Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 3, 2024

Review of Convolutional Neural Network Models and Image Classification

Weiqi Hua[1], Chunzhong Li [1, *], Xinsheng Wang[2]

[1] College of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233030, China
[2] School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China
[*] Corresponding author: Chunzhong Li (Email: czhongli@163.com)

Abstract: With the arrival of the era of big data and the improvement of computing power, deep learning has swept the world. Traditional image classification methods are difficult to deal with the huge image data, and cannot meet the requirements of people on the accuracy and speed of image classification, the image classification method based on convolutional neural network breaks through the bottleneck of the traditional image classification method, and becomes the mainstream algorithm of image classification, how to effectively use convolutional neural network to classify images has become a hot spot of research in the field of computer vision at home and abroad. In this paper, we review the research background, significance and current research status of convolutional neural network model and image classification, study two image classification methods based on ResNet and ShuffleNet, and provide a comprehensive review of the construction methods and characteristics of the two deep convolutional neural network model structures, and finally compare and analyse the performance of the two classification models.

Keywords: Convolutional neural network, Image classification, ResNet, ShuffleNet.

1. Introduction

Image classification is an important research direction in the field of artificial intelligence and an important branch in image processing. It has long attracted the attention of academia and industry, and many research results have been achieved. Image classification is to extract the features of the image, and then according to the category to which the features belong to be classified. Early image classification is to use manual to mark, with the eyes to judge the category of the picture. As you can imagine, this primitive method is very labour-intensive and inefficient, and it is increasingly unable to meet the requirements of the ever-changing new era.

As more and more types of images and more and more complex structures, how to effectively manage images is a very meaningful topic at present. With the advent of the information age, the image classification technology develops as a machine identifies and classifies the objectively existing things by simulating the human visual system. In the past decade or so, as the demand for image processing related technologies has been growing, there have been many significant breakthroughs in the research of image classification. As early as the 1980s, LeGun et al. have proposed the use of convolutional neural networks[1] for image classification. Since then there have been many competitions on image classification, e.g., between 2005 and 2012, including Pattern Analysis, Statistical Modelling, Computational Learning, the Visual Object Class (PASCAL VOC) Challenge, and the ImageNet Large Scale Visual Recognition Challenge[2]. Since Krizhevsky et al. proposed a pioneering AlexNet[3] convolutional neural network model in 2012, the performance of image recognition has been significantly improved, dramatically reducing the error rate of image classification. This was a major advancement in image classification and a pioneering development in the field of deep learning. Since then, various CNN-based image classification methods have continued to appear. To name a few, LeNet[4], GoogLeNet[5], VGGNet[6] and many other convolutional neural network models. The performance of image classification techniques continues to improve, and some methods have even surpassed the human level.

With the rapid development of computers and the great improvement of computing power, deep learning has gradually stepped into our vision. In the field of image classification, the convolutional neural network in deep learning can be very useful. Compared with traditional image classification methods, it no longer needs to manually describe and extract features from the target image, but through the neural network to autonomously learn the features from the training samples, and these features are closely related to the classifier, which is a good solution to the problem of manually extracting features and selecting classifiers[7]. Therefore, it is of great significance to study image classification methods based on convolutional neural network models in various scene applications. Therefore, two image classification methods based on ResNet[8] and ShuffleNet[9] will be discussed and studied in this paper.

2. Relevant Theories

2.1. Basic Theory of Image Classification

Image classification involves extracting features from the input image and categorising them. In image classification, the appropriate dataset is first selected, then the dataset is preprocessed to eliminate the influence of other factors on the features, then the feature information is extracted from the image, and finally, after learning and training, the classification result is finally obtained and assigned a label. The labels are obtained from a predefined set of possible classifications. The main processes of image classification include image preprocessing, image feature description and extraction, and classifier design. Preprocessing includes operations such as image filtering and size normalisation, which are designed to facilitate the subsequent processing of the target image; image features are descriptions of salient features or attributes, and each image has some of its own

features, feature extraction, i.e., according to the features of the image itself, select suitable features and extract them efficiently in accordance with a certain established way of classifying images; a classifier is an algorithm for classifying the target image according to the selected features. Classifier is an algorithm that classifies the target image according to the selected features.

