Industry Science and Engineering Vol. 2 No. 2, 2025

A Study on an Improved Inception Architecture Integrating SE Attention Mechanism for Image Classification

Zihao Yan, Yubo Liu, Yang Zhang, Zhengxin Liu*, Jingjing Hou, Cewu Hang
Xi'an Mingde Institute of Technology, Xi'an, Shaanxi, China
**Corresponding Author*

Abstract: Image classification, as one of the core tasks in computer vision, plays a vital role in many real-world applications such as medical diagnosis, autonomous driving, security surveillance, and industrial inspection. With the development of deep neural networks, the Inception module has become a widely adopted structural unit in image classification networks due to its multi-scale convolutional branches, which offer strong feature extraction capabilities and high computational efficiency. However, traditional Inception structures often suffer from insufficient focus on discriminative regions and redundant features when dealing with complex images, limiting further improvements in classification performance. To address these issues, this paper proposes an improved Inception architecture that integrates the Squeeze-and-Excitation (SE) channel attention mechanism. Based on this enhanced module, a lightweight image classification network named FruitNet is designed. By incorporating SE modules after each level of the Inception modules, the network dynamically recalibrates key channel features, thereby enhancing the model's discriminative power and robustness. To validate the effectiveness of the proposed approach, extensive experiments were conducted on the standard image classification dataset CIFAR-10. The results show that FruitNet significantly improves classification accuracy while maintaining low computational complexity, achieving a well-balanced trade-off between lightweight model design and feature representation capability. This study provides a novel structural perspective and technical support for the design and deployment of efficient image classification models.

Keyword: Image Classification; Inception Module; Squeeze-and-Excitation (SE); Mechanism Lightweight Network; FruitNet

1. Introduction

With the rapid development of artificial intelligence, computer vision, as a significant branch of AI, has demonstrated immense potential across various domains. Among its foundational tasks, image classification remains a key focus and active research area. Traditional image classification methods largely rely on handcrafted features combined with shallow learning models for decision-making [1]. However, such approaches often struggle with complex image backgrounds, diverse object morphologies, and high-dimensional feature spaces.

In recent years, Convolutional Neural Networks (CNNs), with their end-to-end learning capability and powerful feature representation abilities, have emerged as the mainstream approach for image classification [2]. Among numerous CNN architectures, Google's Inception series (e.g., GoogLeNet, Inception-V3) has achieved exceptional performance in various image recognition tasks. This success is attributed to their unique multi-scale convolutional branch structures, which enhance model representation capacity while significantly reducing parameter count and computational overhead [3]. By utilizing parallel operations with convolutional kernels of different sizes, the Inception module jointly models local details and global structures.

However, despite its strong multi-scale feature extraction capabilities, the Inception structure exhibits limited modeling capacity across feature channels. It fails to fully capture dependencies among channels, which can weaken critical features and allow redundant information to persist. To address this limitation, Hu et al. proposed the Squeeze-and-Excitation (SE) attention mechanism. By dynamically weighting feature channels, the

SE mechanism enables models to adaptively adjust channel responses based on input images, substantially enhancing model expressiveness and classification accuracy. Combining the SE mechanism with efficient network architectures has become a crucial direction for improving lightweight model performance.

In light of these advancements, this paper introduces an improved Inception module that integrates the SE channel attention mechanism into the traditional multi-scale feature extraction structure. This combination forms the foundation of FruitNet, an image classification network that balances efficient feature extraction with strengthened inter-channel dependency modeling. The network retains its lightweight advantage while incorporating SE modules after each level of the Inception module to dynamically enhance critical features, thereby improving overall classification performance.

Extensive experiments were conducted on the standard image classification dataset CIFAR-10 to validate the effectiveness of the proposed method. Comparative analyses with various classical models demonstrated that FruitNet significantly improves classification accuracy while maintaining low computational complexity. These results validate the feasibility and practicality of the proposed architecture in lightweight model design.

The research presented in this paper aims to explore a lightweight design approach for image classification networks by integrating attention mechanisms and multi-scale feature extraction structures. It provides theoretical support and model insights for subsequent intelligent vision applications in resource-constrained environments such as mobile and edge computing devices.

