Title of Design

***First Author1), Co-Author2) and Corresponding Author3\*)***

(1line spacing)

**1)Dept. of Inform. Display, Your College, Town/City, Seoul 121-791, Korea**

**2)Dept. of Electric Engineering, Your College, Town/City, MA 02139, U.S.A.**

**3)Dept. of Electric Engineering, Your College, Town/City, MA 02139, U.S.A.**

*E-mail:*

 *(2 line spacing)*

**Abstract:** 5~8 lines of abstraction for the design.

**Keywords**: 3 ~ 5 keywords that describ your design

1. 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.

2. 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 CortexM4 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.

3. Evaluation

We perform extensive design space exploration by varying the number of hidden layer and the number of nodes in each layer. Figure. 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.



1. Example of a figure caption. (figure caption)

4. 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.