Optimization and Performance Analysis of Lightweight Convolutional Neural Networks for Image Classification

Amirul Hafiz*
Universiti Malaysia Terengganu,Kuala Terengganu,Malaysia
Email: amirul.hafiz@umt.edu.my

Abstract:With the rapid development of edge computing and mobile intelligent devices, traditional deep convolutional neural networks (CNNs) suffer from excessive parameter volume, high computational complexity, and poor deployment adaptability on resource-constrained devices. To solve this problem, this paper proposes an improved lightweight neural network model based on MobileNetV2, which optimizes the inverted residual structure and introduces a dynamic channel attention mechanism. The model reduces redundant feature extraction operations while retaining effective image feature information, and balances classification accuracy and computational efficiency. A series of comparative experiments are conducted on the public CIFAR-10 dataset. The experimental results show that the improved model achieves a classification accuracy of 94.21%, with only 2.13M parameters and 0.58G floating-point operations (FLOPs). Compared with the original MobileNetV2, ResNet18 and other classic models, the proposed model has significant advantages in parameter quantity, computational consumption and inference speed, and is more suitable for real-time image classification tasks on edge devices. This study provides an effective technical reference for the lightweight deployment of computer vision algorithms.
Keywords: Lightweight Neural Network; MobileNetV2; Channel Attention; Image Classification; Edge Computing; Computational Optimization

1. Introduction

In recent years, deep learning technology has achieved remarkable breakthroughs in the field of computer vision, and convolutional neural networks have become the core algorithm for image classification, target detection, semantic segmentation and other tasks[1]. Classic deep network models such as ResNet, VGG, and AlexNet have excellent feature extraction capabilities and high classification accuracy[2], but their complex network structures bring a large number of parameters and huge computational overhead[3], which makes it difficult to deploy on embedded devices, mobile terminals and other edge devices with limited computing power, memory and power consumption[4-7].

To adapt to the application requirements of edge intelligent scenarios, lightweight neural network design has become a hot research direction in computer science. Mainstream lightweight models such as MobileNet, ShuffleNet, and SqueezeNet reduce network complexity through depthwise separable convolution[8], channel shuffling, and feature compression respectively. However, most existing lightweight models have obvious defects: shallow network structure leads to insufficient feature extraction capability, and fixed convolution kernel settings cannot adapt to different image feature distributions, resulting in reduced classification accuracy under extreme lightweight conditions[9].

MobileNetV2, as a classic lightweight network, improves the feature utilization efficiency through the inverted residual structure and linear bottleneck layer[10], but it still has redundant channel features and insufficient attention to key feature regions in the feature extraction process. To address the above problems[11], this paper optimizes the MobileNetV2 network structure[12]: firstly, a dynamic channel attention module is embedded into the inverted residual unit to enhance the weight of effective feature channels and suppress invalid noise channels[13]; secondly, the convolution kernel combination mode is adjusted to reduce repeated feature extraction calculations[14-16]; finally, the network regularization strategy is optimized to improve the model generalization ability. Through theoretical derivation and experimental verification, this paper analyzes the performance advantages of the improved model in terms of accuracy, parameters and computational complexity, and verifies its feasibility for edge deployment[17-20].

2. Related Work

Lightweight neural network optimization technologies are mainly divided into three categories: network structure redesign[21-24], model compression, and algorithm optimization. In terms of network structure redesign, Howard et al. proposed MobileNetV1 based on depthwise separable convolution, which splits standard convolution into depthwise convolution and pointwise convolution, effectively reducing model parameters and FLOPs[25-27]. MobileNetV2 further proposes an inverted residual structure, which solves the problem of feature information loss in shallow lightweight networks and improves the network feature expression ability. ShuffleNet realizes channel information interaction through channel shuffling operation, which greatly reduces the computational cost of pointwise convolution, but its feature fusion ability is weak for complex image datasets[28-30].

