Time-Domain Neural Network with Superconducting Single-Flux-Quantum Devices

Tatsuya Hoshino
Department of Electrical Engineering and Computer Science, School of Engineering, Kyushu University
819-0395, Fukuoka
Japan
tatsuya.hoshino@cpc.ait.kyushu-u.ac.jp

Koji Inoue
Department of I&E Visionaries, Faculty of Information Science and Electrical Engineering, Kyushu University
819-0395, Fukuoka
Japan
inoue@ait.kyushu-u.ac.jp

ABSTRACT

Although expanding the scale of Convolutional Neural Network (CNN) improves image recognition accuracy, it raises computational complexity and power consumption. Increasing the number of hidden layers improves the accuracy of identification, but also increases the amount of calculation. ResNet, which recorded the best accuracy[1] in image recognition in 2015, consists of 152 layers, 230 MB of memory for storing weights, and 11.3 billion multiply-accumulate operations are required. Currently, high-performance computer that processes a large amount of computation at high speed performs Deep Learning and it consumes a lot of power. In order to execute Deep Learning on a device such as a smartphone, it is necessary to execute a large amount of computation with low power consumption and low latency. As impact of the end of CMOS miniaturizing, development of software algorithms and hardware accelerators for program is significant to shorten the execution time. The switching time tends to become larger due to an increase in parasitic capacitance between circuits. Latency becomes a serious problem as the clock rate of the processor increases.

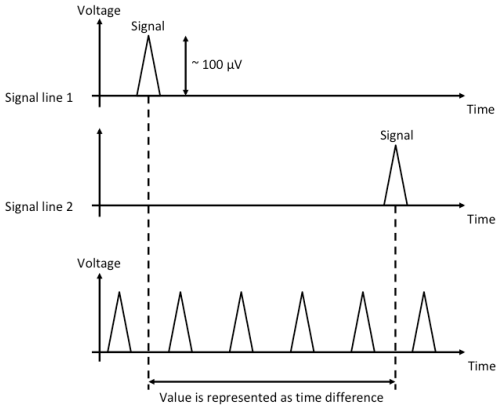

Figure 1: Value is represented by time difference between two signals.

1 INTRODUCTION

As a countermeasure for latency problem, a method of executing dominant multiply-accumulate operation by CNN processing hardware with excellent power efficiency attracts attention, and as one of its implementations, Time Domain Neural Network (TDNN) has been proposed [1]. TDNN adopts time domain analog and digital mixed signal processing for calculation and in this approach, weights are expressed as difference in two lines signal transmission time as Fig. 1. This architecture reduces memory accesses for activations by broadcasting activation to multiple processing elements as Fig. 2. Energy consumption is reduced by adopting the time-domain data transmission because the number of voltage transition is reduced. TDNN binarizes the input to the neural network, while weights are expressed as signal propagation delay. The element manipulating delay is called variable delay cell and implemented by variable resistance. Although TDNN realizes high-energy efficiency, signal propagation based on charging / discharging of capacitance is the basic operation, thus there arises a problem that operating speed is lowered.

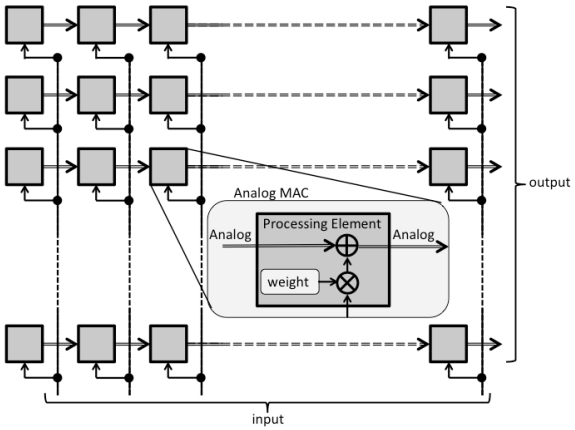

Figure 2: Proposed TDNN coprocessor for convolutional layer in CNN.

[1]ImageNetCompetition in 2015.

