Research on Neural Network in Deep Learning

Chenggang Zhen [a], Jingxue Li [b]

School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China.

[a]zhengcg@163.com, [b]ljx119512@qq.com

Abstract

In recent years, with the rapid development of artificial intelligence, the deep learning algorithm used at the bottom of artificial intelligence has gradually attracted people's attention. Its ultimate goal is to enable machines to have human analysis and learning ability. Deep learning has made remarkable achievements in a large number of machine learning problems, forming a new branch of machine learning that is the most brilliant in the field, which has set off a new climax in the research on theory, method and application of deep learning. Deep learning algorithm is to point to by computer to simulate human brain neurons on the complex data processing, its essence is a kind of hierarchical feature extraction learning process. it implied by constructing multi-layer neural network model, input a large number of training data, such as text, images and sounds to train the model of feature extraction of the optimal parameters, the characteristics of the simple combination of abstract into high-level characteristics, in order to realize the abstract expression of the data or real object. The model ability of deep learning increases exponentially with increasing depth. Now more and more artificial intelligence products based on neural network are widely used in various fields, and people's work and life efficiency has been greatly improved. This paper mainly introduces several common Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) in deep learning, as well as the gradient descent method which is used in Neural Network training.

Keywords

Deep learning, Neural network, Optimization algorithm.

1. Introduction

Deep learning constructs a multi-layer neural network to extract more representative feature information from the original input information, and then maps these feature information to the output to classify and predict the sample data. Compared with traditional machine learning models, deep learning requires more data. The differences between deep learning models and traditional machine learning models are as follows :(1) Deep learning model structure contains deeper neural network, for example, the number of convolution kernel in convolution layer in convolutional neural network is usually multiplied.(2) the deep learning model pays more attention to feature learning. Through feature extraction layer by layer, the features of data samples in the original space are transformed into a new feature space to represent the initial data. In the new feature space, the data will become more abstract, which makes classification or prediction easier. Compared with the manually designed feature extraction method, the data features obtained by deep model learning are more advantageous to data classification and regression.

2. Neural Networks

2.1 Convolutional Neural Networks

Convolution neural networks are usually used in image information processing.The Convolutional neural networks maintain the invariance of image position information by convolving and translating the convolution kernel on the image,the spatial location information in the image can be extracted.At the same time, the convolution kernel has a parameter sharing mechanism that reduces the number of parameters in the neural networks and reduces the computational complexity, which accelerates the training process of the neural networks [1]. Convolutional neural networks use gradient back-transfer to update network model parameters. The optimization method can choose to use Adam, NAG and other update algorithms according to different data requirements.

Convolutional neural networks have five hierarchical structures, namely, input layer, convolution layer, activation function layer, pooling layer and full connection layer. The input layer is the same as the traditional neural network, and is used to send the pre-processed picture data to the neural networks. The pre-processing steps usually have de-mean and decorrelation. The de-mean usually refers to the average brightness value of the removed image. The brightness value information is less important for the classification of the image than the position information in the image. Decorrelation is usually used to remove the correlation between each column of the picture pixel matrix, which facilitates information extraction.

The convolution layer highlights two advantages of convolutional neural networks, namely neuronal local connections and parameter sharing mechanisms. Traditional neural networks typically connect each neuron with pixel information in a picture to process image information, which presents two problems. Firstly, too many parameters in the neural network lead to an increase in the difficulty of model training. Secondly, all pixel information in the picture is input to the network with equal importance, resulting in no use of the location information in the picture. The convolutional neural network extracts the information in the image through the convolution kernel, and each convolution kernel constitutes a local receptive field, as shown by C1 in Fig.1. The size of the convolution kernel is much smaller than the overall size of the image. In the process of extracting the image information from the convolution kernel, each neuron only perceives the local image information, and the partial image information perceived by each neuron can be combined to obtain the overall image information. Compared with the traditional neural network, each neuron is connected with the image information. This local connection can extract the position information in the image, and each neuron shares the parameters, which reduces the number of parameters in the model and accelerates the model training. . In order to extract different kinds of information in the image, multiple convolution kernels may be set, and different convolution kernels extract different feature information of the image. As shown in Fig.1, there are three convolution kernels in the convolution layer C1. The larger the number of convolution kernels, the more sufficient the extracted image information is, and the number of convolution kernels between the convolutional layers is usually a multiple.

