Study on the optimization of CNN based on image identification

Feng Yanyan
College of Electronic Engineering
Guangxi Normal University
Guilin, China
e-mail: 1095115523@qq.com

Zeng Shangyou
College of Electronic Engineering
Guangxi Normal University
Guilin, China
e-mail: zsy@mailbox.gxnu.edu.com

Yang Yuanfei, Zhou Yue, Pan Bing
College of Electronic Engineering
Guangxi Normal University
Guilin, China
e-mail: 287852761@qq.com, 744427475@qq.com, 276696163@qq.com

***Abstract*—The feature extraction method of traditional image classification is difficult to deal with complex image problems. Although feature detector based on convolutional neural network can easily extract image features, many current network models have poor recognition accuracy and too many parameters. This paper proposes a multi-scale dual-channel dimension reduction module (DR module) to extract image features. Based on the AlexNet model, a deep global optimization model (GONET model) is proposed by exploiting the DR module, dropout and global pooling strategy. Compared with the AlexNet model, this model has better recognition performance. The accuracy on the Caltech256 dataset reaches 58.8% with GONET model, exceeding that of the AlexNet model by about 4.0%. The accuracy on the 101_food dataset reaches 69.0% with GONET model, exceeding that of the AlexNet model by about 8.9%. The experimental results show that the GONET model has superior image recognition effect, and it significantly reduces the network parameters while improving the recognition accuracy.**

Keywords-Cnn; Alexnet; dimension reduction; feature extraction; the global pool

I. INTRODUCTION

Artificial neural network is proposed and developed on the basis of modern neuroscience. It aims to design an abstract mathematical model that reflects the structure and function of human brain. Since the emergence of the MP model, significant advances have been made in the research of artificial neural networks. The Convolutional Neural Networks (CNN) [1] is a branch of the artificial neural network. Its neural network structure has multiple hidden layers. and forms a more abstract high-level representation attribute or feature by combining low-level features. It is the current hot spot in the field of speech analysis and image recognition.

Deep Learning [2] is a new field in the study of machine learning. LeNet is the most representative network model of early neural network structure, which was designed and proposed by Yann LeCun in 1998. It is mainly used to identify handwritten digits. Its network structure is simple with convolutional layers, pooling layers and fully connected layers. The three layers are the basic components of modern CNN networks. LeNet is the basis of the network model for small data volume, simple image recognition and classification. AlexNet [3] is the basis of the network model for large data amount and complex pictures. Followed by more complex network models such as GoogLeNet[4], VGGNet[5], Siamese[6], SqueezeNet[7] models are all inseparable from these basic network layer components, so the AlexNet model lays the foundation for deep learning in computer vision and has great research significance.

This paper proposes a dimension reduction and depth global optimization model named GONET based on AlexNet model. Compared with AlexNet model, this model has stronger feature expression ability, higher recognition accuracy, fewer neurons and less storage requirements. Moreover, it presents better performance with fewer resources and calculation.

II. DIMENSIONAL REDUCTION AND DEPTH GLOBAL OPTIMIZATION MODEL

A. The principle of the model

The CNN training process is divided into forward and backward propagation. Forward propagation is used for extracting feature and obtaining image features through a series of transformations such as convolution, down sampling, and so on. Backpropagation uses the traditional BP mechanism to propagate the error forward by layer and use the chain derivation to update the convolution kernel (ie, weight and offset). In the application of machine learning, the convolutional input is usually a multi-dimensional array of data, and the parameters of the convolution kernel are the multidimensional arrays optimized by a learning algorithm. For example, a two-dimensional image I is taken as the input and a two-dimensional kernel K is employed

$$F(i,j) = (K*I)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n), \quad (1)$$

where (i,j) denotes the position index of the picture, m and n represent the distance of the picture translation, and $F(i,j)$ is the output of the convolution layer.

In general, the expression of network features is very vague in the process of performing element-by-element operations in the convolution layers. By introducing non-linear factors and adding an activation layer[8] below

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00041

convolution layers can make up for the insufficiency of the linear network feature expression capability.

Activation layer, the activation function is the commonly used ReLU function. The formula can be described by

$$f(x) = \max(0, x), \tag{2}$$

where x denotes the input and $f(x)$ is the output.

