Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/14/20230761

Research on optimization technology based on mobile terminals convolutional neural networks

Jiahao Zhong

HD NINGBO SCHOOL Ningbo 315000 China

1901261@stu.hdschools.org

Abstract. Convolutional neural networks play a very important role in computer vision, such as image classification, image segmentation, and handwriting recognition have been widely used. In daily life, this technology is used in the photo recognition of e-commerce platforms. However, the timing of the identification process grows into a major problem. Therefore, it is particularly important to reduce the recognition time by optimizing the deep learning model. To solve this problem, two experimental methods are proposed to optimize the volume of the convolutional neural network model. The first is to reduce the size of the model by scaling down the convolutional kernel. The second is to prune the model with L-1 norm to reduce the size of the model and improve the running speed. According to the experimental results, the two experimental methods have achieved remarkable optimization effects. In the first experiment, the method of scaling down convolutional kernel has an important optimization effect for training the deep learning model of small data sets. In another experiment using L-1 pruning algorithm greatly improves the running speed of models by reducing the size of models. To sum up, the optimization method proposed above for the convolutional neural network model on the mobile end can be applied in the field that requires a large amount of image classification, such as delivery package sorting. At the same time, to better improve the performance of the model, it will become feasible to use a variety of optimization methods to tune it.

Keywords: convolutional neural network, deep learning, optimization technique, L-1 pruning.

1. Introduction

Nowadays, AI is becoming more and more popular in different fields of work. Deep learning is becoming more and more important. Convolutional neural network is one of the deep learning models which is widely used. Different from other neural networks, such as cyclic neural networks, used in speech recognition [1]; recursive neural network is used in text classification[2]. Convolutional neural network is often used to train a large number of sample data, and plays a huge role in the field of handwritten text recognition [3] and image classification. With the development and innovation of convolutional neural networks in deep learning, more and more mobile software began to use these deep learning models. The popularization of convolutional neural networks also makes people more dependent on these technologies. At the same time, the application fields of convolutional neural networks gradually began to diversify, and the computational amount of convolutional neural networks became larger than before. In this era when mobile terminals (such as mobile phones and laptops) are the mainstream communication devices, the deep learning model that consumes less memory becomes more important because smaller memory means faster running speed and higher efficiency. Meanwhile, the RAM usage of the whole system will also decrease, and the multitasking capability will be improved accordingly. Therefore, it is an effective method to reduce the amount of computation while maintaining accuracy as much as possible through optimization techniques. From LeNet-5 convolutional neural network model [4] proposed in 1998 to various models today, the main structure of most convolutional neural network models is shown in the figure below: Based on this structure, convolutional neural networks with various characteristics are derived (as shown in Figure 1).

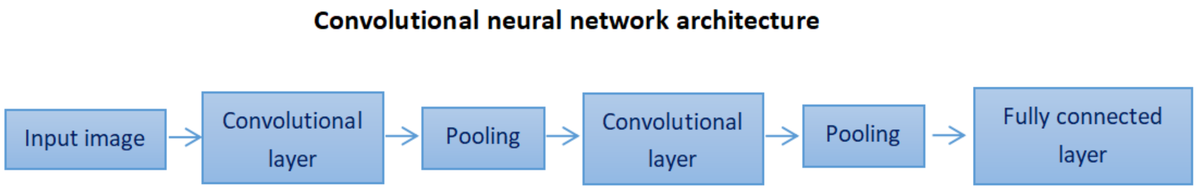

Figure 1. Convolutional neural networks with various characteristics.

With the development of these convolutional neural networks, the accuracy of the models is greatly improved, but the subsequent side effects are reflected in the increase of model size and running speed. As a result, optimization techniques for various fields emerged in an endless stream during this period. For example, as a common structural pruning algorithm, the main principle is to delete redundant channels as the target, and accurately judge the importance of channels [5]. Or for the text detection optimization algorithm, through the mixed pruning method to predict the direction of the picture, in the text area for positioning [6]. This paper mainly covers the task of image classification for small and few categories of training data set, which can be applied to intelligent classification of logistics packages [7], product size qualification detection and other classification work. Taking logistics size sorting as an example, the logistics packages on the conveyor belt are photographed and captured, and then the data set trained according to different needs is forecasted to achieve the effect of intelligent classification. Therefore, this application can greatly improve the efficiency of today's logistics.