In terms of model compression, common methods include pruning, quantization and knowledge distillation. Network pruning removes redundant neurons and channels to simplify the network structure, but excessive pruning will lead to model performance degradation[31]. Quantization compresses model parameters by reducing data precision, which will cause certain feature information loss. Knowledge distillation transfers the knowledge of large pre-trained models to small lightweight models, which improves the accuracy of small models, but increases the pre-training cost of the algorithm[32].

In terms of algorithm optimization, attention mechanism is widely used in lightweight network improvement. SENet realizes channel feature weighting through squeeze-and-excitation module, but its fixed weighting mode cannot adapt to dynamic feature changes of images[33]. ECANet optimizes the channel attention structure and reduces the number of attention parameters, but it lacks adaptive adjustment for different feature layers. Based on the above research, this paper designs a dynamic channel attention module suitable for lightweight networks, which realizes adaptive weighting of feature channels with low computational cost, and completes the lightweight optimization while ensuring classification accuracy[34].

3. Model Principle and Optimization Method

3.1 Basic Principle of MobileNetV2

The core components of MobileNetV2 are depthwise separable convolution and inverted residual structure. Standard convolution integrates spatial feature extraction and channel feature fusion in one step, with high computational complexity[35]. Depthwise separable convolution decomposes the standard convolution operation into two independent steps: depthwise convolution completes spatial feature extraction for each single channel, and pointwise convolution completes channel feature fusion, which greatly reduces computational consumption.

The computational complexity of standard convolution is shown in Formula (1):

$$C_{std} = H \times W \times K^2 \times C_{in} \times C_{out} \quad (1)$$

Where H and W represent the height and width of the feature map, K represents the convolution kernel size, C_{in} represents the number of input channels, and C_{out} represents the number of output channels. The computational complexity of depthwise separable convolution is shown in Formula (2):

$$C_{sep} = H \times W \times K^2 \times C_{in} + H \times W \times C_{in} \times C_{out} \quad (2)$$

The complexity reduction ratio of depthwise separable convolution relative to standard convolution is derived as Formula (3):

$$R = \frac{C_{sep}}{C_{std}} = \frac{1}{C_{out}} + \frac{1}{K^2} \quad (3)$$

It can be seen from the formula that depthwise separable convolution can significantly reduce the computational complexity of the network, especially when the number of output channels is large. The inverted residual structure of MobileNetV2 adopts the design of "expand first and then compress", which expands the channel dimension through pointwise convolution, extracts rich features in high-dimensional space, and then compresses the channel dimension to retain core features, avoiding the feature loss problem of traditional residual networks in low-dimensional space.

3.2 Improved Dynamic Channel Attention Module

Aiming at the problem that MobileNetV2 cannot distinguish the importance of different feature channels, this paper designs a dynamic channel attention (DCA) module. Different from the fixed weighting mode of traditional channel attention, the DCA module can adaptively calculate the channel weight according to the feature distribution of the input image, enhance the response of effective feature channels such as image edges and textures, and suppress the interference of invalid noise channels.

The DCA module first performs global average pooling on the input feature map $F \in R^{H \times W \times C}$ to obtain the global channel feature vector $V \in R^{1 \times 1 \times C}$, as shown in Formula (4):

$$V_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c\,(i, j) \quad (4)$$

Where V_c represents the feature value of the c-th channel, and $F_c(i, j)$ represents the pixel value of the c-th channel at position (i, j). Then, the adaptive weight coefficient is calculated through the lightweight fully connected layer and Sigmoid activation function, and the channel weight vector $W \in R^{1 \times 1 \times C}$ is obtained, as shown in Formula (5):

$$W_c = \sigma\big(FC(V_c)\big) \quad (5)$$

Where σ represents the Sigmoid activation function, and FC represents the lightweight fully connected layer with few parameters. Finally, the original feature map is weighted and fused with the channel weight vector to obtain the optimized feature map F', as shown in Formula (6):