Traditional image classification techniques mainly consist of two parts: feature extraction and classifier learning. Feature extraction is more important than classifier learning in traditional image classification, because when feature extraction does not extract enough feature information, the classification accuracy of the classifier will decrease and the error rate of image classification will become higher. Among feature extraction, feature coding is one of the most studied areas. In feature-based image classification, it can usually be divided into three steps, the first step is the input image, the second step is feature extraction, which processes the input image, and finally outputs the result for classification. This part of feature extraction can be further divided in detail into three stages: descriptor extraction, feature coding and spatial pooling.

With the rapid development of computers and the great improvement of computing power, deep learning has gradually stepped into our vision. In the field of image classification, convolutional neural network in deep learning can be very useful. Compared with traditional image classification methods, CNN-based image classification can automatically learn to extract image features, has a strong feature expression ability, and is an important component of today's image classification tasks. Its classification process consists of preprocessing, feature extraction, learning and training. Among them, its focus part is the training process, which can be divided into forward and back propagation. When training an image, the results obtained from the input image after convolutional layer, pooling layer, and classifier are compared with the target value, after which forward propagation or back propagation is chosen and finally the output is obtained.

2.2. Basic Theory of Convolutional Neural Networks

Convolutional neural networks are similar to traditional artificial neural networks in that they are made up of a number of neurons which are self-optimising and constantly learning. Each neuron is first assigned a feature of the input data and then proceeds to the next operation, and countless neurons form together to form the basis of the neural network.

Neurons in a CNN are typically composed of three dimensions, input height, width and depth. Unlike standard artificial neural networks, the neurons in any given layer of a CNN are connected to only a small portion of the area of the previous layer.The overall architecture of a CNN consists of a convolutional layer, a pooling layer, and a fully connected layer. These layers can be stacked on top of each other, and when stacked together, they make up a CNN architecture. The essence of convolutional neural network is a multi-layer perceptual machine. Compared with fully connected neural network, convolutional neural network can effectively reduce the size of training parameters in the network by setting local receptive fields, weight sharing, pooling layer and other operations, which can greatly reduce the amount of computation and the complexity of the model, and therefore it is especially suitable for use in image recognition and feature extraction tasks as a feature extractor for images.

3. ResNet-based Image Classification

3.1. Feed-forward neural network

Feedforward neural networks are one of the most commonly used function approximation techniques and have been applied to problems arising from a variety of disciplines. Feedforward neural network is a deep learning model with a unidirectional multilayer structure in which each layer packs a number of neurons. Its zero-layer structure is called the input layer, which is the location of the input of the original data, the hidden layer is an intermediate layer or layers after the zero layer, the number of layers of the hidden layer determines the depth of the network, and the hidden layer is followed by the output layer, which represents the classification probability in the classification task.

As shown in Figure 1, this is a typical multi-layer feed-forward neural network, this neural network has four neurons in the input layer, the neurons are connected to each other after passing through the input layer, and then they become five when they reach the first hidden layer, and then after the first hidden layer, the neurons are reduced to three, and the neural network extracts deeper features in these processes, and then finally outputs the classification results through the output layer.

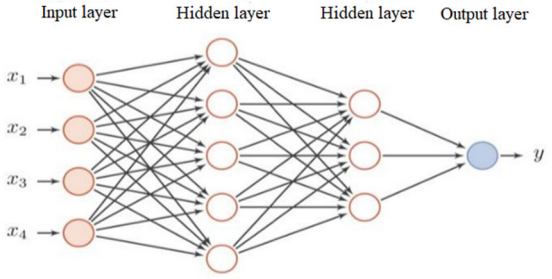

Figure 1. Multilayer feed-forward neural network structure

Feedforward neural network not only possesses the outstanding features of simple structure and wide range of application scenarios, but also it is easy to fit a variety of continuous functions, and its representation ability is very powerful to learn the data laws of any data set. In terms of computation, feedforward neural networks lack rich dynamic behaviour. And from a system perspective, the nonlinear mapping ability of feedforward neural networks is static. However, it has a powerful nonlinear processing capability, the implementation of which is obtained by simple combinatorial mapping of neurons. As far as most of the feedforward neural networks are concerned they are a relatively good learning network with better performance than the usual feedback networks in the field of image classification.

