Embedded Human Activity Recognition with Optimized MLP and CNN

Jiyeong Moon
Duksung Women's University
Seoul, Korea
moon09@duksung.ac.kr

Soojin Lee
Duksung Women's University
Seoul, Korea
sojinlee1004@duksung.ac.kr

Soyeong Lee
Duksung Women's University
Seoul, Korea
dlthdud0112@duksung.ac.kr

Yeji Yang
Duksung Women's University
Seoul, Korea
willy019@duksung.ac.kr

Heerae Lee
Duksung Women's University
Seoul, Korea
dlgmlfo0415@duksung.ac.kr

Seongyeong Kim
Duksung Women's University
Seoul, Korea
sg000828@duksung.ac.kr

***Abstract*—To find the optimal model for HAR (Human Activity Recognition) in resource-constrained IoT (Internet of Things) devices, we basically conduct a system-level design space explorations using MLP (Multi-Layer Perceptron) and CNN (Convolutional Neural Network) models under the limitation of maximum number of parameters around 200,000. We first investigate the subsampling method to reduce the size of the input data as well as to increase the number of training data. Then we explore the design space in terms of the accuracy and the number of parameter used for the model. Finally, we apply the quantization to reduce the memory occupation of the model during the implementation on a target device. Our extensive evaluations demonstrate that the optimized model achieves 98.40% accuracy with 13,640 bytes memory space.**

Keywords— human activity recognition, inference, machine learning, convolutional neural network, multi-layer perceptron

I. INTRODUCTION

The most important thing in embedded HAR is to find a light-weight model that uses small resources while showing the accuracy enough to trust the inference made by the model [1]. To this end, we perform the system-level design space exploration similar to [2]. In addition to exploring the model structure, we explore the input data to further optimize the classification model by preprocessing the input data.

We aim to find the optimal model in two different aspects. The first one is to maximize the recognized accuracy. The other one is to minimize the required computation and memory resources while keeping similar accuracy. To aggressively reduce the memory space requirement, we also apply several popular quantization techniques during the model implementation process. The performance of the model with additional weight reduction is summarized by comparing the accuracy, execution time, and size of the model for each technique.

II. METHODOLOGY

A. Hyper-parmameters Search Space Exploration

We try to find the optimal model suitable for a certain environment in which the model is to be executed through design space exploration. STM32L476RGT6 (ARM Cortex-M4 core at 80MHz) is considered as the target device to execute the optimized classification model. The target device includes 1 Mbytes of flash memory and 128 Kbytes of SRAM. As a result of executing the proposed CNN model with 212,492 parameters on the target board for testing purpose, a total of 849.38 KiBs of flash memory space and a total of 19.74 KiBs of SRAM are required including the library data for calculation. Therefore, the design space exploration is performed under the following limited condition.

- The maximum number of model parameters used for design space exploration is set to 200,000.
- Since it is necessary to consider the weight reduction of the model, MLP and CNN are considered as the types of models to be explored.

B. Data Subsampling

In order to reduce the input data size, we first analyze the characteristics of the 3-axis accelerometer data and the stretch sensor data, and then conduct several methods to finally remove the redundancy among the data. After manually investing the training datasets, we find that there are some missing (or errored) data values at specific locations, which degrades the prediction performance. To mitigate these errors, we exploit subsampling by extracting only odd-numbered data value from the original datasets to reduce the effect of missing values while keeping the unique pattern of the datasets. We confirm that even with subsampled datasets after the retraining, the results are similar to those of learning the entire datasets. This means that the input data size can be reduced while sufficiently reflecting the characteristics of the datasets.

If the odd-numbered subsampling datasets are evaluated with the model trained using the even-numbered subsampling datasets, the accuracy is still closed to 82%. Therefore, we confirm that the odd-numbered subsampling datasets show high similarity with the even-numbered subsampling datasets. This means that both types of datasets can be used as training datasets, which ultimately doubles the number of training datasets. Before using doubled training datasets, the highest accuracy is 96.04% for the optimized CNN, and 95.95% for the optimized MLP. After using doubled training datasets, the optimized CNN and MLP show the accuracy of 98.27% with 4,643 parameters and the accuracy of 98.01% with 9,408 parameters, respectively.

III. PERFORMANCE EVALUATION

We perform extensive design space exploration by varying the number of hidden layer and the number of node in each layer. Fig. 1 shows the results of the design space exploration (total 150 designs) in terms of the accuracy (Y axis) and the number of parameters (X axis) for CNN and MLP models.

The CNN model with the highest accuracy achieves 98.27% accuracy only with 4,643 parameters. The MLP with the highest accuracy achieves 98.01% accuracy with 9,408 parameters. The MLP model that uses the lowest number of parameters among the all MLPs, uses 4,083 parameters but shows 95.95% accuracy.

