ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering (WCSE 2019)
Hong Kong, 15-17 June, 2019, pp. 176-180
doi: 10.18178/wcse.2019.06.026

Research on Image Feature Recognition Based on Convolution-Long Short Term Memory Network

Chao Yu, Jing Zhou [+], Liang Gong , Lei Sun , Pengfei Shi and Xinxin Ou

School of Mathematics and Computer Science, Jianghan University, Wuhan Hubei 430056, China .

Abstract. In order to improve the image recognition rate of Long-Short-Memory-Network, using convolution calculation to reduce dimension and extract feature, which can remove large amounts of redundant information from image samples and accurately extract image features. The recognition rate of image can significantly improve by using classification of serialized image features obtained by compression and dimensionality as the input of LSTM. Firstly, CNN is used to extract the image features accurately. Secondly, the features are serialized into continuous picture bars with strong correlation. LSTM is used to classify the images, thus obtaining a better recognition effect. The experimental results show that the recognition rate of the CNN-LSTM is 30% higher than the basic LSTM, and the recognition rate of the CNN is 5% higher than the basic CNN.

Keywords: Deep Learning, Long Short Term Memory Network,CNN,cifar-10, Image Recognition.

1. Introduction

Convolutional neural network (CNN) is a kind of feedforward neural network with deep structure, which includes convolution or correlation computation. In the 1990s, CNN is proposed by Lecun[1] in his relevant papers, and then it is perfected to propose the LeNet-5[2], which has a good effect in handwriting recognition. Subsequently, Convolutional neural network has been gradually valued, and CNN ushered in a period of rapid development. In 2012, Hinton and his student Alex Krizhevsky propose the AlexNet[3] based on the network structure of LeNet , which achieved great success in image recognition and won the championship of ImageNet competition, then more and more excellent neural networks were put forward. CNN plays an important role in the field of image recognition. Recent years, more and more attention is paid on RNN. RNN is mainly applied in the field of natural language processing, such as machine translation, emotional analysis, intelligent dialogue and so on. LSTM is a kind of recurrent neural network, which can process and recognize serialized data. The features of multigroup images can be divided into multi-sequence data according to row or column, so LSTM can also be applied in image recognition in theory. However, the inaccurate input eigenvalues of LSTM will result in inaccurate serialized classification and low recognition rate. Therefore, this paper provides the method of fusion for CNN and LSTM, that is, CNN is used to extract serialize features of multigroup image accurately, and LSTM is adopted to recognize the serialize features, and then the recognition rate is improved significantly compared with the CNN[4] and LSTM.

2. Convolution and Long Short Term Memory Network

2.1. Convolutional Neural Network Model

Convolutional neural network can optimize the objective function only by training a few parameters, which has shorter training time and higher efficiency than fully connected network. Meanwhile, convolution neural network can avoid the complex pre-processing of the image, and the original image can be directly

[+] Corresponding author. Tel.: +18040535659.
E-mail address: zhou_8132@163.com

input to the Network. The main structure of convolution neural network is the convolution layer, pooling layer and full connection layer. Through the interconnection of convolution layer and pooling layer, the convolution group is formed to extract features layer by layer, and then the classification is completed by the full connection layer.

Convolution layer is the core part of convolutional neural network. Convolution operation is an operation mode in analytical mathematics, which can extract features and recognize specific features of images by using different convolution kernels. Features extracted from different positions of the input image constitute feature maps, and different feature maps can be extracted by different convolution kernels.

The process of extracting features by convolution kernels is shown in the following . The input is a 5×5 matrix, and the corresponding convolution kernel is a 3×3 matrix. Start at the top left corner (0, 0) of the input image, a grid slides once, each position of the convolution core multiplies the data of the corresponding position of the image, and then a 3×3 feature matrix is obtained finally.

Pooling is the operation of down sampling, which compresses images and reduces parameters without affecting image quality. Generally, there are two methods of pooling. One is to select the maximal value of the feature points in the neighborhood and the other is to select the average value of the feature points in the neighborhood when pooling the image. The role of the pooling layer is to reduce the data dimension. For a 64×64 pixel image, a 5×5 convolution core will get a feature map of (64-5+1)×(64-5+1), and 200 samples will get a dimension of 3600×200=720000 for convolution feature map. Excessive data will lead to over-fitting, so it is necessary to down sample the image.

