Theory and Practice of Science and Technology
2026, VOL. 7, NO. 5, 95-99
DOI: 10.47297/taposatWSP2633-456919.20260705

A Review of Model Pruning Methods for Edge-Deployed MobileFaceNet

Jianyong Yang

School of Artificial Intelligence and Information Engineering, Yichun University, Yichun, Jiangxi, 336000, China

ABSTRACT
MobileFaceNet is a lightweight backbone network for face recognition specially designed for mobile and embedded devices, which has been widely used in resource-constrained scenarios such as identity authentication, video surveillance, and wearable devices. To further compress the model size, reduce computational overhead, and accelerate inference speed, model compression and acceleration methods have become a research hotspot, among which network model pruning is the most commonly used. Focusing on pruning methods, this paper systematically reviews the pruning research on MobileFaceNet. It first gives an overview of pruning, then summarizes and analyzes typical existing pruning experiments on MobileFaceNet, and finally prospects the future development directions. This paper can provide a reference for the pruning research of MobileFaceNet deployed on edge devices.

KEYWORDS
Face recognition; MobileFaceNet; Network pruning; Lightweight

1 Introduction

1.1 Research Background

As a mainstream biometric recognition technology, face recognition has been widely applied in public security, financial verification, wearable devices and other fields. Different from traditional large-scale networks, edge deployment requires models to feature small size, low computation, low latency and strong robustness. MobileFaceNet, an improved model based on MobileNetV2. Centered on depthwise separable convolution, inverted residual structure and global depthwise convolution, it achieved 99.55% accuracy on the LFW dataset with only about 0.99M parameters, becoming the preferred lightweight backbone network for embedded and mobile platforms [1]. However, in extremely resource-constrained scenarios such as smart wristbands, low-end MCUs, and miniature cameras, the standard MobileFaceNet still suffers from high computational overhead and large memory occupation. Network model pruning realizes lightweight design by removing redundant weights, convolution kernels, channels or layers in the network with controllable accuracy loss, which is the most direct and effective compression method.

1.2 Research Status and Review Structure

In recent years, numerous pruning studies for MobileFaceNet have emerged. A greedy pruning scheme using Taylor expansion to evaluate the importance of convolution kernels has been adopted, where the least important kernels are iteratively pruned according to a given pruning ratio, followed by network fine-tuning to reduce accuracy loss, so as to strike a balance between pruning rate and recognition accuracy [2]. Channel importance has been evaluated using the L2 norm, with pruning mainly performed on the depthwise and pointwise convolutional layers in the bottleneck structure of MobileFaceNet to remove channels with low contribution [3]. A structured channel pruning approach has been proposed to rank convolution kernels by a Taylor-based criterion, retaining high-weight kernels while removing low-weight ones, with the optimization objective of minimizing the model loss difference before and after pruning, expressed as min_{W'}|C(D|W')−C(D|W)|, where C denotes the loss value, W is the network weight before pruning, W' is the weight after pruning, and D represents the model input and output [4]. On the basis of MobileFaceNet, model lightweighting has been achieved through input size compression and hybrid structured pruning; the SE attention module and central difference convolution (CDC) have been integrated to enhance texture features and illumination robustness; a Fourier frequency-domain auxiliary branch has been introduced for supervision during training, and a dual-scale fusion strategy with 1.0×/4.0× face boxes has been developed. High-precision and low-latency face liveness detection has been realized on the RV1126 embedded platform, providing a complete solution for the optimization and deployment of lightweight face models on embedded devices [5]. A collaborative optimization method combining channel pruning and knowledge distillation has been presented for MobileFaceNet, taking bottleneck inverted residual blocks as pruning targets and calculating channel importance using the product of activation outputs and gradients of 1×1 upsampling convolutional layers to implement structured channel pruning. The original model serves as the teacher network and the pruned model as the student network, and knowledge distillation is conducted with a unit vector Euclidean distance loss function to