However, in the field of deep learning, feedforward neural networks also have flaws. Feedforward neural networks in the training aspect, using the traditional gradient descent method, which makes the training rate is lower, the training time is longer. At the same time, the learning rate chosen by the model is not flexible enough, and the learning rate is closely related to the performance of the neural network, too big or too small will lead to bad consequences.

3.2. ResNet model

The core concept of ResNet is to add a constant mapping path to the neural network, as shown in Figure 2, where the input data is passed through two consecutive network layers to get a non-linearly mapped output, and added to the original input to get the final residual output. That is, the addition of a constant mapping converts the original function $H(x)$ that needs to be learned into $F(x)+x$, and the hypothetical optimisation of $F(x)$ would be much simpler than $H(x)$, which would be the same for both representations but not the same level of difficulty to optimise. The emergence of this model has allowed the network model depth to be unrestricted over a wide range (currently up to 1000 layers or more) and has had a profound impact on the subsequent development of convolutional neural networks. The idea has gone through a short period of time from its creation to its practical application.

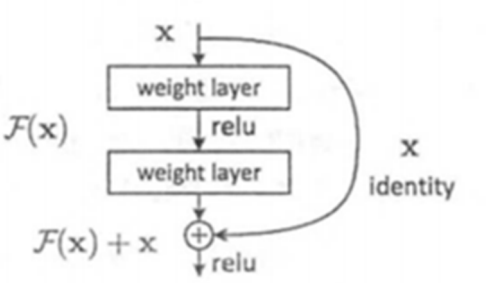

Figure 2. ResNet's residual learning module

In image processing, VLAD (Vector of Locally Aggregated Descriptors)[10] is a representation encoded by a vector of residuals, while Fisher Vector[11] is described as a probabilistic statistical version of VLAD. Residual-based fitting as well as probabilistic distribution modelling is a powerful in image retrieval and classification tasks data shallow characterisation techniques. In vector quantisation, encoding residual vectors will present a more efficient performance than encoding raw vectors directly. In low-level vision tasks, in order to solve Partial Differential Equations (PDES), scientists divide the system into multiple sub-tasks on multiple scales by means of multiple meshes. Or rather, the task is refined step by step to create a different sub-task, each dealing with residual solutions at different scales. Compared to normal solvers, these solutions that employ the residual property converge faster in processing the image, which proves that applying more advanced concepts and using a better processing method can enhance the optimisation well.

While conducting research on residual connectivity, "Highway Network[12]" provides shortcut connections with gating features. This kind of shortcut connections can somehow be beautiful with the idea of residual connections, however, these gating mechanisms are more data dependent and they also have extra parameters, this drawback increases the computational effort and training of the network. On the other hand, when the gating mechanisms are turned off, their network representation has no advantage as in the case of non-residual networks. Unlike this network, the layer-hopping connected constant mapping module in the residual network does not introduce additional training parameters or hyperparameters and is committed to learning the residual function this operation does not ignore the learning of the constant mapping and the Highway Network does not have the advantage of performing as well as the residual network in the deeper network.

4. ShuffleNet-based Image Classification

4.1. Group convolution under channel substitution

Modern convolutional neural networks are generally made up of network layers with the same structure stacked in different ways. In the Xception and ResNeXt models, the 1×1 convolution and deep separable convolution were introduced into the models in order to strike an effective balance between the feature extraction capability of the model and the computational complexity. However, the 1×1 convolution contains significant computational complexity and parameters in both models. For example, in ResNeXt, the only convolutional layer that employs group convolution is 3×3, and the other layers are not equipped for use. So in each residual structure in the ResNeXt model, the 1×1 convolution has 93.4% product. In a small network, the higher complexity of the 1×1 convolution makes the number of feature channels limited and cannot be stacked, which will lead to the performance of the model. Therefore how to balance the accuracy and computational complexity of the model becomes the problem addressed by ShumeNet.

