Research on the Application of Convolutional Neural Networks in the Image Recognition

Gu Hongxian
[1.]School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
[2.]State and Provincial Joint Engineering Lab of Advanced Network, Monitoring and Control
Xi'an, 710021, China
E-mail:15757118020@163.com

Gao Zhiyu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail:516750613@qq.com

Liu Bailin
[1.]School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
[2.]State and Provincial Joint Engineering Lab of Advanced Network, Monitoring and Control
Xi'an, 710021, China
E-mail:498194312@qq.com

Mu Jing
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail:55897293@qq.com

***Abstract*—Over the past few years, benefits from the strong feature extraction advantages of Convolutional Neural Networks (CNN)themselves and the efforts and applicationby researchers making, research work on CNN in the field of image recognition has yielded many results and achieved the best performance in classification and regression tasks. This paper focuses on the improvement and application history ofCNNand summarizes the direction of improvement and optimization of CNN in recent years from the perspective of the structure of CNN themselves and their applications in various fields. Finally, this review is summarized with a further outlook on the development direction of CNN.**

Keywords-Image Recognition; Deep Learning; Machine Learning; Convolutional Neural Networks

I. INTRODUCTION

Since the concept of deep learning was proposed by Hinton et al[1]. In 2006, during more than a decade of development, machine learning is closer to the original goal of "artificial intelligence". Deep learning is a hierarchical machine learning approach that involves multiple levels of nonlinear transformations that learn the inherent laws and representation levels of sample data, and the feature information obtained in the process of learning can help the machine achieve analytical judgments about the data.

Compared to traditional machine learning methods, it has achieved good results in search technology, image recognition, machine translation, natural

DOI: 10.21307/ijanmc-2020-014

language processing, multimedia learning, speech, recommendation and personalization technologies.

With the practice of researchers in various fields, many network models have been proposed, such as DBN (Deep Belief Network), CNN (Convolutional Neural Network), RNN (Recursive Neural Network), etc. The introduction of CNN into the field of image recognition has taken researchers a long time to explore and practice. Image recognition technology originated in the 1940s, when it was not rapidly developed due to inadequate technology and inadequate hardware facilities. It was not until the 1990s that artificial neural networks combined with support vector machines(SVM) facilitated the development of image recognition technology, which was widely used. However, traditional image recognition techniques are based on shallow structural models, which require human pre-processing of the image, resulting in reduced accuracy of image recognition. As computer hardware levels and GPU evolved, researchers began to work on deeper models of network structure, and in 2012, Krizhevsky et al. reduced the error rate of the tested Top-5 to 15.3% in ImageNet's large-scale visual recognition challenge competition based on CNN, 10.9% lower than the error rate of the second-place team's Top-5, showing the great potential of deep models. In the following years, CNN have made leaps and bounds in digital image recognition and processing with their powerful feature extraction capabilities.

II. OVERVIEW OF CNN

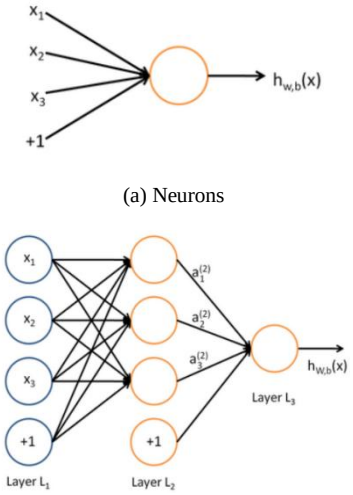
CNN, compared to other network models, are better able to adapt their structures to image structures while extracting features and classifying them, with outstanding performance in image processing. In addition, its weight sharing feature educes the training parameters of the network, which makes the network structure simple and more generalizable, and has become a current research hotspot.

A Development history

The prototype of the CNN is the BP algorithm proposed by Rumelhart in 1986[2].In the 1990s, Lecun proposed the LeNet-5 model[3], which was mainly applied to image classification of handwritten numbers, used the stochastic gradient descent method and reverse propagation method for supervised training of the CNN, and achieved the best recognition results on the MNIST dataset[4], laying the foundation of modern CNN In 2006, Hinton proposed the concept of deep learning in his paper and pointed out that multi-cryptic neural networks have better feature learning capabilities and their complexity in training can be effectively mitigated by layer-by-layer initialization[1]. In the next few years, the development of CNN has also had some achievements, thanks to the substantial update of computer hardware devices and the rapid development of GPU.In 2012, the ImageNet competition, the model based on CNN took the first place with a 10% accuracy rate higher than the second place, and was once again pushed to the deep research boom by scholars.In 2014, the Computer Vision Group of Oxford University and Google DeepMind jointly developed VGGNet[5], and won the first and second place in the ImageNet competition respectively.In 2015, Kaiming He et al. proposed the residual neural network ResNet[6], which solves the problem of deep networks being difficult to train by fitting the residual term with cross-layer connections.Although the number of network layers reaches 152, the complexity is lower and the Top-5 error rate on ImageNet is only 3.57%.

