Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230657

Researches advanced in image recognization based on deep learning

Zhizhi Li

Dalian Jiaotong University, Dalian, Liaoning Province, 116028, China

lizhizhi@st.btbu.edu.cn

Abstract. Image recognition is a fundamental problem in the field of computer vision community, which aims to understand the content of an image to predict the class of the image. Traditional image recognition methods rely heavily on the quality of handcrafted features, and the recognition accuracy and generalization ability cannot meet practical application requirements. Thanks to the speedy progress of deep learning theory and technology, convolutional neural networks can adaptively learn image semantic features through high-dimensional nonlinear changes, which sensibly rises the exactness of image recognition. At present, image recognition has been generally applied to many fields such as commodity circulation, smart retail, pest identification, medical image analysis and so on. In this paper, we introduce the latest research progress in deep learning-based image recognition. Specifically, following the temporal clues of technological development, we first introduce representative image recognition networks, including their design ideas, basic network structures, advantages and disadvantages, etc. Second, we quantitatively compare the performance of different image recognition networks. Finally, we summarize the existing problems in the sphere of image recognition research and discuss its possible future directions.

Keywords: Image Recognization; Deep Learning; Feature Extraction.

1. Introduction

Thanks to the speedy progress of science and technology in contemporary society, people pursue fast, efficient and convenient. The birth and rapid progress of artificial intelligence are the cornerstone behind the achievement of these goals. The image recognition, related to artificial intelligence, has become a technology with high research value, wide application range and great development potential. Image recognition can be understood as the practical application of deep learning algorithms.[1] By using computers to analyze and process images, data can be obtained from the recognition of various types of objects. Deep learning technology analyzes and processes related features of images through multi-layer structures, so as to rises the exactness of image recognition. At present, image recognition technology can be broken down into two sorts: biometric recognition and object recognition. Biometric identification mainly includes face recognition for identity authentication and payment information security[2]; And the classification of plant and animal species in relation to pathology. Object recognition application is relatively wide, mainly used in commodity circulation, intelligent retail; And other communications military and other fields.

DOI: 10.54254/2755-2721/5/20230657

The advancement of image recognition technology began roughly from 1950, which was uesd for the text recognition of image identification, analysing and processing the text data, such as numbers, symbols and so on. From 1965 to the beginning of the 21st century, image recognition technology has transitioned to digital image recognition and processing research. Digital image better solves the problem of data loss in transmission and storage. This step of development process is of great significance for recognition technology. However, The traditional method of extracting features by manually designed extractors requires the understanding of relevant professional knowledge and complex parameter adjustment process. Moreover, each method can only target fixed applications, so the generalization ability and robustness are poor. Since the end of the 20th century, machine learning has emerged and brought great breakthroughs in image recognition, and CV (computer vision) has also made many practical advances. In the following time, deep learning has become the main force in image recognition related techniques. Feature extraction in deep learning is realized based on a large amount of data. A large number of samples are input for learning, so as to obtain deep and dataset specific feature representations. It is more robust, efficient and accurate to express the abstract features extracted by end-to-end [3] deep learning in the dataset, with better generalization ability, but the disadvantage is that the sample set has great influence and requires high computational power.

The steps of image recognition can be roughly summarized as: generate network structure, input data, then train, get the model and use the model for recognition. [4] Neural networks are now the primary tool used in the deep learning of image recognition. Typical neural networks commonly have three layers: input layer, intermediate layer (hidden layer) and output layer. The input layer is mainly used to obtain input information, such as what color are the pixels in a picture. The scale of this layer was decided by the the amount of input information. The hidden layer's primary job is to extract features by modifying the weights so that the neural units respond to a particular pattern. To determine whether neurons in different hidden layers are responding correctly to stimuli, the output layers adjust their weights, and the final output excitability is the result. The convolutional neural network is made up of the input layer, convolution layer, pooling layer, reLU layer, and fully connected layer.[5] The convolutional layer, which houses the majority of the network's computation, is the network's core. The convolutional layer contains the learnable filter, so that when the filter sees the specific type of features, it will be activated. The function of the Relu layer is to retain values with smaller features and discard values with eigenvalues less than 1. By periodically inserting the pooling layer into the continuous convolution layer, the quantity of parameters in the network can be tightly regulated, and the size of the data space can be better decreased. The fully connected layer plays the role of classification, and its core operation is matrix multiplication, mapping the extracted feature space to the sample label space.[6]

