 CMC, vol.57, no.1, pp.11-24, 2018

Improved VGG Model for Road Traffic Sign Recognition

Shuren Zhou[1, 2, *], Wenlong Liang[1, 2], Junguo Li[1, 2] and Jeong-Uk Kim[3]

Abstract: Road traffic sign recognition is an important task in intelligent transportation system. Convolutional neural networks (CNNs) have achieved a breakthrough in computer vision tasks and made great success in traffic sign classification. In this paper, it presents a road traffic sign recognition algorithm based on a convolutional neural network. In natural scenes, traffic signs are disturbed by factors such as illumination, occlusion, missing and deformation, and the accuracy of recognition decreases, this paper proposes a model called Improved VGG (IVGG) inspired by VGG model. The IVGG model includes 9 layers, compared with the original VGG model, it is added max-pooling operation and dropout operation after multiple convolutional layers, to catch the main features and save the training time. The paper proposes the method which adds dropout and Batch Normalization (BN) operations after each fully-connected layer, to further accelerate the model convergence, and then it can get better classification effect. It uses the German Traffic Sign Recognition Benchmark (GTSRB) dataset in the experiment. The IVGG model enhances the recognition rate of traffic signs and robustness by using the data augmentation and transfer learning, and the spent time is also reduced greatly.

Keywords: Intelligent transportation, traffic sign, deep learning, GTSRB, data augmentation.

1 Introduction

Traffic sign is an important part of traffic system, it contains a lot of intuitive and useful information. Traffic sign recognition can provide drivers with safe environment, it also provides some guidance information to solve the traffic jam. At present, many mobile devices based on ARM processor have the function of driving assistance, which can effectively prompt drivers. The traditional traffic sign recognition algorithms are mostly based on the single background, in complex scene, the recognition effect still needs to be improved.

At present, the identification methods of traffic signs can be roughly divided into

[1] Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science & Technology, Changsha, 410114, China.

[2] School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha, 410114, China.

[3] Department of Energy Grid, Sangmyung University, Seoul, 110743, Korea.

[*] Corresponding Author: Shuren Zhou. Email: zsr@csust.edu.cn.

shape-based [Mohammadi and Makui (2016)], color-based [Jung, Lee, Jung et al. (2016)] and convolutional neural network (CNN) [Girshick, Donahue, Darrell et al. (2014)]. In [Lasota and Skoczylas (2016)], traffic sign recognition is done by SIFT feature and descriptor of feature subspace. However, SIFT is based on the point of object appearance and ignores the intrinsic related information of the whole image. Wang et al. [Wang, Yu and Deng (2015)] are based on the central projection features, the trained probabilistic neural network is used to identify traffic signs and compared with SIFT feature recognition. The projection feature in the experiment get high recognition rate, but the center projection feature only describes the shape feature. Sheikh et al. [Sheikh, Kole and Maity (2017)] proposed a color-normalized traffic sign classification model, which simplified the complex color information into 5 classes. For the five basic colors, the M_SVMs [Yan and Du (2009)] is used to identify the traffic signs, which performs better in the coarse classification and less effectively in the sub-classification. In Jin et al. [Jin, Fu and Zhang (2014)], a method based on stochastic gradient descent and convolutional neural network recognition algorithm, the recognition rate is 99.65%, but the training time for the model is up to 50 h. Yin et al. [Yin, Peng, Liu et al. (2015)] proposed rotation-invariant pattern features based on traffic signs, this feature is used as an input to train the artificial neural network, the recognition rate is 98.62%.

After many years of accumulation and perfection, the artificial neural networks have been widely used in various fields and also have achieved remarkable results in the field of image recognition. One of the major advantages is that the original image can be used as an input and the automatic training feature, and further reduce the manual pre-processing. We present an Improved VGG (IVGG) model, and to use the data augmentation and transfer learning strategies to further enhance the traffic sign recognition effect on the GTSRB [Namor, Shehab, Khalife et al. (2011)]. Moreover, the amount of traffic sign information is large, the burden on the processor is relatively large [Cai, Wang, Zheng et al. (2013)]. We try to design the model as simply and practically as possible in order to explore the possibility of porting to mobile devices later.

