

Research of Convolutional Neural Network on Image Classification

Xiyu Lin *

Department of Computer Science, Guangdong University of Finance, Guangzhou, China

* Corresponding author email: 631301120419@mails.cqjtu.edu.cn

Abstract. With the progress of artificial intelligence, technology based on deep learning is becoming more and more mature, and the application of deep convolutional neural network for image classification has become a popular topic for researchers. The number of the structure of deep convolutional neural network for image classification is keep increasing, and its performance is consistently improving, gradually replace that of traditional methods. According to the process of model development and model optimization, this paper divides the convolutional neural network into two models: classical deep convolutional neural network model and attention mechanism deep convolutional neural network model. The construction methods and characteristics of various kinds of deep convolutional neural network models are comprehensively reviewed, and the performance of various classification models is compared and analyzed. Finally, the problems of deep convolutional neural network model are presented.

Keywords: Image Classification Convolutional; Attention Mechanism; Neural Network.

1. Introduction

With the progress of deep learning, many deep learning networks structures have achieved great success in the field of image processing, providing more development space for image classification research. Image classification is one of the most demanded technologies of our time, and it is used in various fields such as healthcare and business. For the super-powerful human visual system, it is easy to identify the category of an image, but for computers, it cannot get the information of the image like human, which is also a very important task in machine learning [1]. Therefore, understanding and making state-of-the-art computer vision models is a necessary field of AI. However, traditional image classification methods have certain limitations, the deep convolutional neural network structure in this field is keep increasing, and it shows better results than that previous image classification methods, so the use of deep convolutional neural networks for image classification has gradually become popular among many researchers [2].

Since the relevant research content is not comprehensive enough and needs further analysis and discussion, this paper summarizes the convolutional neural network research system based on deep learning image classification, roundly reviews the structure methods and features of different deep convolutional neural network structures, and compares and analyzes the performance of various classification models. The experimental results of the convolutional neural network model on the dataset are analyzed, and the core difficulties and future development trends faced by the convolutional neural network model are pointed out to promote the further development of image classification.

2. Main Body

Deep learning has had a lot of success in machine learning as well as other fields, and it is widely used in image classification while being applied to matching, segmentation, detection, and recognition. After more than ten years of development, convolutional neural networks (CNNs) have the ability to solve different problems maturely, avoiding the shortcomings of traditional methods such as low efficiency and high computational cost. According to the development process and direction of DCNN-based image classification research in recent years, this paper divides deep

convolutional neural network models into the following four categories: early deep convolutional neural networks, ResNet-based and DenseNet-based neural networks; and deep convolutional neural network models based on attention mechanisms.

2.1 Early Deep Neural Networks

The LeNet model is the earliest convolutional neural network model, which is mainly used for handwriting recognition in MNIST (modified NIST) datasets, and Figure 1 depicts the model structure. It has three convolutional layers, two pooling layers, and two fully connected layers, each of which has trainable parameters, which lays a foundation for the development of deep convolutional neural networks [3].

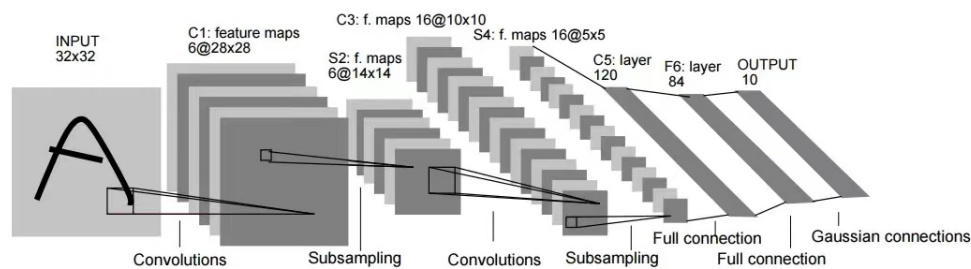


Fig 1. Architecture of LeNet [3]

While LeNet works well on small-scale MNIST datasets, complex image classification tasks require large-scale datasets and more learnable network structures. In 2012, Krizhevsky et al. proposed AlexNet. By using the rectified linear unit (ReLU) as the activation function, the network uses local response normalization (LRN) to decrease the gradient disappearing. The increase of data and Dropout techniques greatly mitigates the problem. Zisserman and Simonyan proposed VGGNet (Visual Geometry Group), inheriting the architecture of LeNet and AlexNet, and the primary improvement is to increase network depth and improve network performance [4].