This article will first introduce alexnet zfnet, VGG, googlenet, and resnet the five representative classic networks [7], then, the existing relevant data sets are introduced through experimental collection and sorting, as well as the indicators of evaluation and classification. Finally, the existing problems and prospects are discussed.

2. Advanced image recognization algorithms

2.1. Alexnet

The AlexNet network structure is a model that ranked first in the 2012 ImageNet competition, named after its first author Alex. The main structure of Alexnet network structure contains eight layers, which are divided into five convolutional layers. Through these five layers, image edges and simple features are extracted from core and complex features. There are three fully connected layers, which output data through RelU and dropout. Rule, as the activation function, removes the negative values in the convolution results and keeps the positive values unchanged, which effectively overcomes the shortcomings such as vanishing gradient and slow convergence rate of Sigmod and Tanh functions. The dropout function can relatively alleviate overfitting, so regularization effect can be achieved to a certain extent. In order to prevent overfitting, Alexnet network structure also uses overlapping

DOI: 10.54254/2755-2721/5/20230657

maximum pooling, that is, the pooling window is bigger than the pooling step size. In addition, this network structure uses two Gpus for calculation, which greatly improves the learning speed compared with one GPU. Meanwhile, the Local Response Normalization (LRN) is proposed in Alexnet network structure for the first time, which highlights features through competition in neuronal reactions, enlarging or narrowing differences, and thus enhances the model's ability to generalize. In terms of data processing, the means of enhancing data are cutting, mirror flipping and scaling, which can increase the model's ability to generalize. At the same time, Alexnet network structure also brings a huge amount of computation. In terms of the improvement of Alexnet, there is room for optimization in terms of network depth, which can also be optimized by optimizing functions such as Relu or adding attention mechanism.

2.2. VGGNet

VGGNet network structure is a model that ranked first in ImageNet for localization task and second in ImageNet for classification task in 2014. It is named after the Visual Geometry Group of Oxford University. Both VGGNet and Alexnet adopt the same size of 2*2 maximum pooling size and 3*3 convolution kernel size, among which VGG16 and VGG19 are the most commonly used ones, but VGGNet is optimized on the basis of Alexnet. Compared with Alexnet network structure [8], VGGNet's strengths are reflected in the quantity of convolutional layers and core size. While each convolution layer in the VGGNet network structure has 2-4 convolutions with a convolution kernel of 3*3, each volume basis in the Alexnet network structure contains just one convolution with a convolution kernel of 7*7. In this way, a larger convolutional layer is replaced by several smaller convolutional layers, reducing the parameters as desired, and simultaneously, the fitting and expression ability of the network is also enhanced by more nonlinear mapping. Multi-Scale method is used to enhance data in VGGNet. The input picture is randomly cut into 224*224 size, which increases the amount of data and achieves a good effect of preventing overfitting. In the process of training, VGGNet first trains the simple A-level network with shallow layers, and then trains the subsequent network by training the data obtained from the A-level network, so as to accelerate the convergence speed. The disadvantage of VGGNet is that there are too many convolutional layers in the middle, which leads to a larger memory footprint.