The error of feature extraction in convolution layer mainly comes from two aspects: (1) The variance of estimation value increases due to the limitation of neighborhood size. (2) The error of convolution parameters results in the deviation of the estimated mean value. Generally, the average pooling operation of feature points in neighborhood can reduce the first kind of error, and retain more background information of image, and the maximum pooling operation of feature points in neighborhood can reduce the second kind of error and retain more texture information.

2.2. Manuscript requirements

Recurrent neural network[5] contains feedback connections, so it has the strong memory function, which can deal with the sequence problems very well. Long short-term memory is one kind of the recurrent neural network, which can deal with sequence problems well. There is the shortcoming of fast weakening for node memory in the common recurrent neural network, which can be overcome by the model of Long short-term memory network [6].

LSTM can be applied in many fields, such as the speech recognition, language translation, image recognition, handwriting recognition and other tasks. LSTM adds the input gate, the output gate and the forget gate based on the RNN, which can realize the selective memory of input, forgetting some unimportant parts and the selective output. The forgetting mechanism of LSTM can avoid the occurrence of gradient explosion[7], the structure is shown in Figure 1.

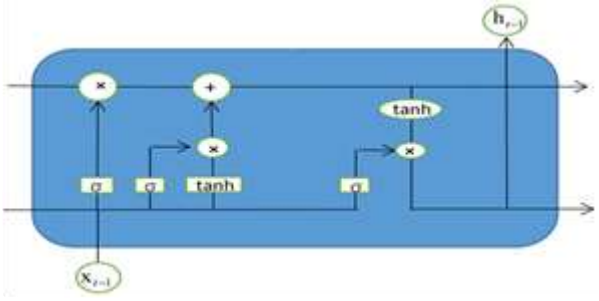

Fig. 1: LSTM schematic diagram

The three control gates-forgetting gate, input gate and output gate are adopted in LSTM[8], which effectively solves the problem of long-term dependence. Forgetting gate determines the information to be

discarded for the model. The gate takes the output h_{t-1} of the previous moment and the input x_t of the current moment as the inputs and outputs a value between 0 and 1. Then the state of completely reserved is expressed by 1 value, and the state of completely discarded is expressed by 0 value. The formula is as follows (1).

$$f_t = (W_f \cdot [h_{t-1}, x_t] + b_f \tag{1}$$

The information will be added to the network is determined by the input gate. And the candidate vectors $\tilde{C}$ is generated, which is updated by the hyperbolic tangent layer through formula (4). The formula is as follows:

$$i_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_t \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

Finally, the output information is selected by the output gate through sigmoid layer, and then it is processed through the hyperbolic tangent layer as shown in formula (5) and formula (6).

$$O_t = \sigma(W_O \cdot [h_{t-1}, x_t] + b_c) \tag{5}$$

$$t_t = O_t * \tanh(C_t) \tag{6}$$

Where the data input at t time is represented by x_t and the output at the previous time is represented by h_{t-1}.The weight to be trained is denoted by W, and the bias value is denoted by b. σ is the sigmoid function.

3. The framework of Convolution-Long Short Term Memory Network

3.1 The Structure of Convolution-Long Short Term Memory Network

Convolutional and long-short term memory network is composed of two convolution layers, two pooling layers and long-short term memory module, as shown in Figure 2. The network input is a three-channel RGB image with the size of M×M. The first layer is the convolution layer with zero padding. The number of convolution cores is N, and the output is N feature maps with the size of M×M. The second pooling layer is used for dimension reduction. The number of convolution cores is still N, and the output is N feature maps of M/2×M/2 size. The third layer is the same convolution layer as the first layer, and the output of N feature maps is invariant to the size of M/2×M/2. The fourth pooling layer outputs the N feature maps of M/4×M/4 size. Then, N feature maps are connected into one feature map and trained in a long-short term memory network.

