Research and Prospect of Image Recognition Based on Convolutional Neural Network

Hanqing Hu1,*, Jin Lyu2 and Xiaolin Yin3
1Beijing Information Science and Technology University, Haidian 100192, China
2Beijing Guozhixue Culture Co., Ltd, Chaoyang 100101, China
3BeiJing GuoXue Times Culture Co., Ltd, Fengtai 100070, China

*Corresponding author: hanqinghu@bistu.edu.cn

Abstract. This paper compares common image recognition technology, introduces three image recognition and classification technologies, and explains the most popular deep learning image recognition algorithm based on CNN in detail. As a new method of ANN, CNN combines ANN and deep learning technology. In the area of image recognition, the advantage of CNN is particularly prominent. Compared with the traditional image processing algorithm, CNN has higher recognition accuracy and speed, and can directly process the original image to avoid the complicated pre-processing image data. CNN is the first choice for image recognition.

Keywords: Image Recognition, Histogram of Direction Gradient, Principal Component Analysis, Convolutional Neural Network

1 Introduction
Image processing is to transform the image into a digital matrix and store it in a computer, and process it with a certain algorithm. The basis of image processing is mathematics, and the main task is the design and implementation of various algorithms. At present, image processing technology has been widely used in biomedicine, communication technology, remote sensing technology, cultural creativity, industrial design and production and many other fields. Image recognition and classification technology is to extract potentially hidden, useful and even unknown knowledge from a large number of images. Image recognition and classification technology is the intersection of data mining and analysis, machine/deep learning, image retrieval and image processing, machine/computer vision, artificial intelligence technology, database/data warehouse and other disciplines [1].

2 Convolutional Neural Network
The purpose of deep learning is to establish a neural network to simulate the work of human brain neurons. Therefore, deep learning theory is closely related to the development of neuroscience. David Hubel and Torsten Wiesel have discovered that the human visual system is hierarchical. Specifically, when a person sees an object, the picture first maps to the retina, then the edge features of the object are extracted from the V1 area, the local features of the object are extracted from the V2 area, the V4 area is responsible for the overall features of the object, and the higher PFC (prefrontal cortex) is used.
for classification. It can be seen from the whole process of information processing in the brain that information features are extracted from the bottom to the top, and the higher up, the more abstract the features are. The whole process is as follows:

Fig. 1 Brain information extraction process

CNN has a structure of convolution, it can effectively reduce the computer memory consumption and the number of network factors of deep-seated network, and alleviate the problem of model over fitting. CNN is impacted by these key actions: local receptive field, share of weight and layer of pooling.

CNN is a multilayer perceptron [2], it is successful because it adopts the way of local connection and weight sharing. In this way, it not only reduces the weight number, makes the network optimization easy, but also reduces the complexity of the model, that is, the risk of over fitting is reduced. The advantage performance in the input of the network is the image more obvious, and makes the image can be directly as network input, to avoid the traditional recognition algorithm of the feature extraction and data reconstruction of the complex process, in the process of image processing has a lot of advantages, such as network to extract image feature including color, shape, texture and topology structure of the image and the processing of two-dimensional image, especially the identification of displacement, zoom and other forms of distortion invariance applications has good robustness and operation efficiency, etc. [3].

There are lots of factors in neural network, which are prone to over fitting and long training time. However, compared with boosting, logistic regression, SVM and other methods based on statistical learning theory (also can be seen as having a layer of hidden nodes or not the learning model with hidden nodes, known as shallow model), has great advantages.

Deep learning’s two main viewpoints[4] are as follows: (1) The multi hidden layer ANN(artificial neural network) has a good ability of the feature learning. The learned data could better react the vitally important characteristics of the data, and make the data visualization or classification better; (2) The difficulty of DNN(deep neural network) in training can be effectively overcome through unsupervised training layer by layer.

Figure 2 shows the changes of the model. With the deepening of the model, the error rate of top-5 also becomes lower and lower, which has been reduced to about 3.5% at present. In the same ImageNet data set, the error rate of human eyes is about 5.1%, which means that the recognition ability of deep learning has surpassed that of human beings [5].

Fig. 2 The top-5 error rate of ILSVRC over the years

3 Algorithm
3.1 Model Introduction

Figure 3 is a simple CNN structure. The original image of the 1st layer is convoluted to obtain the characteristic image of the second layer with a depth of 3. The 2nd level feature map is combined to get the 3rd level feature map with depth of 3. Repeat the above operations to get the characteristic map of the 4th layer with a depth of 5. Finally, five feature maps, namely five matrices, are expanded by lines and connected into vectors. The whole connection layer is BP neural network. Each characteristic graph in the graph can look upon as a neuron arranged in matrix form, analogue to the neuron in BP neural network.

