Plant recognition method based on deep residual network and attention mechanism

Xiaomeng Lu, Ming Zhang

School of Computing, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, 212000, China

674633396@qq.com, 302548426@qq.com

Keywords: plant image recognition, residual network, Feature extraction, Attentional mechanism, Transfer learning

Abstract: This paper introduces an improved version of ResNet18network recognition model, which aims to solve the shortcomings of traditional plant recognition methods, such as the complicated steps of manual feature extraction, long time and low accuracy. The model uses the attention mechanism and transfer learning technology, and combines the SE-Net module to improve the model. Specifically, SE-Net modules are added to each residual block to enhance the weight of useful features and reduce the effect of useless features such as noise, thereby improving feature extraction capabilities and enhancing the robustness of the model. In addition, the model uses the trained parameters on the ImageNet data set to apply to the expanded plant image data set to improve the generalization ability of the model. In order to further reduce the influence of overfitting, the network structure is adjusted, and the batch standardization layer and activation function layer are placed in front of the convolution layer to enhance the regularization effect of the model. The experimental results show that the model performs well in recognition accuracy and has certain guiding significance.

1. Introduction

In recent years, due to the diversification of wild plant species in China and the insufficient public understanding of them, many plants are facing the risk of extreme endangerment, requiring effective conservation measures. Therefore, in order to improve the effectiveness of plant conservation, it is necessary to research an accurate and efficient plant identification method.

This paper proposes a plant recognition model that combines an attention mechanism with an improved residual network structure. Firstly, an attention mechanism is introduced into the residual network to enhance the weights of relevant features, thereby reducing noise in the input data. In addition, this paper improves the traditional "Conv-BN-Activation" structure in convolutional neural networks by changing it to "BN-Activation-Conv" structure. This modification reduces training time and the number of parameters, improves the model's generalization performance, enhances network regularization, and mitigates the impact of overfitting. Moreover, this paper also considers the application of transfer learning [12] to leverage pre-trained model parameters and features to enhance the accuracy of the model. Through experimental verification, our model achieves satisfactory recognition results, demonstrating the feasibility and effectiveness of the proposed method.

2. Basic network model

With the continuous advancement and complexity of deep learning models, traditional Convolutional Neural Networks (CNNs) often encounter the problem of "network degradation." This refers to a phenomenon where, as the network depth increases, the training error of the network actually increases, leading to a decline in network performance[1]. To address this issue, the concept of residual networks[13] was introduced.

Residual networks solve the network degradation problem of traditional convolutional neural networks by adding cross-layer identity links. When the network performance reaches its optimum, the residual mappings in the network become zero, leaving only the identity mappings. This ensures

  DOI: 10.25236/icceme.2023.011

that as the number of network layers increases, the network performance does not degrade, solving the network degradation problem [2]. As plant images are greatly influenced by changes in lighting, posture variations, and occlusions during recognition and classification, and also due to the many classification categories, this paper uses Resnet-18, which has stronger expressive power and feature extraction capabilities, as the base network and makes improvements to it. Residual networks are composed of multiple cascaded residual blocks, each of which consists of batch normalization layers, activation function layers, and convolution layers. In the residual block, identity mappings and residual mappings are used to construct the network [3]. Residual mapping refers to the mapping after some processing of the input, while identity mapping refers to the mapping without any processing of the input. When the residual mapping approaches zero, the network reaches its best state, and as the network depth continues to increase, the network performance will remain optimal. Figure 1 shows a schematic diagram of the structure of a residual block. The residual block contains the identity mapping (curved part in the figure) and the residual mapping (all parts outside the curve). Assuming the optimal solution required is $H(x)=x$, the residual mapping refers to the difference between mapping $H(x)$ and x, represented as $F(x)$, that is $F(x)=H(x)-x$. When $F(x)$ approaches zero infinitely, the network achieves the best state, and the optimal performance of the network will be maintained as the network depth continues to increase. When the input of the residual block is x_n, the output after calculation can be obtained as

$$x_{n+1}=f\left(x_n+F\left(x_n,W_n\right)\right), \tag{1}$$

Wherein $F(\cdot)$ is the residual mapping, W_n is the corresponding weight parameters, and $f(\cdot)$ is the activation function. As seen in Figure 1, there may be dimension mismatches between different residual blocks. In this case, a linear transformation W_n on the identity mapping x_n is needed, and we can get:

$$\mathrm{X}_{n+1}=f\left(W_s x_n+F\left(x_n,W_n\right)\right), \tag{2}$$

where W_s is the weight parameters.