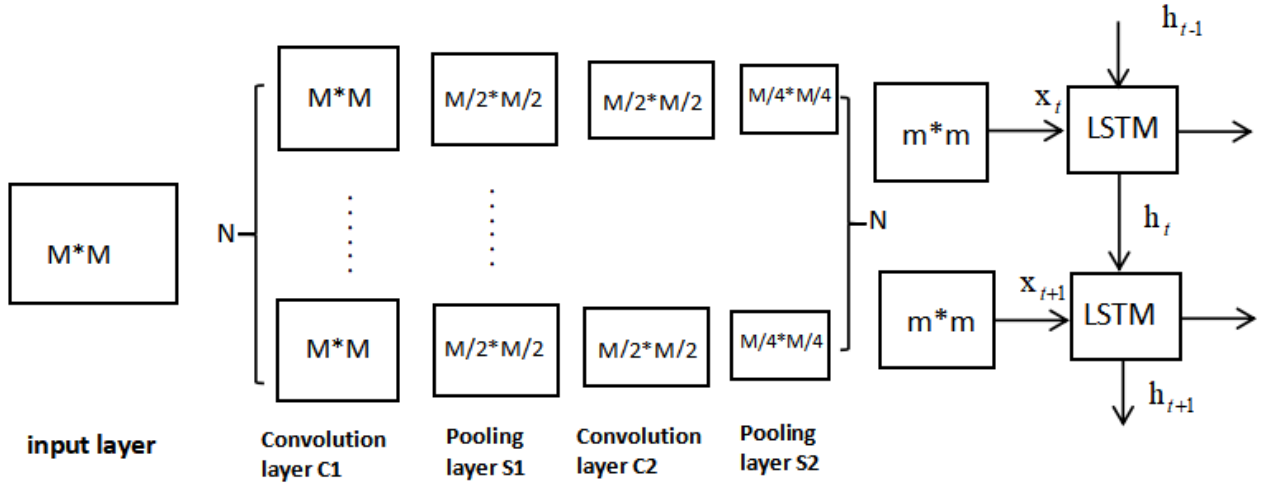

Fig. 2: Structural chart of convolutional-long short term memory network

Before the feature map extracted from convolutional neural network is put into the long-short term memory network, the feature map should be sequential processed[9], that is, each batch of 48×48 feature

map is divided into 48 elements, and each element is an 1×48 array. Each time one row of images in the batch is input as a time series. Finally, the output of the long-short term memory network is divided into 10 categories by the SoftMax function.

3.2 The Structure of Convolution-Long Short Term Memory Network

The competition mechanism is created for the activity of local neurons, so that the value with a larger response becomes larger and other neurons with a smaller feedback are suppressed。At the same time, it enhances the generalization ability of the model, to prevent the over-fitting of the model .The calculation formula is shown in formula (7).

$$b_{x,y}^{i} = a_{x,y}^{i} /(k + \alpha \sum_{j=\max(0,i-N/2)}^{\min(N-1,i+N/2)} (a_{x,y}^{j})^{2})^{\beta} \tag{7}$$

In formula (7), $a_{x,y}^{i}$ is the output for relu function after the ith convolution filter, and it is entered as the input of LRN. $b_{x,y}^{i}$ is the output of LRN. N is the number of convolution kernels in current convolution layer k and n is the local region size of the convolution output. k, α, β is constant[10]。

Regularization is a rule that prevents overfitting by reducing the complexity of the model. A regularization term is added after the loss function, so that the error between the output and the standard result is interfered, and then the trained model cannot be fully fitted with the sample, thus preventing the overfitting. There are mainly two overfitting norms of L1 and L2 norm for regularization. L1 norm is the sum of the absolute values of all weights w, and L2 is the square root of the sum of the squares of all parameters. The mathematical forms of the two forms are shown in equation (8) and equation (9).

$$\| x \|_1 = \sum_i^n | x_i | \tag{8}$$

$$\| x \|_1 = \sqrt{\sum_i^n (x_i)^2} \tag{9}$$

Usually, the L1 norm will lead to sparse solution, and L2 norm leads to the dense solution. L2 norm is usually used to prevent model over-fitting in practice.

4. Experimental Results and Analysis

Cifar-10 data set is used in this paper to train and test the CNN-LSTM network. The cifar-10 data set is collected by Hinton and Alex Krizhevsky for object recognition, which is composed of 60000 color images of 10 classes with size of 32×32 . There are 60000 images for each class, including 50000 training images and 10000 test images.

