An Architectural Study for Inference Coprocessor Core at the Edge in IoT Sensing

Daisuke Watanabe[1], Yuji Yano[2], Shintaro Izumi[2], Hiroshi Kawaguchi[1], Kiyoshi Takeuchi[3], Toshiro Hiramoto[3]
Shoichi Iwai[4], Masami Murakata[5], Masahiko Yoshimoto[2]
1 Graduate School of Science, Technology and Innovation, Kobe University, Kobe, Japan
2 Graduate School of System Informatics, Kobe University, Kobe, Japan
3 Institute of Industrial Science, the University of Tokyo, Tokyo, Japan
4 SALTYSTER, Shiojiri, Japan
5 Device&System Platform Development Center, Kawasaki, Japan
E-mail: watanabe.daisuke@cs28.cs.kobe-u.ac.jp

***Abstract*—In this paper, random forest (RF), convolutional neural network (CNN), and support vector machine (SVM) algorithms are evaluated in terms of accuracy and performance for one-dimensional time series data in the target application fields of wearable healthcare and factory automation, considering field programmable gate array (FPGA) and system-on-a-chip (SoC) implementations. The results show that the RF is an optimal learning/inference algorithm from the viewpoint of energy efficiency and that the CNN is effective for high-precision applications. For example, in arrhythmia detection, the inference accuracies of RF and CNN are 94% and 97%, respectively. In contrast, RF is approximately 4 orders higher in energy efficiency. Next, an architecture for the inference coprocessor core embedded at the edge sensor was proposed, which can efficiently implement the above RF and CNN inference algorithms. An inference-oriented data-path that accelerates CNN computation was proposed, allowing for a faster order of computation with approximately 1/8 power consumption. Additionally, a port-reconfigurable RAM that increases the memory bandwidth for RF and CNN processing was introduced, which doubles the energy efficiency for RF processing. As a result, in arrhythmia detection, heartbeat interval extraction, and human activity classification (three wearable applications), the power consumption of the RF inference coprocessor was estimated to be 0.6 μW, 0.4 μW, and 0.4 μW, respectively, assuming a standard low-power 65-nm CMOS technology.**

Keywords—IoT, machine learning, low power inference, SoC, coprocessor

I. INTRODUCTION

With the development of Internet of Things (IoT), which connects almost everything to the Internet, various types of data analyses can be performed in fields such as human health care, factory operation, and environmental monitoring. Conventionally, the data obtained from IoT sensors are analyzed using large-scale cloud computations. However, the number of IoT devices being used has been rapidly increasing and system schemes that send all the collected data to the cloud for analysis are incapable of detecting instantaneous abnormal values from data changes that occur frequently. Therefore, this becomes a challenge in applications that require real-time performance [1]. In addition, the increase in the power consumption of data centers that collect cloud data as well as the increase in communication traffic for uploading data to the cloud have become significant recent issues.

To solve the above-mentioned problems, an edge computing system that provides edge devices and gateways intelligence and performs cooperative distributed processing using the cloud has been proposed [2]. In particular, the inference function needs to be installed into an edge sensor, which is required to obtain low power characteristics under limited power conditions. Here, a trade-off study was conducted to evaluate the inference accuracy and energy efficiency of the considered machine learning (ML) algorithms suitable for wearable healthcare and factory operations, in which one-dimensional time series data are assumed, and images and sounds are not employed to realize extremely energy-efficient systems. The study is based on the results of field-programmable gate array (FPGA) implementation. Moreover, a system-on-a-chip (SoC) architecture that executes inference was investigated. An inference-oriented circuit was introduced to improve the processing performance, and the power dissipation was also estimated.

This paper is based on results obtained from a project commissioned by New Energy and Industrial Technology Development Organization (NEDO).

II. TRADE-OFF STUDY ON INFERENCE AT THE EDGE

In this study, random forest (RF), support vector machine (SVM), and convolutional neural network (CNN) were selected as the ML algorithms. They were evaluated and compared in terms of accuracy, power consumption, and hardware cost of the inference function at the edge sensor. An overview of each algorithm is presented in Fig. 1. The CNN includes a one-dimensional convolutional layer to handle the one-dimensional time series data. By changing the number of convolutional layers, the number of connected layers, and the number of feature maps of the hidden layers, the inference accuracy, calculation costs, and memory usage can be adjusted. An SVM is used for finding a class classified from a support vector and an input vector at the inference by learning a support vector for the class classification. Calculation costs for the SVM at the inference increase proportionally with the number of dimensions and classification classes of the input vector and the number of input vectors in the learning process. The RF technique is a method of inferring with a plurality of decision trees; here, the final inference result is obtained from each inference result using the majority decision method. It is characterized by low computational costs during inference.