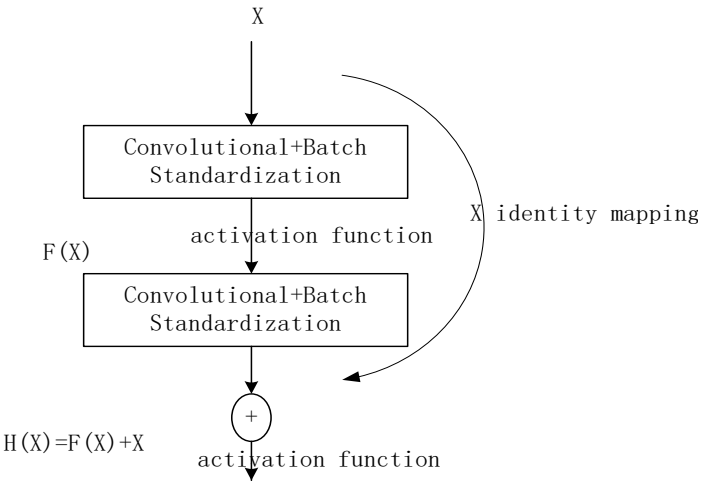


Fig.1 Schematic diagram of residual block structure

The Residual Network model can be divided into 18 layers, 34 layers, 50 layers, 101 layers, and 152 layers, etc., according to the different depths of the network. This article improves based on the ResNet-18.

3. Construction and optimization of network models

3.1 Attention mechanism

The structure of the attention module in the network is shown in Figure 2. In the SE-Net structure, the input data needs to go through squeezing and excitation [4]. Firstly, the input feature maps undergo global average pooling (GAP), extracting M W×H sized feature maps into a real number sequence of length M. This allows each feature map to utilize the context information of other feature maps, thus having a global receptive field, and at the same time, lower-level networks with smaller receptive field sizes can also utilize global information [5]. Then, the real number sequence is sent into two fully connected layers, first reducing dimensions to get a 1×1×(M/r) vector, then using the ReLU[18]activation function to increase dimensions to get a 1×1×M vector, and finally using the Sigmoid activation function to get the weight corresponding to each channel. Finally, each channel is multiplied by its corresponding weight for channel weighting, and the updated feature map is obtained [6]. Here, X represents the input feature map, and Y represents the output feature map updated by the SE-Net. The structure of the residual module after using the attention mechanism is shown in Figure 2.

