ISSN 2522-3488 Vol. 7, Issue 3: 21-27, DOI: 10.25236/IJNDES.2023.070304

Pavement Recognition Based on Improving VGG16 Network Model

Shuoyi Wen[1], Li Yang[1,*], Hailong Duan[1], Tingting Zhang[1]

[1]School of Automation and Electrical Engineering, Tianjin University of Technology and Education, Tianjin, China
[*]Corresponding author

Abstract: *In order to improve the accuracy of pavement recognition, an improved RA-VGG16 network model classification method based on VGG16 is proposed in this paper. The improvements include reducing the number of convolution cores in VGG16 to optimize the network structure, adding the improved residual attention module to achieve the extraction of road notable features, using the global average pooling layer instead of the full connection layer to significantly reduce the network parameters and prevent network over fitting. The experimental results show that the accuracy of the improved VGG16 network model is 99.36%, 15.26% higher than that of the original VGG16, and significantly higher than other models (KNN, Alex Net, VGG13, ResNet50, ResNet101).*

Keywords: *Road surface recognition, VGG16, The residual attention module, Feature extraction*

1. Introduction

With the development of intelligent vehicles, environmental awareness technology serves as the of vehicle driving decision control and path planning, and road type recognition is an important of ecological awareness technology. The purpose of pavement classification is exercised by vision sensors to realize automatic classification of different pavement types such as grassland, muddy and so on. In the research of road type classification and identification, researchers have proposed various solutions using different vision sensors. Pavement classification methods can be divided into 2 one is based on traditional machine learning and the other is based on CNN. Blas et al. Used LBP[1] feature extraction and K-averaged clustering algorithm to identify the ground type and the passable and realized the prediction of different road types. The VGG network including LBP features in[2] has realized part of pavement recognition. A classical neural network structure suggests in [3], but its classification accuracy is poor. In 2010, T. Y. Kim adopted wavelet feature extraction method and selected neural network machine learning method to train the classifier with an average accuracy of 80% [4] .

In this paper, 6 kinds of different pavements are taken as the research object. The VGG16[5]network model is improved and compared with other models to verify the performance of an improved model. enhancing the network, the road type can be recognized quickly and accurately.

2. Related work

2.1. VGG16 network architecture

Common convolution neural networks are Le Net [6], Alex Net [7], VGG [8-9] and Google Net[10]. VGG Net is invoked as a model developed by the Visual Geometry Group of the University of Oxford. The 2014 Image Net Image classification [11] and the positioning challenge, ILSVRC-2014, ranked second the classification task and first in the positioning task. VGG16 is one of the best performing networks VGG Net, and its network structure is shown in Figure 1. As can be observed in Figure 1, VGG16 consists of a convolution layer, a pooling layer, and a fully connected layer. The performance of. For enhancing the network, the road type can be recognized quickly and accurately.

ISSN 2522-3488 Vol. 7, Issue 3: 21-27, DOI: 10.25236/IJNDES.2023.070304

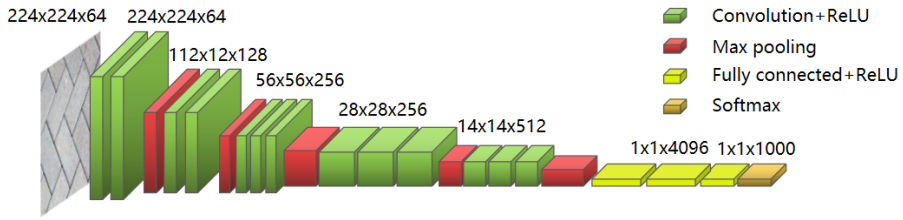

Figure 1: VGG16 network architecture

VGG16 has 16 layers including 13 convolution layers and 3 full connection layers, which can be subdivided into 6 segments, namely 5 segments convolution and 1 full connection. Five segments of convolution are utilized to extract the image features of depressed, medium and high layers, each with 2 or 3 convolution layers. In order to enhance the non-linearity of the network, prevent the gradient from disappearing, reduce the over fitting and improve the training speed of the network. ReLU activation function is used in each convolution layer. In order to capture the details, obtain better nonlinear effects and decrease the number of parameters. The convolution kernel used in each convolution layer makes the network structure more concise. At the tail of 5 convolutions, there is a maximum pooled layer, which can decrease the deviation of the estimated value caused by the convolution layer parameter error and preserve the detail information. The number of neurons in the last three connective layers is 4096, 4096 and 1000 respectively. The number of neurons in the remaining connective layer is decided by the specific task.