B Basic structure and working principle

The basic building blocks of a CNN are also neurons one by one, containing weights with learning abilities and paranoid term constants. When multiple neurons are combined with a hierarchical structure, a neural network model is formed. A figurative representation of both is shown in Figure 1.



(a) Neurons

(b) Neural networks

Figure 1. Neurons and neural networks.

CNNmodels, are neural network modelsthat contain feature extractors consisting of convolutional and pooled layers. A typical network structure is shown in Figure 2, which includes five parts: input layer, convolution layer, pooling layer, full connection layer, and output layer. Among them, the convolutional layer and the pooling layer are the core modules to realize the feature extraction function of convolutional neural networks.

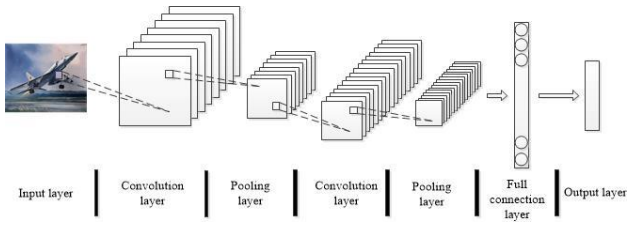


Figure 2. Typical structure of a CNN

CNN is a multilayered supervised learning neural network structure with the following workflow: A series of pre-processing operations are performed on the data at the input level, such as data normalization, de-normalization, etc. Entering the convolutional layer, the image of the input layer is convolved with a convolutional nucleus, and then the activation function outputs a feature extraction diagram of the layer, which is expressed as (1).

$$X_j^{i-1} = \sum_{i \in M_j} X_j^l * W_{ij}^l + b_j^l \quad (1)$$

Wherein, f(·) represents theactivation function,$M_j$represents the set of images participating in the current convolutional layer operation, X_i^{l-1}represents the value of a certain pixel input to the current layer image, $*$represents the convolutional operation, W_{ij}^l represents the weight vector of the l layer convolutional nucleus, and b_j^l represents the paranoid term of the l layer. Currently the commonly used activation functions are: Sigmod function, Tanh function, ReLU function, etc[7].

Connecting the pooled layer after the convolutional layer allows a certain degree of feature invariance and can reduce the amount of data. The input feature map is split into non-overlapping regions in the layer, and for each subregion features are further extracted by pooling operations, the common pooling operations being average pooling and maximum pooling.

After a number of alternating convolutional and pooling layers, multiple sets of highly abstracted feature maps are obtained; then the fully connected layers are entered and the multiple sets of feature maps are combined into one feature map. Then based on business needs, the final output of classification or identification results.

The goal of training is to minimize the loss function $L(W, b)$of the network, so the weights and biases of each layer need to be constantly updated during the training.

Common loss functions are Mean Squared Error (MSE) function, Negative Log Likelihood (Negative LogLikelihood, NLL) function, etc. In practice, in

order to reduce the occurrence of overfitting, the loss function increases the L2 parametre to control the overfitting of the weights and the parameters to control the strength of the overfitting effect.

$$J(W,b) = L(W,b) + \frac{\lambda}{2} W^T W \quad (2)$$

The weights and biases were updated as (3) and (4).

$$W_{ij}^{l} = W_{ij}^{l} - \alpha \frac{\partial}{\partial W_{ij}^{l}} J(W,b) \quad (3)$$

$$b_{i}^{l} = b_{i}^{l} - \alpha \frac{\partial}{\partial b_{ij}^{l}} J(W,b) \quad (4)$$

C Significance of the study

CNNhave been so successful in a number of applications and researchers have moved on to other areas, which brings more challenges. In terms of the CNN itself, that's where the research comes in.

1) Refinement of the theoretical system through domain application effects.

CNN have undergone more than 70 years of bumpy development, from MP models, BP neural networks, to various deep learning networks that are popular nowadays, all of them are judged directly by experimental effects, and there has been no complete set of theories for mathematical verification of these methods. Thus, as the field of application of CNN expands, it will also promote theoretical research in the field of CNN to a certain extent.

2) Facilitate the optimization of neural network structures and extend their application value.