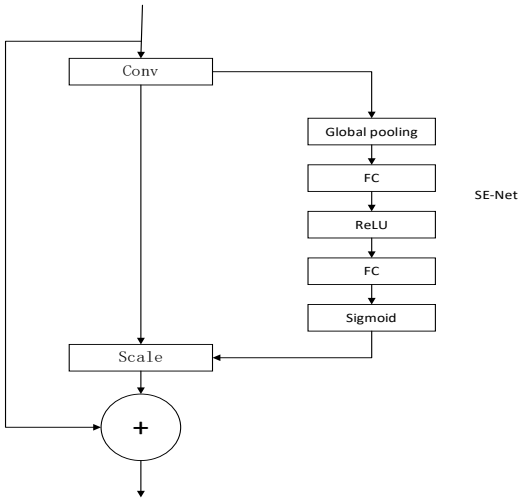


Fig. 2 Residual block diagram with added attention mechanism

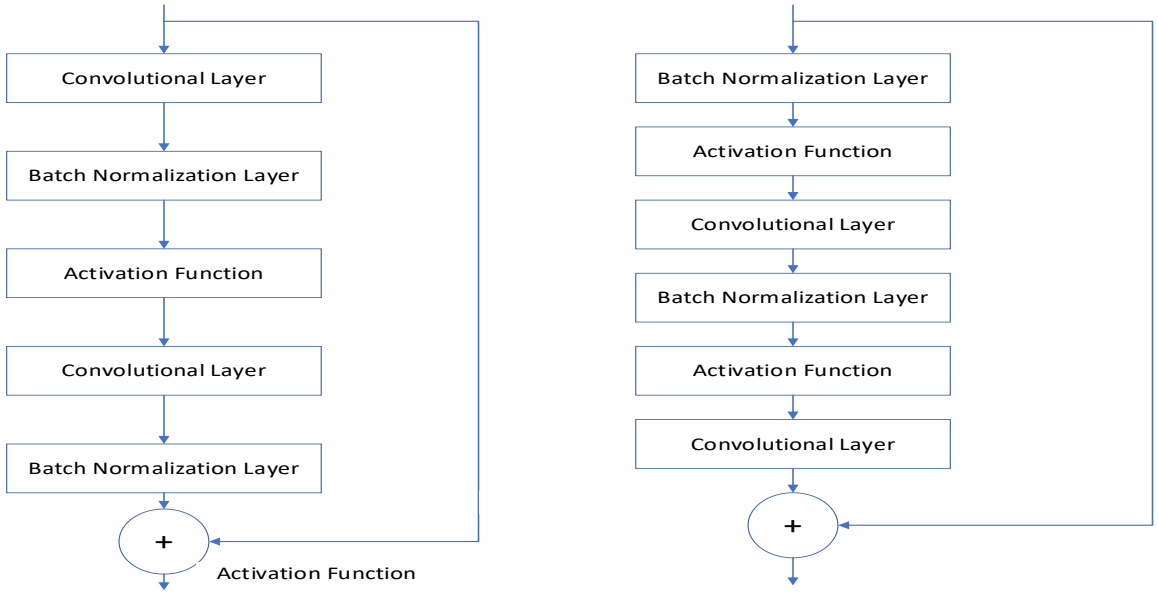
3.2 Transfer learning

The present study adopts a transfer learning approach based on the ResNet18 network. It utilizes the parameters pre-trained on the ImageNet dataset and applies them as the initial parameters for training a small-sized dataset of wild plant images. This method takes advantage of the feature parameters trained on a large-scale network and avoids the issues associated with training a model from scratch on a small dataset. As a result, it improves the classification efficiency and accuracy of the model [7].

3.3 Improvements in network architecture

The process sequence of the traditional residual network ResNet-18 consists of convolutional layers, batch normalization layers, and activation function layers. In this paper, the network structure is adjusted by placing batch normalization and activation functions before the convolutional layers [8]. The input data is first normalized through batch normalization and then activated using an activation function, followed by convolutional operations on the input data. This adjustment is commonly referred to as the "BN-Activation-Conv" structure. In comparison to the traditional convolutional neural network structure "Conv-BN-Activation," this adjustment has the following main effects: 1. Accelerating model convergence: By placing batch normalization before the convolutional layers, the input is normalized before entering the activation function, avoiding disturbances caused by the non-linear properties of the activation function [9]. This facilitates rapid model convergence. 2. Reducing model training time and parameter count: In the traditional structure,

the batch normalization layer requires additional parameters and computations, and mean and variance calculations are performed for each batch, increasing training time and model parameter count. In the "BN-Activation-Conv" structure, the batch normalization layer can be omitted since the input has already been normalized, resulting in a reduction of model parameters and computational requirements[10]. 3. Improving model generalization performance: The batch normalization layer normalizes the features, making the input distribution more stable, thereby enhancing the model's generalization performance. Placing the batch normalization layer before the convolutional layers allows for better control of the input data distribution, thus improving the model's generalization performance [11]. Additionally, experimental results show that the number of convolutional kernels can be reduced without significantly affecting the network's recognition accuracy, resulting in a lighter network and significantly improved training speed [12]. The schematic diagrams of the traditional residual network and the improved residual network structure are shown below.