2.2. Attention mechanisms and full-scale feature fusion

Human visual perception selectively focuses on the object while scanning the global scene, ignoring the useless information to determine and analyze the regions of extreme interest. Although the datasets used throughout this paper are reprocessed, there are still some redundant features in the road surface images. The attention mechanism distributes the weight to represent the attention degree by the way of assigning the weight to the vital information which needs to be given attention. So it can focus on the road area and ignore the superfluous information.

In the field of computer vision, attention mechanisms can be separated into channel attention[12]and spatial attention [13-14] depending on the way in which they function and the dimensions of their characteristics. The channel attention mechanism sets weights for each channel domain to indicate the correlation between the channel information and the final core information. The channel attention module structure diagram is shown in Figure 2. Spatial attention mechanism is mainly used in the longitudinal dimension of image feature map, which can efficiently mine the principal feature information in image and help convolution neural network to capture the target area.

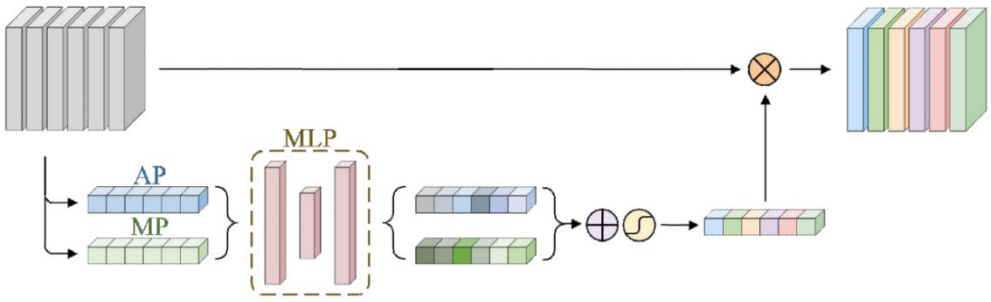

Figure 2: Structure of channel attention module

In most image classification tasks, a convolution neural network is utilized to extract the target features of the input image layer by layer. In input image processing of convolution neural network, the low-level network has smaller receptive field, stronger detail expression and weaker semantic expression, while the deep-level network has the opposite. Therefore, the fusion of multiscale features can take full advantage of image details and high-level semantic information, and finally achieve the accuracy of road classification.

ISSN 2522-3488 Vol. 7, Issue 3: 21-27, DOI: 10.25236/IJNDES.2023.070304

2.3. Improvements to VGG16

Although VGG16 has a simple structure, it contains an astonishing amount of weight, almost 75% of which is fully connected, and it is easy to over-fit when there is little data and low similarity. The weight file size of the model generated by VGG16 is about 500 MB, which occupies a lot of memory and is not conducive to the deployment of the model.

In order to recognize pavement type quickly and accurately, this paper presents a lightweight VGG model based on VGG16 model, as shown in Figure 3. The following adjustments are made.

