A Prediction Scheme in Spiking Neural Network (SNN) Hardware for Ultra-low Power Consumption

Jeonggyu Yang/Kyungpook National University
School of Electronic and Electrical Engineering
Daegu 41566, South Korea
jg.yang@knu.ac.kr

*Taigon Song/Kyungpook National University
School of Electronics Engineering
School of Electronic and Electrical Engineering
Daegu 41566, South Korea
tsong@knu.ac.kr

Abstract— **The tremendous success of the artificial neural networks (ANNs) has led to an increase in the demand for embedded neural network hardware. In this trend, researchers aggressively studied spiking neural network (SNN) architectures due to its advantages in power consumption. Still, better SNN architectures are needed to support more neurons required in low-power systems. Therefore, we propose a new prediction scheme based on neuron potential that significantly reduces power in SNN architectures. Our prediction scheme applies to SNN architectures composed of simple Integrate and Fire (IF) neuron models without leakage. We designed an SNN hardware with our prediction scheme and verified that our scheme reduces -19.75% power consumption with only 0.85% accuracy decay.**

Keywords—SNN hardware; low power SNN; edge device

I. Introduction

Most neural network (NN) services are offered in the form of cloud computing. Users exchange data with servers in real-time, which inevitably leads to stability and security issues. In contrast, edge computing eliminates these problems by performing NN operations on the edge devices themselves. For this reason, NN services are moving toward edge computing. However, edge computing has been having trouble popularizing despite the advantages of edge computing. This is because NN services have become highly complicated that the power required for operations has increased. To resolve these issues, a spiking neural network (SNN) is gaining more attention. SNN is an architecture that mimics the behavior of biological neurons, and this architecture operates based on events only for low power consumption [1], [2], [3].

Designing an SNN architecture requires several consideration points for edge devices. First, digital SNN is typically preferred over analog SNN due to its lower design difficulty. Second, the leaky integrate-and-fire (LIF) model is the most commonly used neuron model, but the implementation of leaky operations requires additional circuitry. Thus, adopting a simple integrated-and-fire (IF) neuron model without leaky operations is beneficial in terms of power. Third, many SNN training methods are not considered reliable when training multi-layered SNNs. However, it is reported that the method of converting the ReLU-based artificial neural network (ANN) to ReLU-based SNN successfully obtains trained SNN parameters with high-performance [3], [4], [5].

Inspired by the fact that ReLU-based SNN is highly advantageous in SNN designs, we propose a novel prediction scheme that significantly reduces the power consumption to these SNNs. We implement our hardware and our prediction scheme in physical layout (Place&Route) and measure the design metrics to prove our prediction scheme useful. To the best of authors' knowledge, our scheme is the first proposed that reduces significant power in SNN architectures based on neuron potential.

II. Proposed prediction scheme

To discuss the advantages of our proposed scheme, we designed a ReLU-based SNN for MNIST classification that stems from the same ANN, which is presented in [3]. The converted SNN is transcribed into RTL, and we applied our prediction scheme to SNN. Our designs are implemented up to P&R for validation.

A. Neural network structure

Our baseline SNN model consists of a size of 784-128-128-10. Since there is a trade-off between accuracy and model size, we chose a relatively small size to obtain a lightweight model suitable for embedded environments. In the process of converting ANN to SNN to obtain SNN's trained parameters, a slight decrease in accuracy may occur. This is because neurons cannot fire multiple spikes at the same time. To solve this problem, we adjusted the range of the activation to within ±1 by using a batch normalization. We also limited the range of weights from 8/16 to +7/16 to apply the weight quantization, and weights are mapped to 8, 7, …, +6, and +7, respectively, before converting to the RTL. Lastly, we removed 85% of the weights by using the pruning technique to make the model lighter [6]. By applying the pruning, 0.75% accuracy decay occurred.

B. Our prediction scheme

In the IF neuron model, SNN's potential generally tends to increase or decrease continuously for the same input. The reason is that IF neurons do not have a leaky operation that converges potential into a resting potential, and the ReLU-based SNN has the characteristic of having a constant firing rate. Also, IF neurons fire only when the potential exceeds the threshold voltage (V_{th}). Here, if a neuron's potential is low enough for a negative value, the possibility of this neuron firing a spike will be very low. Therefore, we propose a new scheme that saves power by gating the clock and input spike train that are entered into neurons with a very low probability of firing spikes. We named the potential at which the neurons gated as the gating potential (V_g) and set the value to -2.5 V_{th}. Fig. 1 (a) shows a case where the neuron's potential continues to increase. Fig. 1.

The EDA tool was supported by the IC Design Edu-cation Center (IDEC), Korea; This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1G1A109470212); This research was results of a study on the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA.

2020 International SoC Design Conference (ISOCC) | 978-1-7281-8331-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISOCC50952.2020.9333106

(b) shows a case where the potential of a neuron continues to decrease. In the case of Fig. 1. (b), there is no output spike, so that we can gate the clock and input spike train at time-step 4.

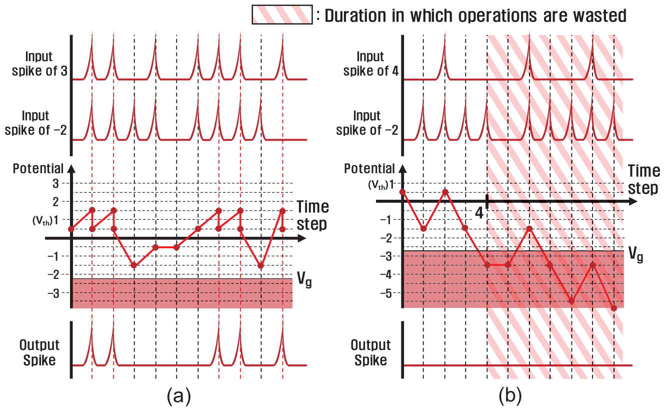

Figure 1. (a): The behavior of a neuron when the potential increases continuously, (b): The behavior of a neuron whe n the potential decreases continuously. In the case of (b), there is no output spike, so that we gate the clock and input spike train without accuracy decay at time-step 4.