2. Research Foundation

2.1 Overview of the Development of Image Classification Techniques

The task of image classification aims to assign input images to predefined categories based on their visual features. It is one of the most fundamental and widely applied tasks in the field of computer vision. Before the advent of deep learning, image classification relied on manually designed feature extraction algorithms such as SIFT, HOG, and LBP. These features were classified using traditional machine learning methods such as Support Vector Machines (SVM) or k-Nearest Neighbors (k-NN). Although these methods achieved moderate success in specific tasks, they generally struggled with limited understanding of image content, weak generalization capabilities, and redundant feature dimensions.

Since AlexNet achieved a breakthrough performance in the 2012 ImageNet competition, convolutional neural network (CNN)-based deep learning methods have rapidly dominated the field of image classification. Subsequent architectures such as VGGNet [4], ResNet [5], DenseNet [6] and MobileNet [7] have not only continuously improved classification accuracy but also made significant advancements in areas such as model architecture design, depth expansion, and parameter optimization. Among these, Google's Inception series networks stand out as a prominent example of lightweight and efficient network design. With their efficient multi-scale convolutional structures, these networks achieve high model accuracy while significantly reducing parameter count and computational costs.

2.2 Inception Structure and Multi-Scale Feature Extraction

The core idea of the Inception module is to apply convolutional kernels of various sizes (e.g., 1×1, 3×3, 5×5) in parallel to the same input feature map, enabling the capture of features at multiple scales. The early Inception v1 architecture introduced 1×1 convolutions to compress input channels, thereby reducing the computational complexity of subsequent 3×3 and 5×5 convolutions. This design effectively controlled the number of parameters and computational cost while maintaining the model's expressive power. Later versions, such as Inception v3 and v4, further optimized the structure by incorporating such techniques as factorized convolutions and batch normalization, improving feature extraction efficiency and classification performance.

Although the Inception architecture excels in multi-scale modeling, its feature fusion process is relatively static, lacking interaction and weight adjustment among branches. This limitation restricts the model's ability to allocate importance across the channel

dimensions effectively, leading to unmitigated redundant information and insufficient emphasis on critical features.

2.3 SE Attention Mechanism and Feature Channel Modeling

The Squeeze-and-Excitation (SE) module, proposed by Hu et al. in 2017 [8], is designed to enhance a model's sensitivity to channel-wise information. Its core idea involves a two-step process: "Squeeze" and "Excitation." In the squeeze step, global average pooling is applied to the feature maps to obtain a global representation for each channel. In the excitation step, a bottleneck structure composed of fully connected layers learns the importance weights of each channel. Finally, through a process known as channel recalibration, the original feature maps are weighted and adjusted to emphasize informative features while suppressing ineffective or redundant ones.

The SE module is characterized by its simple structure, low computational overhead, and strong generalization capability. It can be seamlessly integrated into various mainstream CNN architectures. Extensive experiments have shown that incorporating SE modules consistently improves performance across multiple image recognition tasks. In particular, for architectures with limited channel-wise modeling capabilities—such as Inception—the integration of the SE mechanism can significantly enhance the model's ability to focus on relevant features.

2.4 Lightweight Neural Network Design Trends

With the widespread adoption of artificial intelligence in mobile and embedded devices, lightweight model design has become a major research focus. The development of model architectures with reduced parameters, accelerated computation, and lower resource consumption—while still maintaining high performance—is a critical requirement for practical deployment. To meet this need, researchers have proposed lightweight networks such as MobileNet , ShuffleNet, and EfficientNet, which typically employ such techniques as depthwise separable convolutions, group convolutions, and channel pruning to significantly reduce model complexity.

In addition, combining attention mechanisms with lightweight architectures has emerged as a new trend for further improving model performance. For example, MobileNetV3 integrates the SE module, highlighting the effectiveness of attention mechanisms even in lightweight networks. Striking a balance between performance and complexity within lightweight structures has thus become a key direction in the design of modern image classification models.