3.2 Algorithm Implementation

(1) Convolution

After inputting a picture, it is necessary to convert it into a matrix, and the values of each element of the matrix correspond to the values of each pixel of the picture. If there is a $5 \times 5$ image as fig4, a $3 \times 3$ convolution kernel as fig4 is used as convolution, and a $3 \times 3$ characteristic image could be got. The convolution kernel is also called filter.

The specific operation process is shown in the figure 4 below:

![Convolution operation](image)

Yellow areas represent convolution kernels sliding in the input matrix, each slide to a location, the corresponding digital multiplication and summation to get a element of matrix. Note that dynamic convolution kernels in figure sliding a unit at a time, actually slip amplitude can be adjusted according to need. If the sliding step of the convolution core is greater than 1, it is possible that the convolution core cannot slide the edge of the picture exactly. To solve this problem, we can zero the matrix’s outermost layer and fill the matrix with zero as shown in Fig.5.

![Convolution complement zero](image)

The number of zeros can be set as needed. Zero Padding is a super factor which can be set, but according to the size of the convolution kernel, it needs to be adjusted, the step length and the size of the input matrix, so that the convolution kernel can slide to the edge.

Under normal circumstances, the input image matrix, and the back of the convolution kernel, feature matrix is square, set size of the input matrix for $w$ here, convolution kernels for $k$ size, stride for $s$, layer number of zero padding for $p$, characteristics of figure size after the convolution calculation formula is:

$$w' = \frac{w + 2p - k}{s} + 1$$  \hspace{1cm} (1)

(2) Pooling
Pooling is also called Dwon sampling, as opposed to Up sampling. A pooling layer is generally required to reduce the amount of data for the feature graph obtained by convolution. The pooling operation is shown below:

\[
\begin{array}{cccc}
1 & 1 & 2 & 4 \\
5 & 6 & 7 & 8 \\
3 & 2 & 1 & 0 \\
1 & 2 & 3 & 4 \\
\end{array}
\]

**Fig. 6 Pooling operation**

Like convolution, Pooling also has a sliding kern, which can be called a sliding window. In figure 6, the sliding size window is 2x2 and the step length is 2. For each sliding area, the maximum value is taken as the output. Such operation is called Max Pooling. Mean Pooling can also be used to output the Mean value.

(3) The Whole Connection

The neurons of output layer and each neuron in the input layer are connected by each other. What is the purpose of a full connection? Because the output of traditional network is classification, that is, the probability of several categories or even a number -- category number, the full connection layer is a feature of high purification, which is convenient for the final classifier or regression.

However, there are too many parameters for full connection. The current trend is to avoid full connection as much as possible. One of the mainly methods is global average.

(4) Key Formulas

Input:

\[
V = \text{conv2}(W, X, "valid") + b
\]  

(2)

Output:

\[
Y = \varphi(V)
\]  

(3)

The above formula of input and output is for each of the convolution layer, each layer convolution \( W \) has a different weight matrix, and \( W X, Y \) is matrix form. For the last full connection layer, set as the first \( L \) layer, the output is a vectorial \( y^L \), in the form of expected output is \( d \), has the total error formula.

Total error:

\[
E = \frac{1}{2}||d - y^L||^2_2
\]  

(4)

\( \text{Conv2}() \) is the function of convolution operation in Matlab. The third parameter valid indicates the type of convolution operation. The convolution method introduced above is the valid type. \( W \) convolution kernel matrix, \( X \) as input matrix, \( b \) is biased, \( \varphi(x) \) is the activation function. Total error of \( d, y \) were expected output and network output vector.\( ||x||_2 \) represents the 2-norm of the vector \( x \), the calculation expression is:

\[
||x||_2 = \left( \sum x_i^2 \right)^{\frac{1}{2}}
\]  

(5)

The input and output formulae of the all-connected layer neurons are exactly the same as the BP network. \( \varphi \) is a kind of activation function[5].

### 3.3 Local Connection Properties of CNN

The convolutional neural network belongs to the local connection network, which is based on the deep study of natural images. Natural images have the property of local region stability, and the statistical characteristics of a local region are similar to those of other adjacent local regions. Therefore,
characteristics of a local region learned by neural network from natural images are also suitable for other adjacent local regions of images.

Compared with local connected network and full connected network, weight sharing network has great advantages. The weight sharing network can reduce the number of training parameters, so as to simplify the network structure and have wider application space. If the input image of 1000×1000 is taken as an example, and the number of nodes in the convolutional layer is 106, 1012 weights are required for the full connection. Assuming that the fully connected network adopts a local receptive field of 10×10, the weights need to be reduced to 108. If the system has 100 filters, the weight sharing network can reduce the number of weights to 104 based on the local receptive field. Thus, the weight sharing network and the local connection network can greatly reduce the network parameters, simplify the network structure, and improve the efficiency of image recognition.