$$F'_c = W_c \times F_c \quad (6)$$

3.3 Network Structure Optimization

The improved network takes MobileNetV2 as the basic framework, embeds the DCA module after each inverted residual unit, and optimizes the convolution kernel configuration of the shallow network. The shallow network is responsible for extracting low-level features such as image edges and colors, and uses 3×3 convolution kernels to ensure fine-grained feature extraction; the deep network extracts high-level semantic features, and appropriately reduces the convolution kernel density to reduce redundant calculation. At the same time, the L2 regularization strategy is optimized, and the regularization coefficient is dynamically adjusted according to the network layer depth to suppress overfitting and improve the model generalization performance.

4. Experimental Design and Result Analysis

4.1 Experimental Environment and Dataset

The experiment is based on the PyTorch deep learning framework, the hardware environment is Intel i7-12700H CPU, NVIDIA RTX 3060 GPU, 16G memory, and the software environment is Python 3.9, CUDA 11.6. The dataset adopts the classic CIFAR-10 image classification dataset, which contains 60,000 32×32 color images, divided into 10 categories such as airplanes, cars, and birds, including 50,000 training images and 10,000 test images. The dataset covers rich scene features and is widely used in the performance verification of lightweight image classification models.

In the experiment, the batch size is set to 32, the initial learning rate is 0.01, the learning rate decays by 0.1 every 20 epochs, the total training epochs are 100, and the Adam optimizer is used for network parameter optimization. To ensure the fairness of the experiment, all comparison models adopt the same training hyperparameters and data preprocessing methods.

4.2 Evaluation Indicators

The experiment adopts four core indicators to evaluate model performance: classification accuracy (Accuracy), model parameter quantity (Params), floating-point operations (FLOPs), and single-image inference time (Inference Time). Accuracy reflects the model's feature classification ability, Params and FLOPs reflect the model's lightweight degree and computational complexity, and inference time directly reflects the model's real-time performance on edge devices.

4.3 Comparative Experimental Results

To verify the performance advantages of the improved model, this paper compares it with four classic network models: ResNet18, ShuffleNetV2, original MobileNetV2, and SENet-MobileNetV2. The experimental results are shown in Table 1.

Table 1. Performance comparison among different network models on the CIFAR-10 dataset

Model	Accuracy (%)	Params (M)	FLOPs (G)	Inference Time (ms)
ResNet18	93.57	11.69	1.82	12.36
ShuffleNetV2	92.14	2.26	0.65	8.52
MobileNetV2	93.12	2.38	0.71	9.17
SENet-MobileNetV2	93.68	2.45	0.75	9.63
Proposed Model	94.21	2.13	0.58	7.85

4.4 Result Analysis

It can be seen from Table 1 that the improved model achieves the best comprehensive performance among all comparison models. In terms of classification accuracy, the proposed model reaches 94.21%, which is 1.09% higher than the original MobileNetV2, 0.53% higher than SENet-MobileNetV2, and significantly higher than ResNet18 and ShuffleNetV2. This proves that the dynamic channel attention module can effectively screen effective feature channels, suppress feature noise, and improve the network's feature classification ability.

In terms of lightweight indicators, the parameter quantity of the improved model is only 2.13M, and the FLOPs are 0.58G, which are reduced by 10.5% and 18.3% respectively compared with the original MobileNetV2. The inference time of a single image is 7.85ms, which is 14.4% faster than the original model. This is because the optimized network structure reduces redundant convolution operations, and the lightweight attention module brings almost no additional computational burden, which significantly improves the network inference efficiency.

Compared with ResNet18, the improved model has a slight accuracy advantage, while the parameters and computational complexity are only 18.2% and 31.9% of ResNet18, which fully reflects the lightweight advantages of the proposed model. Compared with ShuffleNetV2, although the parameter quantity is similar, the classification accuracy is 2.07% higher, and the inference speed is faster, which verifies the superiority of the improved algorithm in feature extraction and fusion.