The topic of this research is the optimization technology of convolutional neural network based on mobile terminal. The research is mainly divided into two methods. In the first experimental method, the number of convolution kernels is reduced by different proportions for optimization, so as to carry out three sets of comparison experiments, and then through data analysis, the conclusion is drawn. In the second experimental method, the concept of L-1 norm is introduced, and then pruning is carried out through this method. Finally, the model data before and after pruning are compared to get the results.

This paper is divided into five parts: The first part is the introduction, which briefly introduces the background and construction principle of convolutional neural network, as well as its deployment and application on mobile terminal. The second part is the method introduction, mainly introduces the basic methods and experimental reasons of two kinds of mobile convolutional neural networks. The third part is the results and analysis of the experimental data, the analysis of the data obtained by two different experimental methods. The fourth part is the research discussion, mainly put forward the above experimental methods suggestions and limitations. The fifth part is the summary, including the research results and the prospect of the research.

2. Methods

2.1. *Scaling Reduction of Convolutional Kernels*

In the optimization process of convolutional neural network, there are many ways to reduce the calculation amount of the whole model. Firstly, one of the simple methods is to scale down the number of convolution kernels of convolutional layers in the convolutional neural network. For a data set with a small data size for training, excessive number of convolution kernels will significantly increase the amount of computation and operating memory. Therefore, scaling down the number of convolution kernels can not only greatly reduce the amount of computation and the number of parameters, but also maintain the prediction with little change in the accuracy of the original model. In this experiment, AlexNet was used as the model, and the flower data set with a total amount of 443MB was selected as the data set, among which 197MB of training data was selected for model training. Each group of experiments will reduce the number of convolution nuclei in a specific proportion, and the convolution layer size of the last three fully connected layers will remain unchanged. Because the ReLU activation function (1) in AlexNet's fully connected layer makes the value of a large number of neurons zero.

$$f(x) = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases} \tag{1}$$

Also dropout in between each fully connected layer has deactivated a large number of neurons to prevent overfitting. If the convolutional layer size is reduced on this basis, the excessive loss of neurons may reduce the accuracy of the final model training. The first group of experiments uses AlexNet source code. The second group uses the convolutional kernel number reduced to three-quarters based on AlexNet source code to train the data set. The third group is to reduce the number of convolution kernels by half on the original basis, and the data and the structure of convolutional neural network is shown in Table 1 and Figure 2.

Table 1. Data of experiment.

LAYER	KERNEL SIZE	KERNEL NUMBER		
		Exp1	Exp2	Exp3
CONV1	11	64	48	32
MAXPOOL1	3	None	None	None
CONV2	5	192	144	96
MAXPOOL2	3	None	None	None
CONV3	3	384	288	192
CONV4	3	256	192	128
CONV5	3	256	192	128
MAXPOOL3	3	None	None	None
FC1	4096	None	None	None
FC2	4096	None	None	None
FC3	5 (decided by the kernel number)	None	None	None

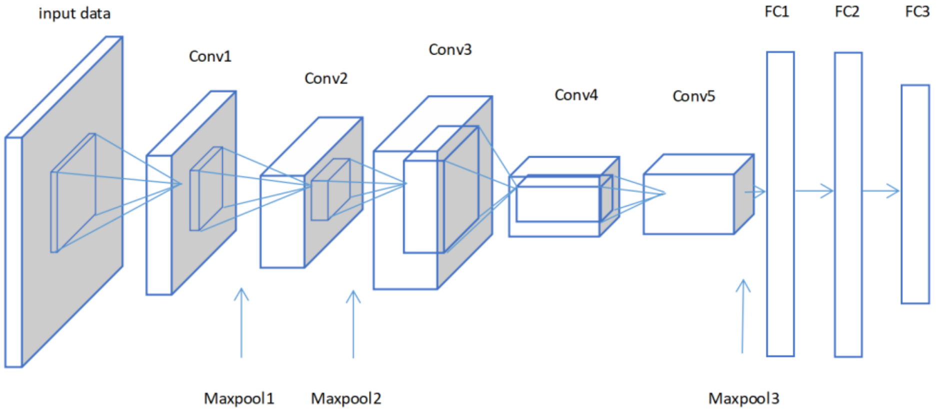

Figure 2. Structure of CNN in experiment.

2.2. *L-1 pruning*

The above steps in experiment show the method of optimizing the size of the model by reducing the number of convolution kernels. In addition, structured pruning of convolutional neural networks is also a mainstream optimization method in the field of pruning optimization.