III. Result

We designed an SNN with 95.97% accuracy by the method of section II-A. Although the accuracy is less than that of state-of-the-art, we judged it to be sufficiently meaningful accuracy in an embedded environment because of the use of various lightweight techniques. Proposed SNN (SNN with our scheme) and naive SNN (SNN without our scheme) were implemented up to P&R, and the implementation results are shown in Fig. 2. Based on these two designs, we measured the accuracy in functional verification, and verified power consumption in the PrimeTime PX based on the switching activity of the VCD file obtained through the MNIST test set.

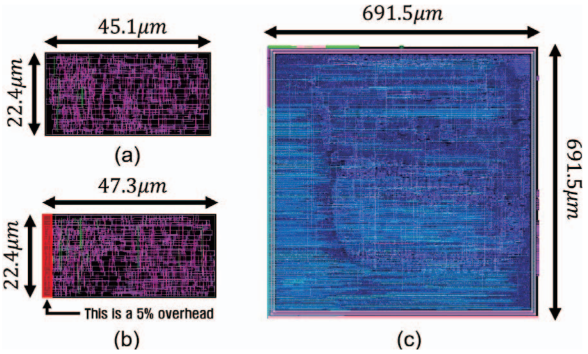

Figure 2. (a): layout of a spiking neuron without our scheme, (b): spiking neuron with our scheme, (c): layout of the entire SNN consisting of (b). The design is made of Silvaco 45nm library, and our scheme involves about 5% area overhead.

Fig. 3. shows the accuracy of SNN according to each time-step. We define time-step as the unit time taken to recognize the number, and the clock period is set to 10ns. The naive SNN converges with 95.97% accuracy as the time-step elapses, and proposed SNN converges with 95.12%. Only 0.85% accuracy loss occurred, which is the result of some unintended neurons gated by our scheme. Fig. 4. shows the power consumption ratio of proposed SNN and naive SNN at each time-step. At the early time-step, the power consumption of proposed SNN uses 9 percent more than naive SNN, and this is due to the overhead required for the gating. However, the power overhead almost disappears five time-steps after. The power consumption of the proposed SNN and naive SNN measured was 31.98mW and 39.85mW after 85 time-step. Over time, total power consumption will eventually converge to 80.25 percent. The longer time-steps, the smaller the impact of overhead. Thus, we anticipate our scheme more effective for complex problems that require a long duration, such as ImageNet.

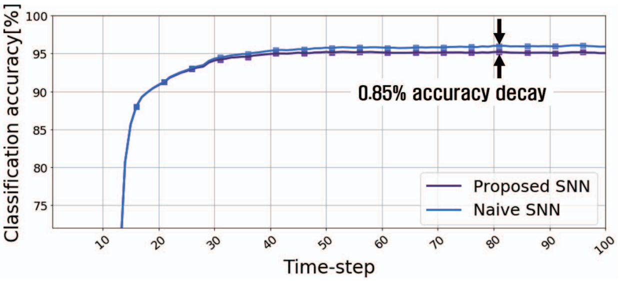

Figure 3. SNN's accuracy according to the time-step. When applying our scheme, there is a decay of accuracy of 0.85%.

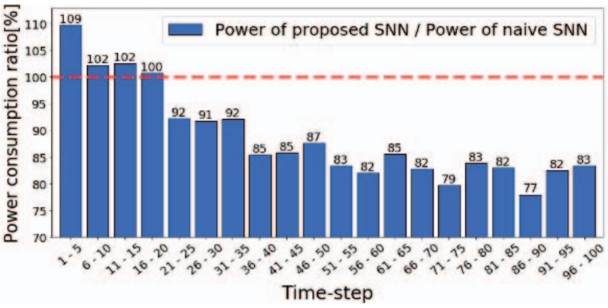

Figure 4. Power consumed by proposed SNN compared to naive SNN at each time-step. At the early time-step, there is about 9% power overhead. However, as the classification time increases, our prediction scheme consumes significantly less power.

IV. Conclusion

We developed a novel prediction scheme based on neuron potential that achieves ultra-low power consumption for SNN hardware. We verified the accuracy and power consumption of our scheme in physical layout, and the results show that our scheme reduces -19.75% power consumption with only 0.85% accuracy decay.

References

[1] Rueckauer Bodo *et al.*, "Conversion of Continuous-Valued Deep Networks to Efficient Event-Driven Networks for Image Classification," Frontiers in Neuroscience, vol. 11, pp. 682, 2017.

[2] F. Akopyan *et al.*, "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 10, pp. 1537-1557, Oct. 2015.

[3] Peter O'Connor and Max Welling, "Deep Spiking Networks", arXiv, 2016.08323, 2016.

[4] P. U. Diehl *et al.*, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," IJCNN, pp. 1-8, 2015.

[5] S. Hwang *et al.*, "System-Level Simulation of Hardware Spiking Neural Network Based on Synaptic Transistors and I&F Neuron Circuits," in IEEE Electron Device Letters, vol. 39, no. 9, pp. 1441-1444, Sept. 2018.

[6] Song Han *et al.*, "Learning both Weights and Connections for Efficient Neural Networks," arXiv, 1506.02626, 2015.