Fruit Image Recognition with Enhanced CBAM Attention Mechanism Based on Resnet18

Wenyan Huang *

Department of Computer Science and Technology, Beijing Jiaotong University, Weihai, 264209, China

* Corresponding Author Email: 23722046@bjtu.edu.cn

Abstract. In the field of computer vision today, the accurate recognition of fruit images is of great significance for practical applications such as intelligent agriculture and food quality inspection. Traditional image recognition models encounter a precision bottleneck when processing fruit images, primarily due to the diversity and complexity of fruit features, particularly when different fruit varieties exhibit similar appearances. This paper focuses on improving the performance of fruit image recognition models and proposes an innovative method based on the ResNet18 architecture, which integrates multiple pooling and spatial attention mechanisms. By constructing a dynamic pooling fusion module, the model can adaptively learn the weights of different pooling methods, thereby effectively alleviating feature conflicts among multiple pooling methods. At the same time, the introduction of a channel-spatial attention dynamic balance mechanism optimises the allocation of attention to features in different dimensions, enhancing the pertinence and effectiveness of feature extraction. Experimental results show that, compared to the original ResNet18 model, the average accuracy of the improved model on the fruit30 dataset has increased significantly from 73.933% to 75.108%. The highest accuracy on the fruit360 dataset has also increased, fully demonstrating the excellent effect of this method in enhancing fruit feature recognition ability and improving the model's generalisation performance. It lays a solid foundation for the broad application of fruit image recognition technology in practical scenarios.

Keywords: Fruit image recognition; ResNet18; Attention mechanism.

1. Introduction

Fruit image classification, as a fundamental technology in agricultural intelligence, has extensive applications in various areas, including automatic sorting, quality inspection, and yield assessment. However, traditional manual recognition methods have several drawbacks, including low efficiency, high costs, intense subjectivity, and a high degree of dependence on personnel's technical expertise [1,2]. This not only leads to resource waste and increased operating costs but also limits the promotion of sorting systems in large-scale production. Especially in commercial applications, insufficient classification speed and accuracy will directly affect the efficiency of the supply chain and economic benefits.

With the development of deep learning, Convolutional Neural Networks (CNNs) have shown significant advantages in image classification. Among them, ResNet18, with its lightweight structure and balanced performance, is widely used in small and medium-sized image recognition tasks [3]. Thanks to the design of residual connections, ResNet18 can not only effectively alleviate the problem of gradient disappearance in deep networks but also achieve high recognition accuracy with low computational cost. However, in the face of practical scenarios with diverse fruit types, similar morphologies, and complex lighting conditions, relying solely on the basic model still has issues such as insufficient feature extraction and inadequate attention to key features, thus affecting the accuracy and robustness of classification.

To address the abovementioned issues, this study introduces an optimised attention mechanism based on ResNet18 and proposes an Enhanced Convolutional Block Attention Module (ECBAM). This module combines channel and spatial dual-attention mechanisms. It introduces four pooling methods (adaptive average pooling, adaptive max pooling, global average pooling, and global max pooling) to capture multi-level features of images. The model can select the most suitable pooling

method by learning the weights of different poolings, achieving dynamic pooling and strengthening the most important features of the image. At the same time, the residual branch is used to retain the original features, thereby avoiding the loss of basic features that can occur when attention is over-focused. Additionally, the coordinate enhancement mechanism is employed to enhance spatial feature perception, thereby making the model more stable in handling fine-grained differences and complex backgrounds.

The significance of this method lies in that it can significantly improve the classification accuracy while maintaining a lightweight structure, meet the deployment requirements of edge devices, and provide the potential to be transferred to other agricultural visual tasks (such as pest and disease detection). Compared to traditional models, ECBAM not only offers noticeable improvements in the depth and breadth of feature extraction but also maintains high performance in small datasets and low-computing-resource environments.