2 Related works

2.1 VGG model

VGG [Sercu, Puhrsch, Kingsbury et al. (2016)] is the champion of the 2014 ImageNet competition, the model is developed by AlexNet [Krizhevsky, Sutskever and Hinton (2012)]. The VGG model has two characteristics: The first is that the convolution kernels are small, most of size are 3×3, and few are 1×1. The convolution operation is accompanied by an activation function, which can identify more abundant features; the second is the small pooling kernel, compared to AlexNet's 3×3 pooling kernel, VGG only has pooling kernel of size 2×2, making the layers deeper, the feature map can be wider. As the convolution kernel focuses on expanding the number of channels, pooling layers focuses on narrowing the width and height, making the model structure deeper and wider, while the increase in computational volume slows down.

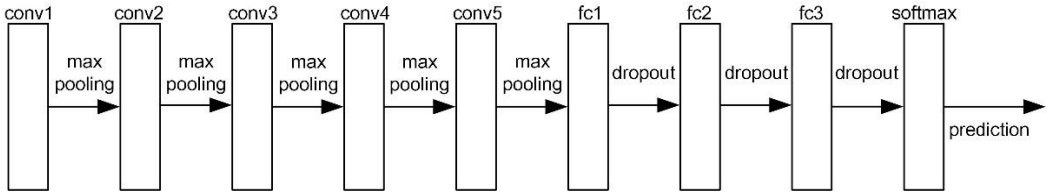

Figure 1: VGG_16 model

VGG_16 is one of the most classic version, as shown in Fig. 1, conv1 and conv2, each contains two convolutional layers; conv3, conv4 and conv5, each contains 3 convolutional layers; fc1, fc2 and fc3, each contains a fully-connected layer; the last one is the image feature classification layer, they have a total of 16 layers. Although VGG_16' effect is better, convergence is faster, but it also has some disadvantages. With the deepening of model level, the training parameters is becoming more complex, it will take a lot of time, what we do is to solve these problems.

2.2 Research of model

In the case of insufficient artificial characteristics, to further improve the effect of the model, the usual measures are:

(1) Artificially increase the training set, through picture flip, noise adding and other methods, to create a new group of data, although this method can improve the effect of the model, the artificial processing efficiency is low, and there are a lot of human errors, the effect of the model will be affected by many uncertain factors.

(2) Regularization method, we do not add new data, but add regular items after loss function to suppress the production of fitting. Although it is beneficial to achieve better results in small dataset, it is necessary to introduce a hyper parameter of manual debugging which will cost a lot of time.

(3) Unsupervised pre-training methods [Erhan, Bengio, Courville et al. (2010)]. It performs unsupervised training through an automatic encoder. In the end, to fine-tune the model, which takes a lot of time, and we need change a lot.

(4) Data augmentation [Houben, Stallkamp, Salmen et al. (2008)] and transfer learning [Shen, Wu and Suk (2017)]. Data augmentation is to generate new datasets by means of image rotation and adding noise. Transfer learning can achieve the same functionality in a new field of the same task after model training for data in similar areas. By studying the data characteristics in the previous fields, the model parameters are applied to the new domain to achieve the same function. Data augmentation eliminates the impact of artificial labeling and human error, while data augmentation can automatically transform the raw data and greatly reduce human consumption. Transfer learning is beneficial to accelerate the convergence of the model, to classify the detailed data and improve the classification effect. The combination of the two can help prevent the model from over fitting and enhance the generalization ability of the model.

In our work, we combine data augment and transfer learning, to further enhance the recognition effect of the model.

3 IVGG model and its method

3.1 IVGG model

In our experiment, we design an IVGG model with only 9 layers. It contains max-pooling and dropout operations. Max-pooling can divide the images into several blocks of the same size, and only takes the largest value in each block. After abandoning other nodes, it maintains the original plane structure invariance, and further reduces the amount of computation.

The main goal when using dropout is to regularize the neural network we are training. The technique consists of dropping neurons randomly with some probability p. Those random modifications of the network's structure are believed to avoid co-adaptation of neurons by making it impossible for two subsequent neurons to rely solely on each other, which is a good way to prevent overfitting.

