Extended Abstracts of the 2019 International Conference on Solid State Devices and Materials, Nagoya, 2019, pp675-676

Investigation of Adaptive Weight Quantization in Application for Convolutional Neuromorphic System Using Gated Schottky Diode

Dongseok Kwon, Suhwan Lim, Jong-Ho Bae, Sung-Tae Lee, Hyungsu Kim, Young-Tak Seo, Gyuho Yeom, Byung-Gook Park, Jong-Ho Lee,

Department of Electrical and Computer Engineering and Inter-University Semiconductor Research Center, Seoul National University, Seoul, 151-742, Republic of Korea
Phone: +82-2-880-1727 E-mail: jhl@snu.ac.kr

Abstract

In this work, convolutional neuromorphic system (CNS) to which is applied adaptive weight quantization considering characteristics of synaptic devices is investigated. A simple way is proposed to compensate the nonlinearity of synaptic devices in application of adaptive weight quantization. In addition, we verify that the CNS using *n*-type gated Schottky diodes performs well even with non-ideal characteristics of synaptic devices such as conductance variation or stuck devices.

1. Introduction

Recently, an artificial intelligence (AI) has been shown remarkable growth with development of algorithms and network structures [1]. Especially, convolutional neural networks (CNNs) have outperformed performance of human in image application. However, CNNs require a large amount of vector-by-matrix multiplications, which causes substantial power consumption and memory bottlenecks in Von Neumann computing system. To overcome these limitations, a neuromorphic system using analog synaptic devices has been regarded as a promising computing system capable of low power consumption and parallel computing. In neuromorphic systems employing off-chip training methods, adaptive weight quantization has been proposed and performed well in MNIST classification even considering the non-idealities of synaptic devices [2, 3]. However, it should be confirmed that the quantization method performs well in a convolutional neuromorphic system (CNS) for more difficult tasks. In this work, the adaptive weight quantization depending on the device characteristics is applied to the trained weights in CNNs for CIFAR-10 data classification. We then investigate whether gated Schottky diodes (GSDs) are suitable for the CNS even in the presence of non-ideal characteristics.

2. Device Structure and Method

Fig. 1(a) shows that schematic cross-sectional view of the GSDs [4]. Electrons or holes can be stored in Si_3N_4 layer of the oxide/nitride/oxide stack, by applying programming or erasing pulses to the bottom gate under the Schottky junction. Then, the GSDs modulate their conductance which is used as a synaptic weight in CNS. Fig. 1(b) shows that measured conductance behavior with respect to the number of identical pulses applied to the *n*-type GSDs. The conductance behavior can be fitted as follows:

$$G(x) = a + \frac{1}{\beta}\ln(x + c), \qquad (1)$$

where G is the conductance of the GSDs, x is the number of pulses, a and c are the fitting parameters, and β is the nonlinearity factor of synaptic devices.

As a baseline network, CNN is designed to classify CIFAR-10 data, as shown in Fig. 2. The activation function of the hidden layers is hard-sigmoid, and that of output layer is softmax. We obtain 90.35% of accuracy with floating point operation using batch normalization and Adam optimizer [5]. The trained weights are then normalized according to predetermined normalization ratio [6]. If the normalization ratio is 99%, for example, 99% of total number of weights are normalized to the maximum conductance of synaptic device, and rest of them (1%) are set to the maximum conductance of the device. The adaptive weight quantization is applied to the normalized weights considering the characteristics of synaptic devices, and can be mapped to the real synaptic devices by adjusting the number of identical pulses applied.

3. Simulation Results

Fig. 3 shows the accuracy versus total conductance level of the synapse device as a parameter of the nonlinear factor when the normalization ratio is 100% in the CNS with adaptive weight quantization. The accuracy of the CNS decreases as the total conductance level decreases, and larger nonlinear factor gives smaller accuracy. Fig. 4 shows the accuracy versus normalization ratio as a parameter of β (nonlinearity factor) at a fixed conductance level of 32. As the nonlinearity factor of synaptic device increases, the 1st quantization boundary, which determines whether the number of pulses applied to the device is 1 or 0, also increases. This increases the probability that the trained weights between $G(0)$ and $G(1)$ will be quantized to zero (=$G(0)$) and consequently degrades the performance of the CNS. To alleviate this problem, we modify the 1st quantization boundary to be decreased from 0.097 to 0.065, for example, as shown in Fig. 5(a). The boundary modification improves the accuracy for the CIFAR-10 classification by compensating for problems due to device nonlinearity (Fig. 5(b)).

When the trained weights are transferred to the synaptic devices, the conductance of devices is often not tuned to the target value of the weights. Assuming that the conductance values of the *n*-type GSDs at a specific target value are distributed in a Gaussian distribution, we evaluate the performance of the CNS for the conductance variation of these devices (Fig. 6(a)). In addition, accuracy of the CNS with respect to the ratio of the number of stuck-at-off state synapses to the total number of synapses is evaluated (Fig. 6(b)). If the stuck-at-off ratio is greater than about 8%, the accuracy is severely degraded.

4. Conclusions

We have investigated the performance of a convolutional neuromorphic system (CNS) with adaptive weight quantization for CIFAR-10 data classification under various conditions. It has been shown that, depending on the increasing nonlinearity factor of synaptic devices, how much synaptic weights are normalized to the maximum weight greatly affects accuracy. When applying the proposed adaptive weight quantization method to a nonlinear synapse device, we modified the quantization boundary to reduce the number of weights between G (0) and G (1) quantized to zero, resulting in improved accuracy. The accuracy of the CNS was also studied in terms of conductance variation and stuck-at-off ratio of synaptic devices.

