Summary of CNN Algorithm for Image Recognition

Ningyuan Feng[a]
School of Data Science, City University of Macau, Kunming, China

Keywords: Convolutional Neural Network, Machine Learning, Deep Learning.

Abstract: With the advancement of technology, the way computers get information is also constantly improving, from the Linux system that can only use code input to the text input of the later Windows system, language character recognition system, to today's more advanced image recognition system. New technologies are constantly emerging to refresh people's understanding of it, but also continue to facilitate people's lives. This paper will analyse the principle and logic of image recognition by computer CNN algorithm from the perspective of a computer, and explain the process of feature extraction, feature analysis, and feature classification of images by hidden layers such as the convolution layer, pooling layer and fully connected layer. At the same time, the paper also studied and explored the advantages and disadvantages of each layer of the hidden layer and tried to put forward corresponding solutions in combination with subsequent studies. For example, the limitation of the receptive field of the convolutional layer led to the decline of robustness and accuracy. Therefore, this shortcoming can be remedied by introducing a residual mechanism or attention mechanism. Finally, a reasonable analysis of the future algorithm direction is made according to the existing research.

1 INTRODUCTION

The key to the long-term survival of human beings in nature is the ability to quickly perceive and understand the environment. This process relies on human's visual system to accurately lock the target, identify the target, and then achieve a thorough understanding and vivid description of the visual scene. Therefore, in the current flourishing of science and technology, if the computer can also apply the magic skills of automatic image recognition like the human visual system, it will bring many earth-shaking changes to human life. For example, when traveling, intelligent navigation can accurately identify road conditions and skilfully avoid congestion. In shopping, the e-commerce platform can realize virtual trying on, breaking the restrictions of both time and space. In the medical field, image recognition identifies areas where problems may occur to help doctors accurately diagnose conditions. All kinds of conveniences make image recognition technology stand out in artificial intelligence, become the focus of attention at present, and become one of the important research directions in the field of artificial intelligence. In this paper, the general structure logic of graph convolutional neural networks will be studied in order to arouse readers' interest and provide guidance for beginners.

2 CORRELATIVE PRINCIPLE OF IMAGE FEATURE RECOGNITION

In today's diversified application fields of science and technology, the practical problems are complicated, which makes it difficult to frame image feature extraction in an accurate and fixed definition category. In fact, the construction of many computer image analysis projects and related algorithms cannot abandon the core point of "features". The algorithm ultimately achieves the ideal effect and successful landing, depending on whether the selected and defined features are accurate and appropriate.

In the process of image processing, the most basic and critical step is feature extraction, which aims to extract key information features of images and then extract geometric parameters and texture features based on shape features to simplify complex images

[a] https://orcid.org/0009-0001-1396-7654

Feng, N.
Summary of CNN Algorithm for Image Recognition.
DOI: 10.5220/0013680700004670
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 182-187
ISBN: 978-989-758-765-8

into a matrix. After completing a series of smoothing processes. Derivative rule will be used to carry out fine operations on the image, so as to successfully mine and calculate the feature information contained in the image. This step will provide strong basic support for subsequent higher-order image processing tasks like image classification and target recognition (Rosenfeld,1969).

3 HISTORY OF CONVOLUTIONAL NEURAL NETWORKS

The basis of a convolutional neural network was proposed in 1998 by Yann LeCun, who was the pioneer to successfully apply the CNN model and successfully build the general framework of the model. However, due to the backward computer technology and the limitation of data at that time, the research results did not get widespread attention. This part of the study continued until 2012 when Krizhevsky et al. trained an 8-layer depth model with ImageNet data. CNN algorithm have attracted much attention in image classification and recognition and achieved great success. However, this algorithm also has a lot of data and relatively low recognition efficiency. Different tasks can not be flexible and other problems. In 2014, the Google team and the team of Oxford University tried to adapt the size of the convolutional layer to the multiple requirements of different environments or tasks, and launched their own models Google Net and VGG Net, which successfully improved the recognition efficiency and accuracy on the basis of reducing parameters (Sun, Xue, Zhang,2020). Subsequently, methods such as residual learning, lightweight design, attention mechanism, self-supervision and integration with other models (transformer) are constantly introduced to improve convolutional neural networks so that they can better serve today's society.