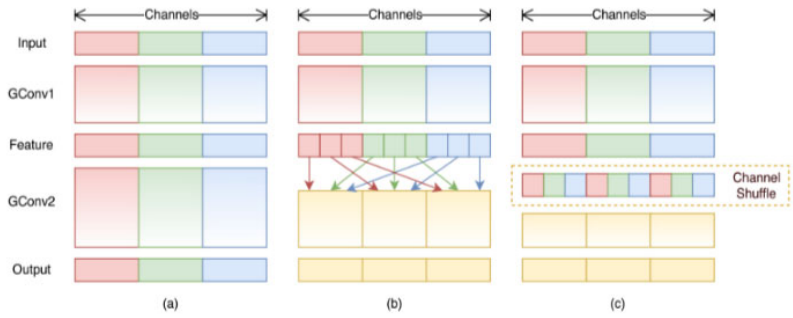

Figure 3. Channel disruption with two stacked group convolutions. (a) Two stacked group convolution layers; (b) Stacked group convolution with channel disruption; (c) Group convolution after channel disruption

In order to do this while satisfying sufficient accuracy and complexity, some ideas have been proposed in ShumeNet. Sparse channel connections are applied in the model, for example by performing group convolution on the 1×1 layers. Previously, the convolution of each group was run only on the corresponding channel with no additional number of parameters, and the computational complexity of the model decreased. However, setting up multiple group convolutions

stacked together for model design purposes inevitably creates an important problem: a channel in the middle layer of the network is only correlated with the input channel corresponding to that channel, and the output value of that channel is then only correlated with its corresponding input channel, and not with the channels of the other groups. The case of two stacked group convolutional layers is shown in Fig. 3(a). It is obvious that the output features from a given group only interact informatively with the features within its corresponding input group, and are completely disconnected from the features between the groups. As a result, the grouping convolution greatly hinders the flow of information and reorganisation between the individual groupings of the channel, largely weakening the expressive power of the network.

Assuming that group convolution can obtain data from different input groups (as shown in Fig. 3(b)), the information between the input and output features of different groups in the model will circulate with each other in time. That is to say, for the feature maps generated in the previous set of layers, the channels in each group are first divided into several subgroups, and then the channel feature information of different subgroups in the previous layer is provided to each group in the next layer, and then the information of the output channels and the input channels of different groups will interact. In order to achieve the goal, the model can be constructed by channel disambiguation operation to realise the vision with as little computational time as possible (Fig. 3(c)): assuming that the input of a convolutional layer is divided into a different group of a, and its corresponding output channels are $a \times b$; firstly, the dimension of the output channels is changed to (a, b) transposed to (b, a) before being flattened and used as inputs to the next layer. Note that even if the number of channels in the two groups is not equal, this implementation is not affected in any way and still achieves its desired effect. In addition, the channel disruption is also differentiable, which means that this disruption operation can be embedded in the network structure to participate in end-to-end learning, reducing the cost of redundant operations.

With the proposed method of channel disruption, building more robust structures using multiple groups of convolutional layers has thus become possible.

4.2. ShuffleNet module

Using the channel disruption approach, a novel ShumeNet network cell has been proposed for designing the structure of the microminiature network model. The structural details of the baseline model can be seen in Fig. 4(a). ShumeNet is different in that, in its residual branch, for the 3×3 convolutional layers, the model applies a depth-separable convolution on the feature map to extract spatial information. Then, the group convolution layer of 1×1 is used to replace the 1×1 convolution to reduce the model computation, and then the output feature map is channel disrupted to achieve the purpose of channel interaction, so far a ShumleNet unit is formed, as shown in Fig. 4(b) The purpose of the second 1×1 convolution is to recover the channel dimensions to match the residual-connected data input dimensions. For simplicity, no additional channel disruption operation is applied after the second 1×1 layer, as the channel exchange in each residual cell is sufficient and no additional operation needs to be introduced. The use of batch normalisation (BN) and nonlinearity Similar to the Xception and ResNeXt models, ShufeNet does not use the ReLU activation function after deep convolution as in the case of residual networks. For the ShumeNet application, only two modifications were made (see Fig. 4(c)): (i) a window of 3 average pooling was added to the residual connections; and (ii) the element summing was replaced with channel splicing, making the channel dimension easy to scale up without adding extra cost.