4.5 Ablation Experiment

To further verify the effectiveness of each optimization module, this paper designs ablation experiments on the CIFAR-10 dataset, and the results are shown in Table 2. The experimental groups include the original MobileNetV2, the model with only channel attention added, the model with only network structure optimized, and the full improved model.

Table 2. Ablation results of DCA module and structure optimization on CIFAR-10 dataset

Experimental Group	DCA Module	Structure Optimization	Accuracy (%)	FLOPs (G)
1 (Baseline)	×	×	93.12	0.71
2	√	×	93.85	0.69
3	×	√	93.46	0.62
4 (Proposed)	√	√	94.21	0.58

The ablation experiment results show that both the DCA module and network structure optimization can improve the model performance. The single DCA module improves the accuracy by 0.73% with negligible computational loss. The single structure optimization reduces FLOPs by 12.7% and improves the accuracy by 0.34%. The combination of the two optimization methods achieves the best performance, which proves the mutual coordination and complementary advantages of the two modules, and verifies the rationality and effectiveness of the improved scheme.

5. Conclusion and Outlook

Aiming at the problems of low accuracy and insufficient computational efficiency of existing lightweight neural networks in image classification tasks, this paper proposes an improved MobileNetV2 model based on dynamic channel attention. By designing a lightweight adaptive channel attention module, the model enhances effective feature expression and suppresses invalid noise interference. By optimizing the network convolution structure and regularization strategy, the redundant calculation of the network is reduced and the generalization ability is improved. Experimental results on the CIFAR-10 dataset show that the improved model has significant advantages in classification accuracy, parameter scale and inference speed compared with classic lightweight and deep network models, and can better adapt to edge deployment scenarios.

In future research, we will further optimize the attention module structure, combine model quantization and pruning technology to achieve more extreme lightweight processing. At the same time, we will apply the improved model to complex scenarios such as target detection and semantic segmentation, and explore the application value of the algorithm in intelligent monitoring, mobile terminal vision and other fields, so as to realize the wide popularization of high-efficiency lightweight computer vision algorithms.

References

[1] Que, H., Gao, H., Shan, W., Liu, M., An, J., Deng, F., ... & Mu, L. (2026). FM-SAM: individual tree crown delineation and classification based on Segmentation Anything Model (SAM) and YOLOv10 in UAV imagery for forest monitoring. Computers and Electronics in Agriculture, 240, 111162.

[2] Zhu, H., Huang, X., Gao, H., Jiang, M., Que, H., & Mu, L. (2025, July). A Wireless Collaborated Inference Acceleration Framework for Plant Disease Recognition. In International Conference on Intelligent Computing (pp. 331-341). Singapore: Springer Nature Singapore.

[3] Liu, M., Que, H., Yang, X., Zhang, K., Yu, Q., Yan, L., ... & Zhou, N. (2023). A Selective Bit Dropping and Encoding Co-Strategy in Image Processing for Low-Power Design in DRAM and SRAM. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 13(1), 48-57.

[4] Gao, H., Que, H., Au, H., Shan, W., Liu, M., Qin, Y., ... & Qiao, F. (2025). SenseExpo: Efficient Autonomous Exploration with Prediction Information from Lightweight Neural Networks. arXiv preprint arXiv:2503.16000.

[5] Wu, H., Liu, J., Zha, Z. J., Chen, Z., & Sun, X. (2019, August). Mutually Reinforced Spatio-Temporal Convolutional Tube for Human Action Recognition. In IJCAI (pp. 968-974).

[6] Li, W. (2026, August). Adaptive Prototypical Contrastive Learning for Time Series Clustering. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 2730-2741).

[7] Han, Z., Chen, W., Han, Y., Mao, R., & Qin, J. (2026). Fast Diversified Top-k Rule Discovery via User-Guided Embeddings. IEEE Transactions on Knowledge and Data Engineering.