Structural pruning mainly includes four methods: filter pruning, channel pruning, shape pruning and block pruning. In deep learning, convolution layer and fully connected layer are two main research directions. For pruning used in the convolutional layer, filter-wise will be used. By reducing the number of convolution kernels, the number of output channels will decrease, which will also affect the number of channels in this convolution layer. As the number of these things decreases, the memory footprint of the entire convolutional neural network will decrease, making it better for mobile applications. For a fully connected layer of pruning, the other direction can also reduce program storage. The main principle is to reduce the low weight multiplications in MAC (multiply-add operation) during join computation. For example, each neuron connected to the next layer of neurons will generate a number. If the absolute value of the number that has been calculated is small (*e.g.* 0.01), it has no real effect on the final value and can be ignored (denoted as 0). This greatly reduces the number of connections between neurons in each layer. In this experiment, ResNet-18 was used as the model and L-1 pruning algorithm was used. L-1 pruning algorithm mainly introduces a scaling factor γ into each channel, which is multiplied by the output of the channel. Then the network weights and these scaling factors are trained jointly, and the latter is sparsely regularized. Finally, small factors were used to trim the channels and fine-tune the network after pruning. The formula obtained is as follows (2):

$$L = \sum_{(x,y)} l(f(x, W), y) + \lambda \sum_{\gamma \in \Gamma} g(\gamma) \tag{2}$$

The first half part of formula is calculating the minimum difference of square. The subsequent part is a lasso regression, which means regularizing the absolute value. The scaling factor acts as a proxy for channel selection. Because they are optimized in conjunction with network weights, the network can automatically identify insignificant channels (channels whose calculated weights are small enough not to significantly affect the final output value) and safely remove them without significantly affecting generalization performance. The main operation of the experiment is as follows. First, in order to adapt to the size of the input image, the average pooling is changed to the global average pooling. Then L-1 pruning algorithm was used for pruning with a sparsity of 0.8.

3. Results analysis

In the experiment of scaling down the number of convolution kernels using AlexNet as the model, the experimental data of three groups including basic computation amount, parameter number, loss value after training and accuracy are shown in Table 2 (3*224*224 input features are used to obtain computation amount and parameter number in all three groups of experiments. The code of pytorch is as follows: torch.randn (4, 3, 224, 224). Conduct training with epoch 10 as the standard)

Table 2. Experiment data.

EXPERIMENT	FLOPS	NUMBER OF PARAMETERS	TRAIN LOSS	ACCURACY
EXP1	2.85863G	61.10084M	0.831	0.690
EXP2	1.72453G	50.58774M	0.834	0.681
EXP3	0.78087G	40.37018M	0.823	0.706

By comparing the three experiments, is proved that scaling down the number of convolution kernels can greatly reduce the amount of computation and parameter. In a third set of experiments, the optimized model required nearly 73 percent less computation than the original model, and the number of parameters also dropped by about 34 percent. After the training of the data set, the lowest loss value remained at about 0.83, while the accuracy remained at 70%.

In another experiment, L-1 pruning algorithm was used to prune ResNet-18. The results showed that the computational energy (FLOPs) decreased from 27215.5M to 2628.7M, the number of parameters decreased from 11.16M to 1.18M, and the average CPU running time decreased from 709ms to more than 208ms, which greatly improved the performance of the model.

4. Discussion

According to the result above, the experimental results are roughly the same as the expected results. The main reason is that the pixel size of the input feature map is small, and the number of reduced convolution kernels can still be used to extract key feature data through multiple convolutions. Therefore, deleting redundant data by this method will not affect the final results because of removing those parameters that are useless (shown as Figure 3).

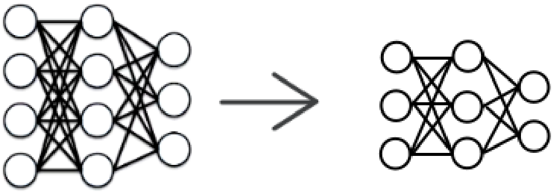

Figure 3. Removing parameters.

However, in the second experiment, the model slimming effect obtained by L-1 pruning was very significant, which was in line with the expected effect. This is because the L-1 operator[8] will converge the weight parameters to 0, and then update the parameters through the gradient descent algorithm[9], and the values of all parameters will be close to zero. Since the gradient descent algorithm will have different gradient signs on both sides of the turning point, all parameters should be regularized from negative to positive and from positive to positive. Then a specific algorithm is used to select a threshold based on the ownership re-weight parameter: keep the parameter larger than the threshold, delete the parameter less than the threshold and change to 0. So, the size of the model was significantly reduced compared to the previous period, but the highest training accuracy was not applied to subsequent model predictions under multiple epoch exercises.