3. Method

3.1 Overview of the FruitNet Architecture

FruitNet is a lightweight network designed to enhance both accuracy and efficiency in image classification. The network is built upon an improved Inception module and integrates the Squeeze-and-Excitation (SE) attention mechanism. By stacking multiple Inception-SE modules, FruitNet achieves an effective combination of multi-scale feature extraction and channel-wise attention, significantly improving the model's ability to capture key features while maintaining low computational complexity. With its simple and efficient architecture, FruitNet is well-suited for deployment in resource-constrained environments and offers strong practical advantages.

The overall architecture of FruitNet consists of the following key components:

Initial Convolution Layer: Extracts low-level features from the input image through convolution operations.

Multiple Improved Inception-SE Modules: These modules extract features at different scales via parallel multi-scale convolution paths, and apply SE modules to recalibrate channel responses, highlighting important features and suppressing redundant information.

Pooling Layer: Reduces the spatial dimensions of the feature maps via max pooling.

Global Average Pooling (GAP) Layer: Compresses the spatial dimensions of high-dimensional feature maps to 1×1, providing the final input for the classification layer.

Fully Connected Layer: Performs the final classification based on the processed features and outputs the predicted class label.

3.2 Improved Inception Module Design

The Inception module itself enhances the network's ability to capture features at multiple scales by employing parallel convolutional branches. Building on this foundation, the improved Inception module proposed in this paper introduces targeted optimizations to further strengthen feature extraction and channel-wise discrimination. The structural design of the improved module is illustrated in Figure 1.

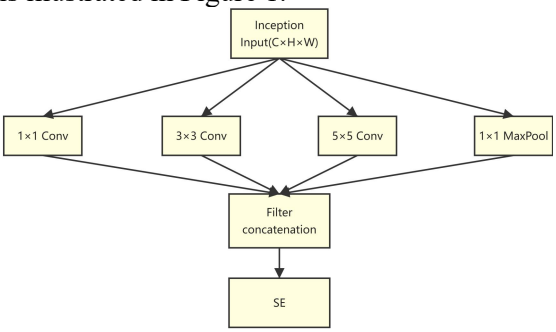

Figure 1. Schematic Diagram of the Improved Inception Module Structure with Integrated SE Attention Mechanism

Specifically, the module consists of four parallel branches. The 1×1 convolution branch captures local information and reduces the dimensionality of input channels, acting as a bottleneck layer to improve computational efficiency. The 3×3 convolution branch focuses on extracting medium-scale features, contributing to the model's ability to detect patterns of moderate complexity. The 5×5 convolution branch captures broader contextual information by leveraging a larger receptive field. In parallel, a max pooling branch followed by a 1×1 convolution is employed to preserve spatial invariance while reducing dimensionality and retaining salient features.

The outputs from all four branches are concatenated along the channel dimension, forming a rich multi-scale feature representation. Following this, a Squeeze-and-Excitation (SE) attention mechanism is applied to recalibrate the channel responses. This dynamic weighting process enhances the activation of important channels while suppressing less informative or redundant ones, thereby improving the model's overall discriminative capacity.

3.3 SE Attention Mechanism Integration Method

The SE module extracts global channel-wise information through global average pooling, followed by two fully connected (FC) layers that learn the inter-channel dependencies. This process outputs a set of channel-wise weights, which are then multiplied with the original feature maps to achieve dynamic channel-wise recalibration. This mechanism is lightweight and efficient, significantly enhancing the model's focus on critical features, and has demonstrated superior performance across various visual tasks.

By integrating the SE module, FruitNet gains the ability to automatically identify and amplify key channels while suppressing irrelevant ones in diverse image features. This selective emphasis effectively enhances the model's discriminative power, leading to improved recognition performance.

3.4 Overall Network Architecture

The overall architecture of FruitNet is depicted in Figure 2, which outlines the network's hierarchical structure from input to final output. The network is composed of several core components designed to balance rich feature extraction with computational efficiency.