(A)Traditional Staggered Network (B)Improved residual network

Fig.3 Schematic diagram of residual network structure

The adjusted network's parameters for each layer are as follows:

(1)Input Layer: Image size of 224x224.

(2)Convolutional Layer: Kernel size of 7x7, 64 convolutional kernels, stride of 2, and BN-Activation-Conv structure.

(3)Max Pooling Layer: Pooling kernel size of 3x3, stride of 2.

(4)First Residual Block: Includes two convolutional layers and a skip connection. Both convolutional layers have a kernel size of 3x3, 64 convolutional kernels, and use the BN-Activation-Conv structure.

(5)Second Residual Block: Includes two convolutional layers and a skip connection. Both convolutional layers have a kernel size of 3x3, 64 convolutional kernels, and use the BN-Activation-Conv structure.

(6)Third Residual Block: Includes two convolutional layers and a skip connection. Both convolutional layers have a kernel size of 3x3, 128 convolutional kernels, and use the BN-Activation-Conv structure.

(7)Fourth Residual Block: Includes two convolutional layers and a skip connection. Both convolutional layers have a kernel size of 3x3, 256 convolutional kernels, and use the BN-Activation-Conv structure.

(8)Average Pooling Layer: Pooling kernel size of 7x7.

(9)Fully Connected Layer: Outputs 62 categories.

(10)Softmax Layer: Normalizes the output, obtaining the probability values for each category.

Please note that the BN-Activation-Conv structure refers to the specific arrangement of batch normalization, activation function, and convolutional layers in that order [14].

3.4 Image Classification Process

After dividing the dataset, a portion of the preprocessed images is used for training the new model, while another portion is used for evaluating the learned features. The basic classification process of the improved network in handling plant images is shown in Figure 4.

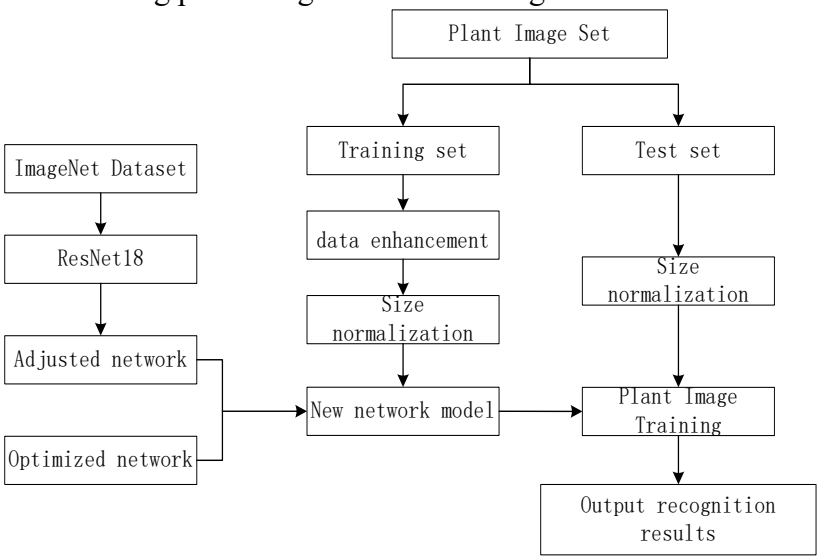


Fig.4 Flowchart of plant classification

4. Experimental Results and Analysis

4.1 Wild Plant Dataset and Its Preprocessing

The plant image dataset used in this study is sourced from the wild plant image dataset on Kaggle. It consists of 62 plant categories, totaling 6558 images, including species such as Fir, Blue Fescue Grass, and Plantain. Due to the relatively small size of the dataset, overfitting may occur, which can affect the model's recognition accuracy [15]. Therefore, we employed various data augmentation techniques, including translation, shearing, flipping, brightness adjustment, and cropping, to augment the dataset[20]. This expanded the original dataset to five times its size, resulting in a total of 32,790 plant images [16]. For experimentation purposes, we divided these images into training and testing sets in a 4:1 ratio, with the training set consisting of 26,232 images and the testing set consisting of 658 images. Finally, we normalized the processed data to images of size 224x224. This data augmentation approach provided crucial support for the effectiveness of the subsequent experiments in this paper [17].