At present,CNN have a place in natural language processing, image recognition, speech processing and other fields, and theirtrend is positive. And the network still has the problem of gradient disappearance, training sample size limit, computing power limit, is the short board of its development. For the problem that networks are difficult to train, an analysis of the problems related to network training is also given in the literature[8], but the solutions given have not become mainstream. Therefore, there is an urgent need to improve the learning capabilities of deep neural networks, so that the networks have better generalization capabilities and can be adapted to more complex application scenarios.

III. IMPROVEMENT AND OPTIMIZATION OF CNN

In recent years, improvements in CNN have been driven primarily by factors such as final detection effectiveness, network operating efficiency, and computational complexity. As of now, the improvement and optimization of CNN are mainly considered in three aspects: network depth, network structure, and network training methods.

A Network depth

Lecun et al. designed the Lenetnetwork[9], which uses alternately connected convolutional and pooled layers, and eventually passes through full-connected layers.There is 5 layers in Lenet and Lenet became the originator of CNN.In 2012, Krizhevsky proposed AlexNet[10], a network model with five convolutional layers, some of which are followed by a pooled layer for downsampling and finally two fully connected layers. The last layer is the softmax output layer, with 1000 nodes, corresponding to 1000 image classifications in the ImageNet graph set and using the Dropout mechanism and ReLu function, which has improved the accuracy and training time. Subsequently, the VGGnet proposed by Karen et al. uses almost all 3 $\times$3 convolutional nuclei, while adding pooled layers after several convolutional layers, instead of pooling immediately after each convolutional layer, to guarantee the depth of the network in many ways.VGGnet[11]demonstrated that increasing the number of network layers is beneficial for improving the accuracy of image classification. This increase is not unlimited and too many layers can create network degradation problems. The number of layers that affect the test results VGGnet was finally determined in two

versions, 16 and 19 layers. In the following years, there were different teams of researchers who proposed GoogleNet (22 layers), ResNet (152 layers) and they all deepened in terms of network depth and got better and better.

In terms of network depth alone, increasing the network level has an effect on the learning effect of convolutional neural networks, which also confirms the need for deeper learning.

B Improvements of the network structure

Improvements to the network structure mainly revolve around the idea of reducing network complexity. In 2014, Google proposed GoogLeNet[12], whose main innovation is the Inception mechanism, which sets up different convolutional cores in the same layer, i.e., multi-scale processing of images, and adding 1*1 convolutional core before 3*3, 5*5 for dimensionality reduction, which reduces parameters and improves the accuracy of image recognition on ImageNet datasets by about 10%.In 2015, Springenberg J T et al. proposed a full convolutional structure in literature[13]. Instead of the classical pairing of alternating convolutional and pooled layers in a classical convolutional neural network, the stride convolutional layer is used instead of the feature extraction layer in this structure. It was found that the error rate of this new network structure was reduced by 10 percentage points compared to traditional convolutional neural networks, and it was found that in some cases, the addition of apooling layer to this network structure resulted in weakened performance. Lin M et al. proposed a network-in-network structure in the literature, Network In Network[14]. is a subversion of the traditional structure of convolutional neural networks. The mesh structure replaces the full connection layer with a global averaging pooling layer, allowing the input feature map to be classified directly at the output, improving performance. However, this structure makes the convergence process longer and, on the whole, is not a very effective network structure.

The Google DeepMind team proposed a self-contained transformation module to flexibly and efficiently extract image invariant features[15], the specific structure of the module is shown in Figure 3, the model named Spatial Transformer is mainly composed of three parts: Localization Network, Grid generator and Sam-pler. Localization Network uses input feature mapping and outputs spatial transformation parameters through multiple implicit layers; Grid generator uses predictive transformation parameters to create sampling grids; Sampler uses feature mapping and sampling grid as inputs to generate output mapping from grid points.

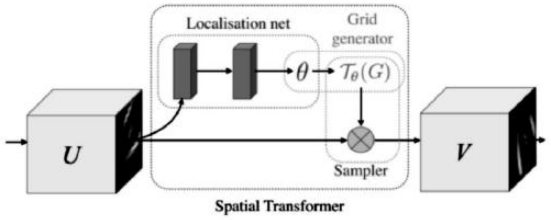


Figure 3. Invariant image feature

This model is used to good effect: it is a self-contained module that can be added anywhere in the network structure (not just the CNN), and there are no limits. It is easy to differentiate and can be used directly for end-to-end training. Its easy-to-differentiate and fast nature allows it to be added to the structure without slowing down the original network. The features extracted by this method are more common than the conventional structures.