Batch normalization Layer (BN)[9] solves the problem of changing the distribution of data in the middle layer between network convolution layers. The input data is normalized to a mean of 0 and the variance is 1. Access the BN transform after the activation layer can reduce the reliance on initialization and improve training speed. The formula is given by

$$\begin{aligned} \mu_B &\leftarrow \frac{1}{m}\sum_{k=1}^{m} x_k \\ \sigma_B^2 &\leftarrow \frac{1}{m}\sum_{k=1}^{m}(x_k - \mu_B)^2 \\ \hat{x}_k &\leftarrow \frac{x_k - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \\ y_k &\leftarrow \gamma \hat{x}_k + \beta \equiv BN_{\gamma,\beta}(x_k) \end{aligned}, \tag{3}$$

Down sampling layer is also known as the pooling layer. The commonly used pooling layers[10] include the largest pooling, average pooling, etc. Pooling synthesizes all the feedback within the adjacent region by synthesizing the k-pixel statistical characteristics of the pooled region instead of a single pixel. For the next layer, the input parameters are reduced by about k times, and the reduction in the input size not only improves the statistical efficiency but also reduces the storage requirements for the parameters. The downsampling formula is described by

$$f(x) = w \cdot \mathrm{down}(x) + b, \tag{4}$$

where w represents the weight, x represents the input, b represents the offset, down(.) represents the down sampling function, and $f(x)$ is the output of equation (2).

Fully connected layer is a process of linear feature mapping which maps all learned distributed feature connections to the sample markup space through matrix vector product operations. The formula is described by

$$f(x) = wx, \tag{5}$$

where w is the weight vector matrix, x is the parameter matrix of the input neuron, and $f(x)$ is the output matrix of the fully connected layer.

After the convolution, activation and pooling layers, all acquired features are connected by a fully connected layer and the output values are passed to the Softmax classifier.

B. Model analysis

In the convolution feature sampling process of traditional CNN network, only one kind of convolution kernel is used in the convolution layer so that the sampling information is single, which easily leads to the loss of characteristic information. This paper proposes a multi-scale dual-channel dimension reduction module (DR module) to replace a single 3*3 convolution as well as reduce convolution parameters at the same time to obtain more detailed feature information. The module is shown in Figure 1.

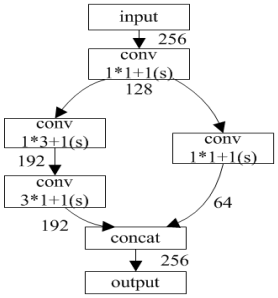

Figure 1. DR module

In order to reduce the number of neurons, it use 1*1 con volution to reduce the picture thickness before the DR modu le enters dual-channel, and then use two channels for feature extraction. Channel 1 uses 1*3 and 3*1 convolution as the e quivalent of 3*3 convolution. To increase the diversity of co nvolution kernel without greatly increasing network paramet ers, channel 2 uses a layer of 1*1 convolution to extract feat ures. The feature maps extracted from the two channels are combined into a new set of feature through cascading operat ions.

For example, taking the input and output as 256 feature maps. The DR module uses a 1*1 convolution to reduce the input feature map thickness by half, and then uses multi-scal e dual-channel feature to extract feature. Channel 1 uses 64 1*1 convolution kernels convolution sampling, channel 2 us es 192 1*3 convolution kernels and 192 3*1 convolution ker nels for convolution operations. Lastly the two-channel outp ut map uses cascade operations to form a new signature. The convolution parameter is 256*1*1*128+128*1*3*192+192 *1*3*192+128*1*1*64=255280. When the conventional 3* 3 convolution kernel has 256 feature maps for input and out put, the convolution layer parameter is 256*3*3*256=58982 4. By contrast, the convolution layer parameters have been g reatly reduced.This article uses the DR module to design thr ee network models.

DRNET, the DR module is used instead of the 3*3 convolution layer in the AlexNet network. After the improvement, the types of network convolution kernels are increased, and the computing complexity of width and local

sensing field is also increased. Moreover, the extracted features are more abundant.

