International Core Journal of Engineering Volume 7 Issue 4, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0017

Improved LeNet-5 Convolutional Neural Network Traffic Sign Recognition

Lifen Li*, Yangrui Wang

School of Computer, North China Electric Power University, Baoding, Hebei 071000, China.

Abstract

In order to improve the speed and accuracy of road traffic sign detection and recognition, the classical convolutional neural network architecture Lenet-5 is proposed to be improved and optimized. LeNet-5 network structure was improved, reducing unnecessary link and calculation to improve the real-time performance of algorithm, and combined with the feature of multi-layer fusion method, will be in the two layers of convolution layer extracts the features of the network are sent to all the connection layer, strengthen the characteristic power of expression, on the basis of meet the real time improve the identification accuracy of the algorithm and introduce Dropout mechanism at the same time, to avoid over fitting, improve the network generalization ability of the model. Experimental results show that the improved Lenet-5 convolutional neural network proposed in this paper achieves the desired effect in terms of accuracy and real-time performance.

Keywords

LeNet-5; Multilayer Feature Fusion; Convolutional Neural Network; Traffic Sign Recognition.

1. Introduction

With the continuous development of the economy, the living standards of our people have gradually improved.Although transportation has greatly facilitated people's travel, due to my country's huge population base, people's increasing demand for transportation has contradicted my country's weak road infrastructure.The emergence of such traffic problems makes the use of ITS (Intelligent Transportation System) to improve traffic problems have received great attention.In the Intelligent Transportation System (ITS), the traffic sign recognition system is an important part of the Advanced Driver Assistance System (ADAS), which provides auxiliary information such as warnings and instructions to the driver.In the real environment, the traffic signs on the road are often due to the complex outdoor environment. Problems such as excessive or insufficient light, obstruction by obstacles, and unstable shooting have a great influence on the recognition of traffic signs, causing vehicles to drive. The safety and orderliness of traffic cannot be guaranteed.

In order to solve this series of problems, scholars at home and abroad are keen on the research of traffic sign detection and recognition. Following this, a series of complex algorithms have been proposed. By segmenting and extracting and integrating three types of features of traffic sign images, LBP, Gabor and HOG. Using SVM and AdaBoost to complete the recognition of traffic signs, but the recognition situation in the real situation is complicated, and the application of traditional algorithms is difficult. Literature [3] integrates HOG, SIFT, LBP features for identification, However, the recognition of the actual situation is more complicated, and it is difficult to use traditional algorithms. HOG, SIFT and LBP functions reflect the characteristics of the road sign image through linear coding. However, due to the large end feature size and the large amount of calculation, the recognition time is difficult to meet the expected requirements. In order to avoid the complexity and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0017

insecurity of manual description, methods based on deep learning have developed rapidly in recent years. In many image classification competitions, the improved deep learning network based on the Convolutional Neural Network (CNN) has shown good performance in recognizing road signs. LeNet-5 [4] is a convolutional neural network proposed by Professor Yanna-LeCuna in 1998. It is one of the most representative experimental systems of early convulsive neural networks. On this basis, this paper proposes a new and improved CNN network model based on the analysis of classic convolutional neural networks such as LeNet-5 and multi-feature fusion technology to identify and classify road signs. The improved CNN network adapts to the number and size of convolution kernels, increases the batch normalization BN (input normalization layer), changes the activation function to a nonlinear RELU function, and adopts the dropout strategy to improve the robustness of the model, Effectively avoiding over-fitting. The model aims to identify traffic signs in a timely and accurate manner, thereby effectively improving driving safety.

2. Methodology

2.1 Basic structure of convolutional neural network

In recent years, because deep CNN can directly input data into the CNN model without any processing, it can automatically learn image features. Convolutional neural networks have been widely used in computer image classification problems due to their local perception and weight sharing characteristics, Has been widely used in image recognition. Convolutional neural network consists of convolutional layer, pooling layer, fully connected layer and softmax layer. It is responsible for receiving the detected images, passing the training results of each round of training set and verification set back to adjust the network structure parameters. The basic structure is shown in Figure 1.