Acknowledgements

This work was supported in part by the Brain Korea 21 Plus Project in 2019 and in part by the National Research Foundation of Korea under Grant NRF-2016M3A7B4909604.

References

[1] S. Yu, *Proc*. IEEE, vol. 106, no 2, p. 260-285, 2018

[2] D. Kwon, et al, IEEE Trans. on Electron Devices, vol. 66, no. 1, pp. 395-491, 2019.

[3] S. Lim, et al, International Joint Conference on Neural Networks (IJCNN), 2019, accepted

[4] J.-H. Bae, et al., IEEE Electron Device Lett., vol. 38, no. 8, pp. 1153-1156, 2017.

[5] S. Ioffe, et al, Arxiv, 2015

[6] B. Rueckauer et al, Front. Neurosci., vol. 11, 2017

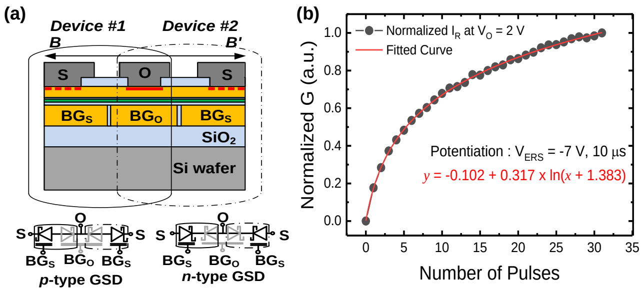

Fig. 1. (a) Schematic cross-sectional view of GSDs. (b) Measured conductance behavior with respect to identical pulses applied to the *n*-type GSDs.

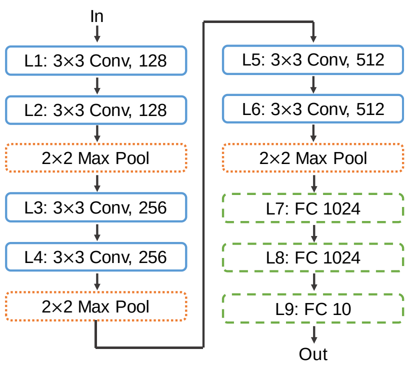

Fig. 2. Structure of convolutional neural networks for CIFAR-10 data classification

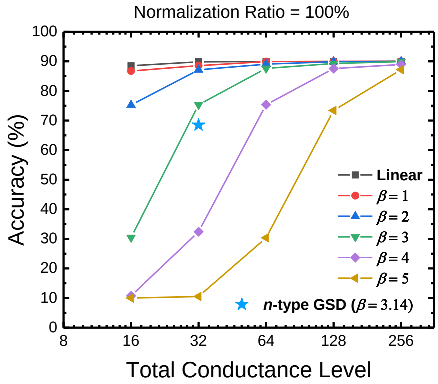

Fig. 3. Accuracy versus total conductance level as a parameter of nonlinearity factor (β) in CNS using adaptive weight quantization. The star symbol represents the accuracy for the *n*-type GSD with a β of 3.14.

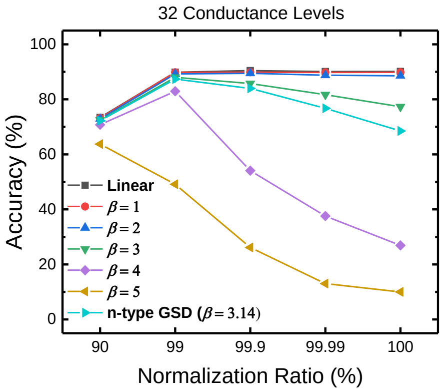

Fig. 4. Accuracy versus normalization ratio as a parameter of β at a fixed conductance level of 32.

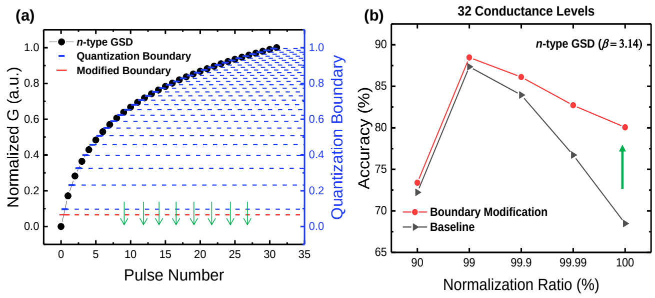

Fig. 5. (a) Normalized G versus the number of pulses to show the effect of the quantization boundary modification. (b) Accuracy versus normalization ratio according to modification of quantization boundary in CNS.

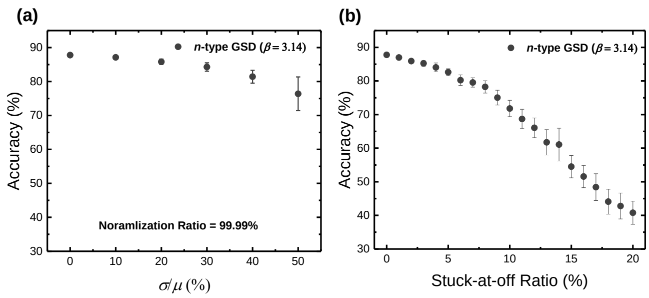

Fig. 6. (a) Accuracy of the CNS with respect to conductance variation (σ/μ). (b) Accuracy of the CNS with stuck-at-off ratio. The error bars indicate one standard deviation of the accuracy in each case.