2 TECHNICAL APPROACH AND FUTURE WORK

To execute TDNN at low cost, we propose implementation by SFQ circuit which is the device enables to low latency and low power consumption signal propagation. Superconductor single-flux-quantum (SFQ) circuit is a practical technology for the general purpose computing. Owing to its high speed and low power features, some researchers have been contributed to develop SFQ microprocessors, SFQ devices, and their design technologies [2-4]. As a result, it has been demonstrated that some SFQ components operate with more than 100 GHz clock frequency. The feature is that SFQ works with direct current power supply for bias current. Its power consumption is small because only impulse voltage is generated in the Josephson junction only for several picoseconds when a flux quantum passes through Josephson device. For the reason, power consumption of SFQ logic gates is 1/1,000 of CMOS logic gates [5].

Since its operation speed exceeds 100 GHz, it can be expected to improve the latency in TDNN significantly. Variable delay cell is a key element to express weights value as time delay. Therefore, method of implementing SFQ variable delay element is one of the essential challenges in SFQ-TDNN implementation. Although the feasibility of variable delay elements in SFQ has been shown, the accuracy of identification, when SFQ variable delay elements are used, has not yet been clarified. In this study, as the first step to the goal that implementing SFQ-TDNN, we investigate the relationship between the range of weights and recognition accuracy when TDNN is applied to the MNIST and Fashion MNIST for the image recognition data set. Therefore, we estimate the relationship between resolution of time and image recognition accuracy using the SFQ variable delay element.

First, in the convolutional neural network, we constructed a measuring environment to simulate the image recognition accuracy at the time of inference in the case of using the SFQ variable delay element for convolutional layer. As a result, changes in the recognition accuracy for the MNIST and Fashion MNIST data sets of the neural network with respect to the number of stages by weight quantization were clarified. From the result, the resolution required for the variable delay element in TDNN is considered and the case of SFQ variable delay element is discussed. We estimated the range of the desired weight from the simulation. As a result of matching the estimated value with the time resolution of the SFQ variable delay element, recognition accuracy of 98.24% for MNIST and 81.66% for Fashion MNIST can be achieved. In future, it is a goal to build up the whole as TDNN using SFQ variable delay elements.

REFERENCES

[1] D. Miyashita, S. Kousai, T. Suzuki and J. Deguchi. 2017. A neuromarphic chip optimized for deep learning and cmos technology with time domain analog and digital mixed-signal processing. *IEEE Journal of Solid-State Circuits*, vol. 52, no.10, pp. 2679-2689. DOI: https://doi.org/10.1109/JSSC.2017.2712626

[2] M. Tanaka, K. Takata, R. Satoh, A. Fujimaki, T. Kawaguchi, Y. Ando, K. Takagi, N. Takagi and N. Yoshikawa. 2014. Design of RSFQ Microprocessors Integrated with RAMs Based on Bit-Serial Processing, *7th Superconducting SFQ VLSI Workshop*.

[3] M. Tanaka, T. Kawamoto, Y. Yamanashi, Y. Kamiya, A. Akimoto, K. Fujiwara, A. Fujimaki, N. Yoshikawa, H. Terai and S. Yorozu. 2006. Design of a pipelined 8-bit serial single-flux-quantum microprocessor with multiple ALUs, *Superconductor Science and Technology*, Vol. 19, No. 5, p. S344.

[4] M. Tanaka, Y. Yamanashi, N. Irie, H. Park, S. Iwasaki, K. Takagi, K. Taketomi, A. Fujimaki, N. Yoshikawa, H. Terai, et al. 2007. Design and implementation of a pipelined 8 bit-serial single-flux-quantum microprocessor with cache memories, *Superconductor Science and Technology*, Vol. 20, No. 11, p. S305.

[5] Likharev, K. K. and Semenov, V. K. 1991. RSFQ logic/memory family: a new Josephson-junction technology for sub-terahertz-clock-frequency digital systems. *IEEE Transactions on Applied Superconductivity*, Vol.1, No.1, pp.3-28. DOI: https://doi.org/10.1109/77.80745