To resolve the problem that the maximum accuracy trained in the second experiment was not being used effectively, the model code should be fine-tuned to record the current training accuracy with each epoch training. If a higher precision occurs in a subsequent epoch training, the current precision overrides the highest precision [10] to ensure that the training results of the highest precision are used. According to the discussion of the above results, the first experimental method is obviously better than the second in the degree of convenience, but in the determination of the reduction scale value, it brings some limitations to this method: Firstly, the degree of reduction depends on the size of the training image. Secondly, there is no specific algorithm to calculate the reduction ratio to achieve the optimal optimization effect, and the relative optimal value can only be selected through multiple debugging and testing.

5. Conclusion

This paper mainly reviews the optimization technology of convolutional neural network based on the mobile terminals. Two experimental results show that the optimization method of scaling down the number of convolution kernels and using L-1 norm has a significant effect on the size optimization of convolutional neural network models. This result can be applied to the field of onboard artificial intelligence. At present, it is difficult for onboard terminals to carry out a large number of computing tasks without specific hardware support, so reducing the size and running speed of AI models can reduce the cost of installing additional processors in vehicles and be more widely deployed in all kinds of vehicles, not limited to household vehicles. In addition, full coverage of application software is another important research goal. Today's deep learning model has been applied to the mobile terminal at scale, but with the increase in complexity, the time to reach the calculation results is getting longer and longer. The popular ChatGPT, for example, takes more than 10 seconds to respond to a complex question. This greatly reduces the user experience, so it is especially important to reduce the run time by reducing the model size.

In the experimental study of this paper, based on the obvious optimization effect of the two experiments, if the two methods are combined, the effect of one plus one may be greater than two, which is more effective than a single optimization method. In conclusion, this expectation will be confirmed in future studies or further studies.

References

[1] Yang Li, Wu Yuqian, Wang Junli, et al. Review of research on cyclic neural networks [J]. Journal of Computer Applications,2018,38(S2):1-6+26

[2] Huang Lei Du Changshun. Text classification based on recursive neural network study [J]. Journal of Beijing chemical university (natural science edition), 2017, 44 (01) : 98-104. The DOI: 10.13543 / j.b HXBZR. 2017.01.017.

[3] Lu Hongtao, Zhang Qinchuan. A review of the application of deep convolutional neural networks in computer vision [J]. Data Acquisition and Processing,2016,31(01):1-17.DOI:10.16337/ J.1004-9037.2016.01.001.

[4] Li Dan, SHEN Xia-jiong, ZHANG Hai-xiang, et al. Convolution neural network algorithm based on Lenet - 5 [J]. Journal of computer age, 2016, No. 290 (8) : 4-6 + 12. DOI: 10.16644 / j.carol carroll nki cn33-1094 / tp. 2016.08.002.

[5] Wang Jiong. Research on Deep Neural network model compression method based on Structured pruning [D]. Nanjing university of posts and telecommunications, 2022. DOI: 10.27251 /, dc nki. GNJDC. 2022.001412.

[6] Zhou Xiangyu, Gao Zhonghe, Zhao Luteyao, et al. Text region detection Algorithm based on Pruning Optimization [J]. Computer and Digital Engineering,2022,50(09):2059-2064.

[7] Li Tianjian, Huang Bin, Liu Jiangyu et al. Convolution neural network object detection algorithm in the application of logistics warehouse [J]. Computer engineering, 2018, 44 (6) : 176-181. The DOI: 10.19678 / j.i SSN. 1000-3428.0047321.

[8] Song Yu, Sun Wen Yun. Edge Preserving image smoothing Algorithm Based on Reweighted l1 norm [J]. Journal of South China University of Technology (Natural Science Edition), 201,49(06):109-121.

[9] Pan Jinjie, Lin Dajun, Luan Haitao. A broadband light source of phase reconstruction method based on neural network [J]. Applied laser, and 2022 (6) : 137-143. The DOI: 10.14128 / j.carol carroll nki. Al. 20224206.137.

[10] Du Xiaohao. Training and Structure optimization method for Convolutional Neural networks and its application [D]. Lanzhou university, 2022. DOI: 10.27204 /, dc nki. Glzhu. 2022.003431.