After each fully-connected layer, we add dropout and Batch Normalization (BN) to further accelerate the model convergence and improve the classification effect. Training convolutional neural networks is complicated by the fact that the distribution of each layer's inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates and careful parameter initialization, and makes it notoriously hard to train models with saturating nonlinearities. The BN is to solve this problem. By normalizing the input of each layer, it ensures that the input data distribution in each layer is stable, thus achieving the purpose of accelerated training. The data augmentation and transfer learning also can accelerate the training of the network model. The improved model is shown in Fig. 2, and the layer information as follows:

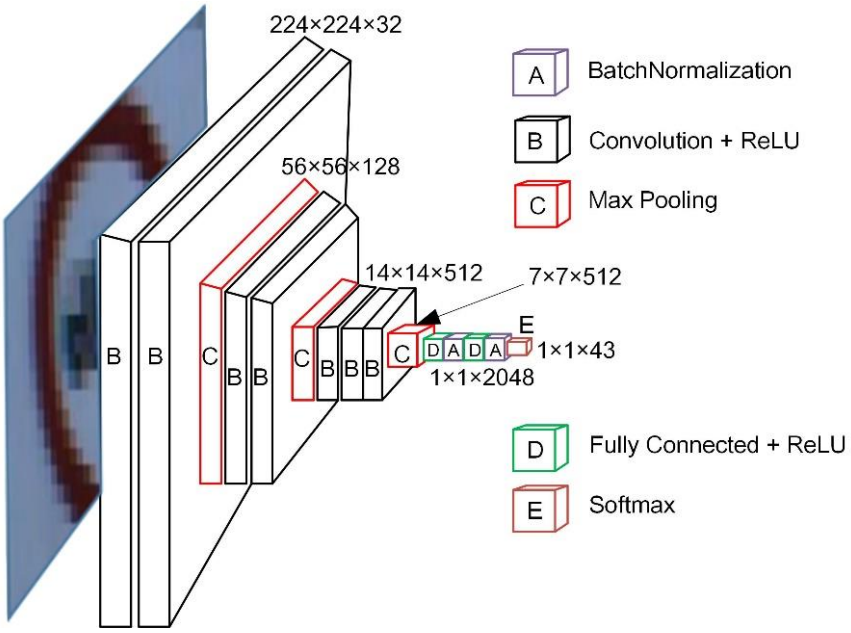

Figure 2: Structure of the IVGG model

Input layer: Loading the image, to produce an output vector as input to the convolutional layer. The model loads the entire 3-channel traffic sign image and the image is automatically scaled to a size of 224×224 to accommodate the convolutional layer of the model.

Convolutional layer: Responsible for feature learning, which consists of a set of feature maps, and a feature map shares a convolution kernel. A learning convolution kernel is convoluted with several previous feature maps, adds the corresponding element and a bias, and then transfers to a non-linear activation function. In our model, we adopt the ReLU function to get a feature map and achieve a feature extraction. This model has a total of 7 convolution kernels of size 3×3, the first two convolutional layers contains 32 feature maps, it followed by a pooling operation. The pooling kernel of size 4×4 with a stride of 4 pixels (this is the distance between the receptive field centers of neighboring neurons in a kernel map), the way is max-pooling. The middle 2 convolutional layers contain the feature map size of 128, followed by the pooling kernel size of 4×4 with a stride of 2 pixels. The last 3 convolutional layers contains feature maps size of 512, followed by a pooling kernel size of 7×7 with a stride of 2 pixels, and each pooling operation followed by a dropout, their parameter is 0.2, and the final dropout is 0.5, the structure is shown in Fig. 3.

Fully-connected layer: They vote for the abstract feature high-dimensional probability map, the probability of achieving each category classification. The last two layers of this model are fully connected, feature map size of 2048, and each fully-connected layer is followed by BN, and BN followed by a dropout with a parameter of 0.5.

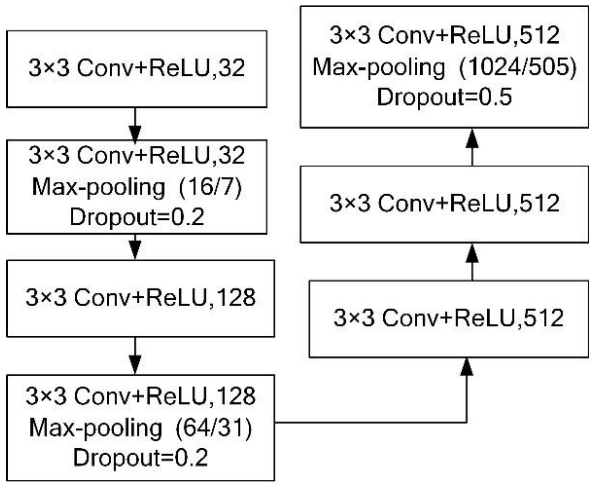

Figure 3: Structure of convolutional layers

3.2 Data augmentation and transfer learning

Due to the lack of a large number of traffic sign data, especially some special traffic signs, it is obvious that the training of the model easily leads to over-fitting problem. Therefore, based on some German traffic sign dataset [Suzuki, Zhang, Homma et al. (2016)], we adopt a combination of two solutions:

One is the data augmentation. We use reflection deformation, the image data for 90 degrees and 180 degrees of rotation, according to the image 0.8 or 2.0 scaling, in the horizontal and vertical mirror and its combination of operations, the data thus expanded by nearly 2 times. Expanding from over 30,000 in 43 categories of signs to nearly 120,000. The samples are shown in Fig. 4.

Figure 4: Traffic sign image data augmentation effect

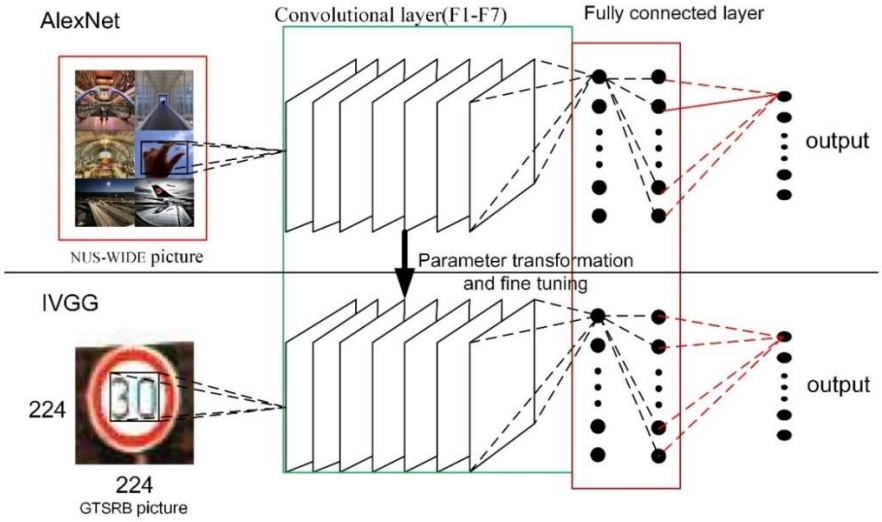

Figure 5: Flow chart of transfer learning

Another one is the transfer learning method, we pre-train existing datasets in order to get the initialization parameters of the model, and migrate to the target dataset for fine-tuning training. Transfer learning can obtain the basic characteristics of dataset classification in pre-training, such as color, texture, etc. We can improve the classification ability of the model. In this paper, we pre-train the Alex Net model using the NUS-WIDE [Chua, Tang, Hong et al. (2009)] (NUS-WIDE is a dataset with a network label tagging, containing 269,648 images from the website, 5,018 different types of tags) dataset, then transfer the model parameters to the IVGG model and fine-tune the training using the German traffic

sign dataset. The transfer learning process is shown in Fig. 5.

In our work, we use data augmentation and transfer learning method to further improve the IVGG model. Data augmentation mainly affects the input stage of the model and enhances the quantity and uniqueness of data. Transfer learning mainly plays an important role in model parameter initialization. Transfer learning can obtain good initialization parameters, which has a great influence on the training of models.

4 Experiments

4.1 Datasets

In our work, we use the dataset from the GTSRB to make experiments. The dataset is obtained by cropping the traffic signs in the natural scene. It contains the sign image under the influence factors such as illumination, occlusion, rotation, etc. The samples of this dataset are shown in Fig. 6.

Figure 6: Samples of the GTSRB

In our experiments, 31,367 images of the dataset and total 43 categories are used. Moreover, data augmentation is adopted to expand the data, then the data volume is increased to 125,648 pieces, and there is no identical image in all dataset.

4.2 Evaluation metric

The main evaluation metrics in this paper are the spent time, $Precision\ rate$, $Recall\ rate$ and $F1\_score\ rate$ of the model. There are only four conditions: The prediction is positive, and actual is positive, called TP (true positive); the prediction is positive, and actual is negative, called FP (false positive); the prediction is negative, and actual is positive, called FN (false negative); the prediction is negative, and actual is negative, called TN (true negative).