[8] Zhang C, Gu X, Bhandari U, Lei J, Guo S, Kourouma M, Karoui A and Yang S (2022) Machine learning prediction of the mechanical properties of refractory multicomponent alloys based on a dataset of phase and first principles simulation. Front. Met. Alloy 1:1036656. doi: 10.3389/ftmal.2022.1036656

[9] Gu, X., & Wang, Q. (2025, July). Pathfence: Reducing cross-path dependencies in microservices. In Proceedings of the 34th International Symposium on High-Performance Parallel and Distributed Computing (pp. 1-13).

[10] Yao, T., Lu, B., Kowarschik, M., Yuan, Y., Zhao, H., Ourselin, S., ... & Qi, P. (2025). Advancing embodied intelligence in robotic-assisted endovascular procedures: A systematic review of ai solutions. IEEE Reviews in Biomedical Engineering.

[11] Yao, T., Lu, B., Kowarschik, M., Yuan, Y., Zhao, H., Ourselin, S., ... & Qi, P. (2025). Advancing embodied intelligence in robotic-assisted endovascular procedures: A systematic review of ai solutions. IEEE Reviews in Biomedical Engineering.

[12] Chaoyong Ma, Nan Si, Xuewei Song, Chen Liang and Haoming Luo, A Bearing Fault Diagnosis Method Based on Concise Empirical Wavelet Transform, International Journal of Comprehensive Engineering，2023, 12(1), 1-6

[13] Yuyan Li, Lin Chen, Mengyu Yang, Qiaomei Li and Taifu Li, Polarization Converters Composed of Metasurfaces with Independent Manipulation of Polarization and Phase, International Journal of Comprehensive Engineering，2023, 12(1), 7-14
[14] Jiaxun Du, Zhenbao Fu, Jiachen Liu, Huaqing Wang, Zhentao Ge and Liuyang Song, Bearing Fault Diagnosis Method Based on Attention Residual Network Under Small Sample Condition, International Journal of Comprehensive Engineering，2023, 12(1), 15-24
[15] Jiawei Lin, Changkun Han, Liuyang Song, Wei Lu and Huaqing Wang, Time Domain Feature Enhancement Model Based on Correlated B-spline Wavelet Base Convolutional Sparse, International Journal of Comprehensive Engineering，2023, 12(1), 25-33
[16] Bing Wang, Haihong Tang, Xiaojia Zu and Wuwei Feng, Application of Parameter Adaptive VMD in Bearing Fault Diagnosis, International Journal of Comprehensive Engineering，2023, 12(1), 34-42
[17] Xiaojia Zu, Haihong Tang and Bing Wang, Research on Intelligent Diagnosis Based on Support Vector Machine for Structural Fault, International Journal of Comprehensive Engineering，2024, 13(1), 1-12.
[18] Xi Qiao, Di Zhao, Nan Si, Jifeng Sui and Yonggang Xu, Weighting Spectrum Editing Based on Amplitude Mapping and Its Application in Composite Fault Diagnosis of Rolling Bearings, International Journal of Comprehensive Engineering，2024, 13(1), 13-21
[19] Xixin Chen, Ming Tu, Lei Su and Ke Li, Rolling Bearing Fault Diagnosis Based on Digital Twin Method, International Journal of Comprehensive Engineering，2024, 13(1), 22-39
[20] Jianheng Liang, Yang Yang, Ran Zhang, Guican Wu, Yuanyuan Yang, TaiFu Li, Honglin Duan and Yuyan Li, Research on Perfume User Experience Evaluation Based on Human Machine Hybrid Generation Adversarial Network, International Journal of Comprehensive Engineering，2024, 13(1), 40-54
[21] Han Feng, Xia Qin and Hongtao Xue, A Novel BN and DST Fusion Framework for Fault Identification of In-wheel Motor, International Journal of Comprehensive Engineering，2024, 13(1), 55-62
[22] Feng Sun, Zhenguo Song, Dinghao Dong, Shuai Shen, Jie Cai, Heng Luo and Jianhui Zhou, A Variational Time-domain Decomposition Method for Vibration Signals Based on Exponential Adaptive Weighted Moving Average Optimization, International Journal of Comprehensive Engineering，2025, 14(1), 1-12
[23] Zhichao Qiu, Zhijia Yan, Jinhua Huang, Yewen Chen, Yaotiao Ren, Shimin Yu and Xuewei Song,Bearing Fault Diagnosis Based on Adaptive Multiple Synchronous Compressed Transform and Twin Network, International Journal of Comprehensive Engineering，2025, 14(1), 13-22
[24] Xuefei Ming, Zhenyu Pan, Gang Wang, Yong Ji, Lei Su and Ke Li, Sparse Wavelet Neural Network with Frequency-Domain Gating for Ultrasonic Vibration-Based Flip-Chip Defect Detection, International Journal of Comprehensive Engineering，2025, 14(1), 23-39
[25] Minggang Tong, Xin Wang, Wenhua Zhang, Qi Hu and Zengwei Lyu, Intelligent Fault Diagnosis of Dynamic Equipment in Public Auxiliary Systems Based on Multimodal Deep Q-Networks, International Journal of Comprehensive Engineering，2025, 14(1), 40-54
[26] Lijun Yan, Chaoyong Ma, Heng Zhang, Mengjie Xie and Rongbin Zhang, Multi-objective Optimization of Solenoid Valves Integrating AHP Decision-making and Orthogonal Design, International Journal of Comprehensive Engineering，2026, 15(1), 1-12