Increasing network depth is typically the most direct technique to boost network performance, but as the network depth increases, the number of parameters increases, the network is more likely to overfitting, and the need of computing resources also has large augment. Szegedy et al. proposed GoogLeNet, uses the Inception-v1 module, adopting sparse links to decline the number of model parameters, ensure the efficiency of computing resources, and improve the result of the network when the depth reaches 22 layers.

To further deepen the network depth, Larsson et al. proposed a fractal network (FractalNet), which contains interactive subpaths of different lengths, but does not contain residual connections, which can generate extremely deep networks by reusing simple scaling rules and achieve good performance.

As the depth of the network adds, the gradient disappearance problem becomes more serious, and the modification of the network becomes harder. To solve this problem, He et al. proposed residual networks (ResNet). ResNet alleviates the problem of vanishing gradients in deep neural networks by introducing a "shortcut" module that includes identity connections so that convolutional layers can be skipped. ResNet is composed of stacked residual blocks. In ResNet, the advanced features are more valid for classification tasks. Veit et al. found that in the case of doubling the number of channels of the ResNet feature map, deleting the building blocks with downsampling in the residual cells will greatly reduce the model classification accuracy, and the use of random deep training of the residual network will alleviate the above situation. In order to spread the pressure that builds up on a single residual cell as a result of down sampling over all residual cells, Han et al. presented a pyramid residual network in which the amount of output pathways for each residual cell is gradually raised.

According to Zhang et al. residual mapping is simple to optimize, as is residual mapping of residual map. Based on this, they added the cross-layer connection step by step on the basis of ResNet to construct a multi-level residual convolutional neural network, so that the higher layer can be transmitted to the low-lever layer, further inhibiting the gradient disappearing problem. The RoR

(residual networks for residual networks) structure is shown in Figure 2. RoR initially provides a cross-layer correlation outside all residual blocks; the residual blocks are then divided into categories according to the type of convolution kernel filter used. And the cross-layer connections are added outside each group of residual blocks; Each set of residual blocks can then be bisected to add cross-layer connections; At the end, the layer-skipping connection in the original residual block, and it is called the last-level layer-skipping connection.

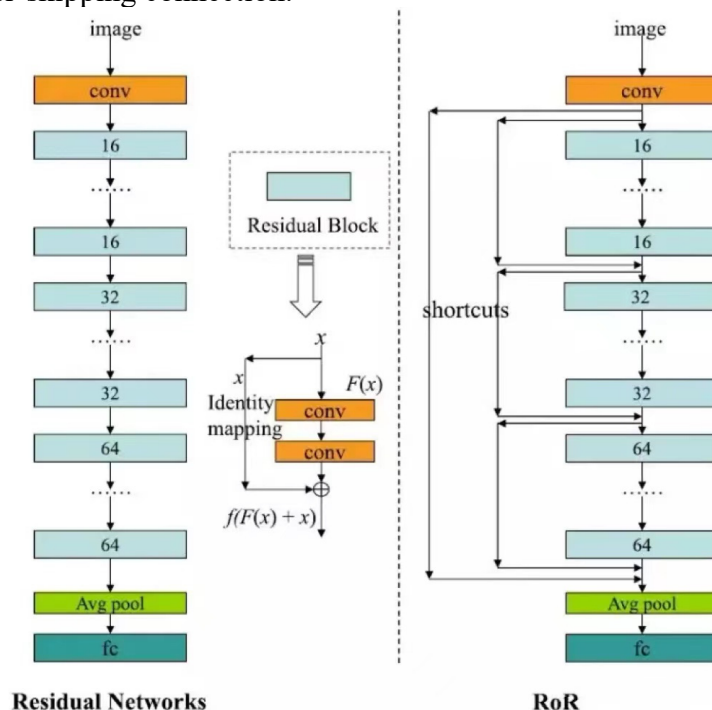


Fig 2. The architecture of RoR [5]