* **Corresponding Author:** Jianyong Yang, 3160722120@qq.com

transfer feature representation knowledge and compensate for accuracy degradation. Experimental results show that the optimized MobileFaceNet reduces parameters by 15% and achieves a 20% pruning ratio of bottleneck modules, with nearly unchanged accuracy on mainstream face recognition datasets such as LFW and AgeDB30, and a 20% improvement in inference speed on Android terminals, effectively realizing model lightweighting and efficient inference for mobile face recognition in resource-constrained scenarios [6]. However, existing research lacks a systematic review dedicated to the independent pruning of MobileFaceNet. Focusing solely on pure pruning methods, this paper reviews the model overview of MobileFaceNet, classification of pruning techniques, pruning implementation flow, experimental comparison of MobileFaceNet-oriented pruning, and future prospects. It clarifies the role and value of pruning in the lightweighting of MobileFaceNet, and points out the limitations and future directions of pruning research on MobileFaceNet.

2 Overview of MobileFaceNet Model

2.1 Design Background and Positioning of MobileFaceNet

Neural networks for general visual recognition tasks such as MobileNetV1, ShuffleNet and MobileNetV2 suffer from large accuracy loss if directly used for face recognition. In response, MobileFaceNet—an improved model based on MobileNetV2 specially designed for face recognition on mobile and embedded platforms—has been proposed to achieve a better balance among small model size, high accuracy and fast inference speed [1].

2.2 Core Structure and Technical Innovations of MobileFaceNet

MobileFaceNet inherits the bottlenecks of MobileNetV2 but reduces the expansion factor of bottlenecks, reducing redundant computations while maintaining feature expression capability. The model replaces the ReLU activation function with the PReLU activation function and adds BN layers to accelerate model convergence and prevent overfitting. It adopts an early downsampling strategy and an early dimension reduction strategy in the last few convolutional layers, and replaces the global average pooling layer of MobileNetV2 with a global depthwise convolutional layer, allowing the network to autonomously learn weights at different positions, making the extracted features more general and global, and effectively improving network accuracy. Table 1 shows the complete structure of MobileFaceNet. Face recognition loss functions improved based on Softmax include Arcface and AdaCos. AdaCos introduces a dynamic adaptive scaling factor sd on the basis of Arcface, facilitating dynamic adjustment of hyperparameters during training [7]. At present, Arcface and AdaCos are mainstream and reliable choices for lightweight face recognition loss functions.

2.3 Advantages and Lightweight Bottlenecks of MobileFaceNet

Compared with models such as Light CNN and MobileNetV2, MobileFaceNet has the following advantages: with 0.99M parameters, the single model size is only 4.8MB; it achieves up to 99.55% accuracy on the LFW dataset and 92.59% TAR on MegaFace (FAR=1e-6), with accuracy comparable to large CNN models such as FaceNet; the fastest inference time on mobile phones is only 18 milliseconds, featuring faster inference speed and higher recognition accuracy with fewer parameters. However, in embedded terminals with extremely low computing power and limited storage resources (such as miniature surveillance and portable identity authentication devices), the parameters and computational complexity of MobileFaceNet still need further compression. Among many model optimization methods, pruning is the most mainstream and direct effective compression method for MobileFaceNet, but a scientific pruning scheme is required for precise pruning to avoid severe accuracy loss caused by inappropriate pruning.

Table 1 MobileFacenet

Input	Operator	t	c	n	s
$112^2 \times 3$	conv3x3	-	64	1	2
$56^2 \times 64$	Depthwise conv3x3	-	64	1	1
$56^2 \times 64$	Bottleneck	2	64	5	2
$28^2 \times 64$	Bottleneck	4	128	1	2
$14^2 \times 128$	Bottleneck	2	128	6	1
$14^2 \times 128$	Bottleneck	4	128	1	2
$7^2 \times 128$	Bottleneck	2	128	2	1
$7^2 \times 128$	Conv1x1	-	512	1	1
$7^2 \times 512$	Linear GDConv7x7	-	512	1	1
$1^2 \times 512$	Linear conv1x1	-	128	1	1

