution kernel is replaced by a larger one to improve the local receptive field of the convolution layer, the number of parameters and computation of the network model will be increased. Since empty convolution can increase the local receptive field of the convolution layer without increasing parameters, it can be considered to replace the standard convolution kernel with the empty convolution kernel with an expansion rate of 2 in the first several convolution layers. This model is called the S-mobilenet model.

3.1 Empty convolution

The void convolution kernel [6], also known as the holed convolution kernel, is a convolution kernel that inserts zero values in the middle of the non-zero values of the convolution filter. The empty convolution kernel can enlarge the receptive field of the convolution kernel without increasing

the number of network model parameters and computation. Empty convolution first applied in the field of image segmentation, because the image segmentation tasks need pixel level category forecast, the output image size with the size of the input image, the depth of the traditional neural network to convolution of the input image and pooling operation, including pooling layer will reduce the size of feature maps, and enlarge the output characteristics of the feeling of the pixels in the figure, Makes the output characteristics of each pixel in the figure reflects the larger receptive field information, then after pooling reverse convolution operation to expand the characteristics of image in figure size, same as the size of the input image, the characteristics of the image size decrease and the process of expanding again, there is the loss of information, makes the network cannot achieve good results, In order to enlarge the receptive field without causing information loss, empty convolution is proposed. Void convolution can not only avoid the problem of information loss caused by the reduction of spatial resolution of feature maps by the pooling layer, but also increase the receptive field like the pooling layer. The perforated convolution kernel inserts zero values in the non-zero values of the convolution kernel to expand the receptive field of the convolution kernel. The size of the expanded receptive field is controlled by the hyperparameter expansion rate, as shown in FIG. 4. FIG. (a) represents the receptive field of the standard 3×3 convolution kernel. FIG. (b) shows that when the expansion rate is 2, the receptive field of 3X3 convolution kernel without filling is 5×5. FIG. (c) shows that when the expansion rate is 3, the receptive field of 3X3 convolution kernel without filling is 7×7. Thus, it can be seen that empty convolution can expand the receptive field of the convolution kernel through the expansion rate, and will not increase the number of parameters of the convolution kernel.



(a) (b) (c)

Figure 4. Empty convolution kernel

3.2 S-mobilenet modelt

Receptive field refers to the area size of each pixel in the output feature map of each layer in the convolutional neural network mapped to the input original image. Only the pixels in this area of the original image will affect the value of this pixel, which is independent of the pixels in other regions. As convolution convolution and pooling layer in the neural network of connections between neurons is local connection, not directly perceived before all information in a layer of output characteristic figure, but as in the network layer of the later is the characteristics of the previous output figure of local connection, the pixel values on the layer of the output in figure mappings on the input image region, the greater the That is, the larger the receptive field is, the larger the range of image information is perceived, and the closer it is to the global receptive field. The smaller the receptive field of a neuron is, the smaller the receptive field is, and the larger the receptive field is, the higher the semantic level is. Based on the expansion of local receptive field to improve the classification accuracy of MobileNets, due to the characteristics of pixels in this picture near the output layer of receptive field close to universal, then small receptive field increases, the extracted features effects were similar, therefore, should be near the through hole in the input layer of the convolution convolution to increase the local receptive field of the convolution layer, Extract better features. In this chapter, according to the difference of the convolution layer where the empty convolution kernel resides, three improved network models are proposed, namely S1-Mobilenet, S2-Mobilenet and S3-Mobilenet.

S1-mobilenet will set the step size of MobileNet's first convolutional layer to 1, and replace the standard convolutional kernel with an empty convolutional kernel with an expansion rate of 2.

Compared with MobileNet, the size of the feature graph output by the convolutional layer of the first layer changes from 112×112 to 224 ×224, as shown in FIG. 5. S2-mobilenet is in the deep convolution layer of MobileNet's second deep detachable convolution layer, the standard convolution kernel is replaced by an empty convolution kernel with an expansion rate of 2, and other layers remain unchanged. In this way, neither the amount of computation nor the number of parameters is increased, nor the size of the output characteristic graph of any layer is changed, as shown in FIG. 6. D3-mobilene sets the step of the first convolutional layer of MobileNet as 1, replaces the standard convolutional kernel with a hole convolutional kernel with an expansion rate of 2, and adds a pooling layer with a step of 2 after the Batch normalization layer [14]. Other layers remain unchanged. See Figure 7.