Obviously given a test set, it has the following relationships:

$$N_{pre} = TP + TN \tag{1}$$

$$N_{total} = TP + TN + FP + FN \quad (2)$$

N_{pre} is the number of the predicted pairs, and N_{total} is the number of all the samples, the $Precision\ rate$, $Recall\ rate$ and $F1\_score\ rate$ can be expressed as:

$$Precision\ rate = TP\ /\ (TP + FP) \quad (3)$$

$$Recall\ rate = TP\ /\ (TP + FP) \quad (4)$$

$$F1\_score\ rate = 2TP(2TP + FN + FP) \quad (5)$$

4.3 Training strategies

All data is randomly divided into three parts: the training set is 50%, the test set is 25%, and the verification set is 25%. The training set is used for model training and parameter learning. The validation set is used for model optimization, and the model is tested in the model training process, then the network fine-tuning is performed according to the model test results. The test set is used to test the identification capability of the model and the generalization ability of the model.

Meanwhile, to verify the effectiveness of data augmentation and transfer learning strategies, we have made a comparison between the VGG_16 model and the IVGG model, which further proves this model is scientific.

4.4 Experimental environment

We adopt Ubuntu14.04, Intel I7 CPU, NVIDIA Quadro K2200 GPU, TensorFlow frame, and others like python3.6 and Anaconda3. It takes about 8 h to train before the data augmentation, while the training time is about 23 h after that. Finally, the experiment code was written and tested in Jupyter Notebook.

4.5 Experimental results analysis

4.5.1 VGG_16 vs. IVGG

VGG_16 compared with the IVGG in this paper, the rate of recognition and loss of sign, to complete an epoch time-consuming, and multiple epoch average time-consuming, from the experiment results we can see:

As shown in Fig. 7, under the test of German traffic sign dataset, the average recognition rate of VGG_16 model is 94% and the average model loss rate is 0.192. While the average recognition rate of IVGG model increases by 5% to 99 %, the degree of model loss decreased by 0.134 to 0.058. It shows that the IVGG model has made great progress compared with VGG_16.

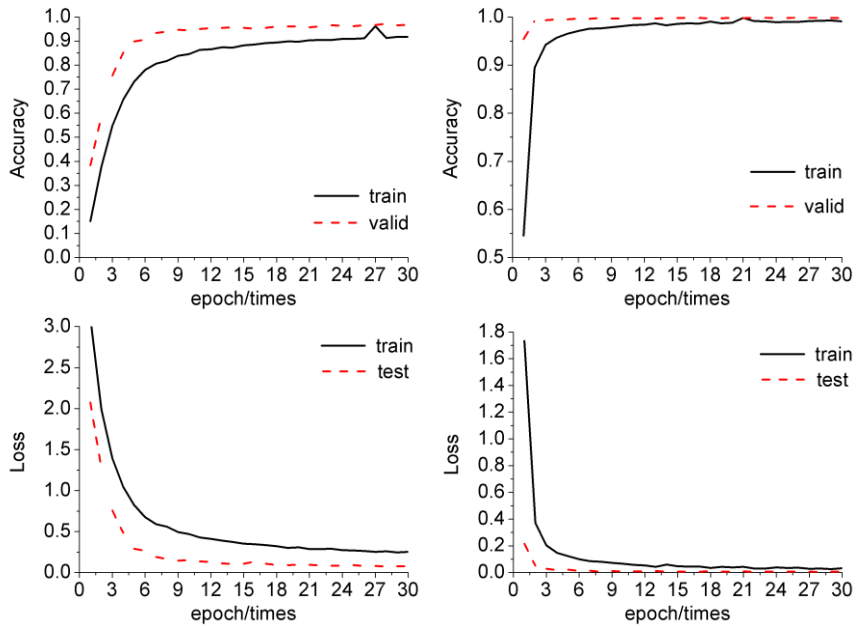

Figure 7: Comparison of accuracy and loss of the VGG_16 (two pictures on the left) and IVGG (two pictures on the right) models

We further compare the performance of VGG_16 with IVGG in $Precision\ rate$, $Recall\ rate$ and $F1\_score\ rate$ which are shown in Fig. 8, Fig. 9 and Fig. 10.