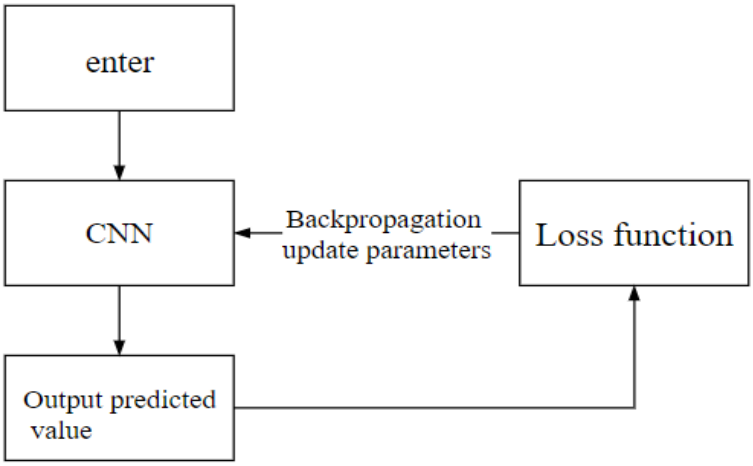

Figure 1. Basic structure of convolutional neural network

The convolutional layer is composed of a series of convolution kernels, which output the input data after the product operation, and the output at this time is the feature map. In convolutional neural networks, convolutional layers are used for feature extraction. After the feature map of the previous layer is input, each convolution kernel is convolved with it. The convolution kernel slides on the feature map with a certain step size. Each slide is performed once for convolution, and finally this layer is obtained. A feature map, so that each feature map establishes a certain relationship with several feature maps in the upper layer. The convolution process is shown in Figure 2. The convolution layer formula is:

$$X_j^l = f(\sum_{i=1} X_i^{l-1} * W_{ij}^l + b_j^l)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0017

The pooling layer, that is, the sampling layer, whose function is to reduce the parameter scale of the neural network, is essentially an image aggregation operation. The data of the feature map after the input image is convolved is often relatively extensive, resulting in excessive calculation. After the features obtained by the convolutional layer are compressed by the pooling layer, the calculation complexity in the subsequent process can be greatly reduced, and redundant features will be removed. The biggest feature of the pooling operation is that it can maintain the translation invariance of the image. Even if the target object in the image has a certain degree of translation or scaling, the pooling operation can still obtain the same pooling characteristics as before the change.

There are two commonly used pooling methods: max-pooling (maximum pooling) and mean-pooling (mean pooling). The main difference between them is that the former uses the maximum value in the selected area as the combined value; the latter uses the average value of the selected area as the pooling value. The two pooling formulas are shown below. Maximum pooling selects the point with the highest element value in the neighborhood, which can store more information about the image texture. The network structure adopted in this paper uses maximum pooling to compress feature.

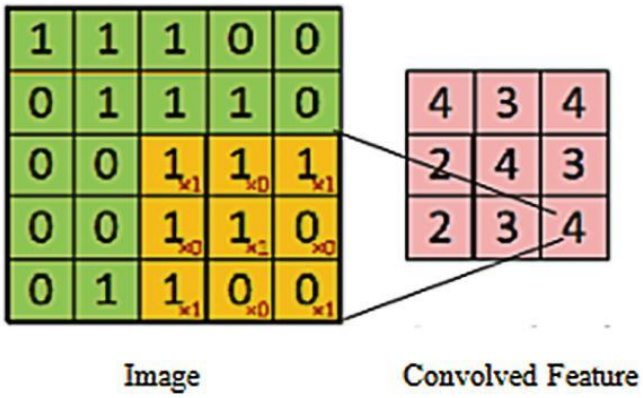

Figure 2. Convolution process

The convolutional neural network enters the connection layer after the convolution and pooling operations are completed. There are two commonly used connection methods: partial connection and full connection. Local connection means that the neurons in the current layer of the network are only connected to some neurons in the upper layer. These neurons are located in the m*m convolution area. Fully connected refers to the neurons in the current layer. All neurons in the previous layer are connected to each other. The parameters of the fully connected method are significantly larger than the parameters of the local combination method. In the convolutional neural network, the fully connected layer usually appears in the last few layers, and the convolutional layer and the pooling layer have class-discriminatory local information and feature weighting.