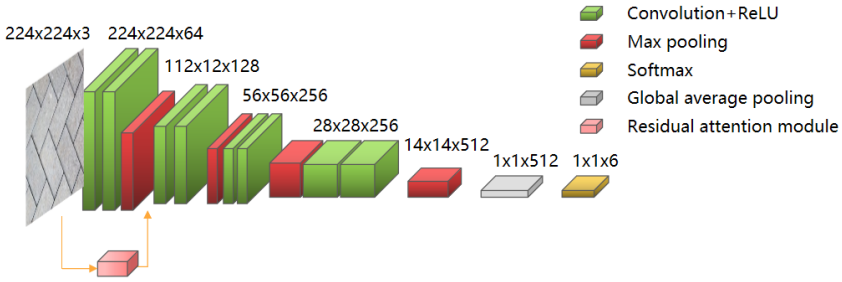

Figure 3: RA-VGG16 network architecture

2.3.1. Reduce the number of convolution codes

In this paper, the pavement dataset is constructed in the laboratory. Most of the images in the dataset are taken by the camera and reprocessed, which greatly reduces the interference of irrelevant feature information and the difficulty of model learning. Therefore, the fifth convolution layer in VGG16 network is removed and the last convolution of the third and fourth convolution layers is deleted in this paper. In this way, each group of convolution has only two convolution stacks, which can optimize the network structure, reduce the complexity of the model and improve the accuracy of pavement classification.

2.3.2. Added improved residual attention module

In this paper, the feature information extracted from multiscale convolution kernels is fused to calculate[15] the weight of the channel of the feature map, realize the assignment of the attention mechanism based on the channel domain, and finally highlight the key channel domain and weaken the non-key channel domain in the network training process. Multiscale convolution kernel is used to optimize the structure of channel attention module as shown in Figure 4.

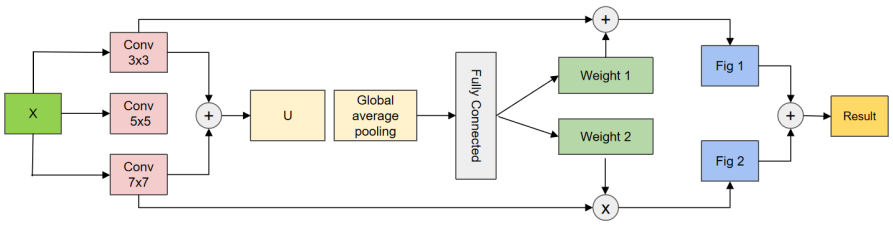

Figure 4: Multiscale Convolution Kernel Optimization Channel Attention Module Structure

The Residual Block is the most basic and core component of the residual network. The basic structure of the residual block is shown in Figure 5.

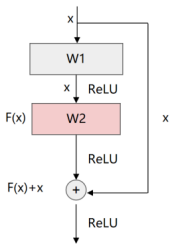

Figure 5: Residual block structures

ISSN 2522-3488 Vol. 7, Issue 3: 21-27, DOI: 10.25236/IJNDES.2023.070304

Where the input to the residue block is x, $F(x)$ is the sum and $H(x)$ is the output of the residue block, the residue block can be represented by formula (1):

$$H(x) = F(x) + x \tag{1}$$

The curve on the right side of the residue block in Figure 3 is the residue jump connection for identity mapping, x is the input data of the residue block, W_1 and W_2 are the convolution operation,σ is the nonlinear activation function ReLU, and the expression of the residue block is the formula (2):

$$F(x) = W_2\sigma(W_1x) \tag{2}$$

The residual block makes the training residual mapping $F(x)$instead of direct fitting $H(x)$ by jumping connection, and the network only needs to solve the problem of minimizing the function $F(x) = H(x) - x$. The model combines the advantages of attention mechanism and residual block. The operation principle of residual attention module is presented in figure 6.

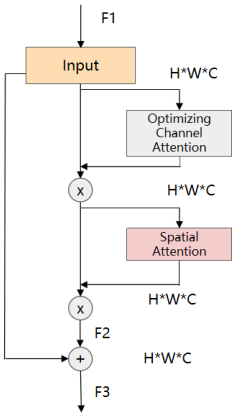

Figure 6: Improved residual attention module diagram in this article

The residual attention module integrates channel attention submodule and spatial attention submodule of multiscale convolution kernel optimization. The input F_1 is the upper output feature diagram, with two branches: one based on the residual-thought jump connection, and the other through two attention modules. Firstly, feature responses to different semantic information are obtained through the channel attention, and then the feature map F_2 of attention enhancement is obtained through the spatial features of key areas of the road surface in the spatial attention capture feature map. The addition of F_1 and F_2 gives the residual attention output feature map F_3 .