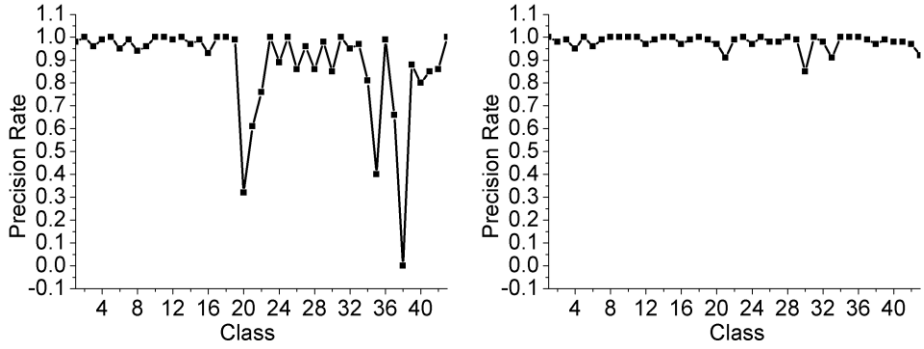

Figure 8: Comparison of VGG_16 (left) and IVGG (right) $Precision\ rate$

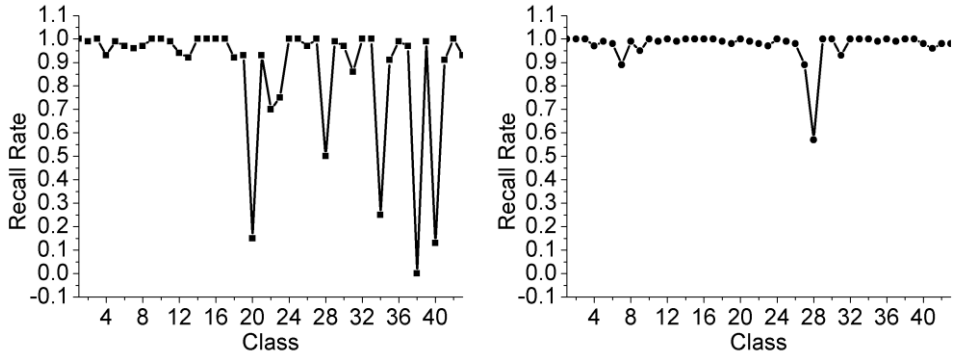

Figure 9: Comparison of VGG_16 (left) and IVGG (right) $Recall\ rate$

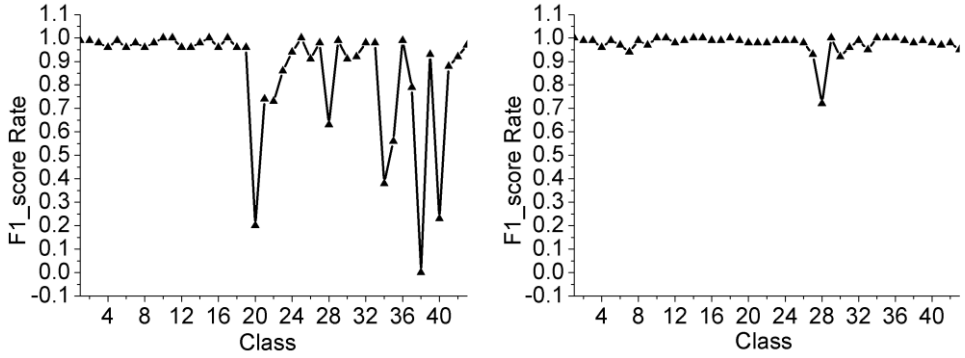

Figure 10: Comparison of VGG_16 (left) and IVGG (right) $F1\_score\ rate$

As we can see from the Fig. 8, Fig. 9 and Fig. 10, the IVGG model performs better than the VGG_16 model in the test of 43 category of traffic signs, with the $Precision\ rate$, $Recall\ rate$ and $F1\_score\ rate$, especially in some categories, such as the 20th, 28th, 32th and 38th categories, where the $Recall\ rate$ is the most obvious and the $Precision\ rate$ is higher.

In contrasting the model time-consuming, we compare the time which takes for each epoch to be calculated by the model, as shown in Fig. 11 below:

In our experiments, the number of images trained in each iteration is 600 and a total of 30 epochs. Fig. 11 shows the spent time of epochs when all 3,000 data samples are compared. It can be seen clearly that the IVGG compared to VGG_16 each epoch about 10 s faster, the statistical analysis of the method of the epoch is about 63 s each time.

We compare the VGG_16 with the IVGG by using the GTSRB. By the test of the 43 different traffic sign datasets, we find that the IVGG model has obvious advantages not only in $Precision\ rate$, $Recall\ rate$ and $F1\_score\ rate$, but also has high recognition effect and generalization ability.