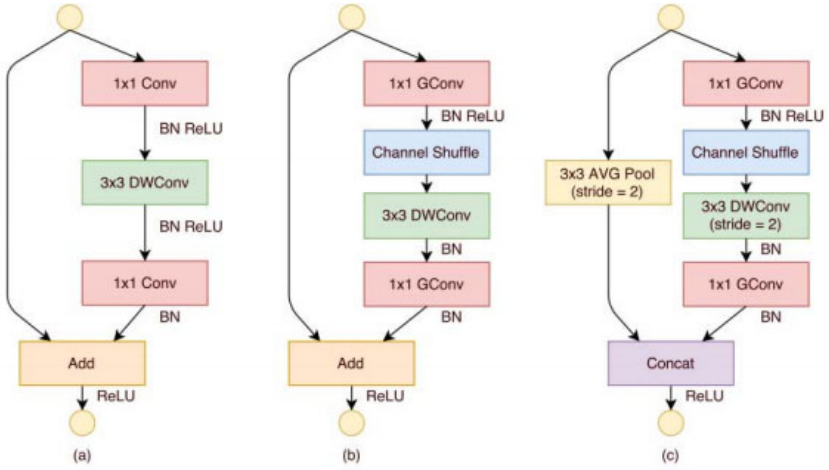

Figure 4. ShufeNet network units (a) Bottleneck unit with deeply separable convolution; (b) ShufeNet unit with 1×1 group convolution and channel disambiguation; (c) ShufeNet with step size 2

The 1×1 group convolution employs channel disruption, allowing the modular components of the ShufeNet model to be combined efficiently. Compared to ResNet and ResNeXt, the structure of ShumeNet has lower complexity for the same setup. For example, ShufheNet requires fewer floating point operations (fops) compared to ResNet and ResNeXt for the same setup of input size and number of bottleneck channels. Alternatively, ShumleNet is able to use feature maps with a larger number of channels in the same computational budget. Since smaller networks generally do not have more channels to process information, the advantage of ShufeNet is quite important for them.

In ShumeNet, only the bottleneck feature maps are deeply convolved. This is mainly due to the fact that small mobile devices have poor memory access rates compared to other intensive computing operations due to their own device parameter limitations, so deep convolution complexity is difficult to accomplish effectively on mobile devices despite

the fact that it has been theoretically reduced to a very low level. This shortcoming is also present in ShufneNet, which has a TensorFlow-based runtime library. In the ShumeNet unit, deep separable convolutional operations are performed only on the bottleneck layer, which minimises redundant model computation overhead.

4.3. Overall structure of the ShuffleNet network

ShufieNet network, in terms of model structure design, was chosen to design a network structure with better performance in order to make the model faster and model size reduction. The details of the ShumleNet model structure are illustrated in Fig. 5 The network is mainly composed of a stack of ShumleNet network units which are mainly composed of three stages. The first convolution in each stage is chosen with a step size of 2, and the subsequent steps are changed according to the training process. The other parameters in that stage are not adjusted, and for the next stage, the number of channels of the output data is doubled compared to that of the input data, which is caused by the final channel splicing operation of the ShumeNet network cells. The number of bottleneck channels for each ShumeNet network is set to 1/4 of the output channels, which is intended to reduce the network parameters and make the model more simplified, but it is mentioned in ShuffleNet that more hyper-parameter tuning may give better results.