From the perspective of the development process, fruit image classification technology has evolved from manual feature extraction and shallow classifiers to automatic feature learning in deep convolutional networks, and then to the introduction of attention mechanisms to enhance the focusing ability of models. In recent years, research trends have primarily focused on the parallel optimisation of lightweight and high-precision models, offering more feasible solutions for agricultural intelligence.

This paper introduces the structure and key module design of the proposed model, analyses its experimental results on the Fruit30 and Fruit360 datasets, and verifies its performance improvement and application value through comparative experiments and ablation experiments, providing a reference for improving subsequent agricultural visual recognition models.

2. Model Structure and Key Module Design

2.1. Model Basic Network Architecture

This study adopts the ResNet18 network structure as the basic architecture. As can be seen from the data in Table 1, compared with models such as AlexNet, VGG16, ResNet50, and MobileNetV2, the main advantage of ResNet18 in fruit image recognition tasks is its balanced performance [3-6]. It achieves relatively high recognition accuracy while having a small number of parameters and a low computational cost. It is easy to train and flexible to deploy. It performs excellently in recognising common fruits. It can not only effectively learn image features through residual connections but also avoid the problems of overfitting and high computational cost in deep networks, thus achieving a high cost-performance ratio.

Table 1. Comparison of Fruit Image Recognition Performance Across Network Models [7]

Network model	Parameter quantity	Computational volume(FLOPs)	Typical accuracy (Fruit-360)
AlexNet	60M	0.7G	85%-88%
VGG16	138M	15.3G	92%-94%
ResNet18	11M	1.8G	95%-97%
ResNet50	25M	4.1G	96%-98%
MobileNetV2	3.4M	0.3G	93%-95%

2.2. Model Training Parameters

This model uses the Adam optimiser with an initial learning rate of 0.001 and a batch size of 128[8]. The cross-entropy loss function is used, with a label smoothing coefficient of 0.1. The model is trained for 300 epochs on the Fruit30 dataset and 10 epochs on the Fruit dataset [9, 10]. Random horizontal flipping is used to enhance the robustness of the data. The mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225] based on the ImageNet dataset are used as normalisation parameters [11]. The training is carried out using an RTX 4090 GPU. The training and testing are based on the

fruit30_split and fruit360 datasets. The baseline algorithm and Fruit30 dataset are sourced from open-source content on GitHub [8]. The fruit30_split dataset contains a total of 5000 images of 30 different common fruits, and the fruit360 dataset contains approximately 80,000 images of 131 fruit categories. The images are preprocessed to adjust the resolution to 224×224. The photos are pre-divided into training and test sets in a 4:1 ratio. The scale of these datasets is moderate, the training time cost is acceptable, and the trained model can achieve relatively high accuracy. In terms of fruit categories, the fruit30 dataset contains 5-6 similar fruits, while the fruit360 dataset contains more similar categories, which is conducive to testing the model's ability to recognise similar fruits.

2.3. Model Optimisation Parameters

The enhanced attention module is improved in two attention dimensions: channel and space. In the channel attention part, the original residual block is replaced with a residual block with an improved attention module. The model selects the pooling method with the most significant weight among four pooling methods: AdaptiveAvgPool2d, AdaptiveMaxPool2d, global max pooling (gmp_pool), and global average pooling (gap_pool). Compared with the single pooling method in traditional channel attention, this model can select the pooling method that best reflects the image features by learning the weights of the four poolings.

In the spatial attention part, the features are subjected to average and max pooling in the channel dimension and then concatenated. Convolutional kernels of different sizes are used to extract multi-scale features. When coordinate enhancement is enabled, additional horizontal and vertical coordinate information is added, and the number of channels is adjusted accordingly. The model autonomously learns to perform dynamic weight allocation and activation.

The overall network comprises four residual groups, each containing two residual blocks, thereby maintaining the original ResNet18 network's overall structure to adapt to the feature extraction requirements of fruit images.