3 Overview of Pruning

3.1 Classification of Pruning Techniques

Network pruning can be divided into pre-training pruning, in-training pruning and post-training pruning according to

pruning timing. For lightweight networks such as MobileFaceNet, post-training pruning is the most mainstream choice in academia and industry, so only post-training pruning is highlighted here. Network pruning can be divided into unstructured pruning and structured pruning according to pruning granularity. Unstructured pruning takes weights and connections as pruning objects, such as the OBD method and the OBS method improved on the basis of OBD [8-9]. After unstructured pruning, some weight parameters in the network are set to zero, resulting in irregularity of non-zero weights and sparsity of weight matrices, making it difficult to achieve hardware acceleration and affecting subsequent optimization and application. Due to a series of problems of unstructured pruning, it is usually not used as a mainstream pruning method, while structured pruning that maintains the complete model structure has become a widely used method at present. Structured pruning can be divided into channel pruning, convolution kernel pruning and layer pruning. Channel pruning first evaluates network channels and then trims poorly evaluated channels. Convolution kernel pruning removes redundant convolution kernels and their corresponding feature maps. Layer pruning takes layers in the network as pruning objects to directly reduce the number of network layers, suitable for deep models.

3.2 Pruning Implementation Process

3.2.1 Pre-pruning Preparation

First, clarify the pruning objectives of MobileFaceNet, such as compressing the model as much as possible under a certain accuracy loss threshold; load the pre-trained MobileFaceNet model; analyze the structural characteristics of the model; prepare datasets required for pruning; determine accuracy evaluation indicators (e. g., test accuracy on LFW dataset), inference speed evaluation indicators (e.g., fps), etc., to compare models before and after pruning.

3.2.2 Pruning Object Evaluation

According to the selected pruning criteria (e. g., Taylor expansion of loss function, L2 norm), evaluate and rank the importance of prunable units of MobileFaceNet, and then delete unimportant units according to the importance ranking. Therefore, selecting appropriate pruning criteria is particularly important, which directly affects the subsequent pruning and fine-tuning process.

3.2.3 Pruning

Accurately delete redundant units according to the evaluation results, combined with pruning strategies and pruning granularity; follow the structural constraints of MobileFaceNet and avoid some special units. Structured pruning ensures the complete structure of the pruned model and adapts to hardware acceleration.

3.2.4 Fine-tuning Process

The pruned MobileFaceNet model may suffer from accuracy degradation due to the removal of some convolution kernels or channels. At this time, fine-tuning is needed to restore accuracy. Select appropriate training parameters (e.g., learning rate, optimizer) for fine-tuning on the validation set (part of the original dataset), focusing on optimizing layers with discontinuous parameters after pruning (e.g., subsequent convolutional layers after channel cropping); over-pruning can be suppressed by adjusting pruning hyperparameters (e.g., t and s in the FOAD method) to balance the lightweight effect and accuracy loss.

4 Analysis of MobileFaceNet Pruning Experiments

4.1 Typical Pruning Method Experiments

To clearly present the actual effects (compression effect, accuracy retention and inference performance) of different pruning schemes in the lightweight transformation of MobileFaceNet, mainstream structured pruning methods are summarized. Performance indicators include pruning/compression rate, accuracy change and speed improvement, with accuracy tested on LFW, AgeDB30 or CelebASpoof datasets, as shown in Table 2.

4.2 Comprehensive Analysis of Pruning Experiments

It can be seen from Table 2 that structured pruning is the mainstream method for current MobileFaceNet model compression, which can achieve a compression rate of 36%~62%, effectively reducing model memory occupation and computing resource consumption, laying a foundation for edge deployment. The accuracy loss is generally controlled within 0.5%, and the inference speed after pruning is increased by 27.8%~1.8 times, meeting the real-time processing requirements of mobile, embedded devices and video surveillance scenarios. According to the structure of MobileFaceNet, the beginning and last few layers of the network (especially the global depthwise convolutional layer) are generally not pruned, and the rest of the structure can be pruned according to different specific compression requirements. Whether the scenario focuses on real-time performance or low power consumption, the lightweight effect brought by pruning is the basic condition for model deployment.