In the process of continuous improvement, it is found that the CNN algorithm can learn image features layer by layer. General features such as edges, corners, and textures are extracted from the bottom layer. On this basis, the top layer combines specific features for specific tasks. It is like simulating the hierarchical information processing mechanism of the human brain, mining image features directly from the original pixels. It can be seen from this that the team improved this model in three main ways: First, the deep network is trained directly on the data set to be classified, and the increase of CNN depth and width can improve the classification performance. For example, Simonyan et al. proposed a 19-layer VGG-19 model and extended the depth of the original model with a small convolutional filter kernel (3×3), which is convenient for practice. Second, inspired by the Hebbian principle and multi-scale processing, Szegedy et al. proposed a 22-layer Google Net, which was stacked with multiple Inception models and used convolution kernels of different band sizes to capture multi-scale visual features and adapt to the apparent multi-scale characteristics of image objects. Third, for different classification tasks, the model trained by Zhou et al on Places has an excellent effect on scene classification. With the continuous improvement of these three approaches, image recognition of convolutional neural networks is becoming more and more mature, from only recognizing static pictures at the beginning to gradually recognizing general parts of videos to today's video precision recognition, such as recognizing finger movements, recognizing high-altitude projectiles and so on. All kinds of conveniences greatly facilitate people's lives, and also make the image recognition function of convolutional neural networks attract attention and research in various fields (Bhatt, Patel, Talsania,2021).

4 DEFINITION AND COMPOSITION LOGIC OF CONVOLUTIONAL NEURAL NETWORKS

4.1 Definition of Convolutional Neural Networks

After understanding the relevant definition of image recognition and the development of its main basic algorithms. Let's talk about the logic and composition of the definition of convolutional neural networks. As a very representative algorithm in deep learning, the convolutional neural network belongs to the category of the feedforward neural network, which has a unique convolutional computing mechanism and depth hierarchy. It has a strong representation learning ability and can carry out accurate classification operations on input information according to its own hierarchical structure. This kind of algorithm simplifies the analysis and classification of different images through its own construction and finally gives the recognized results (Srivastava, Divekar, Anilkumar,2021).

4.2 Convolutional Neural Networks Are Composed to Run Logic

Convolutional neural networks are divided into three layers: input layer, hidden layer and output layer. The hidden layer is the focus of this algorithm. After processing and analysing the image, the content and classification of the image can be accurately identified. This layer is divided into three parts: the convolution layer, the pooling layer, and the fully connected layer. Next, this paper discusses the content and purpose of these layers respectively and introduces the initial advantages and disadvantages of the algorithm as well as the improvement methods and effects of the model later:

4.2.1 Convolution Layer

When the input layer receives the image content, the convolutional layer starts the feature extraction of the input data. The convolution layer contains many convolution nuclei, and each element of the convolution kernel corresponds to a weight coefficient and a deviation quantity, respectively, which is similar to the setting of neurons in a feedforward neural network. Multiple neurons in adjacent regions are connected. The size of this region is determined by the size of the convolution kernel, also known as the "acceptance field", which is responsible for sensing, recognizing, and processing information from the corresponding region. In short, the layer algorithm uses the receptive field to map each pixel of the input image accurately and systematically and submits the processing results of each neuron to the next layer of pooling.