As shown in Figure 1, the enhanced attention module is inserted after performing convolution and normalisation operations on the residual blocks. The parameters of the module are dynamically adjusted according to the number of feature channels. When the number of channels is less than 128, a feature compression ratio of 16 times is adopted, and two convolutional kernels of 3×3 and 5×5 are used. When the number of channels exceeds 128, the compression ratio is increased to 32 times, and three convolutional kernels of sizes 3×3, 5×5, and 7×7 are used. At the same time, the coordinated information enhancement function is enabled.

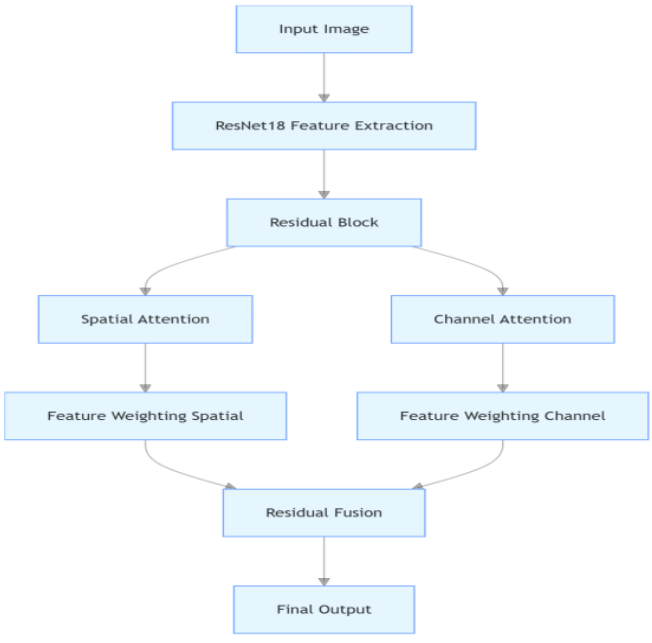

Fig. 1 Flowchart of the Optimised Fruit Recognition Model with Enhanced Attention Mechanism

3. Experimental Design

This study designed a comparative experiment using the original ResNet18 model and an ablation experiment with a model that only utilises a simple attention module with a single pooling method. The experiments aim to verify the performance improvement of the model with the enhanced attention module under the same number of training epochs.

The experiments were carried out using a computing instance provided by the cloud GPU platform Featurize, which is equipped with an RTX 4090 graphics card (24GB VRAM). The instance has 16 × AMD EPYC 9354 multi-core CPUs (central frequency 2.3 GHz), 16GB of memory, and 700GB of disk storage space. The model was developed and trained on a Linux system using the Python ecosystem, with Python version 3.11.8 and TensorFlow version 2.16.1. The core framework of the neural network is PyTorch v2.2.2, and its automatic differentiation and distributed training features are conducive to improving the model iteration speed. Orchvision v26.1.0 was used to load and preprocess the image data. Evaluation metrics, such as model accuracy and F1 score, were calculated using scikit-learn, and the experimental results were visualised using seaborn and matplotlib. All experiments adopted random initialisation.

In the comparative experiment, the original ResNet18 model and this model (ECBAM) were trained for 300 epochs and then tested on the test set. The recognition accuracy, recall rate, F1 score, and information of misclassified images of each fruit category were recorded, and tables and confusion matrices were drawn. The F1 score is the harmonic mean of the recognition accuracy and recall rate, providing a more comprehensive reflection of the model's performance.

In the ablation experiment, the model, which used only one pooling method (global average pooling), was trained for 300 epochs and then tested on the test set. The recognition accuracy, recall rate, F1 score, and information of misclassified images of each fruit category were recorded, and tables and confusion matrices were drawn.

To verify the optimisation of the model, this study trained and tested it using the fruit360 dataset, and statistically analysed the highest accuracy achieved by the model compared to the original ResNet18 model.