2.3.3. Removal of the entire connection layer

Although CNN network has useful application in many aspects, it has too many parameters and too much calculation, and the full connection layer is the main reason. VGG16 gets 25088 neurons by flattening the final convolution pooled feature map in one dimension, and adds two hidden layers containing 4096 neurons. As for the six road classification problems, the number of neurons is too redundant, which easily leads to the model overfitting. Therefore, the global average pooling technique (Global Average Pooling, GAP) is used to replace the detailed connection layer to compress the feature map of the network convolution output directly to the vector.

The global average pooling reduction process is shown in Figure 7. The main idea is to output N characteristic graphs through a network front-end operation, calculate an average of all pixels of each characteristic graph, i.e.n characteristic vectors.

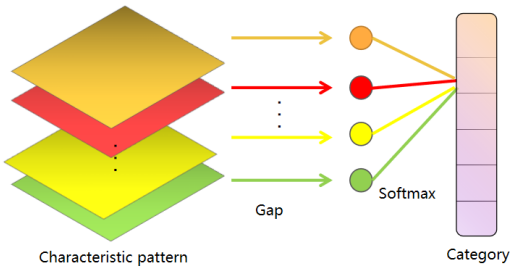

Figure 7: Global average Pooling

ISSN 2522-3488 Vol. 7, Issue 3: 21-27, DOI: 10.25236/IJNDES.2023.070304

3. Experimental results and analysis

3.1. Data sources

This dataset consists of 6 kinds of pavement types: asphalt pavement, snow pavement, gravel pavement, grassland pavement, muddy pavement and block pavement. As shown in Figure 8, part of the pavement data set is pictured. The dataset comprises of 7000, of which 6000 are training sets and 1000 are testing sets.

Figure 8: Partial Data Set Pictures

3.2. Experimental platform and parameter setting

With the Win10 operating system, PyTorch uses a dynamic computational graph mechanism that makes it easy to debug this article's process, so this article uses the PyTorch framework for experimentation and GPU for accelerated model training. Use the computer configuration as CPU ES 2670, RAM 128 GB, video card NVIDIA Tesla V100 16 GB.

The model iteration number is 200, the batch size is set to 8, the learning rate is set to 0.0002, and the cross-entropy loss function is used. The standard form is as following:

$$\text{Loss} = -\sum_{i=0}^{C-1} y_i \ln(p_i) = -\ln(p_c) \tag{3}$$

In the formula:$p = [p_0, \cdots, p_{C-1}]$indicates a probability distribution; Each element p_i represents the probability that the sample belongs to Category i;$y = [y_0, \cdots, y_{C-1}]$is the representation of the sample label.

3.3. Comparative models

Compare the accuracy of the RA-VGG16 and VGG16 base models with the Alex Net, VGG13, ResNet50, and ResNet101 networks.

KNN: The KNN algorithm is a simple but often used algorithm. The common operation process is based on the LBP image feature extraction and the use of PCA reduction in the KNN image classification.

Alex Net: One of the most epochal convolution networks, consisting of five convolution layers and three fully connected layers.

VGG13: VGG is part of the most classical convolution neural networks. VGG13 has 13 layers, including 10 convolution layers and three fully connected layers.

ResNet50: The emergence of Res Net [16] is a milestone in the evolution of convolution neural networks for more layers. The number of layers of Res Net depends on the number of residual connection used in the network, and the 50-layer Res Net is defined as Res Net 50.

ResNet101: Defined as ResNet101 when Res Net has 101 layer residual connection.