A. Wearable Healthcare Applications

In this study, we consider three applications for wearable healthcare: arrhythmia detection, heartbeat interval extraction, and human activity classification.

a) Arrhythmia detection

Portable devices that detect arrhythmia from electro-cardiograms (ECGs) are still under development, and improvement of detection accuracy in this application is a

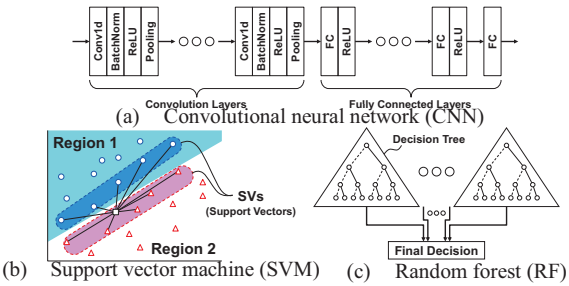


Fig. 1. Three machine learning algorithms for inference at the edge

challenge. The MIT-BIH arrhythmia database was used as a dataset for arrhythmia detection [3]. This dataset consists of ECGs that include data from healthy subjects as well as patients suffering from arrhythmias. It contains two-channel ECGs obtained from 48 people with a sampling rate of 360 Hz. In this study, data of 40 people were extracted and used for processing in the ML algorithms. The data were divided equally into the training and inference datasets. A four-class classification problem was defined as follows: normal, fibrillation, ventricular escape beat, and supraventricular ectopic beat. In the evaluation, RF and CNN were implemented in several configurations. However, as the number of calculations in SVM increased, as shown in Fig. 2, it was excluded from the target algorithm. The inference accuracy was 95.2% in RF and 97.7% in CNN.

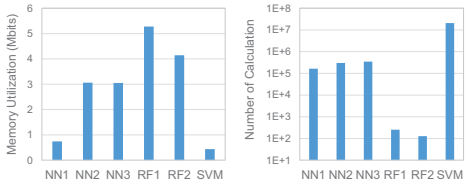


Fig. 2. Memory usage and calculation costs for inference

To estimate the performance and energy consumption of the inference engine, it was implemented on an FPGA circuit using a Xilinx FPGA board. Fig. 3 shows the performance and energy consumption estimation results. It can be observed that RF is advantageous in terms of both performance and energy consumption.

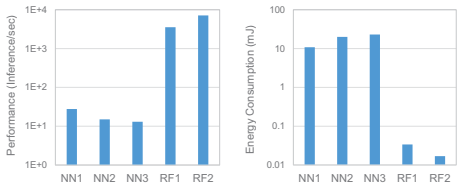


Fig. 3. Performance and energy consumption (FPGA)

Fig. 4 shows the relationship between the inference accuracy and energy efficiency. The accuracy of the CNN is approximately 97%, while that of RF is inferior (94%); however, it is seen that RF is approximately three orders of magnitude more energy efficient than the CNN.

b) Heartbeat interval extraction

Another application of health management using wearable sensors is the heart rate variability (HRV) analysis. HRV analysis can be used in predicting and identifying various diseases and determining the stress state by analyzing the time variation in the heartbeat interval of a human [4]. In this study,

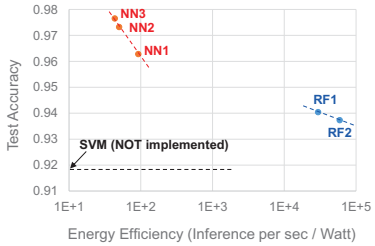


Fig. 4. Performance evaluations for three inference algorithms

a public dataset for heart rate monitoring was used for testing [5]. The dataset includes one-channel ECG, two-channel photoplethysmography (PPG), and three-axis acceleration data recorded at a sampling rate of 125 Hz. The heartbeat interval was inferred using two channels of PPG. Considering inference at the edge, the heartbeat interval was divided into 16 classes of five samples to reduce the complexity of the problem. For the CNN, three models (NN1, NN2, and NN3) were implemented to optimize the number of calculations and memory usage. The obtained inference accuracies were 98% for RF; and 93%, 99%, and 100% for NN1, NN2, and NN3 (for the CNN), respectively.