4. Experimental Results

Table 2. Accuracy Comparison of Original ResNet18 and Optimised Model

Network model	The first test accuracy	The second test accuracy	The third test accuracy	Average accuracy
Original ResNet18 model	72.820%	75.974%	73.006%	73.933%
This model	75.603%	76.160%	73.562%	75.108%

Table 3. Accuracy Comparison of Single-Channel Attention Model and Optimised Model

Network model	The first test accuracy	The second test accuracy	The third test accuracy	The fourth test accuracy	The fifth test accuracy	The sixth test accuracy	Average accuracy
A model of single-channel attention	74.304%	73.933%	76.252%	72.820%	75.881%	76.438%	74.938%
This model	75.325%	78.749%	74.675%	75.046%	76.160%	76.438%	76.066%

The results of the ablation experiment are shown in Table 3. Each model was tested 6 times. The average accuracy of the single-pooling model is 74.938%, and the average accuracy of this model is 76.066%.

The experiments obtained result tables and confusion matrices for each test. The following is a representative set of results.

The test accuracy of this model on the test set after 300 epochs of training is shown in Table 2. The average test accuracy of the original ResNet18 model is 73.933%, and the accuracies of the three

tests are 72.820%, 75.974%, and 73.006% respectively. The average test accuracy of this model is 75.108%, and the accuracies of the three tests are 75.603%, 76.160%, and 73.562% respectively.

5. Experimental Analysis

Table 4. Model Deployment and Inference Statistics

Item	Details
Device Used	cuda
Model Weight Loading Status	Successfully loaded (full model mode)
Test Set Info	1078 samples, 30 categories
Output Dimension Validation	Matches dataset (30 classes)
Accuracy	76.438%
Avg. Single-batch Inference Time	0.0422s
Avg. Single-image Inference Time	0.001330s
Images Processed per Second	751.82

Table 5. Best Test Accuracy of Original ResNet18 and Optimised Model

Network model	Best test accuracy
Original ResNet18 model	72.820%
This model	75.603%

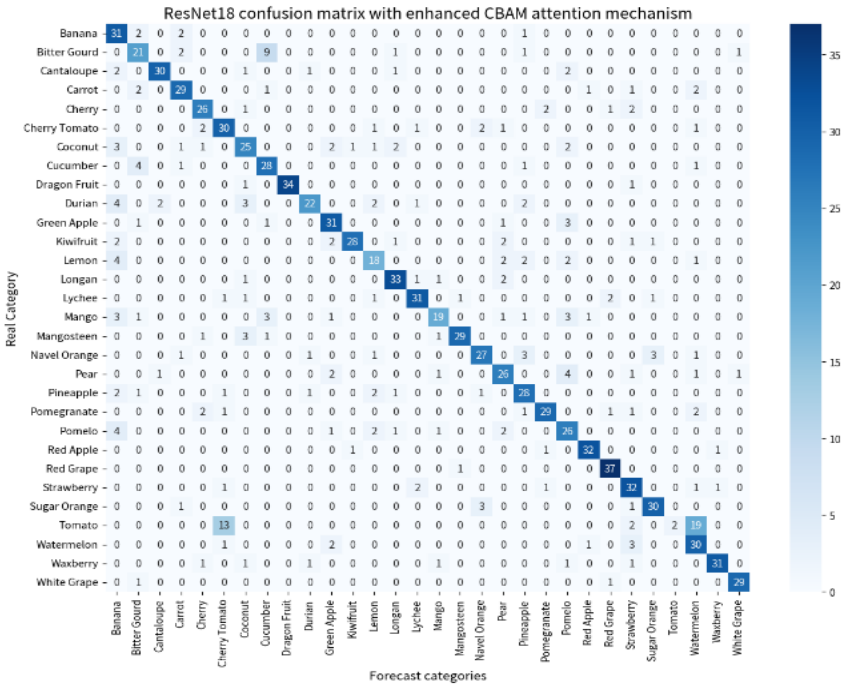

Fig. 2 ResNet18 confusion matrix with enhanced CBAM attention mechanism

The data in Table 3 shows that, after 300 epochs of training, the average test accuracy of this model on the test set is approximately 1% higher than that of the original model.