2.3. ResNet

The ResNet network architecture was the winner of ImageNet in 2015, with much fewer parameters and a lower error rate than VGGNet. It was Residual Unit trained 152 layers of neural network structure, so it was named ResNet. At that time, the importance of network depth to the model has been confirmed, so most of them blindly increase the network depth. But in fact, the deeper the network accuracy is not higher, the result is quite different from the expected. The ResNet team refers to this phenomenon as "degradation," and ResNet addresses the error caused by the training set from the root of the accuracy degradation. As a result, the residuals' successful introduction demonstrates the mapping discrepancy between the anticipated output and the actual input.[9] When the residual element is backpropagated, the gradient can be directly transferred to the previous layer, which improves the vanishing gradient problem by optimizing from other perspectives. However, the issue of vanishing gradient is not resolved by increasing network depth, and network optimization is still challenging. Therefore, a series of residual network variants are generated, which can be roughly seen as four methods: basing on deep residual network optimization, using new training methods, basing on increasing width, and using new dimensions. While ResNet's residual theory is powerful in many ways, it still has a significant drawback: deep Web training often takes weeks. Therefore, the cost of ResNet application in the real world is very high.

2.4. GoogLeNet

As the name suggests, GoogLeNet was designed by network engineers from Google. Compared with previous network structures, GoogLeNet's network structure uses a new deep learning model. In

DOI: 10.54254/2755-2721/5/20230657

Alexnet and VGG network structures, the network depth is increased to improve the training effect, while GoogLeNet uses a deep neural network model based on Inception module. And gain laurels in the classification task of ImageNet in 2014. The core is the Inception module, whose essence is to package many convolution and pooling operations together to form an independent module, and then use the Inceptin module to build the network. The GoogLeNet network structure uses many 1*1 convolutions, which can reduce the dimension and limit the size of the network. Inception structure optimizes both the depth and width of the network, that is to say, units in each stage will increase when the computational complexity is out of control. GoogLeNet is built through the Inception module, so the optimization and improvement mainly focus on the internal optimization of the Incetion module. In the subsequent research and optimization of Inception, Inception2 is presented successively and Batch is proposed to replace dropout and LRN. Inception3 introduces Factorization, which takes two one-dimensional convolution instead of one two-dimensional convolution. Inception4 studied Inception Module and Residual Connection, and ResNet can greatly accelerate training.

2.5. MobileNet

MobileNet is different from other network models. Almost all traditional convolutional neural networks pursue network depth to enhance the accuracy of network models. However, in the meantime, a large and complex network will occupy a large amount of memory, which is difficult to be applied to real scenes. Therefore, in mobile or embedded devices, only the trained complex model can be compressed and reused, or a small but refined model can be designed for use. MobileNet was born out of the latter. The essence of MobileNet is depth-separable convolution, and its operation include Depthwise Convolution and Pointwise Convolution. The difference between Depthwise Convolution and standard convolution is that different input channels in Depthwise Convolution correspond to different convolution kernels, while the convolution kernels in standard convolution are on all input channels. Pointwise Convolution convolution is all using a 1×1 convolution kernel. Compared with VGG and GoogleNet, MobileNet has absolute advantages in the amount of computation and parameters, although the accuracy is slightly higher or lower.

3. Experiments and performance analysis

3.1. Common datastes

This paper mainly introduces three classic datasets: ImageNet, Mnist and Cifar.[10] ImageNet is a computer vision dataset created by a team led by Professor Feifei Li at Stanford University. The data set is a large image data set created to help the development of computer image recognition technology. It contains 14,197,122 images and 21,841 Synset indexes. The images in the data set include most of the scenes and physical categories encountered in life. The annual ILSVRC image recognition competition uses the ImageNet dataset, and this dataset has been reached and used to evaluate the performance of image classification algorithms. The reason why the ImageNet dataset is much more difficult to identify than CIFAR-10 is that dataset A has more images and categories, higher resolution, and more irrelevant noise and variation in the images.