Table 2 Typical Pruning Experiments on MobileFaceNet

Core Pruning Method	Pruning Object	Total Pruning/ Compression Rate	Accuracy Change (LFW/AgeDB/ CelebA-Spoof Test Sets)	Speed/Performance Improvement	Core Strategy
Structured pruning based on first-order Taylor expansion	Bottleneck convolution kernels	62% (volume compressed to 38%)	LFW:99.67%→ 99.45%	Inference speed increased by 1.8 times	Pruning MobileFaceNet model embedded with SE module
Structured pruning based on first-order Taylor expansion	Bottleneck convolution kernels	51% (volume halved)	Accuracy: 99.48%→ 99.12%, decreased by 0.36%	FPS: 15→31, meeting real-time requirements	Pruning MobileFaceNet in real-time video surveillance
Hybrid structured pruning based on BN scaling factor for ordinary convolution and symmetric KL divergence for layer-by-layer convolution	Ordinary convolution channels + layer-by-layer convolution kernels	Channel pruning 48.7%, parameter reduction 56.5%	TPR (FPR=1e-2) ↓ 1%, TPR (FPR=1e-4) ↑ 11%, AUC/EER basically unchanged	Inference speed increased by 52.8%	Dual-criterion hybrid pruning + iterative fine-tuning + overfitting mitigation
Channel-level structured pruning evaluated by L2 norm	Channels of partial convolutional layers (depthwise + pointwise convolution) in CNN	Pruning rate 50%, parameter reduction 36%	LFW: 99.69%→ 99.62%, decreased by 0.07%	Not explicitly given	Pruning MobileFaceNet model embedded with ECA module
Structured pruning based on gradient×activation importance	Bottleneck channels	Channel pruning 20%, parameter reduction 15%	LFW: 98.4%→ 98.3%, decreased by 0.10%	Mobile inference: 18ms→ 13ms, improved by 27.8%	Prune MobileFaceNet and restore accuracy via knowledge distillation; actual accuracy of pruned model should be lower than 98.3%

5 Existing Research Bottlenecks and Future Prospects

5.1 Bottlenecks of MobileFaceNet-oriented Pruning

5.1.1 Insufficient Redundancy Mining in Structured Pruning

Most existing structured pruning methods adopt general evaluation criteria, which are insufficiently adapted to the exclusive structural characteristics of MobileFaceNet, such as the lack of pruning rules for global depthwise convolution, early dimension reduction layers and PReLU activation layers, failing to fully mine potential redundancy in the model and limiting further improvement of compression rate.

5.1.2 Low Automation and Adaptability

The existing pruning process still relies heavily on manual design, including pruning criterion selection, pruning strategy setting, pruning ratio setting, fine-tuning strategy configuration, etc., which are highly dependent on manual experience and experimental adjustment, with weak generalization ability. There is a lack of pruning paradigm that can automatically make decisions according to model structure, task objectives and hardware constraints, making it difficult to achieve integrated adaptive sparsification aware of tasks, data and hardware, limiting the efficient deployment of models in complex edge scenarios.

5.2 Future Development Directions of MobileFaceNet-oriented Pruning

5.2.1 Customized Structured Pruning for the Model

Aim at the characteristics of MobileFaceNet such as inverted residual module, global depthwise convolution, early dimension reduction structure and PReLU activation, design more accurate structured pruning rules to fully mine internal redundancy of the model,and further explore the possibility of higher compression rate of MobileFaceNet on the premise of ensuring accuracy.

5.2.2 Research on Automated Pruning for MobileFaceNet

Traditional pruning methods rely too much on manual experience and experimental adjustment, leading to insufficient generalization ability. In the future, pruning technology will move towards a new paradigm of automation, adaptability and end-to-end optimization, and will become increasingly mature. For lightweight face recognition models