Table 4 reflects the test results of this model on the test set. The average accuracy in recognising 30 categories of fruits reaches 75%, indicating that the model is suitable for general fruit classification scenarios. At the same time, compared to the model using a single pooling method (adaptive average pooling), the average accuracy of this model increases by about 1%, indicating that multi-pooling has a stronger ability to extract features. Table 4 records the processing speed of this model. Table 5 shows that this model can achieve a higher accuracy limit on large datasets such as fruit360, improving the recognition ability of existing models. The confusion matrix in Figure 2 further verifies

the characteristics of this model in the test: it achieves higher recognition accuracy for categories with distinct features, unique shapes, and colours. This further confirms that the enhanced attention module can strengthen the learning of key features.

6. Conclusion

This model demonstrates higher accuracy and innovatively integrates the design of an attention mechanism and residual fusion. By learning weights to select the most suitable pooling method among the four pooling methods, it is beneficial to compress channel information from different angles. It uses three types of convolution kernels for multi-scale convolution and implements coordinate enhancement to capture the spatial differences of fruits at a finer granularity. The model independently learns the weights of the pooling results through a Multi-Layer Perceptron (MLP), and the learned parameters enable dynamic weight learning, allowing the model to capture image features more accurately. This essentially solves the problems of insufficient feature learning and feature loss. At the same time, a residual branch is introduced into the attention module, and the dual-path structure of attention weighting and residual compensation addresses the issue of basic feature loss caused by the attention mechanism's overemphasis on key features.

Even with small datasets with low training costs, this model still shows excellent performance and improves the accuracy of fruit image recognition tasks. Moreover, it shows extremely high accuracy on larger datasets. This method has achieved specific results in fruit image recognition tasks using lightweight models, but it still faces some challenges. The model has a relatively limited ability to recognise highly similar samples, such as tomatoes and cherry tomatoes, which is a limitation inherent to the model's depth. Additionally, the recognition accuracy of this model decreases when processing images with multiple occlusions. The model will be optimised from the following aspects:

To address the issue of similar samples being easily confused, the attention mechanism can be optimised, and a fine-grained feature enhancement module can be introduced to focus on fine-grained differences beyond colour, thereby enhancing the ability to recognise similar samples. To improve the upper limit of the model's accuracy, in terms of dataset selection, samples with extreme lighting and occlusions can be added to enhance robustness. In commercial applications, model quantisation technology can be combined to conduct FPGA deployment experiments, test the inference efficiency of edge devices, and explore commercial value.

References

[1] Jia Weikuan, Tian Yuyu, Luo Rong, Zhang Zhonghua, Lian Jian, Zheng Yuanjie. Detection and segmentation of overlapped fruits using an optimised mask R-CNN application in an apple harvesting robot. Computers and Electronics in Agriculture, 2020, 172: 105380.

[2] Gill Harmandeep Singh, Khalaf Osamah Ibrahim, Alotaibi Youseef, Alghamdi Saleh, Alassery Fawaz. Fruit image classification using deep learning. Computers, Materials & Continua, 2022, 71(3): 4599–4613.

[3] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770–778.

[4] Krizhevsky Alex, Sutskever Ilya, Hinton Geoffrey E. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (NeurIPS), 2012, 25: 1097–1105.

[5] Simonyan Karen, Zisserman Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[6] Sandler Mark, Howard Andrew, Zhu Menglong, Zhmoginov Andrey, Chen Liang-Chieh. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 4510–4520.

[7] Liu Kai. Comparison of different convolutional neural network models on Fruit 360 dataset. Highlights in Science Engineering and Technology, 2023, 34: 85–94.

[8] Kingma Diederik P, Ba Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[9] TommyZihao. Train_Custom_Dataset: A repository for dataset annotation, training, evaluation, testing, and deployment of AI algorithms. GitHub, 2024.

[10] Muresan Horea, Oltean Mihai. Fruit recognition from images using deep learning. Acta Universitatis Sapientiae, Informatica, 2018, 10(1): 26–42.

[11] Deng Jia, Dong Wei, Socher Richard, Li Li-Jia, Li Kai, Fei-Fei Li. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248–255.