Created by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, CIFAR is often referred to as CIFAR-10 because the dataset was originally divided into 10 categories: aircraft, deer, Dog, Frog, horse, truck, Bird, cat, boat, and car. 60,000 32*32 color photos in total are contained in CIFAR without any type of overlap (including 10000 test images and 50000 training images). Then came CIFAR-100, with 100 categories and 600 images for each category, each with two labels for its class and superclass.

The Mixed National Institute of Standards and Technology (MNIST) dataset database has a training set of 60,000 samples and a test set of 10,000 samples collected and curated by the National Institute of Standards and Technology (NIST). The main purpose of this dataset is to recognize the function of handwritten digits through machine learning. MNIST image data is 28px by 28px

DOI: 10.54254/2755-2721/5/20230657

grayscale image, each pixel value between 0 and 255， and each image is labelled accordingly "0", "1", "2", "3", "4", "5".

3.2. Evaluation index

In image recognition tasks, the quality of a model and method is evaluated and determined by the corresponding indicators. Accuracy, confusion matrix, precision, recall, PR curve, AP, mAP, F value, ROC curve, and AUC value are standard classification job indicators. [11]

Precision, which is the most direct and simple index, is usually used to understand the percentage of the correctly categorized samples among all samples.

The confusion matrix is drawn as a matrix if the horizontal axis is the number of categories predicted by the model and the vertical axis is the number of true labels. Some other evaluation indexes can be derived from the confusion matrix, such as precision (also known as accuracy) and recall. Precision represents how many of the results predicted to be positive samples are correctly classified; The recall ratio represents how many of the true positive sample results are correctly classified. Taking the recall and precision as the horizontal and vertical axes to make the curve is the P-R curve. The larger the enclosing area below the curve, the higher the AP value, indicating the better classifier. The F value represents the harmonic average of precision and recall. According to the learner's prediction results, the samples are sorted and then used as positive examples for prediction, which is ROC. Each time, the values of TPR and FPR are calculated, and the "ROC curve "is obtained by taking them as horizontal and vertical coordinates respectively, and the area enclosed by the curve is the AUC.[12] The basic concepts used in the preceding indicators are as follows: True positive, or TP. Thus, the prediction made by the positive sample was accurate. False positive, or FP. Thus, the prediction made by the positive sample is incorrect. True Negative (TN) Thus, the prediction made by the positive sample was accurate. False Negative, or FN Thus, the prediction made by the negative sample is incorrect.

3.3. Performance comparison

We first report the model training time and Top-5 error rate of different representation CNN models on the ImageNet dataset in Tabel 1. [13]

Tabel 1. Training time and Top-5 error rate of different representation CNN models.

Year	Methods	Top-5 error rate	Training time
1998	LeNet	--	--
2012	AlexNet	15.3%	Five to six days(Two GTX 580GPU)
2013	ZFNet	14.8%	Twelve days (GTX 580 GPU)
2014	GoogleNet	6.7%	Week (few high-end GPUs)
2014	VGGNet	7.3%	Either Two to three weeks (4 Nvidia Titan Black GPUs)
2015	ResNet	3.6%	Either Two to three weeks (8 GPU machines)

With the emergence and optimization of new network structures, the error rate for TOP-5 is getting lower and lower, but the cost of training models is increasing, mainly because of the increase of network depth and complexity of network structures. Different parameters are considered, including the quantity of convolutional layers, the quantity of steps, the quantity of fully connected layers and the total quantity. The following Table 2 [13] displays the precise structure. From the table data can be analyzed, the network convolutional layer is gradually increasing from top to bottom, which means that the deeper the network is, the more complex the model is, so the volume and calculation amount

DOI: 10.54254/2755-2721/5/20230657

are also increasing. The difference of GoogelNet is that it uses Inception structure, and the packaged independent modules reduce the dimension and limit the network size.

Tabel 2. Structure details of different representation CNN models.