The activation function layer limits the output of the neural networks by setting thresholds. In the image processing process, insufficient feature intensity of image information of a certain part will cause the activation function output to be zero, and the picture information of this part will not be collected by the networks. Ordinary binary classification neural networks usually use the Sigmoid activation function. Enter a real number and the Sigmoid function outputs a number in the range 0 to 1. The closer the output is to 0, the greater the probability that a neuron is not activated, and the closer the output is to 1, the greater the probability that a neuron in this part will be activated. The Sigmoid function has two shortcomings. One is that the maximum gradient value of the Sigmoid function is 0.25. After the chain derivation, the gradient disappears easily; the second is that the output of the Sigmoid is not 0 mean. In the parameter update process, it is easy to cause the parameters to be updated only in one direction, and it is not easy to converge. The activation function commonly used in convolutional neural networks is Relu, which is characterized by an output value of 0 when the input value is less than 0, and is output as it is when the input value is greater than 0, and the gradient

value is always 1. It avoids the problem of gradient disappearance and explosion, and the convergence speed is faster, and the calculation is more convenient. Since the input is less than 0, the output is 0, causing some neurons to be inactive, so the Leakrelu and elu activation functions have been proposed. These two activation functions solve the problems caused by the Relu zero interval and also include the advantages of Relu. Elu combines two functions, Sigmoid and Relu. The left part has convergence, which suppresses noise and abnormal data. The right part does not have convergence, and the output of the elu function is close to zero mean, so that the neural networks can converge quickly. The activation function used in the LSTM network is the tanh hyperbolic tangent function, which compresses the input real number between -1 and 1. Therefore, the function outputs a zero mean and the network can converge quickly. The disadvantage is that the gradient of the function is less than 1, which causes the gradient to disappear easily during the reverse transfer and the parameters are not updated.

Convolutional neural networks usually perform batch normalization operations on data before the activation function. The purpose is to limit the data range after convolution to the unsaturated region of the activation function, so that the information after convolution is completely transmitted. After the activation function is usually the pooling layer, the pooling layer is usually used to extract the local mean or maximum value of the image, and is divided into the mean pooling layer and the maximum pooling layer according to the calculated values. Most commonly used are the maximum pooling layer.

After the pooling layer is generally a fully connected layer, which is used to synthesize the deep information extracted by the convolutional neural network, and finally maps the original picture information to the output of the neural network for image classification or regression.

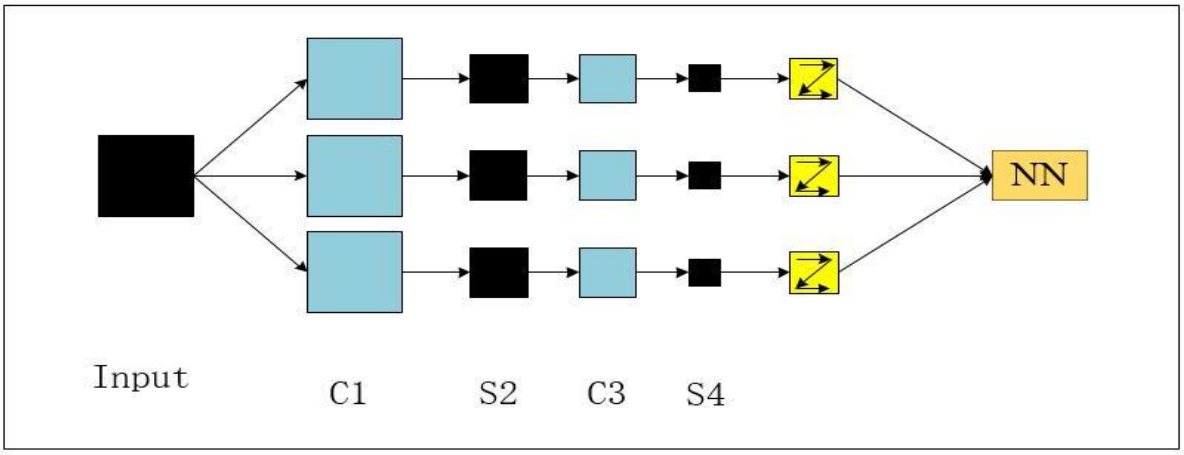

Fig.1 Concept demonstration of CNN

2.2 Recurrent Neural Network

The premise of convolutional neural networks and artificial neural networks is that the elements are independent of each other, and the input and output are independent. There is no necessary connection between before and after, such as cats and dogs. But in the real world, there are many elements that are connected to each other. For example, changes in stocks over time, changes in stock prices are closely related to changes in data before the current time. The current price is inferred from the previous content. The difference between a recurrent neural network and a convolutional neural network and an artificial neural network is that the essence of a recurrent neural network is the ability to have memory like a human being. Its output depends on the current input and the memory before the input. RNN can use its internal memory to process input sequences of any timing, which makes it easier to process some data.