3.4. Ablation experiments

In this paper, ablation experiments are designed in order to verify that the performance of VGG16 network is improved by several improvements on road type recognition. The model after reducing the

ISSN 2522-3488 Vol. 7, Issue 3: 21-27, DOI: 10.25236/IJNDES.2023.070304

number of convolution cores is VGG16-KER. The model after adding improved residual attention module is VGG16-RESA, and the model after replacing the full connection layer with global average pooling is VGG16-GAP. The ablation accuracy curve on the test set is shown in Figure 9.

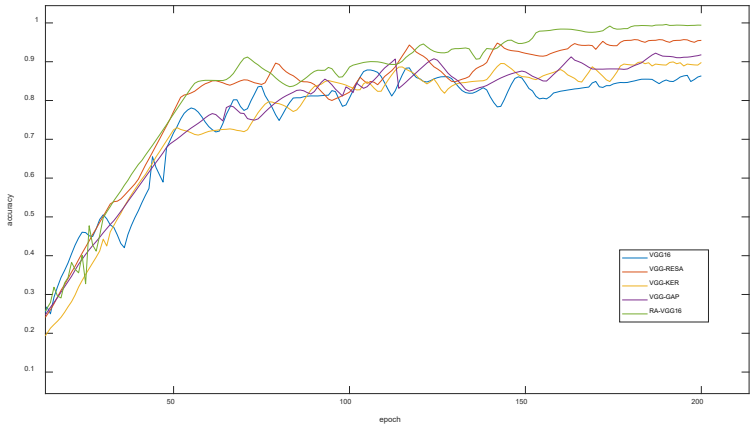

Figure 9: Accuracy curves of ablation experiments

From Figure 9, we can see that the VGG16 network accuracy curve can only be stabilized at about 84% after a zigzag increase, mainly because the model is too complex for road images. Compared with the other two improvements, the accuracy of the model is improved by 5.4% when the global average pooling technique is used to replace the full connection layer.

3.5. Comparison of different models

To verify the validity of RA-VGG16, KNN, Alex Net, VGG13, ResNet50, ResNet101 models were selected to train on the pavement dataset and the accuracy of the test set was accounted for. The training result curve and the accuracy of each model are shown in Figure 10. KNN belongs to the classification model in traditional machine learning, can only extract image edge, color, texture and other shallow features, and the final accuracy rate is only 79.08%; Alex Net model was put forward earlier, and its convolution kernel only convolved with a certain part of the feature map, the model generalization ability decreased, the final accuracy rate is 79.5%; although VGG13 and VGG16 have a slight difference in the number of convolution layers, because they have a large number of parameters in the full connection layer, the accuracy rate is not high, and is less than 85%; because there are residual modules in Res Net Resolution model and ResNet101 model, the test accuracy rate is 91.85% and 92.74% respectively, but these two network layers are deep, which is not conducive to migration and practical application, so the improved residual attention module is introduced into VGG16. Experiments show that the proposed RA-VGG16 has higher recognition accuracy and faster convergence speed, and is more suitable for road image classification.

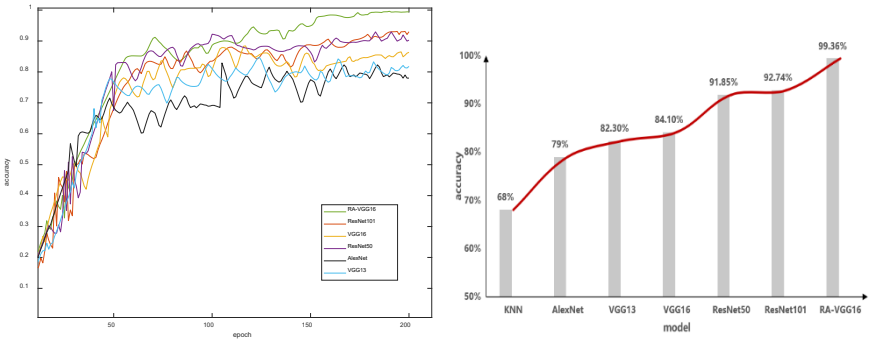

Figure 10: Accuracy of models in road test sets

As shown in Figure 11, RA-VGG16 is an obfuscation matrix for the classification accuracy of 6 types of pavement images, in which the real class of behavior is listed as a prediction class. The picture shows 6 kinds of confusion matrix, which is easy to be confused between the gravel pavement and block pavement, and the muddy pavement and ice-snow pavement.