such as MobileFaceNet, automated pruning technology may achieve a breakthrough in maintaining accuracy under high pruning rate of MobileFaceNet. The directions of automated pruning mainly include designing an adaptive pruning framework based on reinforcement learning, task-aware dynamic sparsification and end-to-end optimization of differentiable pruning technology [10]. Therefore, the following research can be carried out on MobileFaceNet in the future: 1. Based on the GRPO reinforcement learning framework, introduce dynamic scaling factors into the BN layers of MobileFaceNet and combine channel attention mechanism to build an adaptive evaluation system for channel importance of convolutional layers, then construct a multi-dimensional state vector containing layer structure features, input the constructed multi-dimensional state vector into the policy network (agent of GRPO reinforcement learning), and finally generate pruning actions and pruning rates for each layer in the network by the policy network [11]. 2. A task-aware adaptive sparsification mechanism integrated with face recognition task properties can significantly boost edge-side inference efficiency. 3. PDAS, an automated pruning algorithm based on data flow. Combined with the data flow characteristics between neural network layers, this method proposes an alternating search-update mechanism. To find the appropriate network layer width, a differentiable neural architecture search algorithm is used to learn the probability distribution, and the idea of genetic algorithm is adopted to continuously update the search space composed of candidate layer widths, improving pruning efficiency while ensuring search efficiency. For the more complex inverted residual block structure, a constrained progressive search strategy is proposed. Determine the number of channels of each layer in stages according to data flow characteristics, and compress the network model progressively to avoid additional convolution alignment overhead caused by dimension mismatch of residual connections due to pruning [12].

6 Conclusion

MobileFaceNet is cropped and optimized on the basis of MobileNetV2. With innovations such as replacing the global average pooling layer with a global depthwise convolutional layer and adopting an early dimension reduction strategy, its accuracy is effectively improved, making it a mainstream lightweight face recognition model for mobile and embedded platforms. Pruning technology is a direct and effective method to compress MobileFaceNet. The pruning of MobileFaceNet is based on structured pruning, and the technical classification, implementation process and fine-tuning process of structured pruning are summarized. From the summary of typical experiments on MobileFaceNet, structured pruning can achieve 36%-62% model compression of MobileFaceNet on the premise of controlling accuracy loss within 0.5%, and the actual edge inference speed is increased by 27.8%~1.8 times, meeting the requirements of edge deployment. However, existing studies have bottlenecks such as insufficient redundancy mining in structured pruning and low automation and adaptability. In the future, designing customized pruning schemes for MobileFaceNet and researching automated pruning for MobileFaceNet are expected to obtain MobileFaceNet models with faster inference speed and high accuracy retention under higher pruning rates, which can be fully adapted to extremely resource-constrained platforms.

References

[1] Chen, S., Liu, Y., Gao, X., & Han, Z. (2018). Mobilefacenets: efficient cnns for accurate real-time face verification on mobile devices.

[2] Li, H. (2020). Design and Implementation of Lightweight Face Recognition System Based on MobileFaceNet (Master's thesis, Southwest University).

[3] Cui, T. W. (2024). Research on Face Recognition System Based on Edge Intelligence (Master's thesis, Shaanxi University of Science and Technology).

[4] Zhao, J. (2023). Research on face recognition algorithm based on deep learning in video surveillance application. Electronic Design Engineering, 31(13), 182-186.

[5] Wang, Z. Y. (2022). Research and Implementation of Embedded Face Liveness Detection Technology (Doctoral dissertation, Huazhong University of Science and Technology).

[6] Liu, L. F., Li, L., & Wang, Q. F. (2023). A face recognition model compression method and face recognition method (CN Patent No. 117037245A).

[7] Zhang, Z. H., & Wang, R. (2020). Improved face recognition method based on MobileFaceNet network. Journal of Beijing University of Aeronautics and Astronautics, 46(09), 1756-1762.

[8] Cun, Y. L., Denker, J. S., & Solla, S. A. (1989). Optimal brain damage. Advances in Neural Information Processing Systems, 2.

[9] Hassibi, B., Stork, D. G., & Wolff, G. J. (1993). Optimal brain surgeon and general network pruning. IEEE.

[10] Wang, L., Song, Q. R., Geng, S. C., & Luan, Z. Z. (2026). A survey on neural network filter pruning techniques. Computer Engineering and Applications, 62(2), 1–25.

[11] Wang, P., Xu, M. Y., Zhao, H. R., Zhang, M. Y., Lv, J. R., Ding, C., & Liu, T. C. (2025). An automated neural network pruning method based on GRPO reinforcement learning (CN Patent No. 121052319A).

[12] Chen, Y. H. (2022). Research on Automated Pruning of Deep Neural Networks Based on Data Flow (Master's thesis, Huazhong University of Science and Technology).