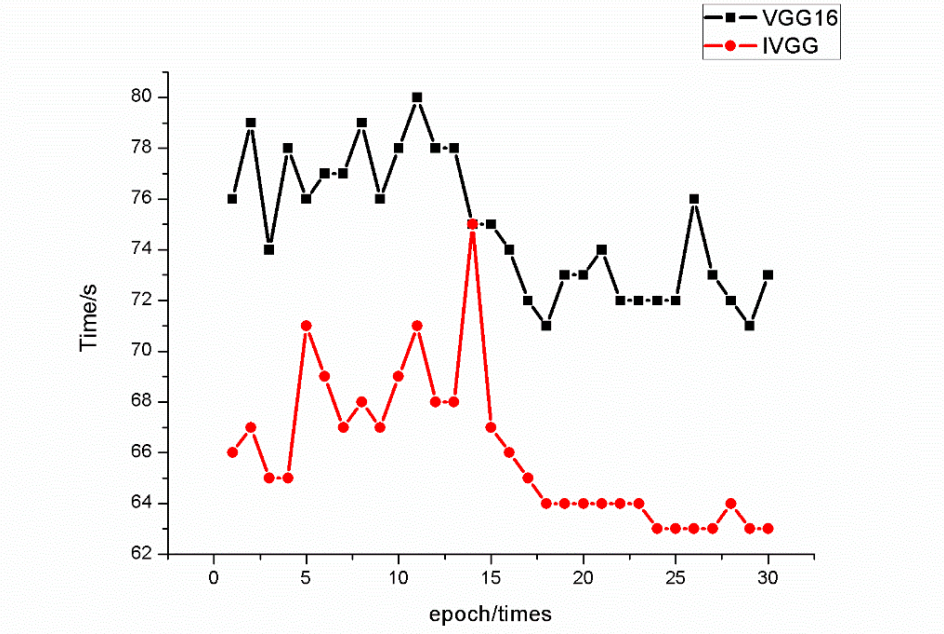

Figure 11: VGG_16 *vs*. IVGG training time

4.5.2 Methods comparison

From the Tab. 1, our method is just lower than the method in Jin et al. [Jin, Fu and Zhang (2014)]. We find this method through by data augmentation, it will get better effect. But the training parameters reaches to 1,162,284, and the training time is up to 50 h, our training time only takes 11 h.

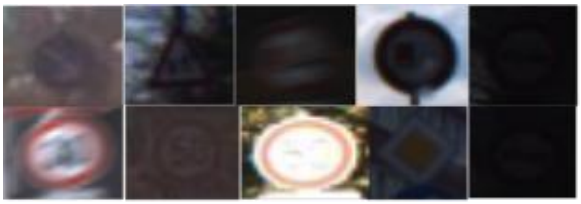

Figure 12: Identifies the erroneous part of the samples

However, the IVGG model still needs to be further improved in some aspects. As shown in Fig. 12, most of the sign images which are erroneous are particularly affected by light, especially when the light is insufficient. Because of movements, the traffic signs are blurred, and the recognition rate has dropped a lot, and the distance of the sign also influences on the recognition.

Table 1: Traffic sign recognition rate in various methods

Comparison	Method	Recognition rate
[Jin, Fu and Zhang (2014)]	Random gradient descending with CNN	99.65%
[Zeng, Xu, Shen et al. (2017)]	Multi-Scale CNN (official)	98.31%
[Sermanet and Lecun (2011)]	Multi-Scale CNN (best)	98.97%
[Aghdam, Heravi and Puig (2016)]	Linear Discriminant Analysis (LDA)	95.68%
[Luo, Yang, Tong et al. (2017)]	CNN for Single CNN (best)	98.80%
Ours	Improved VGG	99.00%

5 Conclusion

In this work, we have presented an Improved VGG (IVGG) model. By using data augmentation and transfer learning methods, it accelerates the network training and convergence. In the German traffic sign dataset test, our recognition rate is up to 99.00%. Our method is much better than the multi-scale convolutional network. But the IVGG model is still deficient in recognizing the traffic signs with dark background and blurred images. We will focus on how to enhance the identification of traffic signs under the darkness and motion blur background in the future work.

Acknowledgment: This work was supported by the Scientific Research Fund of Hunan Provincial Education Department of China (Project No. 17A007); and the Teaching Reform and Research Project of Hunan Province of China (Project No. JG1615).