DONET, generally, the bigger the convolution kernel, the larger the field of perception[11]. Accordingly, you can see more information in the picture and get more features. Large convolution kernel means that the computer needs to bear more computing load, which is not conducive to the increase the depth of network model. In this paper, under the condition that the image size and the number of output feature graphs are not changed, the deep optimization is carried out by using the equivalent effect of the 11*11 convolution kernel and the series 7*7 and 3*3 convolution kernel. Replace the 11*11 convolution layer with 7*7 and 3*3 convolution kernels, and then replace the 3*3 convolution layer with DR module. Compared with AlexNet model, DONET model has a deeper network level, more kinds of convolution kernels and more comprehensive information collection.

GONET removes three fully connected layers of the DONET model and replaces it with 1*1 convolution and global pooling[12], and then transforms the full connected structure into a sparse connected structure. Finally the global optimization model (GONET) is obtained. The architecture is shown in TABLE Ⅰ. The classification problem of CNN is generally connected with the full connection layer after quantifying the pixel value of the feature graph of the last convolution layer, and classifywith a Softmax layer. However, due to the excessive number of parameters in the full-connection layer network, overfitting is likely to occur during training. In terms of hardware implementation, too many parameters will result in lower distributed training efficiency, which imposes a heavy burden on the network bandwidth required for data transmission. It can also cause CNN to be difficult to operate on limited-capacity FPGA hardware. The GONET model removes the full connected layers, uses dropout after Maxpool6 to prevent overfitting, and replaces it with 1*1 convolution and global pooling.In addition, it is capable of converting the full connection structure into a sparse connection structure to improve the generalization ability of the network and reduce network parameters.

TABLE I. GONET STRUCTURE

Layer name	Output size	Filter size/stride	pad
Input image	224*224*3		
Conv1	111*111*96	7*7/2	0
Maxpool1	55*55*96	3*3/2	0
Conv2	55*55*256	5*5/2	1
DR module3	27*27*256		0
Maxpool3	13*13*256	3*3/2	0
DR module4	13*13*384		1
DR module5	13*13*384		1
DR module6	13*13*256		1
Maxpool6	6*6*256	3*3/2	0
Conv7	6*6*256/101	1*1/1	0
Avgpool8	1*1*256/101	6*6/1	0

III. EXPERIMENTAL RESULTS

A. Experiment Settings

The framework that the experiment depends on is the caffe[13] deep learning framework. The computer is configured with i7-6700K quad-core CPU, Ubuntu 14.04 operating system, 32GB memory, and NVIDIA-GTx 1070 GPU.

The dataset includes Caltech256 and 101_food. The Caltech256 data set contains 256 categories of pictures. Each type of picture is randomly divided into a training set and a test set is in a ratio of 4:1. 23,919 training pictures and 5,862 test pictures are obtained. The 101_food dataset contains 101,000 food images of 101 categories, 1000 foods of each category, of which 75,750 training images and 25,250 test images.

The image is preprocessed before training. First, the size of all images is scaled to 256*256. Using data amplification techniques during training, 224*224 pixel image blocks are taken from the upper left, lower left, upper right, lower right, and middle of each image respectively, and then flip it horizontally for a total of 10 new images. After data amplification, the size of the entire training set is expanded to 10 times of the original size, and then all images are subtracted from the mean value.

All network batch sizes are set to 50, and each iteration is tested for 500 times. The parameters of the two datasets solver files are set as follows. The initial learning rate is set to 0.005. Multistep algorithm isadopted, gamma is 0.1, display is100, caltech256 attenuated every 24,000 iterations, twice, 101_food attenuated every 40,000 iterations, and three times in total. Caltech256 iterates 60000 times to generate the final model, and101_food iterates 150,000 times to generate the final model, respectively.

B. Analysis of experimental results

In order to verify the performance of the GONET model, the experiments of AlexNet model, DRNET model, DONET model and GONET model are performed on two data sets. Analyze the network training log file and get TABLE Ⅱ and TABLE Ⅲ.