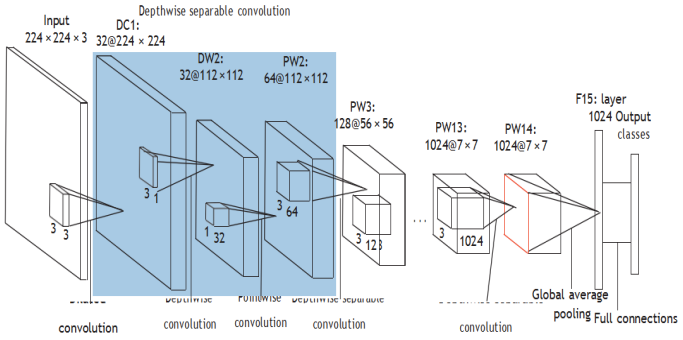


Figure 5. S1-mobilenet network structure

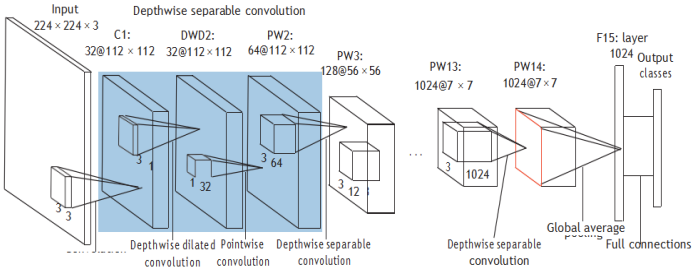


Figure 6. S2-mobilenet network structure

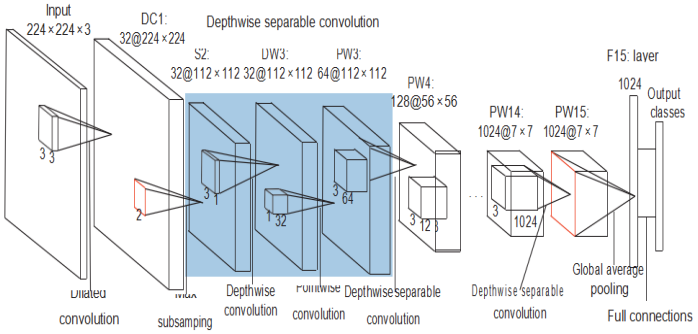


Figure 7. S3-mobilenet network structure

4. Experiments and Result Analysis

4.1 As one of the most common visualization methods at present

Heatmap is widely used in various big data analysis scenarios due to its rich color changes and vivid information expression. Therefore, this paper chooses to draw Heatmap in the process of testing the model. After the model is trained, we call the model to test the data of the verification set. After the test, the results are displayed on the Heatmap. In order to prove the effectiveness of the improved Mobilenet model, this experiment compared the classification results of S-Mobilenet and Mobilenet models on Caltech-101[7], Tubingen animal classification database and Caltech-256 [8] dataset. According to the number of training samples, set different epoch to reduce the learning rate. Weight

initialization using the Xavier initialization method, can determine the random distribution range of parameters according to the number of input and output of each layer. It is a uniform distribution with zero initial deviation. In order to verify the effectiveness of the improved model, the experiments were divided into five groups, and the results were analyzed and compared under the premise of the same operating environment and hyperparameter values. The first group used a Heatmap, a lightweight neural network, to classify the images. The second group used the standard MobileNets neural network structure to classify the images. The third group uses the improved S1-Mobilenet neural network structure for image classification. The fourth group used the improved S2-Mobilenet neural network structure to classify images. The fifth group used the improved S3-Mobilenet neural network structure for image classification.

4.2 Analysis of experimental results

Table 1 shows the comparison of classification accuracy of five classification methods on caltech-101 data set. As can be seen from Table 1, after 40,000-60,000 iterations, the accuracy of the four classification models has reached a balance. The accuracy of MobileNet model and the three improved models is about 25% higher than the Heatmap model, and the accuracy of the three improved models is about 0.5%-2.7% higher than the MobileNet model. Among them, s1-Mobilenet model can improve 0.91%, S2-Mobilenet model can improve 1.22%, S3-Mobilenet model can improve 2.51%, and the final classification accuracy is 78.71%.