In the Shufienet unit, the number of groups g controls the sparsity of the convolution's connections over the channels. Figure 5 discusses various numbers of groups (all approximately 140 MFLOPs) that adjust the output channels to ensure that the overall computational complexity is essentially the same. It is clear that, for a given complexity constraint, a larger number of groups will increase the number of output channels, which will result in more convolutional filters being generated, thus allowing more information to be encoded, although this will also result in degradation of a single convolutional filter due to its respective input channel.Shumenet examines the effect of different computational constraints on this number of channels. Customising the network to the desired complexity only requires applying the scale factor s to the number of channels.

Layer	Output size	KSize	Stride	Repeat	Output channels (g groups)				
					$g=1$	$g=2$	$g=3$	$g=4$	$g=8$
Image	224 × 224				3	3	3	3	3
Conv1	112 × 112	3 × 3	2	1	24	24	24	24	24
MaxPool	56 × 56	3 × 3	2						
Stage2	28 × 28		2	1	144	200	240	272	384
	28 × 28		1	3	144	200	240	272	384
Stage3	14 × 14		2	1	288	400	480	544	768
	14 × 14		1	7	288	400	480	544	768
Stage4	7 × 7		2	1	576	800	960	1088	1536
	7 × 7		1	3	576	800	960	1088	1536
GlobalPool	1 × 1	7 × 7							
FC					1000	1000	1000	1000	1000
Complexity					143M	140M	137M	133M	137M

Figure 5. Shuffle network structure

5. Results

5.1. Image classification based on ResNet model

With the increase of the number of layers of the deep network, better feature learning can be carried out, but it also leads to the re-propagation process in the network can become unstable. The design of ResNet model makes the internal structure of the network model has the ability of certain constant mapping, and solves the problem of gradient explosion or gradient disappearance by adding the residual connection, which enhances the stability in the propagation of the network, and effectively accelerates the convergence of the network, and the degradation problem of the deep network can be solved as a result.

5.1.1. Comparison of classification performance based on different networks

The classification performance of the ResNet model will be evaluated on a flower dataset, as shown in Fig. 6 for some of the images. The selected flower dataset is nearly 4000 images in total, with a total size of about 200M, which can be classified into 5 different types of flower images, namely sunflower, tulip, rose, dandelion, and daisy, and each type contains 600~900 images. The size size of the pictures in the training set in the dataset is not uniform, the common size is 320×240 or 180×240, etc., and the size of the pictures ranges from 30 to 150kb, and the data is not pure, which is mixed with some other pictures. For subsequent experiments, the flower dataset is divided into two parts: flower_train and flower_test.

Figure 6. Selected data sets

We validate the effect of the number of ResNet layers on classification performance by choosing ResNet-18, ResNet-34, ResNet-50, ResNet101, ResNet-152,the number of training validation rounds is 20, and the classification accuracy metrics are denoted by the total accuracy OA.

Table 1. Comparison of classification performance with different network depths

layers	acc%	Flops/10^9
ResNet-18	93.22	1.8
ResNet-34	93.42	3.6
ResNet-50	94.01	3.8
ResNet-101	94.55	7.6
ResNet-152	94.63	11.3

From Table 1, it can be seen that as the number of convolutional layers increases, the overall accuracy also increases, and it can be concluded that the accuracy of ResNet based image classification increases as the number of convolutional layers increases. This is due to the fact that as the number of layers increases, the depth of the network and the performance ability of the model gets better and the overall accuracy is higher. Also, the floating point computing data in the table shows an increase in computing complexity with increasing number of convolutional layers without any decrease in accuracy. This is because as the depth of the network gets deeper, the number of layers increases and the computational complexity required increases.

5.2. Image classification based on ShuffleNet model

With the development of deep learning, the network structure of CNNs is getting deeper and deeper, and most high-precision neural networks need to undergo billions of computations, which makes most network models difficult to use on devices with low computational power. The Shufflenet model, on the other hand, pursues optimal accuracy under a relatively small budget of computational resources by using channel disruption, which allows the input and output channel information flows to interact. For the same computational complexity budget, the Shufflenet model allows more channels to be used than other common architectures, helping to encode more information. Especially when used in small networks, it has better performance, which makes it popular for small devices.