TABLE II. EACH MODEL ON CALTECH256 PERFORMANCE

model	parameters(MB)	accuracy(%)
Alexnet	231.7	52.8
Alexnet+BN	231.7	54.8
DRNET	228.8	55.0
DONET	229.6	57.7
GONET	7.5	58.8

TABLE III. EACH MODEL ON 101 _FOOD PERFORMANCE

model	parameters(MB)	accuracy(%)
Alexnet	229.1	56.7
Alexnet+BN	229.1	59.7
DRNET	226.2	60.6
DONET	227.1	67.0
GONET	7.4	69.0
ZFnet	288.6	56.7

Through the experimental data of the two datasets, the GONET model has better performance than several classic models. It has a higher recognition rate and fewer parameters. In addition, compared with Alexnet, these datas proves that DR module can deduct parameters and rich extraction features. The DONET model has a deeper level and the recognition rate is higher than DRNET model, but the number of parameters increases. In the 5*5 convolution layer, the size of DONET input feature map is 55*55, DRNET is 27*27, and DONET is two time that of the DRNET model, which indicates the increase of the parameters. The GONET model brings a better recognition rate than the GONET model by the dropout strategy and global pooling. At the same time, the greatly reduced parameters make the lightweight network.

The experimental results show that the GONET model proposed in this paper not only improves the recognition rate by 4.0% and 8.9% over AlexNet on two data sets, but also reduces the model parameters by more than 30 times than that of AlexNet. The light weight network is realized, which reduces the difficulty of hardware application to some extent. Moreover, it improves the training efficiency and reduces the burden of data transmission. The final analysis illustrates that the GONET model proposed in this paper has the following characteristics:

1) By using the multi-scale dual-channel DR module, more image hiding features can be extracted than a single convolution, synchronously, certain network parameters can be reduced by controlling the number of output maps of each convolution layer.

2) GONET model is reduced by more than 30 times than the parameter of AlexNet so that the model size is only more than 7 MB. It is convenient for network hardware application and brings better recognition rate than AlexNet.

IV. CONCLUSION

In this paper, a GONET model based on dimension reduction and depth global optimization is proposed according to the mainstream model AlexNet. Performance tests are conducted on two datasets, caltech256 and 101_food. The experimental results show that the performance of GONET model is better than AlexNet, which improves the recognition rate as well as reduces the parameters. The next goal is to further optimize the network and try to reduce the training time. At the same time, testing and optimization should be carried out on a larger data set to compare the performance of the network and increase the generalization ability of the network.

REFERENCES

[1] Chen Y H,Krishna T, Emer J S, et al. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks[J]. IEEE Journal of Solid-State Circuits, 2016, 52(1).

[2] Deng L, Yu D. Deep Learning: Methods and Applications[J]. Foundations & Trends in Signal Processing, 2014, 7(3):197-387.

[3] A.Krizhevsky, I. Sutskever, and G.Hinton. ImagNet classification with deep convolutional neural networks, Alex Krizhevsky etc, NIPS 2012.

[4] C.Szegedy,W.Liu, Y.jia,P.Sermanet,Sermanet,S.Reed, D.Anguelov,D. Erhan, V.Vanhoucke, and A.Rabinovich. Going deeper with convolutions.in CVPR, 2015.

[5] K.Simonyan and A.Zisserman. Very deep convolution networks for large-scale image recognition. In ICLR, 2015.

[6] Learning a similarity metric discriminatively, with application to face verification, Sumit Chopra, Raia Hadsell, Yann LeCun, 2005.

[7] Shafiee M J, Li F, Chwyl B, et al. SquishedNets: Squishing SqueezeNet further for edge device scenarios via deep evolutionary synthesis[J]. 2017.

[8] Van der Schaft, A.J. L2-gain analysis of nonlinear systems and nonlinear state-feedback H∞ control[J]. IEEE Trans.automat.contr, 2016, 37(6):770-784.

[9] Simon M, Rodner E, Denzler J. ImageNet pre-trained models with batch normalization[J]. 2016.

[10] Cherian A, Koniusz P, Gould S. Higher-Order Pooling of CNN Features via Kernel Linearization for Action Recognition[C]// Applications of Computer Vision. IEEE, 2017.

[11] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition[J]. 2014:512-519.

[12] Lin M.Chen Q, Yan S. Network in Network[J]. arXiv preprint arXiv: 1312.4400, 2013.

[13] Yangqing Jia,EvanShelhamer, Caffe Tutorial, http://caffe.berkeleyvision.org/tutorial/, 2016.