ISSN 2522-3488 Vol. 7, Issue 3: 21-27, DOI: 10.25236/IJNDES.2023.070304

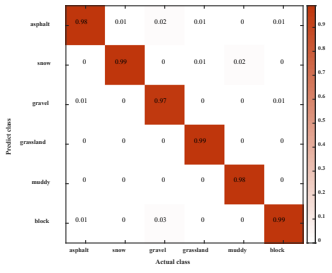

Figure 11: RA-VGG16 confusion matrix

4. Conclusion

In this paper, the improved RA-VGG16 network model based on VGG16 is applicable to the pavement image dataset established in the laboratory to realize the classification of 6 kinds of pavement. Through the ablation experiments, it is found that the reduced convolution kernel can optimize the network structure and increase the accuracy of the model. Through the combination of these three improvements, the requirements of rapid and accurate identification of road type are achieved. Compared with other models and methods, the accuracy of RA-VGG16 network is 99.36%, which is higher than other models tested, KNN (68%), Alex Net 79% VGG13 (82.30%), VGG16 (84.10%), ResNet50 (91.85%) and ResNet101 (92.74%).

References

[1] Blas M R, Agrawal M, Sundaresan A, et al. Fast color/texture segmentation for outdoor robots [C] //Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on. IEEE, 2008: 4078-4085.

[2] Liang Minjian. Key Technology of Visual Perception of Intelligent Vehicle [D]. Riving Environment. Xi'an: Chang'an University, 2017.

[3] KUEHNLE A, BUR {HOUT W. Image-based winter road conditionrecognition C}//Applications of Advanced Technologies inTransportation, ASCE, 1998.

[4] Kim T Y, Sung G Y, Lyou J. Robust terrain classification by introducing environmental sensors [C] //Safety Security and Rescue Robotics (SSRR), 2010 IEEE International Workshop on. IEEE, 2010: 1-6.

[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]. International Conference on Learning Representations, 2015.

[6] Wu Z, Chen H, Lei Y. Recognizing Non-Collaborative Radio Station Communication Behaviors Using an Ameliorated LeNet [J]. Sensors, 2020, 20(15): 4320.

[7] He M, Zhao X, Lu Y, et al. An improved AlexNet model for automated skeletal maturity assessment using hand X-ray images [J]. Future Generation Computer Systems, 2021, 121: 106-113.

[8] Qiu J, Lu X, Wang X, et al. Research on Rice Disease Identification Model Based on Migration Learning in VGG Network[C]//IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2021, 680(1): 012087.

[9] Qu Z, Mei J, Liu L, et al. Crack Detection of Concrete Pavement With Cross-Entropy Loss Function and Improved VGG16 Network Model[J]. IEEE Access, 2020, PP(99):1-1.

[10] Huang Xuehua, Chen Weihong, Yang Wangdong. Improved Algorithm Based on The Deep Integration of Googlenet and Residual Neural Network[J]. Journal of Physics: Conference Series, 2021, 1757(1).

[11] Moon J, Hossain M B, Chon K H. AR and ARMA model order selection for time-series modeling with ImageNet classification [J]. Signal Processing, 2021, 183:108026.

[12] Mao Z , Zhao C , Zheng Y , et al. Research on detection method of pavement diseases based on Unmanned Aerial Vehicle (UAV)[C]// 2020 International Conference on Image, Video Processing and Artificial Intelligence. SPIE, 2020.

[13] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[J]. 2018.

[14] Liu Z, DuJ, Wang M et al. ADCM: Attention Dropout Convolutional Module[J]. Neurocomputing, 2020, 394.

[15] Li X, Wang W, Hu X, et al. Selective Kernel Networks[J]. 2019.

[16] He K, Zhanu X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770-778.