Methods	Convolutional Layers	Fully Connected Layers	Total MACs	Total Weights
LeNet5	2	2	2.3M	431k
AlexNet	5	3	724M	61M
VGG-16	16	3	15.5G	138M
GoogleNet	21	1	1.43G	7M
ResNet-50	50	1	3.9G	25.5M

Table 3 [14] is a comparison of some of the best models in terms of errors, network parameters, and maximum number of connected layers. According to the previous table, with the innovation of network structure, the error rate of the model on TOP-5 is gradually decreasing. At the beginning, some models' error rates are lowered primarily by adding more network layers, which increases the model's accuracy, while the network depth will be saturated, and blindly deepening the network will lead to network degradation. Therefore, the subsequent network model is more by enlarging the network width, or optimizing and using other methods to improve so as to reduce the error rate. The complexity of network structure will naturally lead to the increase of convolutional layers, which will greatly increase the computational load. VGGNet is the most obvious because its middle layer contains a large number of convolutional layers. In GoogleNet and resNet, the computation is optimized due to the Inception module. At the same time, the use of overlapping small convolutional layers instead of large convolutional layers also reduces the parameters.[15]

Tabel 3. Error rate of different CNN models, MACS and related parameters in the Top-5%.

	LeNet-5	AlexNet	OverFeat (fast)	VGG16	GooleLe Net	ResNet-50(v1)
Top-5 errors	n/a	16.4	14.2	7.4	6.7	5.3
Input size	28x28	227x227	231x231	224x224	224x224	224x224
Number of Conv Layers	2	5	5	16	21	50
Filter Size	5	3,5,11	3,7	3	1,3,5,7	1,3,7
Number of Feature Maps	1,6	3-256	3-1024	3-512	3-1024	3-1024
Stride	1	1,4	1,4	1	1,2	1,2
Number of Weight	26k	2.3M	16M	14.7M	6.0M	23.5M
Number of MACs	1.9M	666M	2.67G	15.3G	1.43G	3.86G
Number of FC layers	2	3	3	3	1	1
Number of Weights	406K	58.6M	130M	124M	1M	1M
Number of MACs	405K	58.6M	130M	124M	1M	1M
Total Weights	431K	61M	146M	138M	7M	25.5M
Total MACs	2.3M	724M	2.8G	15.5G	1.43G	3.9G

4. Discussion

Although a lot of work has improved the accuracy of image recognition, there are still some problems to be solved.

(1) Network structure lightweighting. People are more interested in the practical use of these technologies as deep learning networks and object detection, picture recognition, and classification

DOI: 10.54254/2755-2721/5/20230657

technologies grow and become more advanced. At present, some technologies have been gradually applied to every bit of life. However, in most cases, there is the problem of lightweight on mobile. Mobile or embedded devices are limited by small memory, power consumption, and low processor performance. Conventional high-precision models simply cannot be used on these devices. Therefore, the lightweight network develops rapidly. At present, the main means are lightweight design, knowledge distillation, pruning and other ways to process the model. MobileNet is also a network architecture designed for lightweight networks.

(2) Unsupervised Learning. In machine learning, most people need to manually input data, classify and label the data. In such large-scale data learning, the act of training a machine learning model is supervision. However, the cost of manually labeling the data is too high, and it is difficult for the machine to learn the unlabeled data by itself. This is the unsupervised problem. In unsupervised problems, clustering is often used, where things with the same or close similarity are grouped together by calculation. The development of unsupervised learning will make a great contribution to the detection of some malware or human input errors. Perhaps in the future, unsupervised learning will have a breakthrough when it comes to analyzing information not just by "looking" but by giving machines other sensory functions similar to humans.

(3) Overfitting and underfitting. In image recognition, the learning of image data is too good or too poor will lead to problems. When the model overlearns the data, it will also learn the features of the noise data, so that in the subsequent training, it will not be able to learn the data correctly and do the correct classification. Such a situation is called overfitting. However, when the model does not capture any data well in the learning process, so that the model is almost in invalid learning, this situation is underfitting. Today's neural networks are powerful enough that the problem of underfitting is relatively rare and easy to solve. Overfitting is a major concern, and overfitting is most commonly solved by adding training data, improving the model to reduce the number of layers or parameters, and removing data noise.

