A lightweight deep-learning radar gesture recognition based on a structured pruning-NAS

Eungang Son[1] , Seungeon Song[2], and Jonghun Lee*[1,2]

{silv93, sesong, jhlee}@dgist.ac.kr
[1] Department of Interdisciplinary Engineering, Graduate school, DGIST, Daegu, 42988, Korea
[2] Institute of Research, DGIST, Daegu, 42988, Korea

***Abstract*— This paper proposes a structured pruning-network architecture search (NAS) algorithm for a lightweight deep-learning radar foot gesture recognition in a conventional lightweight deep-learning models to quantitatively evaluate its performance. Our goal is to recognize foot gestures using a CW radar, generate their STFT unique signatures, and build a foot gesture recognition system that could be implemented on an edge device. The proposed scheme shows that model size and FLOPs were reduced, and a sub-optimal lightweight model for a foot gesture recognition device based on MobileNet was obtained with a slight decrease in accuracy.**

Keywords — Gesture Recognition, Foot Gesture, radar, STFT, Network Pruning, lightweight network, MobileNet

I. Introduction

Radar sensors for gesture recognition have many advantages over conventional vision sensors, addressing issues such as high cost, a challenging depth estimation, the immunity to an external perturbation such as weather, light, and vibration and so on While radar-based gesture recognitions employing deep learning techniques has been studied, most of them have predominantly relied on huge and deep neural networks [1-4], which may not be suitable for edge device deployment, such as smart trunk systems or smart buildings, where a radar based foot gesture recognition is a promising applications. This paper aims to develop a lightweight neural network for a radar-based gesture recognition for a real-time smart standalone device, achieving enhanced accuracy while minimizing computational power. This is done through integrating lightweight neural networks and network pruning techniques.

II. Related Works

Gesture recognition has been a prominent challenge within the field of computer vision even before the rise of deep learning, and various methods have been researched and developed for gesture sensing without of deep learning techniques [5]. Building upon these classical efforts, the post-golden era of deep learning has seen further endeavors to enhance gesture sensing accuracy by integrating deep learning techniques. Concurrently, various research aimed at developing lightweight deep learning models specifically for gesture recognition have been flourishing. M. Zhang et al. have designed a lightweight network deployable on ARM devices for a hand gesture recognition [6], and B. Leelakittisin et al. has proposed a more lightweight CNN network using the Joint Classification with Averaging Probability technique for a hand gesture recognition [7]. In addition to gesture recognition, there is a vigorous exploration of method to lightweight deep learning itself, with focusing on techniques deployable in mobile and touch-based sensing applications [8].

A. A Lightweght CNN

Lightweight deep-learning networks are architecturally designed for mobile and edge devices to enable the utilization of deep learning. In contrast to deep and heavy models like VITs (Vision Transformer), lightweight architectures sacrifice some accuracy but offer compact model sizes and reduced computational demands. Examples of such lightweight networks include MobileNet [9-11], EfficientNet [12-13], and SqueezeNet [14]. In this paper, we investigate that these architectures facilitate efficient deployment on resource-constrained platforms while catering to the needs of mobile and edge-based deep learning applications.

B. Neural Network pruning

The Lottery Ticket Hypothesis, introduced by Frankle et al. [15], has paved the way for developing diverse network pruning techniques. These methods encompass Structured Pruning [16-18], Unstructured Pruning [19], Quantization [20], Knowledge Distillation [21], and Neural Architecture Search (NAS) [22], among others.

III. A Radar-based Foot gesture Recognition System

A. A Continuous Wave(CW) radar

A CW radar is an effective method for non-contact gesture recognition thanks to its high sensitivity and robustness to environmental variations. It provides sensitive detection of target movement by utilizing high-frequency radio waves. Furthermore, due to its simple hardware architecture and signal processing, it is suitable to gesture recognition edge devices in terms of resource efficiency. A CW radar system can be expressed as equation.(1).

$$X(t) = A* \exp(j2p\pi ft) \ (\text{for } 0 \leqslant t \leqslant T), \qquad (1).$$

Here, $A*$ represents the amplitude of the transmitted signal, and f denotes the operating frequency in a time-limited time domain T. The received signal of a CW radar reflected by any motion is given by Equation (2). The received signal suffers from the frequency shift caused by the Doppler effect due to a motion, as well as the amplitude and phase affected by the Radar Cross Section (RCS) of a moving object.