The feature map size and channel number of each group of residual blocks in the RoR network remain unchanged, and at the beginning of the next set of residual blocks, the number of channels is increased while the size of the feature map is decreased, which makes the transmission of feature information in the network incoherent, and it may cause loss of some useful information related to prediction, and limit the classification performance of the network. In view of this problem, Zhang et al. proposed a pyramid multi-level residuals convolutional neural network.

2.2 DenseNet-based Neural Networks

In 2017, inspired by the idea of identity/skipconnections in ResNet, Huang et al. proposed densely connected convolutional networks (DenseNet), which use a simple connection mode that connects all layers directly. Based on that, the researchers studied the way DenseNet paths are connected, and proposed a series of DenseNet variants.

DenseNet consists of dense blocks with a dense block structure shown in Figure 3, which directly connects all layers in a feedforward way, each layer takes input from previous layers and transmits its output to the back layer, which enhances feature reuse and mitigates the problem of disappearing gradients. Unlike residual networks, DenseNet does not use summation to pass the feature map, but combines the feature map by combining channels.

After study the advantages and disadvantages of the topology of ResNet and DenseNet, it can be seen that ResNet could enhance feature reuse and DenseNet supported feature exploration. To benefit from both topologies, Chen et al. propose a dual-path network consisting of small modules similar to bottleneck layers.

DenseNet connects each layer to its predecessor for feature reuse, but the network creates redundancy when the front layer's features are not required by the back layer. In view of this, Huang et al. proposed ConDenseNet, which incorporates learning grouped convolution and dense joins.

Experiments have shown that CondenseNets are more efficient than the most advanced compact type convolutional networks at the period and can be used in mobile electronic devices.

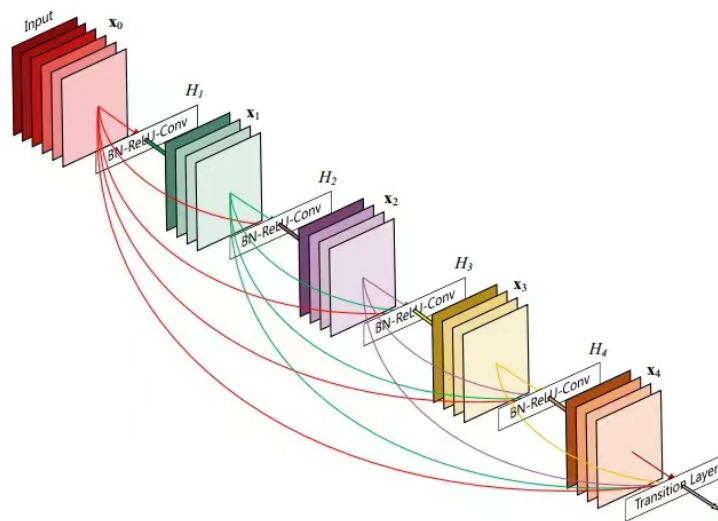


Fig 3. Dense block [6]

2.3 Neural Networks based on Attention Networks

Another focus is the attention mechanism of convolutional neural network research, and the convolutional neural network using the attention mechanism mainly emphasizes the useful part, which can be some spaces, some channels or particular layers, etc., weakening the part of little value. The main attention mechanisms can be divided into spatial domain, channel domain and layer domain attention mechanism.

Wang et al. proposed a convolutional neural network using attention mechanism, which is constructed by stacking various attention modules. In each attention module, the branch of mask adopts a bottom-up, top-down feed-forward model, and expands the feed-forward and feedback attention steps into a feed-forward process to make the network end-to-end training, and generate soft weights of features at the same time.