(4) Degradation. Although the deeper the network structure, the better the performance and the higher the accuracy, the gradient correlation is getting worse and worse. At present, it is not infinite positive correlation, and the accuracy will decrease greatly after reaching a certain limit, even worse than the shallow network. And it doesn't happen because the gradient disappears. This is the network degradation, and the network degradation is not the problem of the network structure itself, but the training method is not ideal. Currently, the best way to reduce the number of layers and parameter values for network degradation is the residual structure introduced with ResNet.

5. Conclusion

This paper first introduces five classic network structures, Alexnet, VGGNet, ResNet, GoogLeNet, and MobileNet, respectively introduces the unique advantages and disadvantages of each network structure, and compares the data results. Their development basically runs through the main process of convolutional neural network structure, each of which plays a role that can not be underestimated. Then, the relevant information of three classical datasets, ImageNet, Mnist and Cifar, and the related indicators and concepts of image classification are introduced. Finally, four common problems in deep learning networks are introduced and common solutions are provided. They are lightweight problem, unsupervised problem, overfitting and underfitting problem, and network degradation problem.

References

[1] Sultana F, Sufian A, Dutta P. Advancements in image classification using convolutional neural network[C]//2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). IEEE, 2018: 122-129.

[2] Islam M, Nooruddin S, Karray F, et al. Human Activity Recognition Using Tools of Convolutional Neural Networks: A State of the Art Review, Data Sets, Challenges and Future Prospects[J]. arXiv preprint arXiv:2202.03274, 2022.

[3] Rahim R, Nadeem S. End-to-end trained CNN encoder-decoder networks for image

DOI: 10.54254/2755-2721/5/20230657

steganography[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018: 0-0.

[4] Li Y. Research and application of deep learning in image recognition[C]//2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA). IEEE, 2022: 994-999.

[5] Agarwal T, Mittal H. Performance comparison of deep neural networks on image datasets[C]//2019 Twelfth International Conference on Contemporary Computing (IC3). IEEE, 2019: 1-6.

[6] Upreti A. Convolutional Neural Network (CNN). A Comprehensive Overview[J]. 2022.

[7] Xuan X, Zhang X, Kwon O H, et al. VAC-CNN: A Visual Analytics System for Comparative Studies of Deep Convolutional Neural Networks[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(6): 2326-2337.

[8] Yu W, Yang K, Bai Y, et al. Visualizing and comparing AlexNet and VGG using deconvolutional layers[C]//Proceedings of the 33 rd International Conference on Machine Learning. 2016.

[9] Hayou S, Clerico E, He B, et al. Stable resnet[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2021: 1324-1332.

[10] Wu W, Pan Y. Adaptive Modular Convolutional Neural Network for Image Recognition[J]. Sensors, 2022, 22(15): 5488.

[11] Sun R, Li D, Liang S, et al. The global landscape of neural networks: An overview[J]. IEEE Signal Processing Magazine, 2020, 37(5): 95-108.

[12] Trottier L, Giguere P, Chaib-Draa B. Parametric exponential linear unit for deep convolutional neural networks[C]//2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2017: 207-214.

[13] Patel S. A comprehensive analysis of Convolutional Neural Network models[J]. International Journal of Advanced Science and Technology, 2020, 29(4): 771-777.

[14] Alom M Z, Taha T M, Yakopcic C, et al. The history began from alexnet: A comprehensive survey on deep learning approaches[J]. arXiv preprint arXiv:1803.01164, 2018.

[15] Wang W, Yang Y, Wang X, et al. Development of convolutional neural network and its application in image classification: a survey[J]. Optical Engineering, 2019, 58(4): 040901.