Shanshan Xu[16] proposed an algorithm for progressively extending the network structure of a convolutional neural network to optimally adjust the network structure to make the network suitable for real-world problems, and at the same time proposed an improved feature extraction method to realize that feature extraction does not occur on its own in the network, and applied this method to handwritten number recognition, and the experimental results

showed that the recognition accuracy and recognition efficiency were higher than other algorithms using the classification method of convolutional neural networks.

C Methods of network training

1) The dropout mechanism[17] was added to the eliminate overfitting.

The Dropout proposed by Hinton et al. effectively improves the generalization performance of the network by randomly ignoring a certain percentage of node responses during training compared to traditional fully connected neural networks. However, the performance improvement of Dropout for convolutional neural networks is not significant, mainly because convolutional neural networks greatly reduce the number of training parameters compared to fully connected networks due to the weight sharing properties of convolutional nuclei, which itself avoids the more severe overfitting phenomenon. Based on the idea of Dropout, Wan et al[18]proposed the Drop Connect approach. Unlike Dropout which ignores some of the node responses of the full connection layer, Drop Connect randomly disconnects a certain percentage of the neural network convolutional layer. Forconvolutional neural networks, Drop Connect, which acts on the convolutional layer, has a much stronger past-fitting capability than Dropout, which acts on the full-connect layer.

Although the functional layer of Dropout and Drop Connect are different, the underlying principle is to increase the sparsity or randomness of network connections in network training in order to eliminate overfitting and improve network generalization capabilities.

2) Training methods using knowledge transfer

Overfitting and gradient dispersion are prone to occur when training against conventional convolutional neural networks. A training strategy using knowledge migration was proposed by Rocco et al. in the literature[19]. Pre-training (Pre-training with soft target, PST) is first performed on the soft target (soft target is a class distribution containing information between sample classes), and the migration of the soft target from the source model to the target model in the same domain allows more supervisory information to be obtained from a limited sample than from a single tag, solving the problem of missing samples. Then the target model adjacent convolutional layer is divided into a module to learn the low-level features of the source model in a modular way, similar to DBN's layer-by-layer pre-training strategy, and the combination of MMT and PST, the sample class information and low-level features of the two knowledge migration at the same time, so that the model convergence to a better position, and then use the SGD algorithm fine-tuning, so that the generalization performance is greatly improved.

Ⅳ. Application of CNN in Image Recognition

Image classification is an image processing technique that identifies different things by the characteristic information given by the image. With the rise of machine learning, automatic image classification techniques have been applied in various development fields. Cao used the classical convolutional neural network VGG-16 as a prototype in the literature[20], and added a multi-scale sampling layer at the end of the convolutional part, so that the model can input any size of images for training and testing, while reducing the number of neurons in the full connection layer, which improves the training speed of the model while ensuring accuracy, and applies it to the problem of multi-attribute classification of human faces. In the literature[21], Wenxu Shi proposed a CNN-based multiscale approach combined with a feature extraction algorithm for reverse convolutional networks (MSDCNN) and classified adenocarcinoma pathology images. Classification experiments performed on adenocarcinoma pathology cell images showed that the MSDCNN algorithm improved the classification

accuracy of the final convolutional feature scale by about 14% over the conventional CNN algorithm and about 1.2% over the classification accuracy of the fusion network model approach, which is also based on multi-scale features. In the literature[22], Chunlei Zhang proposed a parallel network model based on convolutional neural networks for military target image classification. The method uses two edge detection operators to extract the target image features separately and then input them into the convolutional neural network for deep feature extraction, which improves the classification recognition rate by 1.2% and reaches 97% compared to theconventional convolutional neural network. The theoretical analysis and experimental data illustrate that the model enables the effective differentiation of military target image data and is important for the accurate provision of military operational information.

Target detection is a fundamental problem in the field of machine vision as well as artificial intelligence, whose main goal is to pinpoint the category and location border information of various targets in an image. Target detection algorithms based on convolutional neural networks, such as RCNN, Fast RCNN, Faster RCNN, Mask RCNN, etc., are widely used in security monitoring, intelligent transportation and image retrieval and other fields. A target detection algorithm based on multi-scale feature extraction was proposed by Jianghao Rao in the literature[23] and applied to the detection of infrared pedestrian small targets with better results than conventional networks. A Mask-RCNN-based method for building target detection was proposed by Dajun Li et al. in the literature[24]. In the literature[25], Ding Peng "fine-tuned" a variety of mainstream depth convolutional neural networks based on Faster RCNN detectors on two classical data sets for target detection of optical remote sensing images. In response to the problem of target detection in traffic roads, Zhang Qi et al. proposed a traffic target detection method based on anchor point clustering, all-anchor point training strategy and reinforced intersection and merger ratio (SIoU) in the literature[26].