$$Y(t) = A^{\#} \exp(j2\pi(f\text{-}\Delta f)\, t + \phi^{\#}), \qquad (2).$$

Here, $A^{\#}$ represents the magnitude of the received signal reflected by a motion in a CW radar, Δf denotes the Doppler frequency shift caused by a motion, and Δt and ϕ' respectively

Corresponding Author
Prof. Jonghun Lee(jhlee@dgist.ac.kr): DGIST

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

denote the round-trip propagation time between a foot and the radar, and the phase variation due to a motion. The output signal of CW radar caused by a foot motion is a baseband beat signal extracted through complex mixing using a mixer device and a low-pass filter (LPF). Equation (3) represents the beat signal S*(t)* corresponding to a foot motion.

$$S(t)=A\ exp\ (j2\pi\Delta f\ t + \phi) + N_0 \qquad (3).$$

Here, *A* represents the product of the amplitudes of the received and transmitted signals in the CW radar.

B. The Short Time Fourier Transform (STFT)

The input signal of deep learning models is a radar feature signal capable of distinguishing various gestures. The time-frequency spectrogram of the radar beat signal, which corresponds to each foot gesture, is used as a radar feature signal. To extract its radar feature signal each foot gesture, the STFT is applied to the radar beat signal, obtaining a radar signature corresponding to each foot gesture in the time-frequency domain. Equation (4) represents the calculation formula for the STFT of the radar beat signal caused by gestures.

$$F(t', u) = \int S(t)w(t'-t)exp(-j2\pi tu)\ dt', \qquad (4).$$

Here, *w* indicates window function, where *t'* is time axis and u is frequency axis in spectrogram. This equation implies that for every foot gesture input, a unique spectrogram exists that can be applied for a foot gesture recognition.

IV. Experiments

In our previous research, a foot -gesture recognition employing machine-learning like Support Vector Machine(SVM) and deep-learning such as AlexNet, GoogleNet, and ResNet has been developed [23]. This paper intends to develop a radar based foot gesture recognition based on a lightweight deep-learning model by using hybrid network pruning scheme of a deep learning model. The MobileNetV3 small and MobileNetV2 [10-11] are used as backbone networks for a foot-gesture recognition.

A. Data Gathering and Preprocessing

A Continuous Wave (CW) radar is used for a foot gesture recognition, which is operating at a frequency of 24GHz, with a bandwidth of 300MHz, maximum output power of 1mW, and a maximum beam width of 120 degrees. Radar data for four types of foot gestures (kick, swing, slide, and tap) were obtained. The acquired radar beat signals were transformed into their radar feature images of size 227×227 through the STFT. A total of 3,500 images were collected, comprising 600 images for training and 100 images for testing per class. Additionally, 20% of the training data was allocated for validation purposes. Fig 1 shows its typical example image of each class.

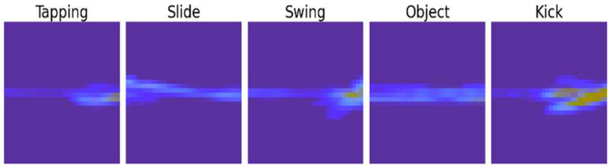


Fig. 1. Radar STFT signatures for 4 kinds of foot gestures and 1 object

B. A lightweight foot-gesture recognition based on Mobile Deep-learning Network

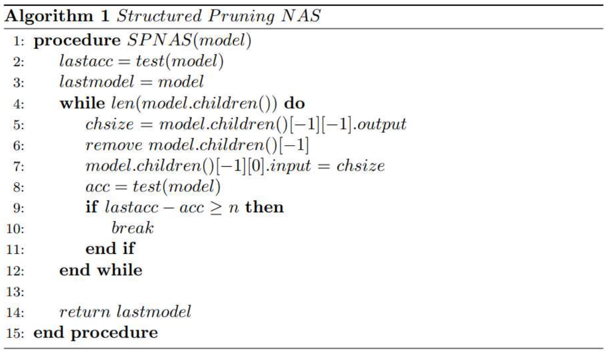
Each backbone network is used for a lightweight foot-gesture recognition. During training each backbone network, Adam optimizer and Cross-Entropy loss are used as the hyperparameters, with an appropriate number of epochs given for the network to sufficiently learn. Plus, the training and validation accuracies and losses were monitored to prevent occurrences of overfitting and underfitting. Table 1 shows the training result of the Backbone deep-learning network.