2.2 LeNet-5

LeNet-5 is a classic convolutional neural network structure, which has greatly promoted the development of CNN. This network model was proposed by Le Cun [5]. The model is aimed at handwriting recognition with extremely high accuracy. The LeNet-5 network structure has a total of 7 layers, including two sets of convolutional pooling layers and 3 fully connected layers.

The first layer of LeNet-5 is the convolutional layer Conv1, the size of the convolution kernel is 5×5, and the number of convolution kernels is 6.The second layer is the pooling layer Pool2, which can output 6 feature maps of 14×14; the third layer is the convolution layer Conv3, the size of the convolution kernel is 5×5, and the number of convolution kernels is 16, and we get 16 10×10 feature maps; the fourth layer is the pooling layer Pool4, and 16 5×5 feature maps are obtained; the 5th, 6th, and 7th layers are all fully connected layers, using Sigmoid activation functions, of which Fc1, Fc2

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0017

The number of nodes in Fc3 and Fc3 are 120, 84, and 10 respectively. The number of nodes in Fc3 is also the number of handwritten digit set minist categories. However, because the LeNet-5 model is aimed at the recognition of handwritten digits, there are fewer recognition categories, and the recognition rate in the recognition of traffic signs is not satisfactory, and the loss value is also high, so it needs to be adjusted and improved [6-7], It is suitable for the recognition of more types of traffic signs.

2.3 Improved Convolutional Neural Network

Traditional convolutional neural networks are generally single-scale features and follow a feedforward hierarchical structure. Each layer only receives the output of the previous layer. Sermanet et al. [8] used global and local features of two different scales in CNN-based classification. Compared with single-scale CNN, multi-scale CNN inputs the output of each pooling layer to the end fully connected layer for Classification and achieved good results. In the problem of traffic sign recognition, the difference in feature difference may not be obvious, which is mainly reflected in the difference in the local pattern of the main part of the figure. Therefore, in the recognition process, both global and local features are very important. It can be better to consider both global and local features. More accurate recognition of traffic signs. The multi-layer feature fusion method can solve this problem well. In the convolutional neural network, the input image is extracted after the convolution layer convolution, and after the visualization process, it can be clearly seen that each layer is extracted The features of each layer have different performances. With the deepening of the convolutional layer, the extracted features are gradually abstracted. In order to make full use of the features extracted from each layer, this article will adopt a multi-layer feature fusion method, that is, in LeNet Based on the -5 model, in addition to outputting the features extracted by the last convolutional layer of the convolutional neural network to the classifier, the features proposed by other convolutional layers will also be fed to the classifier.

In order to identify road signs more accurately and faster in practical applications, the architecture of the LeNet-5 model has been improved as follows: (1) The traditional LeNet-5 network has fewer convolution kernels per layer. If it is used for 62 types of road sign data For classification, it is impossible to fully distinguish the rich features of the target. Therefore, the number of convolution kernels should be changed reasonably. (2) The size of the convolution kernel in the convolutional neural network is usually set to 3×3, 5×5, and 7×7 respectively. In this algorithm, a 5×5 size convolution kernel is used. (3) The traditional LeNet-5 network uses Subsampling for downsampling. Few people have adopted it now. The method in this paper is improved to max pooling. (4) The LeNet-5 network uses the sigmoid activation function, but due to the complexity of the calculation When the input range of the network is wide, the neuron gradient will reach zero, which will affect the back propagation and cause the neural network to fail to train. Compared with the Sigmoid activation function, the ReLU activation function [9] may be more suitable for fitting training data, which helps to propagate the gradient to the next network during the backpropagation process and speed up the convergence of the network model. The number of calculations of the ReLU activation function is relatively small, and only one arithmetic operation is required, which can shorten the learning time of the network. Therefore, the activation function here is changed to a non-linear ReLU function. The formula is as follows: $f = max(o, x)$