Ⅴ. CONCLUSION

This paper describes the basic structure of classical CNN from the development of CNN, briefly analyzes the features of CNN that have been improved and optimized in the development process, and finally elaborates on the wide application of CNN in the field of image classification and target detection. CNN have been developed to date and occupy an important position in the field of image recognition.

In terms of current trends, CNN will continue to evolve and makeCNNsuitable for various application scenarios, such as 3D CNN facing video understanding. There are also challenges, such as limited data sets, network generalization performance, robustness to be improved, and high training costs. The aforementioned issues will be a direct driver for the future development of convolutional neural networks and will directly contribute to the further deepening of artificial intelligence.

ACKNOWLEDGMENT

This work is supported by the Natural science foundation of Shaanxi province(2019JM-603).

REFERENCES

[1] Hinton G E, Salakhutdinov R R , Reducing the dimensionality of data with neural networks[J].Science, 2006,313(5786):504-507.

[2] Rumelhart D E, Hinton G E, Williams RJ. Learning representations by back-propagating errors[J].Nature, 1986, 323:533-538.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, november 1998.

[4] LeCun Y, Cortea C. MNIST handwritten digit database[EB/OL].http://yann.lecun.com/exdb/mnist, 2010.

[5] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition.

[6] He K, Zhang x, Ren S, et al. Deep residual learning for image recognition[EB/OL].http://arxiv.org/abs/1512.03385,2015.

[7] Gu Jiu-Xiang, Wang Zhen-Hua, Jason Kuen,etal. Recentadvances in convolutional neural networks. arXiv:1512.07180v5,2017.

[8] X. Glorot, Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. AISTATS, 2010

[9] Lecun Y, Boser B, Denker J S, etal. Back propagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4):541-551.

[10] Krizhevsky A, Sutskever I,Hinton G E.Imagenet classifi-cation with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems, 2012 : 1097-1105.

[11] Bengio Y, Simard P, Frasconi P.Learning long-term depen-dencies with gradient descent is difficult[J].IEEE Transac-tions on Neural Networks, 1994, 5(2):157-166.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going Deeper with Convolutions, Arxiv Link: http://arxiv.org/abs/1409.4842.

[13] Rougui J E, Istrate D, Souidene W, et al. Audio Based Sur-veillance for Cognitive Assitance Using a CMT Microphonewithin Socially Assistive Living [J]. 2009, 2009 (2009):2547-2550.]

[14] Lin, Min, Qiang Chen, and Shuicheng Yan. Network in network. arXiv preprint arXiv:1312.4400

[15] Zang D, Chai Z, Zhang J, et al. Vehicle license plate recog-nition using visual attention model and deep learning[J].Journal of Electronic Imaging, 2015, 24(3):033001.

[16] Xv Shanshan. Research and application of convolutional neural network [D]. Nanjing forestry university, 2013.

[17] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaption of feature detectors[R /OL].[2015-10-26]. http: / / arxiv. org / pdf /1207. 0580v1.

[18] Wan L, Zeiler M, Zhang S, et al. Regularization of neural networks using drop connect [C] / / Proceedings of the 2013 Inter-national Conference on Machine Learning. New York: ACM Press, 2013: 1058－1066.

[19] Luo Ke, Zhou Anzhong,Luo Xiao. A Convolutional neural network training strategy using knowledge transfer[J]. Control and decision-making, 2019, 034 (003) : 511-518.

[20] Cao Ge. Research on face image classification application based on deep convolutional neural network [D]. Jilin university, 2019.

[21] Shi Wenxu, Jiang Jinhong, Bao Shengli. Application of improved convolutional neural network in pathological image classification of adenocarcinoma[J]. Science, technology and engineering, 209,19 (35) : 279-285.

[22] Zhang Chun-lei. Military target image classification technology based on parallel convolutional neural network[J]. Electronic design engineering, 209,27 (08) : 76-80.

[23] Rao Jianghao. Target identification, positioning and detection based on CNN integration [D]. University of Chinese academy of sciences (institute of optoelectronic technology, Chinese academy of sciences), 2018.

[24] Li Dajun, He Weilong, Guo Bingxuan,Li Maosen, Chen Minqiang. Building target detection algorithm based on Mask - RCNN[J]. Science of surveying and mapping, 202,44 (10) : 172-180.

[25] DingPeng. Research on optical remote sensing target detection technology based on deep convolutional neural network [D]. University of Chinese academy of sciences (changchun institute of optics, precision mechanics and physics, cas), 2019.

[26] Zhang Qi, Ding Xintao, Wang Wanjun, SZhou Wen. Traffic target detection method based on Faster RCNN [J]. Journal of west anhui university, 202,35 (05) : 50-55.