[27] Jiaqi Guan, Jiachen Liu, Huaqing Wang, Liuyang Song, Hassan Sajjad and Zhilong Gao, Research on Rolling Bearing Remaining Useful Life Prediction Method Based on Knowledge Distillationnternational, Journal of Comprehensive Engineering，2026, 15(1), 13-23

[28] Tian Teng and Xiao Zhang, Design Principles and Feasibility Assessment of an Intelligent Spare-Parts Shelving System for Shipboard Storage Environments, Journal of Comprehensive Engineering，2026, 15(1), 24-30

[29] Guo, Z., Fu, J., Bhanushali, S., Zeng, Z., Banerjee, I., & Sanyal, A. (2026, May). Balancing Speed and Accuracy for Robust Analog-Mixed Signal Circuit Design using Closed-Loop Reinforcement Learning with Ensemble Neural Network Surrogates. In 2026 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 803-807). IEEE.

[30] Quan, Q., Gao, Y., & Wang, Q. (2025). The impact of teacher emotional support on students' engagement in AI-mediated English learning environments: The mediating role of resilience and self-efficacy. Acta Psychologica, 260, 105766.

[31] Wang, Z., Yu, J., Liu, H., Zheng, Z., Jin, Y., Li, S., ... & Zhang, K. (2025, July). Generative Music Models’ Alignment with Professional and Amateur Users’ Expectations. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 6909-6920).

[32] Xu, Y., Cui, J., Zhu, Z., Shang, H., Luan, S., Xu, M., ... & Zhu, S. (2026). Wam-flow: Parallel coarse-to-fine motion planning via discrete flow matching for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24918-24928).

[33] Zheng, X., Hu, S., Dwyer, V., Barrett, L., & Derakhshani, M. (2026). Joint attention mechanism learning to facilitate opto-physiological monitoring during physical activity. Biomedical Signal Processing and Control, 113, 108949.

[34] Dong, Guangyuan, Hong, Ziwei, and Liu, Shenghao, et al. Dynamic Interaction-Aware and Causality-Disentangled Framework for Multimodal Sentiment Analysis. arXiv preprint arXiv:2605.30994, 2026.

[35] Su, J., He, C., Zhu, F., Xu, X., Guan, D., & Si, C. (2024, September). Hpff: Hierarchical locally supervised learning with patch feature fusion. In European Conference on Computer Vision (pp. 293-309). Cham: Springer Nature Switzerland.