Hu et al. focused on channel relations and proposed a squeeze-and-excitation block (SE block). As shown in Figure 4, the SE block recalibrates the feature response of the channel by modeling the dependencies between channels in an explicitly way, that is, selectively excites features that are useful for classification and suppresses features that are less useful. With minimal computation and parameter involvement, SENet can significantly improve the performance of most existing CNNs. The module first compresses the global spatial information into the channel descriptor, then produces the weight of each channel characteristic via the gating mechanism, and multiplies the weight by the module's input to finish the recalibration of channel characteristic. Using the SE block yields good results only with the price of slightly increases the model's complexity and computation.

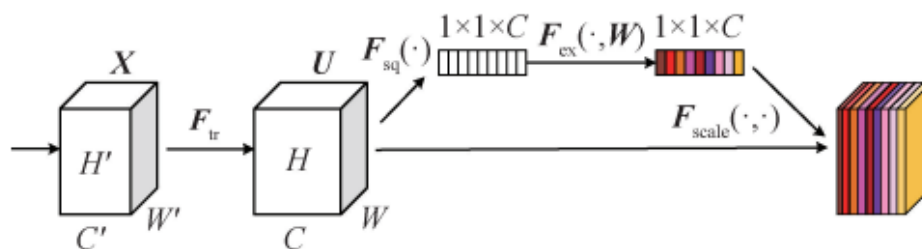


Fig 4. SE block [7]

In SE blocks, there is a lack of consideration for the relevance of spatial domain features. In view of this problem, Woo et al. proposed a convolutional block attention module (CBAM), which includes two parts: channel attention module and spatial attention module, leading the export of convolutional block attention module can be obtained, spatial domain characteristic correlation and channel domain characteristic correlation are took into account at the same time.

Inspired by SE blocks, Gao et al. propose a new GSoP module and introduce it from the lower grade to the upper one, which captures global second-order statistics along the channel aspect or spatial aspect, which can be conveniently put into extant networks and make progress in its result with a small cost.

Convolutional operations are local operations with limitations such as low computational efficiency, difficult optimization, and difficult modeling of multi-hop dependencies. In response to this problem, Wang et al. proposed non-local blocks, using non-local operations to capture remote dependencies. As a response to a location, this operation returns the weighted sum of the individual location characteristics of the input feature map.

Zhang et al. investigated the DenseNet attention mechanism. First, they built a channel with reweight DenseNet (CFR-DenseNet), used the extrusion excitation module to recalibrate channel features, then constructed an inter-layer feature reweight DenseNet, ILFRDenseNet), using the double extrusion excitation module to recalibrate the interlayer features, and finally fusing CFR-DenseNet and ILFR-DenseNet to construct a multiple feature reweight DenseNet (MFRDenseNet).

3. Analysis of Experimental Results

Image classification models are mainly tested on the ImageNet dataset. This article collates the experimental results of each model on this dataset (Table 1), where the ImageNet dataset is the version used by ILSVRC in 2012. The experimental results are derived from the corresponding references, and if there are no corresponding experimental results in the references, they are indicated by "-". The description of the experimental results mainly includes the error rate of image classification and the number of parameters [7,8]. In the early stage of the construction of convolutional neural networks, researchers are committed to making progress to the classification accuracy of the network, and when the development of DCNNs is becoming more and more mature, to save the time and memory use of training tests, deep convolutional neural networks are not only required to have high image classification accuracy, but also require them to have a small number of parameters. In addition to accuracy and the number of parameters, there are floating-point operando's (FLOPs) that measure the advantages and disadvantages of lightweight networks.