4.2 Experimental Environment and Parameter Settings

The experiments were conducted on a Windows 10 system with an Intel(R) Core(TM) i5-10210U CPU and 8GB of RAM. The NVIDIA GeForce MX250 GPU was used for acceleration. The experiments were implemented using Python programming language, Tensorflow 2.4.0 deep learning framework, and the Jupyter Notebook development platform [18].

A comprehensive comparison and evaluation of popular neural networks were conducted, and based on the high overall score, ResNet18 was selected as the base network model for improvement and training [19].

4.3 Experimental Results and Analysis

4.3.1 The impact of improved strategies on experimental results.

(1) Self-comparison experiment of ResNet18 network.

The objective of this study was to investigate the impact of fine-tuning models on plant image recognition. To achieve this, we trained the original ResNet18 model, the fine-tuned ResNet18 model, the ResNet18 model with added attention mechanism, and the model with added attention mechanism and adjusted network. We used the cross-entropy loss function for training and evaluated the model performance based on the maximum test accuracy after 150 iterations. Table 1 presents the comparison of model accuracy and loss [20].

Tab.1 Improved model training results comparision

model	precision /%	loss
ResNet18	78.6	1.4
ResNet18+attention mechanism	84.2	1.2
Resnet18 for improving the network	79.4	0.5
Improving the Resnet18+attention mechanism of the network	86.6	0.4

From Table 1, it can be seen that both improvement schemes have an impact on the accuracy and loss results of the test set. Compared to the original ResNet18 model, the model's accuracy is improved by 5.6% with the addition of attention mechanism and by 0.8% with network adjustment. Additionally, the model's loss is reduced in both cases. The main reasons are as follows: Firstly, the attention mechanism enhances the weights of useful feature information and weakens the weights of irrelevant information, thereby improving the expressive power of the network model and reducing the influence of noise and other interfering factors on the model's recognition performance, thus enhancing the model's classification ability . Secondly, in the BN-Activation-Conv structure, since both batch normalization and activation function are applied before the convolutional layer, the gradients can be naturally propagated, which helps improve the training efficiency and performance of the model. Furthermore, the improved network strengthens the regularization of the model, reducing the impact of overfitting and increasing the model's expressive power, thus improving the accuracy of recognition.Figure 5 shows that, at the same number of iterations, the improved network structure has already reached stability, while the unimproved network structure is still converging. This indicates that the improved network structure achieves faster convergence and can reach better training results more quickly. The above conclusions demonstrate the feasibility of the optimization methods proposed in this paper for plant recognition problems. The final improved model in this study achieved a testing accuracy of 86.6%, which is an 8.0% improvement over the original model. The loss was reduced by 1.0, indicating good recognition and generalization capabilities, making it suitable for classifying images of wild plants .

(2) Comparative experiment with other network models

To verify the performance of the improved network architecture, the popular networks VGGNet16 and VGGNet19 were selected as control groups. After 150 iterations, the training results are shown in the table 2.

Tab.2 Comparision of training results of different models

model	precision /%	loss
VGGNet16	70.0	1.5
VGGNet19	67.7	1.5
ResNet18	78.6	1.4
Improved Resnet18+attention mechanism for the network	86.6	0.8

As can be seen from Table 2, the improved ResNet18 network converges faster and achieves a higher accuracy after stabilization compared to the control group network, demonstrating the effectiveness of this network. The specific testing accuracy of the model as it varies with the number of iterations is shown in Figure 5.