Considering an increase in memory usage in RF implementation, a model was established to perform the final classification using only the top three classes of probability estimates when outputting the decision tree inference results (TABLE I.). Fig. 5 shows the evaluation results of inference accuracy and energy efficiency estimation. Although the accuracy of CNN is close to 98% and that of RF is 95%, the energy efficiency of RF is nearly two orders of magnitude higher than that of the CNN.

c) Human activity classification

To prevent and obtain countermeasures for lifestyle-related diseases, it is important to collect and analyze biological signal data from humans in routine life and to also provide feedback on the analyzed results [6]. One of the widely used data analysis items is activity estimation. Several studies have been conducted on human activity classification for activity estimation using acceleration data [7]. In this study, wearable sensors are used to measure the PPG and acceleration data. The HRV estimated from PPG and the acceleration data are learned by ML algorithms, and human activity classification is performed by inference at the edge. Sensor data were measured mainly for several male and female subjects in their 20s using an experimental kit that can collect 125 Hz PPG and 16 Hz three-axis acceleration data. The PPG data were extracted using peak-to-peak time intervals instead of dealing with raw waveforms. The noise in the low-frequency region such as the digital circuit (DC) component was removed through a high-pass filter for the three-axis acceleration.

TABLE I. MEMORY USAGE AND CALCULATION COSTS

	Conv. Layers	FC Layers	TREEs	Calculations	Memory Utilization (bit)
Heartbeat interval extraction					
NN1	2	2	-	8296	105344
NN2	2	2	-	11368	204672
NN3	3	2	-	26832	382208
RF1	-	-	8	< 96	6455744
RF2	-	-	8	< 96	3397760
Human activity classification					
NN4	4	3	-	23160	353632
RF3	-	-	30	< 300	913376

All the data collected as a ML dataset were divided in such a way that the training and inference datasets contained 50% of data without distinction between subjects. Seven classes of data were collected for human activity classification classes defined in the evaluation: sitting, walking slowly, walking fast, running, cycling, walking upstairs, and vacuum cleaning. The results showed the inference accuracy to be 84% in RF and 93% in CNN. TABLE I. summarizes memory usage and the number of calculations for each algorithm. Fig. 5 shows a graph of the inference accuracy and energy efficiency. In terms of accuracy, RF is significantly inferior to CNN, whereas the energy efficiency is one-order higher in RF.

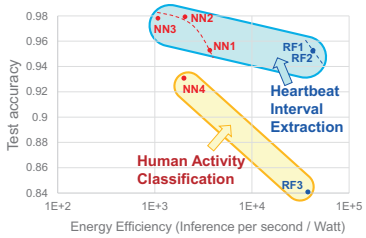


Fig. 5. Performance evaluations for two inference algorithms

B. *Factory Automation Application*

As an application for factory operations, we worked on a defective product pre-judgment system using various sensor data in the production line [8]. In factories, the quality of products that are manufactured and shipped is monitored by conducting inspections at various stages of the manufacturing process, thereby resulting in rejected products. A rejected product found downstream in the process contains more loaded parts as compared to those upstream, which leads to a greater loss. Therefore, it is important to identify defective products as upstream as possible. A sensor was installed in each process of the production line, and the inference accuracy was verified by comparing the inference results with the actual quality inspection results. The results showed that the inference accuracy was 95% for RF and 98% for CNN. Fig. 6 compares the processing performance of RF and CNN after FPGA implementation. It is observed that the RF has a higher inference speed and consumes lesser energy as compared to the CNN.

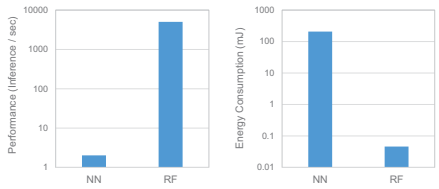


Fig. 6. Performance and energy consumption (FPGA)

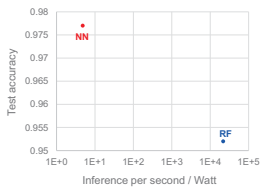


Fig. 7. Performance evaluations for two inference algorithms

Fig. 7 shows the relationship between the inference accuracy and energy efficiency. It is seen that the RF is approximately three orders of magnitude more energy efficient than the CNN.

From the discussion based on FPGA implementation in Section II, it is concluded that the RF algorithm is more energy-efficient and therefore more suitable for low power applications. In contrast, the CNN is more suitable for applications that require higher accuracy. The next section will discuss a VLSI architecture for an inference coprocessor at the edge sensor, considering a system scheme to be learned by the cloud and gateway and to be inferred by an edge sensor.

III. SOC ARCHITECTURE FOR RANDOM FOREST AND CNN

The edge sensor node operates under the battery drive or energy harvester drive condition in both wearable and factory applications, and the power supply function is severely limited. Therefore, the power domain is divided into circuit blocks inside the SoC. It is necessary to reduce the power consumption by disconnecting the power supply to standby circuits.