In order to prevent the model from overfitting on a limited data set and accelerate the network convergence, it is necessary to normalize the data after convolution, and then introduce the BN layer, and introduce the Dropout layer in the fully connected layer [10-11], Dropout is A method of randomly deleting neurons during the learning process. During the training process, neurons in the hidden layer are randomly selected and then removed. The deleted neurons no longer transmit signals, as shown in Figure 4. During training, each time data is passed, the neuron to be deleted is randomly selected. Then, during the test, although all neuron signals are transmitted, the output of each neuron must be multiplied by the deletion ratio during training before output. The principle of Dropout is shown in Figure 3.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0017

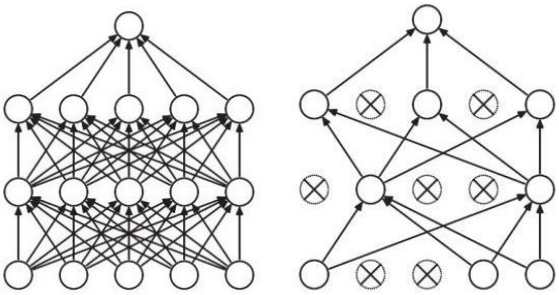

Figure 3. Dropout randomly deletes neurons

3. Results and discussion

3.1 Data set and preprocessing

This experimental data uses the German Traffic Signs Data Set (GTSRB) [12], which contains 43 types of signs, 6 categories: speed limit, indication, warning, ban, lifted ban, other types of signs. The entire traffic sign data set contains 51839 traffic signs collected in the natural environment. The specific information of the data set is shown in Table 1.

Table 1. GTSRB information

Number	Item	Data
1	Number of training examples	34799
2	Number of validation examples	4410
3	Number of testing examples	12630
4	Image data shape	32×32×3
5	Number of classes	43

By visualizing the data set information, the category distribution of the training set, validation set and test set can be obtained. As shown in Figure 3 and Figure 4, it can be seen that the number of various types of images in the training set of the GTSRB data set is obviously unbalanced: some types Has fewer than 200 pictures, and some have more than 2,000 pictures. This means that our model may be biased towards over-representative types, especially when its predictions are uncertain, so data augmentation is needed to alleviate this problem. The training set samples after data enhancement are evenly distributed, as shown in Figure 6, which increases the amount of training data and improves the generalization ability and robustness of the model. Before training the convolutional neural network, the input image needs to be preprocessed, including grayscale processing and size normalization. This paper uses bilinear interpolation to normalize the size of all traffic sign images to 32×32 pixels.

Before entering the convolutional neural network, it is also necessary to perform preliminary processing on the input image, including grayscale processing and size normalization. In this article, the two-line interpolation method is used to normalize the size of all road sign images to 32-x32 pixels.