TABLE I. Result of backbone network training

MobileNet Version	Accuracy	Precision	Recall	F1	Flops	Capacity
V3	0.906	0.918	0.906	0.906	71M	5.93 MB
V2	0.944	0.945	0.944	0.944	373M	8.74 MB

C. Pruning Based NAS

A pruning-based NAS is more straightforward and convenient compared to conventional NAS methods that employ reinforcement learning to explore network components within the search space [22]. This is because a pruning-based NAS leverages existing architectures by extracting required network structures from them, rather than performing a complete search. In this paper, focusing on the backbone, structured pruning is applied to discover a sub-optimal network structure suitable for an edge device deployment. The pruning method involves cutting a redundant network from the top-level architectures of the model with keeping the structure of the backbone. Fig 2 shows the structured pruning NAS algorithm,



```
Algorithm 1 Structured Pruning NAS
procedure $SPNAS(model)$
    $lastacc = test(model)$
    $lastmodel = model$
    while $len(model.children())$ do
        $chsize = model.children()[-1][-1].output$
        remove $model.children()[-1]$
        $model.children()[-1][0].input = chsize$
        $acc = test(model)$
        if $lastacc - acc \geq n$ then
            break
        end if
    end while

    return lastmodel
end procedure
```

Fig. 2. The proposed structured Pruning NAS Algorithm

Our proposed Structured pruning NAS is performing iteratively removing architectural elements starting from the top-level structures and progressing towards those closer to the classifier, until all the structures have been eliminated. A series of necessary network blocks are obtained until a predetermined threshold of accuracy and loss is satisfied. These top-level structures can be adjusted based on the model's characteristics, allowing for further exploration of lower-level structures or by criteria considering the system and performance requirement at the block level.

Table 2 shows the performance evaluation and hardware requirement of a foot-gesture recognition based on the proposed structured pruning NAS algorithm.

TABLE II. RESULT OF BACKBONE NETWORK TRAINING

MobileNet Version	Accuracy	Precision	Recall	F1	Flops	Capacity
V3	0.938	0.941	0.938	0.937	36M	0.78 MB
V2	0.95	0.951	0.95	0.950	172M	0.58 MB

A significant enhancement in overall both model size and FLOPs is obtained, even though a slight tradeoff in terms of accuracy.

V. COCNLUSION

This paper collects foot gestures data using a Continuous Wave (CW) radar and transforms their collected data into their radar signatures by means of the Short-Time Fourier Transform (STFT) processing. These radar signatures are used as a dataset for the selected lightweight backbone networks, including MobileNetV3, MobileNetV2. Structured pruning NAS of the multiple layer architectures of in conventional lightweight backbone networks results in even more lightweight models due to the reduction of their network multiple layers. As an experimental result, the proposed MobileNetV2 exhibits a notable performance, reducing the model size to 15% of the backbone's size and lowering FLOPs by 42.9 with a nearly same accuracy. We will leverage quantization techniques to secure even more lightweight gesture recognition models.

ACKNOWLEDGMENT

This research was partially supported by the Korea Institute for Advancement of Technology (KIAT) and funded by the Ministry of Trade, Industry and Energy (MOTIE) of the Korean Government (No. P0018599).

REFERENCES

[1] H. Gao and C. Li, "Automated Violin Bowing Gesture Recognition Using FMCW-Radar and Machine Learning," in IEEE Sensors Journal, vol. 23, no. 9, pp. 9262-9270, 1 May1, 2023, doi: 10.1109/JSEN.2023.3263513.

[2] L. Qiao, Z. Li, B. Xiao, Y. Shu, W. Li and X. Gao, "Gesture-ProxylessNAS: A Lightweight Network for Mid-Air Gesture Recognition Based on UWB Radar," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, doi: 10.1109/JSTARS.2023.3274830..