References

Aghdam, H. H.; Heravi, E. J.; Puig, D. (2016): Recognizing traffic signs using a practical deep neural network. *Robot 2015: Second Iberian Robotics Conference*, vol. 417, pp. 399-410.

Cai, Z.; Wang, Z.; Zheng, K.; Cao, J. (2013): A distributed TCAM coprocessor architecture for integrated longest prefix matching, policy filtering, and content filtering. *IEEE Transactions on Computers*, vol. 62, no. 3, pp. 417-427.

Chua, T. S.; Tang, J.; Hong, R.; Li, H.; Luo, Z. et al. (2009): NUS-WIDE: A real-world web image database from national university of Singapore. *ACM International Conference on Image and Video Retrieval*, pp. 48.

Erhan, D.; Bengio, Y.; Courville, A.; Manzagol, P. A.; Vincent, P. et al. (2010): Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, vol. 11, no. 3, pp. 625-660.

Glorot, X.; Bengio, Y. (2010): Understanding the difficulty of training deep feedforward

neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249-256.

Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. (2014): Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587.

Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; Igel, C. (2008): Detection of traffic signs in real-world images: The German traffic sign detection benchmark. *IEEE International Joint Conference on Neural Networks*, pp. 1-8.

Jin, J.; Fu, K.; Zhang, C. (2014): Traffic sign recognition with hinge loss trained convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1991-2000.

Jung, S.; Lee, U.; Jung, J.; Shim, D. H. (2016): Real-time traffic sign recognition system with deep convolutional neural network. *IEEE International Conference on Ubiquitous Robots and Ambient Intelligence*, pp. 31-34.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012): ImageNet classification with deep convolutional neural networks. *International Conference on Neural Information Processing Systems*, vol. 60, pp. 1097-1105.

Lasota, M; Skoczylas, M. (2016): Recognition of multiple traffic signs using keypoints feature detectors. *IEEE International Conference and Exposition on Electrical and Power Engineering*, pp. 535-540.

Luo, H.; Yang, Y.; Tong, B.; Wu, F.; Fan, B. (2017): Traffic sign recognition using a multi-task convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, vol. 99, pp. 1-12.

Mohammadi, S. E.; Makui, A. (2016): Multi-attribute group decision making approach based on interval-valued intuitionistic fuzzy sets and evidential reasoning methodology. *Soft Computing*, vol. 21, no. 17, pp. 5061-5080.

Namor, A. F. D. D.; Shehab, M.; Khalife, R.; Abbas, I. (2011): The German traffic sign recognition benchmark: A multi-class classification competition. *IEEE International Joint Conference on Neural Networks*, vol. 3, pp. 1453-1460.

Shen, D.; Wu, G.; Suk, H. I. (2017): Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221-248.

Sercu, T.; Puhrsch, C.; Kingsbury, B.; LeCun, Y. (2016): Very deep multilingual convolutional neural networks for LVCSR. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4955-4959.

Suzuki, S.; Zhang, X.; Homma, N.; Ichiji, K.; Sugita, N. et al. (2016): Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis. *IEEE Society of Instrument and Control Engineers of Japan*, pp. 1382-1386.

Sermanet, P.; Lecun, Y. (2011): Traffic sign recognition with multi-scale convolutional networks. *IEEE International Joint Conference on Neural Networks*, vol. 42, pp. 2809-2813.

Sheikh, M. A. A.; Kole, A.; Maity, T. (2017): Traffic sign detection and classification using colour feature and neural network. *IEEE International Conference on Intelligent*

Control Power and Instrumentation, pp. 307-311.

Wang, Q.; Yu, H.; Deng, D. (2015): Automatic recognition for mechanical images based on sparse non-negative matrix factorization and probabilistic neural networks. *IEEE International Conference on Mechatronics and Automation*, pp. 2408-2413.

Yan, Z.; Du, P. (2009): Generalization performance analysis of M-SVMs. *Journal of Data Acquisition & Processing*, vol. 24, no. 4, pp. 469-475.

Yin, S.; Peng, O.; Liu, L.; Guo, Y.; Wei, S. (2015): Fast traffic sign recognition with a rotation invariant binary pattern based feature. *Sensors*, vol. 15, no. 1, pp. 2161-2180.

Zeng, Y.; Xu, X.; Shen, D.; Fang, Y.; Xiao, Z. (2017): Traffic sign recognition using kernel extreme learning machines with deep perceptual features. *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1647-1653.