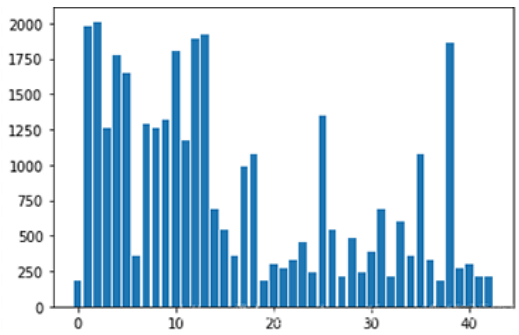

Figure 4. Training set category distribution

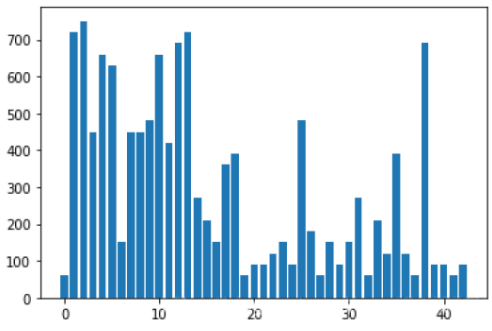

Figure 5. Test set category distribution

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0017

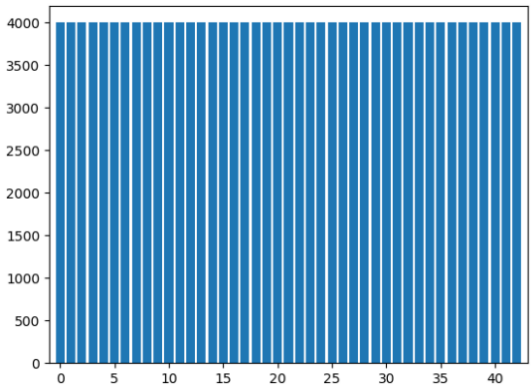

Figure 6. Test set category distribution after data enhancement

3.2 Experiment

The computer environment used in the experiment is Intel(R) Core (TM)i7-9750H, the CPU frequency is 2.6GHz, the memory is 16G, the operating system is Windows10, the graphics card is configured as NVIDIA Ge Force GTX 1080Ti, the video memory is 4G, and the depth The learning framework is Tensorflow, and Python is the programming language.

After a 32×32 image is input, it is convolved through the first convolution layer. The size of the convolution kernel is 5×5, and the number of convolution kernels is 64. The second convolution The layer uses 128 5×5 convolution kernels, the maximum pooling layer window of the two layers is 2×2, and the span is 2. The first layer of convolutional layer pooled features, after 4×4 padding is the maximum pooling of 4, the fully connected layer is connected to the second layer pooled features, the second layer pooling layer features After the same maximum pooling and connection, the obtained multi-scale features are connected in series, and then input to the fully connected layer (the number of neurons is 256); the fully connected layer is followed by the softmax output layer, and the number of neurons is 43.

The parameters of this experiment are set as follows: the initial learning rate is 0.001, when the number of epochs exceeds 50, the learning rate becomes 0.0001. The size of BatchSize is 200, and the total number of epochs is 150. The keep_prob parameter of dropout is 0.5 in the fully connected layer.

Regarding the choice of the convolution kernel size and activation function in the network model, whether to add a Dropout layer, and whether to use a multi-scale feature fusion method, this study did the following comparative experiments, and finally determined the improved convolutional neural network model.

3.2.1. Improved activation function comparison

Inappropriate activation function will cause gradient saturation, gradient disappearance and other phenomena. The experimental results of the activation function comparison are shown in Table 2. From the results, it can be seen that the accuracy of the ReLU activation function is higher than the sigmoid activation function. It also avoids the situation where the gradient disappears and cannot be trained, so it makes sense to improve the activation function.

Table 2. The activation function selects the experimental results

activation function	Iteration time	Recognition accuracy
sigmoid	0.916	95.17
ReLU	0.925	98.48

3.2.2. Convolution kernel size selection comparison

Regarding the choice of the size of the convolution kernel, this experiment compared the influence of the three convolution kernel sizes of 3×3, 5×5, and 7×7 on the network structure. The results are

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0017

shown in Table 3. It can be concluded that although 5× The training time of 5 convolution kernel size is not the shortest, but its recognition rate is higher than the other two convolution kernel sizes, and the loss is also low. The overall performance is better than 3×3, and 7×7 convolution kernel sizes.