[3] Z. Zhang, Z. Tian and M. Zhou, "Latern: Dynamic Continuous Hand Gesture Recognition Using FMCW Radar Sensor," in IEEE Sensors Journal, vol. 18, no. 8, pp. 3278-3289, 15 April15, 2018, doi: 10.1109/JSEN.2018.2808688.

[4] H. Gao, C. Williams, V. G. Rizzi Varela and C. Li, "Violin Gesture Recognition Using FMCW Radars," 2023 IEEE Topical Conference on Wireless Sensors and Sensor Networks, Las Vegas, NV, USA, 2023, pp. 13-15, doi: 10.1109/WiSNeT56959.2023.10046213.

[5] S. Wu, Z. Li, S. Li, Q. Liu, and W. Wu, "An overview of gesture recognition," Proc. SPIE 12609, International Conference on Computer Application and Information Security (ICCAIS 2022), 1260926 (21 March 2023); https://doi.org/10.1117/12.2671842

[6] M. Zhang, Z. Zhou, T. Wang, and W. Zhou, "A Lightweight Network Deployed on ARM Devices for Hand Gesture Recognition," in IEEE Access, vol. 11, pp. 45493-45503, 2023, doi: 10.1109/ACCESS.2023.3273713.

[7] B. Leelakittisin, M. Trakulruangroj, S. Sangnark, T. Wilaiprasitporn and T. Sudhawiyangkul, "Enhanced Lightweight CNN Using Joint Classification with Averaging Probability for sEMG-Based Subject-Independent Hand Gesture Recognition," in IEEE Sensors Journal, doi: 10.1109/JSEN.2023.3296649.

[8] O. Durmaz Incel and S. Ö. Bursa, "On-Device Deep Learning for Mobile and Wearable Sensing Applications: A Review," in IEEE Sensors Journal, vol. 23, no. 6, pp. 5501-5512, 15 March15, 2023, doi: 10.1109/JSEN.2023.3240854.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861 [cs.CV], 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1704.04861

[10] Sandler, Mark & Howard, Andrew & Zhu, Menglong & Zhmoginov, Andrey & Chen, Liang-Chieh. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 4510-4520. 10.1109/CVPR.2018.00474.

[11] A. Howard et al., "Searching for MobileNetV3," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314-1324. doi: 10.48550/arXiv.1905.02244

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv:1905.11946 [cs.LG], 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1905.11946.

[13] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," arXiv:2104.00298 [cs.CV], 1 Apr. 2021, last revised 23 Jun. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2104.00298.

[14] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," arXiv:1602.07360 [cs.CV], 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1602.07360

[15] J. Frankle and M. Carbin, "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks," in Proceedings of the International Conference on Learning Representations (ICLR), 2019. doi: 10.48550/arXiv.1803.03635

[16] Y. He, X. Zhang and J. Sun, "Channel Pruning for Accelerating Very Deep Neural Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 1398-1406, doi: 10.1109/ICCV.2017.155.

[17] 7C. -H. Tu, J. -H. Lee, Y. -M. Chan and C. -S. Chen, "Pruning Depthwise Separable Convolutions for MobileNet Compression," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207259.

[18] Yuefu Zhou, Ya Zhang, Yanfeng Wang, Qi Tian, "Accelerate CNN via Recursive Bayesian Pruning", Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3306-3315

[19] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the Value of Network Pruning," arXiv preprint arXiv:1810.05270, 2018.

[20] D. Gleich, P. Planinsic, B. Gergic and Z. Cucej, "Progressive space frequency quantization for SAR data compression," in IEEE Transactions on Geoscience and Remote Sensing, vol. 40, no. 1, pp. 3-10, Jan. 2002, doi: 10.1109/36.981344.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," in Proceedings of the Neural Information Processing Systems (NIPS) Deep Learning Workshop, 2014. doi: 10.48550/arXiv.1503.02531

[22] T. Elsken, J. H. Metzen, and F. Hutter, "Neural Architecture Search: A Survey," in Journal of Machine Learning Research, vol. 20, pp. 1-21, 2019. doi: 10.48550/arXiv.1808.05377

[23] Seungeon Song, Bongseok Kim, Sangdong Kim, Jonghoon Lee, "Foot Gesture Recognition Using High-Compression Radar Signature Image and Deep Learning," Sensors, vol. 21, no. 11, article 3937, 2021.