5.2.1. Effect of Group Size on Classification Performance

The classification performance of the ShuffleNet model was also evaluated on the floral dataset mentioned above. In order to assess the importance of 1×1 group convolution for the ShuffleNet model, we verified the effect of ShuffleNet group size g on the classification performance when the number of experimental training rounds was chosen to be 30, and the number of groups chosen was g = {1 2 3 4 8},the classification accuracy metric was expressed as the total accuracy OA.

Table 2. Effect of different group sizes on classification performance

ShuffleNet groups	acc%	Flops/10^9
g=1	93.24	0.143
g=2	94.31	0.140
g=3	94.33	0.137
g=4	94.43	0.133
g=8	94.35	0.130

From Table 2, it can be seen that the overall accuracy increases with the increase in the number of groups g, which indicates that the accuracy increases with the increase in the number of groups. At the same time, the floating point operation data in the table also shows that the operation complexity decreases with the increase in the number of groups g without any decrease in accuracy. Table 2 also shows that when the number of groups g becomes relatively large, the overall accuracy of classification decreases, but the results are still better than without channel disruption.

5.2.2. Ablation experiments with ShuffleNet modules

In order to verify the importance of channel Shuffle in the ShuffleNet model, we conducted experiments on it. We operate on two different groups g with and without channel Shuffle. The classification accuracy metrics of the experimental results are expressed as the overall accuracy OA.

Table 3. Ablation studies on whether channels are Shuffle or not with different ShuffleNet modules

ShuffleNet groups	acc error (%,no shuffle)	acc error (%,shuffle)
ShuffleNet(g=3)	94.33	32.6
ShuffleNet(g=8)	94.35	32.4

From the results, it can be seen that when the number of groups g is 3, the overall accuracy of the ShuffleNet model with channel Shuffle is much higher than that of the ShuffleNet model without channel Shuffle. This indicates that whether or not channel Shuffle is performed has a great impact on the classification performance, and pure group convolution hinders the interaction of channel information between groups, so channel information flows better after channel Shuffle is performed, and the classification performance is relatively better.

6. Conclusion

With the arrival of the intelligent information age, computer hardware is constantly upgraded, software algorithms are constantly updated, the field of artificial intelligence is developing rapidly, and image classification technology based on deep neural networks has been more widely used. This paper reviews the research background, significance and current status of convolutional neural network and image classification technology, and well implements image classification based on ResNet and image classification based on ShuffleNet, and carries out comparative experiments on their classification performance, and investigates the classification performance of image classification based on ResNet in different network depths, and the classification performance of image classification based on ShuffleNet in different group sizes. classification in different group sizes, and also verified the importance of channel Shuffle in ShuffleNet model.

Acknowledgment

This work is supported by the Natural Science Fund for Colleges and Universities, Department of Education of Anhui Province (KJ2021A0481), and Anhui University of Finance and Economics Graduate Student Research and Innovation Fund Project (ACYC2023173).

References

[1] LeCun Y, Learning invariant feature hierarchies[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 496-505.

[2] Deng I, Dong W, Socher R, et al. Image Net: A Largescale Hierarchical Image Database[C]//Computer Vision and Pattern Recognition. IEEE Conference on. IEEE, 2009: 248-255.

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017.60(6): 84-90.

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Pro-ceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] Szegedy C, Wei L, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015, 24(1):205-211.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014. 36(1):231-235.

[7] Larochelle H, Mandel M, Pascanu R, et al. Leaming algorithms for the classification restricted Boltzmann machine[J]. The Journal of Machine Learning Research, 2012, 13(1): 643-669.

[8] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016:770-778.

[9] Zhang X, Zhou X, Lin M, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices[J]. 2017.

[10] Arandjelovic R, Zisserman A. All about VLAD[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.2013:1578-1585.

[11] Sanchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: Theory and practice[J]. International journal of computer vision, 2013, 105(3): 222-245.

[12] Srivastava R K, Greff K, Schmidhuber J. Highway networks[J]. arXiv preprint arXiv:1505.00387,2015.