Table 3. Experimental results of convolution kernel size

convolution kernel size	Iteration time (s)	loss	Recognition accuracy (%)
3×3	4578	0.0941	96.54
5×5	3872	0.0730	98.48
7×7	3169	0.0751	97.81

3.2.3. Overall improvement comparison

It can be seen from Table 5 that the method in this paper integrates multi-scale features and improves the traditional LeNet-5 network. The recognition accuracy rate is about 3% higher than that of the traditional LeNet-5 network, indicating the generality of the network model in this paper. Better performance, stronger ability to express the characteristics of traffic signs.

Table 4. The results are compared with other algorithms

algorithms	Training accuracy (%)	Test accuracy (%)
LeNet-5	97.47	95.45
Single-scale features	98.13	96.53
Method of this article	98.82	98.48

4. Conclusion

In order to detect and recognize road signs in real time and accurately, an improved LeNet-5 model of classical neural network structure is proposed and applied to road signs detection. In order to meet the requirements of real-time recognition of road signs, the three fully connected layers in the traditional LeNet-5 network have been improved and optimized to significantly increase the recognition speed; in order to improve the recognition accuracy and reduce the loss in the recognition process, the traditional activation function will LeNet-5 in LeNet-5 is changed to the ReLU activation function, and the multi-layer feature fusion method is introduced to make full use of the features extracted by each layer of CNN; to avoid over-fitting, a loss mechanism is introduced, which effectively improves the road signs The detection and recognition effect. The experimental results show that the accuracy of the proposed method in road sign recognition is 98,82%, and the test accuracy is 98,48%, which is in line with the original intention of improvement, and meets the requirements of recognition accuracy and real-time. It has laid a good foundation for further research on road sign recognition and unmanned driving projects.

References

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 1: 886-893.

[2] Schügerl P, Sorschag R, Bailer W, et al. Object re-detection using SIFT and MPEG-7 color descriptors [C]// International Workshop on Multimedia Content Analysis and Mining. Springer, Berlin, Heidelberg, 2007: 305-314..

[3] ZHU Y, WANG X, YAO C, et al. Traffic sign classification using two-layer image representation[C]// 2013 IEEE International Conference on Image Processing. Melbourne: IEEE, 2013: 3755-3759.

[4] Li Xinye, HUANG Teng. Fine Vehicle Model Identification based on multi-scale leaping-layer convolutional neural network [J]. Science, technology and engineering, 2017, 17(11):246-249.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202104_7(4).0017

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

[6] Zhou Feiyan, Jin Linpeng, Dong Jun. Research review of convolutional neural networks. Acta Computerica Sinica,2017, 40(6):1229–1251.

[7] Wu Yangyang, PENG Guangde, Wu Xiangfei. An improved convolution neural network image recognition method based on LenET-5, Information and computer,

[8] Zhu Ying, TAO Jiyu, Ling Li. Research on Traffic Sign Recognition Algorithm Based on Improved Lenet-5 [J]. Application of Microcomputer,

[9] SERMANET P, LECUN Y.Traffic sign recognition with multi-scale convolutional networks[C]//The 2011International Joint Conference on Neural Networks (IJCNN) .Washington DC:IEEE Computer Society, 2011:2809-2813.

[10] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. ar Xiv preprint ar Xiv:1207.0580, 2012.

[11] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[12] [Krizhevsky, Alex, Sutskever, Ilya, Hinton, Geoffrey E. Image Net Classification with Deep Convolutional Neural Networks[J]. Communications of the ACM,2017,60(6):84-90.

[13] Namor A F D D, Shehab M, Khalife R, et al. The German Traffic Sign Recognition Benchmark:A multi-class classification competition[C]// International Joint Conference on Neural Networks. IEEE, 2011: 1453-1460.

[14] Fan Xing, Zhao Xiangmo, Liu Zhanwen, et al. Traffic Sign Recognition Method based on multi-scale convolutional neural network [J]. Modern electronic technology,2019,42(15): 134- 138.