The network structure of RNN is shown in Fig.2 below:

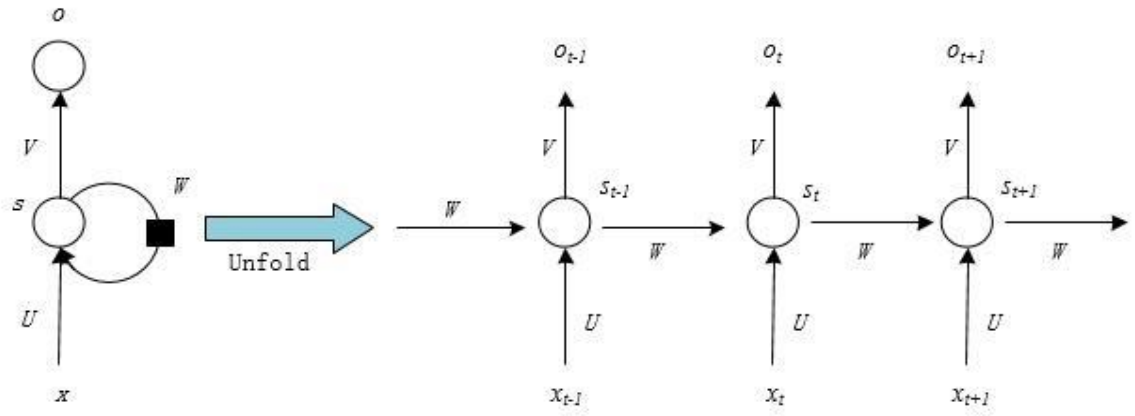

Fig.2 RNN network structure

Each of the circles can be regarded as a unit, and each unit does the same thing, and the left half is folded and integrated by the right half. RNN is essentially a reuse of a unit structure.

RNN is a sequence-to-sequence model [2], X_t: the input at time t, O_t: the output at time t, S_t: the memory at time t. Because the output of the present moment is determined by memory and the input of the present moment. For example,the knowledge of a second-year graduate student is a combination of what the graduate student learned in the second year (current input) and what the graduate student learned in the first year and before (memory). RNN excels at processing predicted sequence data, which is a sequence, a specific order in which one thing follows another, and its internal state can show dynamic temporal behavior.So define the basis of RNN, as shown in Equation 1:

$$S_t = f(U * X_t + W * S_{t-1}) \tag{1}$$

The LSTM network is a special type of Recursive Neural Network (RNN), which is proposed to solve the problem of gradient dispersion of the RNN model [3]. As shown in Fig.3, the LSTM network takes continuous data X_t as an input, and the state vector h_t of each step depends on the state h_{t-1} of the previous time and the input X_t of the current time. The data is input into the LSTM network,the LSTM memory cells first discard a part of the input data through the activation function, that is, there is a 'forget gate' mechanism in the LSTM memory cells. Then, when updating the cell state, LSTM loads a part of the information in the input data into its own memory cells through the Sigmoid function, and outputs a part of the data in the memory cells, and processes the input data through continuous updating of the cell state.And the LSTM memory cell update formula is as shown in Equation 2.Among them, C_{t-1} is the old memory, C_t is the new knowledge accumulation, and C_t is the final new memory.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2}$$

It can be concluded from Equation 4 that when training is performed on the LSTM network, the reverse transmission of the error is reversed through two lines. Even if one of the gradients is close to zero, the presence of the other term does not cause the overall gradient value to become zero. This mechanism can avoid the situation where the gagradient disappears.