In this paper, 64 convolution kernels with size of 5×5 are adopted in CNN-LSTM network model. The data set of Cifar10 is clipped to 24×24 and used as the input of the network model. Due to the padding, the output of the first convolution is still 64 feature maps with size of 24×24. After pooling, 64 feature maps with size of 24×24 were reduced to 64 feature maps with size of 12×12, and then local normalization was carried out. The operation of the second convolution layer is similar to the first layer, applying the 64 convolution filters, the 64 feature maps with size of 6×6 were obtained. And then reshape the 64 feature maps with size of 6×6 to one 48×48 feature map. Then 128 feature maps of a batch were divided into 48 time-sequences by row to input, and the size of each sequence is 128×48. Finally, the SoftMax function are adopted to classify the features and output the corresponding categories.

In this paper, the cifar-10 data set was classified by the network models of regularized CNN, normalized CNN, GRU, LSTM and the CNN-LSTM network. The experimental results are shown in table 1.

As can be seen from table 1, the highest accuracy of recurrent neural network is only 47%, and that of convolutional neural network is only 72%. The accuracy can reach 78%, when the CNN-LSTM network is adopted. Therefore, the recognition rate of the proposed network structure is higher than that of the traditional single network structure (CNN, LSTM). Compared with the traditional CNN and RNN neural network, the CNN-LSTM model extracts the features of images through the convolutional module, and the noise is eliminated. At the same time, it also relies on the unique sequence processing ability of the cyclic neural network. Therefore, the accuracy of image recognition is higher.

Table 1: Comparison of recognition rates of various methods on cifar10 dataset

convolutional neural network	local normalization	72%
	without local normalization	69%
recurrent neural network	GRU	46%
	LSTM	47%
convolution-long short term memory network	two layer convolution-LSTM	78%

5. Conclusion

The accuracy of convolution neural network and recurrent neural network can reach more than 98% when they are used to recognize the Mnist handwritten digital dataset. But when they are applied to the cifar-10 dataset with higher dimension and more samples, the accuracy of single convolution neural network and single recurrent neural network are difficult to improve. The accuracy of LSTM and GRU are also less than 50%, but the recognition rate of convolution-long short term memory network proposed in this paper is significantly higher than those of these single network structures due to the combination of the feature extraction ability of convolutional neural network and the ability of recurrent neural network to process sequential tasks, and then the high recognition rate can be achieved.

6. Acknowledgments

This paper is sponsored by Provincial Teaching Reform Project of Hubei province in 2017 (2017301).

7. References

[1] Y. LeCun, B. Boser, J. S. Denker, et al. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989, 1(4): 541-551.

[2] Y. L. Lecun, L. Bottou , Y. Bengio , et al. Gradient-Based Learning Applied to Document Recognition. *Proc. of the IEEE*, 1998, 86(11):2278-2324.

[3] Xiang Chang, Ming Yang. Graph classification performance based on improved convolutional neural network . *Journal of Chongqing University of Technology*, 2017, 31 (3): 110-115.

[4] A. Krizhevsky, I.Sutskever, GE. Hinton.Imagenet classification with deep convolutional neural networks,2012, 25(2)1097-1105.

[5] W. Yao, Z. Zeng, C. Lian, et al. Training enhanced reservoir computing predictor for landslide displacement. *Engineering Geology*, 2015, 188：101-109.

[6] S. Hochreiter, J. Schmidhuber. Long Short-Term Memory.*Neural Computation*, 1997,9(8):1735-1780.

[7] R. Pascanu, T. Mikolov, Y. Bengio. On the Difficulty of Training Recurrent Neural Networks.*Proc of the 30th International Conference on Machine Learning*.Atlanta,USA:ICML,2013,52(3):1310-1318.

[8] MF Perelló, J. L. Coloma ,N. Masoller , et al. Intravenous ferrous sucrose versus placebo in addition to oral iron therapy for the treatment of severe postpartum anaemia: a randomised controlled trial. *Bjog An International Journal of Obstetrics & Gynaecology*, 2014, 121(6):706-713.

[9] Jinhong Li, Introduction, Principle and Advanced Practice of TensorFlow for Deep Learning . *Machinery Industry Press* 2018.1:225-239.

[10] Ran Peng, Deep Convolution Neural Network of Softmax Classifier and Its Application in Face Recognition *Journal of Shanghai University(Natural Science Edition)*, 2018, 03 (3): 353-366.