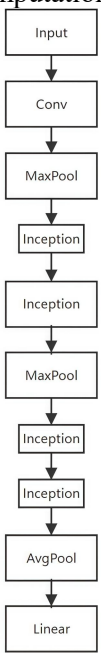

Figure 2. Schematic Diagram of the Overall FruitNet Architecture

At the input stage, the model applies a 7×7 convolutional layer with a stride of 2, which serves as the initial feature extractor. This layer captures low-level spatial information and reduces the resolution of the input image, preparing it for deeper processing.

Following this, the network incorporates four

improved Inception-SE modules, organized into two blocks. Each Inception-SE module enhances the representational power of the network by extracting features at multiple spatial scales and recalibrating channel responses through the Squeeze-and-Excitation mechanism. This attention module dynamically adjusts the contribution of each channel based on global context, enabling the model to focus on the most informative features.

To progressively reduce the spatial dimensions and aggregate higher-level semantic information, a max pooling layer is inserted after every two Inception-SE modules. This pooling operation compresses the feature maps, mitigates overfitting, and improves computational efficiency.

Near the output end of the architecture, a global average pooling (GAP) layer is employed. GAP aggregates each feature map into a single scalar by averaging all spatial locations, thereby producing a compact and translation-invariant global descriptor for each channel.

Finally, the output of the GAP layer is passed into a fully connected linear layer, which maps the feature vector to class probabilities and generates the final classification output.

This hierarchical, multi-stage design—coupled with dynamic channel-wise attention—enables FruitNet to efficiently capture and prioritize relevant features while keeping the number of parameters and computation requirements low. These characteristics make it particularly well-suited for deployment in resource-constrained environments such as mobile or embedded platforms.

4. Experiments and Results Analysis

This chapter aims to systematically validate the effectiveness of FruitNet in the task of fruit image classification. The experimental content includes dataset description, experimental setup, performance evaluation metrics, comparative analysis of experimental results, and visualization analysis, providing a comprehensive assessment of FruitNet's classification capability and generalization performance.

4.1 Dataset Introduction

The experiment uses the publicly available image classification dataset CIFAR-10, which contains 60,000 color images of size 32×32 across 10 categories, with 50,000 images for training and 10,000 for testing. The dataset is widely used for evaluating the performance of image classification algorithms.

4.2 Data Augmentation Strategy

To improve the model's generalization ability, data augmentation operations such as random cropping and horizontal flipping are applied to the input images during training. Specifically:

Random cropping with 4-pixel padding around the borders;

Random horizontal flipping;

Normalization (based on the mean and standard deviation of CIFAR-10 statistics).

4.3 Experimental Setup

The training process is conducted on a GPU device that supports CUDA. Cross-entropy loss (CrossEntropyLoss) is used as the optimization objective, and the Adam optimizer is selected with an initial learning rate of 0.001. The training is carried out for a total of 20 epochs.

During training, data is loaded in batches using train_loader, and the tqdm library is used to display the training progress in real time, facilitating the monitoring of changes in loss values and accuracy throughout the training process.

4.4 Performance Evaluation Metrics

To comprehensively evaluate the model's performance, the following metrics are used in this study:

Accuracy: The ratio of correctly classified samples to the total number of samples.

Precision: The proportion of predicted positive cases that are actually positive.

Recall: The proportion of actual positive cases that are correctly identified.

F1-score: The harmonic mean of precision and recall.

Parameters: The total number of trainable parameters in the model.

Inference Time: The average forward propagation time per image.

5. Results and Discussion

5.1 Classification Performance Analysis

During the training process on the CIFAR-10 dataset, the FruitNet model exhibited stable

and steadily improving classification performance. As illustrated in Figure 3, which shows the training accuracy curve over epochs, the model's accuracy gradually increased with the number of training epochs and began to converge around epoch 15. Ultimately, the model achieved an accuracy of 84% on the test set, demonstrating effective learning and strong generalization capabilities.

An analysis of the accuracy and loss curves confirms that FruitNet not only attains high classification accuracy but also maintains a stable training trajectory with rapid convergence. These results highlight the effectiveness and robustness of the proposed network architecture.