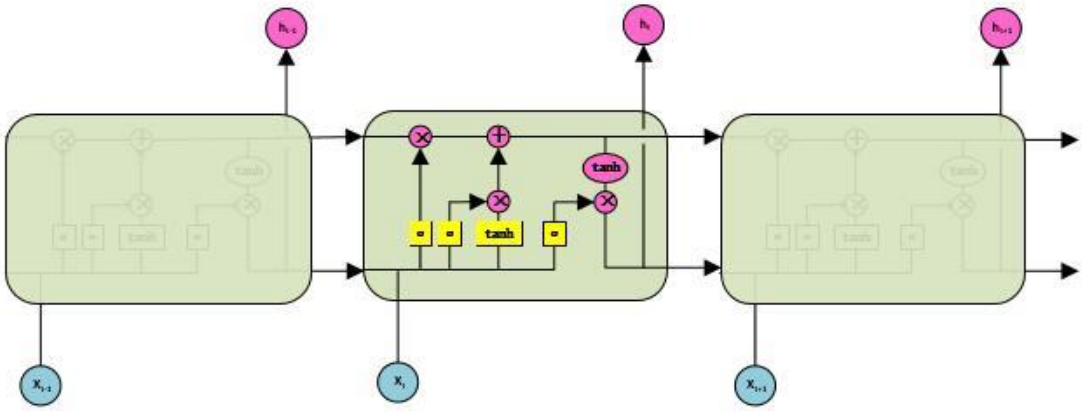

Fig.3 LSTM network structure

3. Gradient Descent

In the neural network, we measure the difference between the network output and the actual real value through the loss function. The process of model training is the process of decreasing the value of the loss function. In order to enable the loss function to converge to the global minimum point, the usual iterative methods are the Newton method and the gradient descent method. In deep learning, the gradient descent method is used more. The principle is that the direction of the loss function is the gradient direction of the function parameter calculated by the loss function value. According to the different ways of calculating the loss function, it can be divided into Stochastic Gradient Descent (SGD) and Batch Gradient Descent (BGD).

3.1 Batch Gradient Descent （BGD）

Random gradient descent Each time a sample data is input, the gradient value of the model parameters is calculated once [4]. Its update formula is shown in Equation 3. Since the SGD update method is based on each input sample data, and its update frequency is fast, only a small amount of training data is needed to converge the parameters and the convergence speed is fast. The disadvantage is that the SGD does not always go in the right direction every time it is updated. Although the overall direction is the direction in which the loss function drops, there may be oscillations at some sample points. Therefore, the accuracy of the training process will decrease, and sometimes it will converge to the local best, but this local best advantage may be better.

$$\theta = \theta - \eta \bullet \nabla_{\theta} J(\theta) \tag{3}$$

3.2 Stochastic Gradient Descent (SGD)

Batch gradient descent uses the full data set to calculate the gradient of the loss function for the parameter when calculating the gradient, reducing the variance of the gradient update [5]. This allows the loss function to always converge in the right direction and easily converge to the full minimum. The disadvantage is that the computational complexity is large, and it is very difficult to encounter a large number of data sets, and new data cannot be added in real time. The update method is shown in Equation 4.

$$\theta = \theta - \eta \bullet \nabla_{\theta} J(\theta; x^{(i)}; y^{(i)}) \tag{4}$$

4. Conclusion

The deep learning model maps initial information to high levels for analysis through complex linear and nonlinear operations. Convolutional Neural Networks (CNN) and Recurrent Neural Network

(RNN) are typical architectures. They are usually composed of a series of neural layers, and the output of each layer is the input of the next layer. This paper introduces the workflow of the convolutional neural network by expounding the local connection and weight sharing in the convolutional neural network, that is, the convolution cts the features in the image from the local information of the original image. The recurrent neural network can find out the connections existing between the data by learning the context of the input data, and finally map to the output data. Both networks can be iterated by gradient descent. Different activation functions in the actual training process will result in different convergence speeds of the network.

References

[1] Lecun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition [J] . Neural Computation, 2014, 1 (4) :541-551.

[2] BENGIO Y, SIMARD P, FRASCONI P. Learning long-term dependencies with gradient descent is difficult［J］. IEEE Transactions on Neural Networks, 1994, 5(2) : 157 －166.

[3] HOCHREITER S, SCHMIDHUBER [J].Long short-term memory［J］.Neural Computation,1997, 9(8) : 1735 －1780.

[4] S. Ruder, An overview of gradient descent optimization algorithms, Jun 2017.

[5] Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J] . Proceedings of the IEEE, 1998, 86 (11) :2278-2324.