The original convolution layer can make good use of the convolution kernel to process graph structure data and mine the relationship between nodes in the graph. Even the graph convolution layer can be calculated using the same convolution kernel on different nodes, which realizes parameter sharing. This greatly reduces the number of parameters in the model, reducing computational costs and the risk of overfitting. However, there are also many problems, such as in the initial stage of graph convolutional neural networks, the receptive field of the convolutional layer is relatively small, and only the information of the nearby neighbours around the nodes can be captured. For some tasks that require long-distance information dependence, a single convolutional layer may not be able to model well, and multiple convolutional layers need to be stacked to expand the receptive field. However, unlimited expansion of the receptive field will increase the hardware requirements and reduce the efficiency and accuracy of recognition. Moreover, the initial convolutional layer relies on nodes to locate the target region of the image. It is difficult to capture the global structure of a graph directly. For some tasks that require global information to make decisions, additional mechanisms or modules may be required to supplement the global information.

In order to solve the problem of limited receptive field of view, the improved model ResGCN introduces residual connection so that the input information can directly skip some convolutional layers and add the output of the subsequent layer. In this way, multiple convolutional layers can be stacked to enlarge the receptive field while retaining the original information in the propagation process, which solves the problems of increasing hardware requirements and decreasing recognition efficiency and accuracy caused by simply stacking convolutional layers. At the same time, the model GAT also introduces the attention mechanism. When calculating the features of nodes, the model can assign weights to the neighbours of nodes so that the model can better capture the global information of images and enable the images to perfectly map the required images into the system through the convolutional layer (Meng, Meng, Gao,2020)

4.2.2 Pooled Horizon

After the feature collection on the previous level, the algorithm maps the features of the convolutional layer to the pooling layer for feature selection and information filtering. The pooling layer is also preconfigured to replace all the results of each isolated point in the feature map with the feature map statistics of the adjacent regions of the point. For example, the value of a single pixel on the original feature map is reassigned based on the mean, maximum, or other statistical indicator of the adjacent area. The steps of selecting the pooling region in the pooling layer are similar to those of scanning the feature map of the convolution kernel, which are controlled by the pooling size, step size and filling to ensure that the pooling process can cover the entire feature map uniformly and completely.

The initial pooling layer can down-sample the graph data and appropriately adjust the number of nodes or feature dimensions according to the requirements of the task and even retain the key features of the picture data, omitting some details. It highlights the main structure and features of the graph to improve its generalization ability while ensuring that the requirements of the task are met so that the model can process large-scale graph data more quickly or effectively. In this way, although the main framework and structure of the general diagram can be preserved, the pooling operation is essentially a down-sampling process, so some information will inevitably be lost during the pooling process. Moreover, it is not difficult to find that the importance of the pooling layer to the features in the diagram lacks the weight comparison, which may seriously cause the loss of important features, thus affecting the normal progress of the following steps. In particular, the accuracy and performance of the model may be affected when the structure of some complex and detailed graph data is rich. In addition, due to technical limitations at that time, the initial pooling operation often only focused on the feature aggregation of local areas, and the ability to capture the global structure information of the graph was limited. For some tasks requiring global information to make decisions, the analysis results might be biased due to the inability to process and analyse all the images.

For this reason, the model was optimized and the ASA Pooling was used to solve the problem of information loss. Its adaptive structure can learn a soft distribution matrix, and the nodes can be allocated to different clusters for pooling, which can more effectively retain the structure and feature information of the graph and reduce information loss. Of course, the Diff Pool method preserves important nodes and their surroundings through a clustering algorithm to reduce information loss of important nodes as much as possible, but the overall effect is more accurate with ASA Pooling. Faced with the problem of information analysis of global structure, people put forward the Eigen Pooling algorithm, which uses Laplacian matrix to pool the features of the global structure information of a graph, reasonably split the graph into subgraphs, and better extract the features and learn the representation of the global information through their respective analysis (Ranjan, Sanyal, Talukdar,2022).