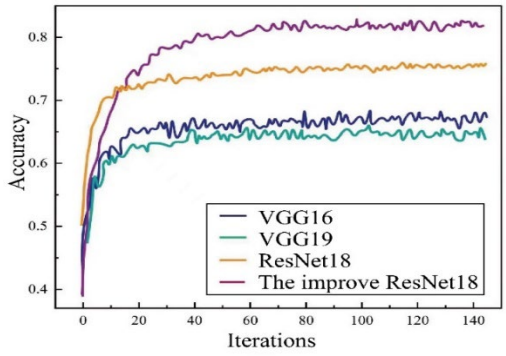


Fig.5 Comparison curves of different network test accuracy

References

[1] Zhang, Y., Du, H., Jin, X., et al. Diversity and Geographic Distribution of Wild Orchid Species in China. Chinese Science Bulletin, 2015, 60(2): 179-188.

[2] Ma, Y., Li, Y. The Theory of Virtual Value and Analysis of Modernity. Fudan Journal (Social Sciences Edition), 2012(1): 103-110.

[3] Zhang, K., Su, D., Wang, P., et al. Progress and Application of Deep Learning Technology in Image Dense Matching. Science Technology and Engineering, 2020, 20(30): 12268-12275.

[4] He, L., Shao, Z., Zhang, J., et al. A Review of Behavior Recognition Algorithms Based on Deep Learning. Computer Science, 2020, S1: 139-147.

[5] GUYERDE, MILESGE, GAULTNEYLD, et al.Application of machine to shape analysis in leaf and plant identification[J].Transactions of the Asae, 1993, 36(1):163-171.

[6] Du, J., Zhai, C. Plant Image Recognition Method Based on Gabor Texture Feature [C]//Proceedings of the 14th National Conference on Image and Graphics. Beijing: China Society for Image and Graphics, 2008: 246-250.

[7] Zhang, N. Research on Plant Leaf Recognition Algorithm Based on Image Analysis [D]. Beijing: School of Information, Beijing Forestry University, 2013.

[8] Li, P., Zhang, B., Zhang, S. Plant Identification Method Based on Leaf Image Processing and Sparse Representation. Jiangsu Agricultural Sciences, 2016, 44(9): 364-367.

[9] Wen, C., Lou, Y., Zhang, X., et al. Plant Recognition Method Based on Improved Dense Capsule Network Model. Journal of Agricultural Engineering, 2020, 36(8): 143-155.

[10] Cao, X., Sun, W., Zhu, Y., et al. Plant Image Recognition Based on Science Prioritization Strategy. Journal of Computer Applications, 2018, 38(11): 3241-3245.

[11] Yu, H., Ma, J., Zhang, Y. Plant Leaf Recognition Model Based on Dual-path Convolutional Neural Network. Journal of Beijing Forestry University, 2018, 40(12): 132-137.

[12] PAN S J, QIANG Y.A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10) : 1345-1359.

[13] HE K,ZHANG X,REN S,et al. Deep Residual Learning for Image Recognition[C]//Computer Vision and Pattern Recognition.Piscataway,NJ:IEEE,2016:770-778.

[14] HU J, SHEN L, SUN G.Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:7132-77141.

[15] ZHU H Y, XIE C, FEI Y Q, et al.Attention mechanisms in cnn-based single image super-resolution :a brief review and a new perspective[J].Electronics, 2021, 10(10):1187--11188.

[16] Liu, X., Li, Y., Liu, L., et al. Improved YOLOV3 Target Recognition Algorithm Embedded with

SE-Net Structure. Computer Engineering, 2019, 45(11): 243-2248.

[17] LIN M, CHEN Q, YAN S.Network In Network[J].arXiv preprint arXiv: 1312.4400, 2014.

[18] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2):1097-1105.

[19] JIA D, WEI D, SOCHE R R, et al.ImageNet: A Large-scale Hierarchical Image Database[C]//IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248-255 .

[20] Jiang, Y., Zhang, H., Chen, L., et al. Image Data Augmentation Algorithm Based on Convolutional Neural Network. Computer Engineering and Science, 2019, 41(11): 10.