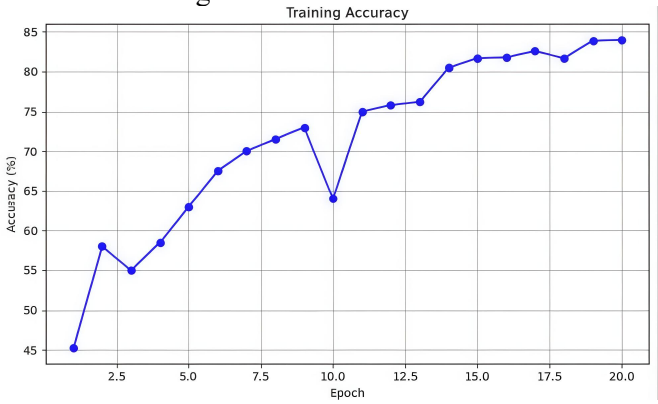

Figure 3. FruitNet Training Accuracy Curve

5.2 Feature Map Visualization

To intuitively illustrate the feature extraction capability of the improved Inception module, we visualized the feature maps from intermediate layers of the network. As shown in Figure 4, the original image serves as the input for feature extraction. The corresponding output in Figure 5 presents the feature maps generated by the network, revealing how key image characteristics are captured during processing.

Figure 4. Original Image

Figure 5. Feature Image

The visualized feature maps clearly demonstrate that the network effectively extracts low-level visual features such as edges, textures, and local patterns. The presence of multiple convolutional branches within the Inception module provides rich multi-scale representations, enabling the model to focus on both fine-grained and broader contextual information. This capability significantly enhances the model's discriminative power and contributes to its overall classification performance.

6. Conclusion and Future Work

This paper proposes an improved Inception module integrated with the Squeeze-and-Excitation (SE) attention mechanism. Based on this enhancement, a lightweight image classification network called FruitNet is constructed. The architecture effectively combines the multi-scale feature extraction capability of the Inception module with the channel-wise adaptive weighting provided by the SE mechanism, significantly enhancing the model’s focus on discriminative regions. Experimental results on the CIFAR-10 dataset show that FruitNet achieves high classification accuracy and fast convergence, while maintaining low computational complexity, thereby demonstrating the effectiveness and practical value of the proposed approach.

The main contributions of this study are as follows:

Designed an Inception module integrated with a channel attention mechanism, significantly enhancing the model's ability to focus on key information;

Built a lightweight and deployable image classification model, FruitNet, suitable for resource-constrained real-world application scenarios;

Conducted systematic experiments on standard image classification tasks, with empirical results showing that the model has strong generalization ability and robustness.

Despite achieving promising results, several areas remain for further investigation. Future work can focus on the following aspects:

Model model compression and structural optimization: Integrating lightweight architectures such as MobileNet and ShuffleNet to further reduce the number of parameters and inference time, aiming to meet the deployment requirements of mobile and edge computing devices;

Transfer learning and cross-domain adaptation: Conducting transfer learning studies on larger-scale or domain-specific datasets (e.g., medical imaging, crop disease identification) to improve the model’s adaptability and practical utility;

Extension to multi-task and multi-label learning: Expanding the current architecture to more complex visual tasks such as multi-label image classification, object detection, and semantic segmentation, to explore its generalizability;

Fusion and evolution of attention mechanisms: Further exploring the performance differences and complementarity among different attention mechanisms such as SE, CBAM , and ECA , and attempting to integrate multiple attention strategies to enhance the model’s representational capacity.

In summary, while maintaining efficient computational performance, FruitNet demonstrates strong classification capability and expansion potential, and is expected to play a broader role in various image recognition tasks in the future.

Acknowledgments

This paper is supported by Xi'an Mingde Institute of Technology (No.s202413894069).

Points to note

References

[1] Lowe D G. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91-110.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90.

[3] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 2818-2826.

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition// International Conference on Learning Representations. San Diego, CA, USA: ICLR, 2015: 1-14.

[5] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.

[6] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 4700-4708.

[7] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.