Fig. 8 presents an overview of the proposed SoC architecture. Here, the inference coprocessor operates as one of the peripheral circuits of the microcontroller unit (MCU). The DSP accelerator that contains the multiplier/ accumulator circuit takes over the processing of the MCU and executes arithmetic processing in a short period with high efficiency. The local memory, which stores the CNN weight and bias parameters and temporarily stores the feature map, is integrated in the inference coprocessor. If the local memory is shared with the MCU cache memory, the performance decreases owing to memory access arbitration in the arbiter circuit, as the memory access bandwidth required for the inference process is higher than that for memory access from the MCU. Additionally, the power consumption can be reduced by placing the memory close to the DSP accelerator to enable frequent memory access that occurs in the inference process. Therefore, the efficiency of the edge SoC can be increased by separating the memory from the inference coprocessor.

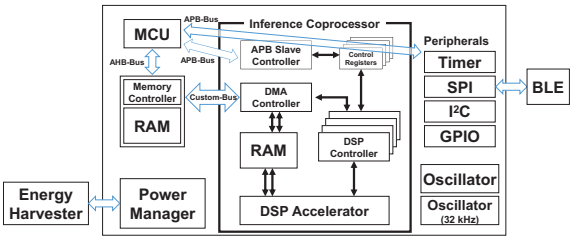


Fig. 8. SoC architecture with inference engine

To support the neural network-based inference, 16-bit floating point (FP16) and 32-bit floating point (FP32) data formats were handled. These comprise a sign bit, an exponent field, and a mantissa field, where the mantissa represents only the decimal part after the decimal point. The mantissa scale is 24 bits for FP32, and 11 bits for FP16. Fig. 9 shows the multiplier/accumulator (MAC) core circuit that saves circuit resources for inference at the edge. By integrating four twelve-bit multiplication circuits, one FP32 multiplication and two FP16 multiplications can be realized using the same circuit resources. Fig. 10 shows the configuration of the DSP accelerator with the MAC core circuit. The DSP accelerator is optimized for digital signal processing and inference processing of one-dimensional time series data frequently handled by edge sensor nodes, and it is utilized to perform four types of calculations such as product-sum / sum-product /

product / sum at high speed. By sharing circuit resources and controlling data paths with the use of a switching network, highly efficient calculations are executed. The FP16 processing in the proposed circuit is 8 times more efficient than conventional FP32 processing. The processing performance can be improved by increasing the number of parallel processing circuits installed, but the memory bandwidth is determined by the RAM data bus width, clock frequency, and the number of RAM blocks mounted. To improve performance, the memory bandwidth must be improved. However, an increase in memory bandwidth leads to an increase in RAM operating power consumption, mounting area, and standby power consumption. Therefore, the memory bandwidth must be optimally designed according to the application. In the proposed SoC, it is possible to switch the RAM configuration between single- and dual-ports, and the configuration can enable parallel search for decision trees of RF and can double the energy efficiency for RF processing (Fig. 11).

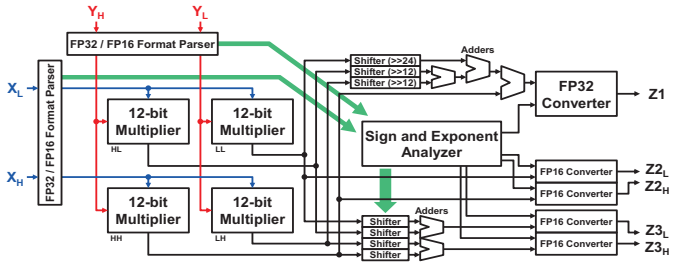


Fig. 9. FP16/FP32 MAC core circuit

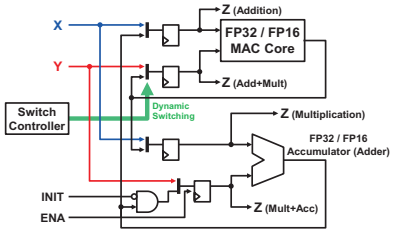
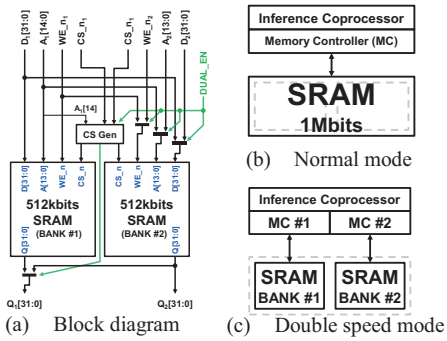