Table 1. The error rate of the basic DCNNs on ImageNet datasets

Model	Top-1 Validation set (Test set) error rate/%	Top-5 Validation set (Test set) error rate/%
AlexNet (single network)	40.70(-)	18.20(-)
AlexNet (five networks)	38.10(-)	16.40(16.40)
GoogLeNet (single network)	-	7.90(7.90)
GoogLeNet (five networks)	-	6.67(6.67)
VGGNet (single network)	24.40(-)	7.10(7.00)
VGGNet (five networks)	23.70(-)	6.80(6.80)
FractalNet	24.12(-)	7.39(-)

P.S.: "-" indicates that there is no data in the corresponding document.

As shown in Table 1, in the basic network model, from LeNet to GoogLeNet, the image classification accuracy is getting higher and higher, which is because the deeper the network, the more efficient the extracted features. As the network deepens, the gradient disappearing problem becomes more and more serious, ResNet uses the cross-layer connection to make the high-level features of the network pass to the lower layer, alleviating the gradient disappearing problem and

greatly improving the performance of the classification task. To continue modify the experimental result of residual networks, the researchers proposed a series of variants of residual networks. Han et al. constructed a pyramid residual network by linearly and gradually increasing the channels amount of the feature map in the residual block. Zhang et al. built RoR by adding cross-layer connections step by step on the basis of the original residual network. Zhang et al. constructed a pyramid RoR by linearly and stepwise increasing the number of channels in each residual block on the basis of RoR. These methods show the importance of improving the flow of information from layer to layer, as well as the importance of incorporating short paths between layers. The experimental results of ResNet and its variants on the ImageNet, CIFAR, and SVHN datasets are shown in Tables 2 and 3.

Table 2. The error rate and parameter amounts of the Residual families on ImageNet datasets

Model	ImageNet		The number of parameters/M
	Top-1 error rate/%	Top-5 error rate/%	
ResNet-152	19.38	4.49	-
Pyramidal Residual Networks-200	19.20	4.70	116.40
RoR-3-152	20.55	5.14	-

P.S.: " - " indicates that there is no data in the corresponding document.

Table 3. The error rate and parameter amounts of the Residual families on CIFAR and SVHN

Model	CIFAR-10/%	CIFAR-100/%	SVHN/%	Parameters/M
ResNet-110	6.43	-	-	1.70
Pyramidal Residual Networks-272	3.31	16.35	-	26.00
Pyramidal RoR+	2.96	16.40	1.59	38.00

P.S.: " - " indicates that there is no data in the corresponding document.

DenseNet proposed by Huang et al. further improves the result of deep convolutional neural networks, it is because that in each dense block of DenseNet, the feature map of each layer is closely linked, that is the layer connected to the front one, and its own feature map is transmitted to all the back layers, thus further improving the propagation of information flow between layers than ResNet, once again highlighting the importance of short paths [9]. Subsequently, a series of DenseNet variants were proposed, continuing to improve the presentation of DenseNet. The performance of the DenseNet family on various datasets are demonstrated in Table 4.

Table 4. The error rate of the DenseNet families on different datasets

Model	ImageNet		CIFAR-10/%	CIFAR-100/%	SVHN/%
	Top-1 error rate/%	Top-5 error rate/%			
DenseNet	20.85	5.30	3.46	17.18	1.59
DPN	18.55	4.16	-	-	-
CondenseNet	26.20	8.30	3.46	17.55	-

P.S.: " - " indicates that there is no data in the corresponding document.

The DCNNs using the attention mechanism also achieves good performance. It maintains or even surpass the accuracy of the classical DCNNs, and the number of parameters is also significantly decline. The experimental results of a deep convolutional neural network applying attention mechanisms on CIFAR are shown in Table 5. For example, the SE module proposed by Hu et al. models the dependence between channel features, and the MFR-DenseNet proposed by Zhang et al. first performs channel feature recalibration and then performs interlayer feature recalibration. It can be witness from the above methods that the attention mechanism essentially emphasizes the useful parts and ignores the parts that are less useful, so that the features extracted by the model are more effective [10].