4.2.3 Fully Connected Layer

The fully connected layer in the convolutional neural network is very similar to the hidden layer in the traditional feedforward neural network, both in function and structure. The fully connected layer is at the end of the hidden layer of the convolutional neural network, and the signal transmission direction is relatively simple and only transmits signals to other fully connected layers. From the perspective of representation learning, the convolutional layer and the pooling layer in the convolutional neural network are mainly responsible for feature extraction of the input data. In contrast, the core function of the fully connected layer is to combine the features extracted from the previous convolution layer and pooling layer to generate the output result. In other words, the fully connected layer itself does not focus on the ability of feature extraction but focuses on how to skilfully integrate the extracted features. After this series of operations, the output of the entire network is finally generated. For example, when you encounter a 5×5×16 feature graph, it means that it has 5 pixel units in each direction of length and width and has 16 channels. Global mean pooling processes each of these 16 channels separately. Global mean pooling will return a vector of 16 where each element is 5×5, step size 5, and mean pooling without padding. When the fully connected layer receives the features extracted by the previous convolution layer and pooling layer, it will improve the feature fusion of all nodes and comprehensively consider the global information to provide comprehensive feature display for the final classification task (Alzubaidi, Zhang, Humaidi, 2021) (Sun, Xue, Zhang, 2019).

Because of its simple structure and strong versatility, this layer is easy to combine and integrate with other types of neural network layers or models, and it is also easy to integrate with other machine learning or deep learning models to expand the function and application scope of the model. However, the fully connected layer itself needs to receive a large amount of information from the convolutional layer and the pooled layer, which consumes a lot of time and resources in model training and reasoning. If too many models are added, the number of parameters may be too large, and the whole fully connected layer will overfit the training data, which will lead to poor learning performance of the entire convolutional network.

In order to reduce the parameters that need to be calculated. Low-rank Approximation technology is proposed to decompose the weight matrix of the fully connected layer into the product of two low-rank matrices. So that fewer parameters can be used to approximate the original weight matrix, thus reducing the amount of calculation and the number of parameters of the model. Of course, some people also proposed the method of Sparse FC, which uses sparse connections to build a fully connected layer so that each neuron needs to be connected to the upper layer of neurons to only connect the part, which can also reduce the number of parameters and reduce the demand for resources. However, although the latter method is simple, it will abandon the accuracy to some extent. Therefore, we should choose different models to solve the problem according to the needs so as to better identify and analyse the image (Astrid, Lee,2018).

4.2.4 Activation Function

The activation function does not run through the entire hidden layer like other layers. In the forward propagation process of CNN, after receiving the input from the previous layer, each neuron first performs linear combination operations (for example, the convolution operation between the convolution kernel and the input feature graph in the convolution layer, and the multiplication operation between the weight matrix and the input vector in the fully connected layer) to obtain an intermediate result, which will be used as the Input to the activation function. The activation function runs a nonlinear transformation of this input to produce the final neuronal output, which acts as the input to the next layer of neurons (Alzubaidi, Zhang, Humaidi, 2021) (Sun, Xue, Zhang, 2019).

According to different needs, the choice of activation function is also different. Here are two more common functions:

In the analysis, if it is only necessary to compare neuronal data or find the maximum value of data, ReLU function is generally used, which can make part of neuronal output become 0, simplify the complexity of the model, reduce the risk of overfitting, and speed up the calculation speed of the model (Ide, Kurita, 2017). However, in the initial function, all output values will be greater than or equal to 0, and 0 cannot be used as the centre, which may lead to an uneven distribution of data received by the later layer. In order to solve this defect, when the input value of the function is less than 0, it will be multiplied by a constant (generally 0.01), so that the function has a certain gradient, so that the model can better update the weight during the training process and improve the robustness of the model.

If the task needs to represent probability or binary classification tasks, simple comparison and finding the maximum function cannot be applied. So the Sigmoid function is generally selected, which maps the data of neurons to a relatively stable space, so that the function curve is very smooth and the probability distribution is also very intuitive. But initially, the function controlled the output value between 0 and 1 in order to stabilize the calculation easily. Therefore, the derivative of the function will approach 0 no matter the value of the neuronal input is very large or very small, which makes it more difficult to train the gradient disappearance model. Later, in the process of continuous improvement, a parameter β was introduced into the Swish function. While retaining the advantages of the smooth function curve of the original function Sigmoid, parameters could be introduced to control the flow of information, which improved the expression ability of the model and achieved good results (Mesran, Yahya, Nugroho, 2024).