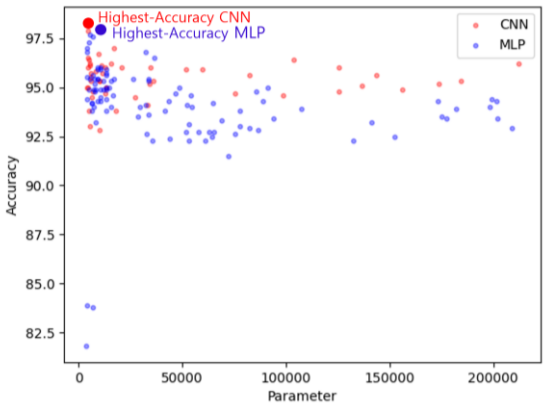

Fig. 1 Performance of various MLP and CNN models

Table I and Figs. 2 to 3 present the details of the models that show the highest accuracy or that use the lowest memory.

TABLE I. HIGHEST-ACCURACY AND LOWEST-PARAMETER MODELS

Model	Accuracy	Parameters
Highest-accuracy CNN	98.27	4,643
Highest-accuracy MLP	98.01	9,408
Lowest-parameter MLP	95.95	4,083

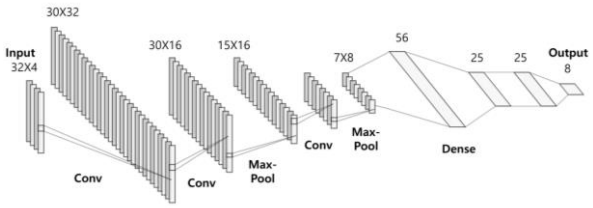

Fig. 2 The highest-accuracy CNN architecture

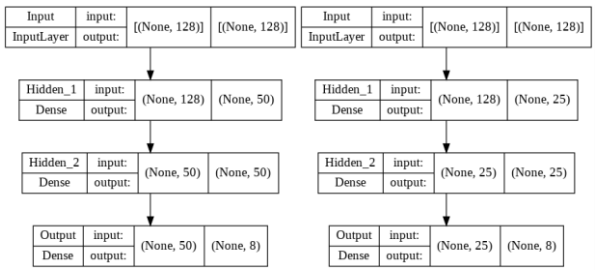

Fig. 3 The highest-accuracy MLP (left) and the lowest-parameter MLP (right) models

IV. IMPLEMENTATION-LEVEL OPTIMIZATION

From the extensive design space explorations, we conduct the optimized CNN and MLP models that show the highest accuracy or lowest number of parameters. In this section, we further decrease the memory requirements by applying three popular quantization methods [3], considering the target device.

- Float 16 quantization: convert weights to float 16
- Dynamic range quantization: convert weights to 8- bit
- Float fallback: convert weights and variable data to 8-bit

A. Performance of Quantized Model

Table II shows the quantization results for the optimized CNN model, in terms of accuracy, execution time, and memory requirement on the target device. Due to the space limitation, we only present the quantization results for optimized CNN that shows the highest accuracy among the results of the entire design space exploration. The highest-accuracy CNN with float 16 quantization is able to reduce the model size by 0.68 times compared to that without quantization. We confirm that there is no accuracy degradation. Dynamic range quantization can reduce the model size slightly more than float 16 quantization. However, it slightly degrades the accuracy and increases the execution time significantly. Finally, float fallback quantization reduces the model size by 0.56 without the accuracy degradation while it reduces the execution time to 350.2 us which is only 31% of the non-quantization model.

TABLE II. PERFORMANCE OF QUANTIZED HIGHEST-ACCURACY CNN MODEL

Model	Accuracy	Execution time (ms)	Size (byte)
Non-Quantization	98.27	1.1277	24,508
Float 16 Quantization	98.31	1.2727	16,548
Dynamic Range Quantization	98.19	2.2280	16,152
Float Fallback Quantization	98.40	0.3502	13,640

V. CONCLUSIONS

In order to obtain the optimized HAR design targeting for STM32L476RGT6 device, we used a system-level optimization which mainly consists of input data preprocessing, CNN and MLP design space exploration, and quantization. For preprocessing, subsampling was applied to reduce the input data size while keeping similar accuracy. It also doubled the training data which finally enhances the recognition accuracy. Optimized model through the design space exploration and quantization shows 98.40% accuracy, 350.2us execution time, and 13,640 bytes memory spaces.

REFERENCES

[1] Xu, Yang, and Ting Ting Qiu. "Human Activity Recognition and Embedded Application Based on Convolutional Neural Network." Journal of Artificial Intelligence and Technology, pp. 51-60, December 2020.

[2] Feng, Kaijie, Xiaoya Fan, Jianfeng An, Chuxi Li, Kaiyue Di, and Jiangfei Li. "ACDSE: A Design Space Exploration Method for CNN Accelerator Based on Adaptive Compression Mechanism." ACM Transactions on Embedded Computing Systems, pp. 1-25 June 2022.

[3] "Model Optimization | TensorFlow Lite." n.d. TensorFlow. https://www.tensorflow.org/lite/performance/model_optimization.