Table 5. The error rate and parameter amount of the DCNNs using attention mechanism on CIFAR

Model	ImageNet		The parameters/M	CIFAR-10/%	CIFAR-100/%	The parameters/M
	Top-1 error rate/%	Top-5 error rate/%				
Residual attention network	19.50	4.80	51.30	3.90	20.45	8.60
SENet	16.88	3.58	-	-	-	-
ResNeXt101+CBAM	21.07	5.59	48.96	-	-	-
MFR-DenseNet	-	-	-	3.57	18.27	14.20
GSoP-Net2	21.19	5.64	58.70	-	18.85	3.60

P.S.: "-" indicates that there is no data in the corresponding document.

4. Conclusion

Deep convolutional neural network is one of the information processing systems. As an important method of deep learning, DCNNs are widely applied in lots of computer vision missions, especially image classification tasks, which use convolution to automatically extract features and judge image categories. Image classification model is becoming more and more mature, this paper focuses on the deep convolutional neural network model construction of image classification, first introduces the basic model from LeNet to GoogLeNet, ResNet family, DenseNet family, and deep convolutional neural network model using attention mechanism, and finally summarizes and analyzes the experimental results of all models. Deep convolutional neural networks shine in image classification tasks, but they are incomparable with the human nervous system, still far away from reaching the level of human cognition, and have the following difficulties and challenges:

1) Deep convolutional neural networks for image classification are supervised learning, while the human brain has the ability to think, which is more unsupervised learning. Therefore, how to make the model like the human brain have the ability to think may be the future research direction, and unsupervised learning is the future development prospect.

2) Deep convolutional neural network models are challenging to apply to mobile devices. That is because they need to maintain model accuracy while considering the size and speed of the model. The design of a deep convolutional neural network structure for mobile devices needs further research.

3) In recent years, improving the efficiency of models has become a research hotspot, but most of the research is based on optimizing indirect metrics to simplify the network, such as MACC (memory accesses) values, FLOPs, number of weights. Although indirect indicators are easy to calculate and integrate into the optimization framework, the optimization of indirect indicators does not necessarily reduce direct indicators, such as model speed and network resource consumption, and the difference between direct indicators and indirect indicators is also related to the platform.

References

- [1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.
- [2] Zhang K, Feng X H, Guo Y R, Su Y K, Zhao K, Zhao Z B, Ma Z Y and Ding Q L, Overview of deep convolutional neural networks for image classification. Journal of Image and Graphics, 2021,26 (10): 2305-2325.
- [3] Yong L I , Lin X Z , Jiang M Y . Facial Expression Recognition with Cross-connect LeNet-5 Network. Acta Automatica Sinica, 2018. 9: 3024-3033.
- [4] Ghiasi G, Lin T Y and Le Q V. NAS-FPN: learning scalable feature pyramid architecture for object detection. Proceedings of 2019 IEEE/ CVF Conference on Computer Vision and Pattern Recognition, 2019: 7036-7045.

- [5] Zhang K, Guo L R, Gao C and Zhao Z B, Pyramidal RoR for image classification. *Cluster Computing*, 2019, 22(2): 5115-5125.
- [6] Guo Y R, Zhang K, Wang X S, Yuan J S, Zhao Z B and Ma Z Y, Image classification method based on end-to-end dual feature reweight DenseNet. *Journal of Image and Graphics*, 2020, 25 (3): 486- 497.
- [7] Kajkamhaeng S, Chantrapornchai C. SE-SqueezeNet: SqueezeNet extension with squeeze-and-excitation block. *International Journal of Computational Science and Engineering*, 2021(2):24.
- [8] Wang F, Jiang M Q, Qian C, Yang S, Li C, Zhang H G, Wang X G and Tang X O, Residual attention network for image classification. *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, 2017:6450- 6458.
- [9] Yang Y B, Zhong Z S, Shen T C and Lin Z C, Convolutional neural networks with alternately updated clique, *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018: 2413-2422.
- [10] Szegedy C, Ioffe S, Vanhoucke V and Alemi A A, Inception-v4, Inception-ResNet and the impact of residual connections on learning. *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. San Francisco, 2016: 4278-428.