Of course, in order to adapt to different needs in different environments, the graph convolutional neural network also introduced functions such as Soft plus, Mish, Tanh, etc., to better analyse different images according to the goal and achieve satisfactory results as possible (Jiang, Xie, Zhang, 2022).

5 CONCLUSION

In this paper, the CNN algorithm and its construction definition are analysed, and the advantages and disadvantages of the more important levels or structures are analysed, as well as the latter's changes to their ideas. However, due to the limitation of permissions, much of the newly proposed confidential content cannot be collected.

In summary, although the current CNN algorithm is more advanced and perfect, there are also more challenges. In terms of overfitting, insufficient amount of training data, too many model parameters or complex structures, and too many training iterations will cause the model to overfit the noise and details in the training data and perform poorly on the new data. In terms of computational efficiency, the CNN model contains a large number of convolutional layers, pooling layers, and fully connected layers that require massive multiplication and addition operations, which requires a huge amount of computation, a large number of resources and a long running time. In terms of scalability, the traditional CNN architecture usually has a fixed hierarchical structure and connection mode. In the face of different types of tasks or data, a large number of modifications and adjustments to the network structure may be needed to achieve good results, and there may be lack of sufficient flexibility.

The future of image recognition is full of unlimited potential and opportunities. As technology continues to advance, on the one hand, it may be possible to gradually advocate lightweight models in the future, using separable convolution structures to facilitate multi-schedule pruning or quantization operations to reduce the number of parameters and calculations.

On the other hand, you may also try to use the trained model to integrate with other architectures, such as Transformer, RNN, GRU. Improve the efficiency and robustness of both sides, and even eventually combine hardware and algorithms to improve the performance of both sides and serve society in more fields.

REFERENCES

Alzubaidi, L., Zhang, J., Humaidi, A.J., et al., 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8, pp.1-74.

Astrid, M., Lee, S.I., 2018. Deep compression of convolutional neural networks with low ‐ rank approximation. ETRI Journal, 40(4), pp.421-434.

Bhatt, D., Patel, C., Talsania, H., et al., 2021. CNN variants for computer vision: History, architecture, application, challenges and future scope. Electronics, 10(20), p.2470.

Ide, H., Kurita, T., 2017. Improvement of learning for CNN with ReLU activation by sparse regularization. In 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, pp.2684-2691.

Jiang, Y., Xie, J., Zhang, D., 2022. An adaptive offset activation function for CNN image classification tasks. Electronics, 11(22), p.3799.

Meng, Y., Meng, W., Gao, D., et al., 2020. Regression of instance boundary by aggregated CNN and GCN. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. Springer International Publishing, pp.190-207.

Mesran, M., Yahya, S.R., Nugroho, F., et al., 2024. Investigating the Impact of ReLU and Sigmoid Activation Functions on Animal Classification Using CNN Models. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 8(1), pp.111-118.

Ranjan, E., Sanyal, S., Talukdar, P., 2020. Asap: Adaptive structure aware pooling for learning hierarchical graph representations. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), pp.5470-5477.

Rosenfeld, A., 1969. Picture processing by computer. ACM Computing Surveys (CSUR), 1(3), pp.147-176.

Srivastava, S., Divekar, A.V., Anilkumar, C., et al., 2021. Comparative analysis of deep learning image detection algorithms. Journal of Big Data, 8(1), p.66.

Sun, Y., Xue, B., Zhang, M., et al., 2019. Completely automated CNN architecture design based on blocks. IEEE Transactions on Neural Networks and Learning Systems, 31(4), pp.1242-1254.

Sun, Y., Xue, B., Zhang, M., et al., 2020. Automatically designing CNN architectures using the genetic algorithm for image classification. IEEE Transactions on Cybernetics, 50(9), pp.3840-3854.