4 Application Scenarios

CNN are widely used in the field of images, are currently less interpretable, just like a "black box". Although they are constantly evolving, it is hard to say why the model performs so well. Convolutional neural network has got big achievements in the field of image processing and recognition. On the international standard ImageNet dataset, lots of successful models are based on convolutional neural network. One of the advantages of the CNN technology algorithm is that it can directly analyze the original image and avoids the original image’s complicated pre-processing[6]. Every major breakthrough in image recognition has involved the use of CNN as well as many derived with this network model, can be directly to the image data as input, not only without manual for image preprocessor and extra complex operations, such as feature extraction, and with its unique way of fine-grained feature extraction, makes the image processing to reach the level of human.

With the further development of random technology, CNN has been a large number of applications applied in face recognition, education image processing, intelligent driving, intelligent security, text recognition, human-computer interaction, image search, and intelligent home[7][8].

The strength of CNN is that it can map low-dimensional low-level features to high-dimensional high-level features. Therefore, all data that satisfy local correlation can be theoretically processed by CNN, such as voice and text. In terms of natural language processing, CNN can be used to do some basic tasks, such as part-of-speech tagging, entity recognition, text classification, etc., as well as some cutting-edge tasks, such as machine translation, chatbot, etc.

5 Summary and Outlook

In recent years, the image recognition technology is increasingly hot, every year at the rocket speed to update the new technology and achievements. CNN has turned into the first solution for image classification. Its image recognition accuracy is so high that it can be used in a wide range of applications across different platforms, such as smartphones, security systems and driver assistance systems.

With the improvement of CNN depth and network structure, the recognition accuracy and speed of CNN in image recognition have been improved, and the field of image recognition has been gradually expanded with increasingly powerful functions. It is very difficult to determine which network structure to use, how many layers to use, how many neurons to use is suitable. Detailed knowledge is still needed to select reasonable values, such as learning rate, regularization intensity, etc., and the application cost is high. Moreover, due to the lack of universality of network structure, there are great limitations in solving problems. It is found that the CNN has a very broad application prospect in the area of image recognition. Therefore, it is necessary to research all the image recognition system, to optimize the network structure and the depth of CNN.

1) In the process of image recognition, whether the filter size is appropriate or not has a direct impact on the training process and recognition accuracy. Therefore, in order to make the image recognition result better, it is necessary to select the filter with the most appropriate size.
(2) When using CNN for image recognition, the network depth of different problem selections is often different, and the depth value needs to be determined by manual pre-selection and experimentation, which limits the universality of the network structure. Therefore, in practical applications, we need to select an approximate general network structure depth for specific image recognition problems.

(3) The application of CNN in image recognition mainly achieves the application effect through data set training, which has great limitations for different data sets. Therefore, the CNN needs to train the existing data set according to different problem sets in order to get the same result. If the distribution of the training data set is different from that of the test data set, it is difficult for the convolutional neural network to obtain a good recognition result.

In the image processing based on CNN, a complete normal theory has not been formed. At present, many identification systems design the depth and level of the network based on a specific database, and find the best parameters and optimization algorithm through continuous exploration. Human reasons are relatively prominent, and there is no systematic theoretical explanation of factors affecting the recognition effect of the CNN. In particular, when classifying and recognizing natural images, the selection of the initial state parameters of the CNN and the optimization algorithm will have a big effect on the network training [9]. If your validation set is more accurate than your training set, the model is under-fitted. Overfitting occurs when your model overfits the training set. Selection will result in the network not working, or it may over-fitting, under-fitting and any other problems.

We need to understand deeply CNN meanings and roles of each part, adjust various parameters to optimize and deepen the network, make the network to obtain more information, or even to add our innovation in the structure to solve various problem.

Acknowledgments
This research was supported by National Key R&D Program of China (Grant No.2017YFB1400400).

References
[1] Meng-xue Xu: An overview of image recognition technology based on deep learning. Computer Products and Circulation 1, 213-213(2019).
[2] JING L, CHENG J H, SHI J Y, et al: Improvement. 2012.
[3] Liang Chang et al.: Convolutional neural networks in image comprehension. Automation 9, 1300-1312(2016).
[4] HINTON G E et al.: A fast learning algorithm for deep belief nets. Neural Computation 7, 1527-1554(2014).
[5] Xian-chang Chen: Deep learning algorithm and application research based on convolutional neural network. Zhejiang Gongshang University,2014.
[6] Lin Zhang et al.: A review of research on convolutional neural networks. Chinese Journal of Computers 40, 1229-1251(2017).
[7] Quan-Sen Sun et al.: A new method of feature fusion and its application in image recognition. Pattern Recognition 12, 2437-2448(2005).
[8] Misgana Negassi et al.: Application of artificial neural networks for automated analysis of cystoscopic images. World Journal of Urology 1, 2020.
[9] XU K, FENG D, MI H, et al.: Mixup-Based Acoustic Scene Classification Using Multi-channel. Convolutional Neural Network, 2018.