Fig. 10. DSP accelerator optimized to inference at the edge



(a) Block diagram (b) Normal mode (c) Double speed mode

Fig. 11. SP/DP SRAM control circuit for inference at the edge

The current fluctuation in an SoC equipped with an inference coprocessor is illustrated in Fig. 12. It can be divided into three main parts: a sensing period to collect data from an external sensor device, an inference processing period using an inference engine, and a communication period to transmit the inference result. During the sensing period, most of the circuit blocks, other than the memory that holds the data, are in the standby state; thus,, it is possible to reduce the power consumption by using the MCU sleep mode, powering off the circuit block, and shutting down the clock. During the inference processing period, peripheral circuits such as I/O can be stopped, and the current dissipated in the MAC core operation and memory access significantly affect the power consumption during this period. During the communication period, the RF circuit, MCU, and memory section operate, and the maximum current consumption occurs instantaneously.

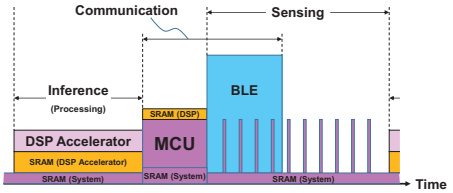
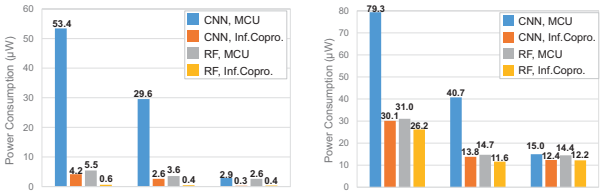


Fig. 12. Breakdown of the current consumption in SoC

IV. PERFORMANCE EVALUATION

With a focus on wearable applications, the power reduction due to the proposed inference coprocessor was estimated, as shown in Fig. 13. Here, a standard low-power 65 nm CMOS process was assumed as the fabrication technology. By implementing the inference function with the coprocessor, it can be seen that CNN power can be reduced to approximately 1/10–1/12, and RF power can be reduced to about 1/7–1/9. In each application of arrhythmia detection, heartbeat interval extraction, and human activity classification, the CNN inference power is 4.2 µW, 2.6 µW, and 0.3 µW, respectively and RF inference power was 0.6 µW, 0.4 µW, and 0.4µW, respectively. Furthermore, the RF power of the SoC which integrated the inference coprocessor was 26.2 µW, 11.6 µW, and 12.2 µW for the above three applications, respectively. Here, the power in the normal MCU, when the low leakage process was not used, is included.



(a) Coprocessor (b) SoC (excluding BLE)

Fig. 13. Power consumption of edge SoC with the inference coprocessor

V. CONCLUSIONS

Three learning inference algorithms were evaluated in terms of the accuracy and performance of the edge sensor inference coprocessor. The results showed that RF is an optimal learning /inference algorithm with respect to energy efficiency, and CNN is effective for high-precision applications. Based on the evaluation results, the architecture of the inference coprocessor core (including an inference-oriented data path and bank reconfiguration RAM) embedded in the edge sensor SoC was proposed. As a result, for three type of healthcare applications, RF inference power was estimated to be 0.6µW-0.4µW, and the power during CNN inference was estimated to be 4.2µW-0.3µW, assuming a standard 65 nm CMOS technology. If a low-leakage process and a hardware sequencer are employed, the MCU power is reduced by approximately one order of magnitude; therefore, the power of the edge sensor SoC is expected to decrease further.

REFERENCES

[1] F. Samie, L. Bauer, and J. Henkel, “From cloud down to things: An overview of machine learning in internet of things,” IEEE Internet of Things Journal, 2019.

[2] Shi, W., Cao, J., Zhang, Q., et al: ‘Edge computing: vision and challenges’, IEEE Internet Things J., 2016, 3, (5), pp. 637–646.

[3] Moody GB, Mark RG. ”The impact of the MIT-BIH Arrhythmia Database." IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209)

[4] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5] Zhilin Zhang, et al. ‘‘TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise,’’IEEE Trans. on Biomedical Engineering, vol. 62, no. 2, pp. 522-531, February 2015.

[6] Murakami, Haruka, et al. “Accuracy of wearable devices for estimating total energy expenditure: comparison with metabolic chamber and doubly labeled water method.” JAMA internal medicine 176.5 (2016): 702-703.

[7] J. Parkka, et al. "Activity Classification Using Realistic Data from Wearable Sensors", IEEE Trans. Information Technology in Biomedicine, vol. 10, no. 1, pp. 119-128, Jan. 2006.

[8] B. Chen, et al. "Smart factory of Industry 4.0: Key technologies application case and challenges" in IEEE Access.