Table 1. Comparison on caltech-101 data sets

Model	Number of iterations				
	40000	45000	50000	55000	60000
Heatmap	53.5	53.5	53.45	53.41	53.48
MobileNet	76.71	76.67	76.64	76.72	76.63
S1-mobilenet	77.41	77.45	77.52	77.48	77.48
S2-mobilenet	77.65	77.81	77.76	77.69	77.78
S3-mobilenet	78.64	78.61	78.57	78.56	78.71

Table 2 compares the accuracy of five classification methods on caltech-256 data set. As can be seen from Table 2, the accuracy of the five classification models has improved after 40,000-60,000 iterations, while the accuracy of the other four groups is balanced. The accuracy of MobileNet and the three improved models is much higher than that of Heatmap model, and the accuracy of the three improved models is about 0.6%-1.9% higher than that of MobileNet model. Among them, s1-Mobilenet model can improve 1.97%, S2-Mobilenet model can improve 0.63%, S3-Mobilenet model can improve 1.88%, and the final classification accuracy is 65.91%.

Table 2. Comparison on caltech-256 data sets

Model	Number of iterations				
	40000	45000	50000	55000	60000
Heatmapt	41.49	43.03	43.37	43.59	44.06
MobileNet	64.49	64.59	64.54	64.65	64.51
S1-mobilenet	65.76	65.72	65.81	65.91	65.85
S2-mobilenet	66.18	66.04	65.93	65.82	65.91
S3-mobilenet	64.93	64.94	64.89	65.12	65.15

Table 3 compares the classification accuracy of five classification methods in the dataset of Uebingen Animals. As can be seen from Table 3, after 40,000-60,000 iterations of the five classification models, the accuracy of Heatmap is improved to a certain extent and finally reaches a balance of 74.12% at about 60,000 iterations. The other four groups were balanced, with little change in accuracy. The accuracy of MobileNet and the three improved models is much higher than that of Heatmap. The accuracy of improved S-Mobilenet is about 0.3-0.92% higher than that of MobileNet. Among them, s1-Mobilenet model can be improved by 0.94%, S2-Mobilenet model can be improved

by 0.5%, S3-Mobilenet model can be improved by 0.92%, and the final classification accuracy is 92.80%.

Table 3. Comparison on a dataset of Uebingen animals (21)

Model	Number of iterations				
	40000	45000	50000	55000	60000
Heatmap	72.64	72.3	73.28	73.51	74.12
MobileNet	91.61	91.66	91.63	91.59	91.66
S1-mobilenet	92.48	92.44	92.51	92.39	92.42
S2-mobilenet	92.01	92.05	92.09	92.05	92.01
S3-mobilenet	92.86	92.73	92.81	92.78	92.80

5. Conclusions

In this paper, we introduce a grid classification method based on the S-Mobilnet model of local receptive field expansion. Based on the property that the local receptive field of the convolutional kernel can be increased without adding parameters, the local receptive field of the convolutional kernel of the convolutional immaturus aurantii is introduced into a convolutional layer instead of the common convolutional kernel to enlarge the local receptive field of the convolutional kernel. Based on MobileNets, three sub-models of S-mobilenet model are proposed, and the performance of the improved model is analyzed. Experimental results show that the improved S-mobilenet has better classification accuracy in experimental data sets.

Acknowledgements

This work is supported by Zhuzhou Social Science project (Grant Nos. ZZSK2021151), Natural Science Foundation of Hunan Province (Grant No.2020JJ7052), Construction Project of Railway Communication and Information Professional Resource Database of Hunan Province (Hunan Finance Education Guide (2019) 49).

References

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[2] Wang Naiyan, Yeung D Y. Learning a deep compact image representa- tion for visual tracking [C]/ / Proc of the 26th International Confe- rence on Neural Information Processing Systems. [S.l.]: Curran Associates Inc., 2013:809-817.

[3] Wan Ji,Wang Dayong, Hoi S C H, et al. Deep learning for content- based image retrieval: a comprehensive study [C]/ / Proc of the 22nd ACMInternational Conference on Multimedia.New York:ACMPress, 2014: 157-166.

[4] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[5] Iandola F N, Han S, Moskewicz MW, et al.Heatmap: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size [EB/ OL]. (2016-11-04). https: //arxiv.org/abs/1602. 07360.

[6] Livny Y, Yan Feilong, Chen Baoquan. Automatic reconstruction of tree skeletal structures from point clouds [J]. Association for Computing Machinery, 2010: article No.15 1.

[7] Howard A G, Zhu Menglong, Chen Bo, et al.MobileNets: efficient con- volutional neural networks for mobile vision applications [EB/OL]. (2017-04-17). https: //arxiv.org/abs/1704. 04861.

[8] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]/ / Proc of the 32nd In- ternational Conference on International Conference on Machine Lear- ning.2015: 448-456.