Extended Abstracts of the 2019 International Conference on Solid State Devices and Materials, Nagoya, 2019, pp395-396

Analyzation of pruning in spiking neural networks trained by STDP and Back propagation

Seongbin Oh[1], Soochang Lee[1], Jang-Saeng Kim[1], Byung-Gook Park[1] and Jong-Ho Lee[1]

[1] Seoul National University Electrical and Computer Engineering Department
Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
Phone: +82-2-880-7285 E-mail: jhl@snu.ac.kr

Abstract

Pruning is an effective technique to make network sparse. In this paper, we construct SNNs with NOR flash type synaptic devices and analyze the effect of pruning. The results show trade-off between performance and power efficiency in network, but in the optimized point, power can be reduced without significant accuracy loss. Moreover, we also propose asymmetric pruning that can reduce power consumptions more efficiently.

1. Introduction

Recently, spiking neural networks (SNN) have been widely studied to compose low power cognitive systems [1]. SNN mimics the energy efficient system, human brain, and processes data through spikes, which is superior to conventional deep learning in terms of power consumption. The size of the network is getting bigger and more complex to achieve higher accuracy [2]. Therefore, research on pruning that constructs a sparse network by excluding non-critical synapses is actively studied [3]. In this paper, we analyze the effect of pruning in SNN composed of fabricated synaptic devices.

2. Hardware neural network

A synaptic device based on NOR flash memory was reported to compose artificial synaptic array in [4]. As shown in Fig. 1 (c), an asymmetric floating-gate acts as a charge storage layer to shift the threshold voltage of the device. The device is compatible with CMOS process, and is able to be operated at a low voltage due to its geometrical characteristics. As shown in Fig. 1 (d), the device is erased and programmed by the voltage difference between gate and source, which corresponds to the long-term potentiation (LTP) and depression (LTD) in neural systems. The network composed of the artificial synapses compute vector-matrix-multiplication very efficiently by biasing the input to gate and summing the current to the drain.

In the network, synaptic weight can be trained by various methods. First, weights are optimized by learning using back-propagation and transferred to the array. Only inference process is conducted by SNN. The method, also known as off-chip learning, has the advantage that it can reduce power consumption without degradation of the performance of the network. On the other hand, networks can be trained through bio-inspired method, STDP. In this paper, simplified STDP is studied to train network reliably [5]. The learning process is depicted in Fig. 2 (a). If a neuron fires, the feedback signal is applied to the source line. As like in Fig. 2 (b), the weight of synapses contributed to the neuron's fire are increased and the others are decreased [6]. This method has a big advantage that the network can self-cluster similar patterns without supervision. The performance of this unsupervised STDP based SNN is shown in Fig. 3(a).

3. Analyzation on pruning

After training process, non-critical synapses are pruned for getting sparsity in the network. In conventional studies, synapses whose weights are lower than pruning threshold are completely excluded. However, in this paper, the weights are set to the minimum value and continuously included to learning. The process of training is shown in Fig. 3 (b) and (c). Through the retraining process between pruning, the accuracy is recovered from distorted weight distribution.

The effect of pruning in the network is shown in Fig. 4. We compare the recognition accuracy and the number of spikes required for inferencing MNIST data set. The number of spikes is a parameter indicating the power used in inference. Fig. 4 (a) and (b) shows the results for the network trained by STDP. In this case, only the excitatory synapses were considered. As the pruning threshold increases, both accuracy and the number of spikes tend to decrease. The best accuracy shows 89.4% in accuracy, but pruning results in 0.86% drop in accuracy with 1.2 times reduction in energy consumption. With further pruning, accuracy dropped in 5.84% and power is decreased 2.33 times. The results for the SNN trained by back propagation (BP) are shown in Fig. 4 (c) and (d). In this case, excitatory and inhibitory synapses are pruned equally. Likewise, the best accuracy was 97.78%, but pruning caused an accuracy drop of 0.57%, resulting in a 1.84 times power reduction. With further pruning, power is reduced 2.02 times with an accuracy drop of 1.12%. Compared to the network trained by STDP, it shows relatively less degradation of accuracy with pruning.

In SNN trained by BP, inhibitory synapses help to reduce the number of spikes by discharging I&F circuits. In other words, pruning inhibitory synapses less than excitatory synapses can give more reduction of energy consumption. The effect of asymmetric pruning is analyzed in Fig. 5. Asymmetric pruning results in degradation of the network's performance. Especially, excessive asymmetric pruning makes it impossible to fire, so that the network shows poor accuracy. However, proper asymmetric pruning reduces the spike number in both layers without significant loss of accuracy. In Fig. 6, comparison between two types of pruning is made. Even on networks with similar accuracy, the asymmetric pruning

consumes much less power in the inference process. It indicates that reduction of energy consumption can be achieved more effectively by adopting asymmetric pruning.

4. Conclusion

In this paper, we have analyzed the effect of pruning in SNN composed with asymmetric floating-gate based synaptic devices. The network was trained by two manners, STDP and BP. The accuracy and the spike numbers show trade-off relationship with pruning threshold, but an optimization point can be found. Also, we proposed asymmetric pruning. Setting different pruning thresholds in excitatory and inhibitory parts of the synapses enabled more effective pruning.

Acknowledgements

This work was supported by the MOTE (10080583) and KSRC support program for the development of the future semiconductor device and the Brain Korea 21 Plus project

References

[1] S. Yu *et al.*, 2012 Int. Electron Devices Meeting **10** (2012) 4
[2] P. U. Diehl *et al.*, Front. Comput. Neurosci. **9** (2015) 99.
[3] N. Rathi *et al.*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (2019) 38(4).
[4] C.H. Kim *et al.*, IEEE. Trans. Electron Devices **65** (2018).
[5] D Querlioz *et al.*, Proc. Int. Jt. Conf. Neural. Netw. (2011) 65.
[6] S. Lee *et al.*, J. Nanosci. Nano tecnol. **19** (2019) 10.

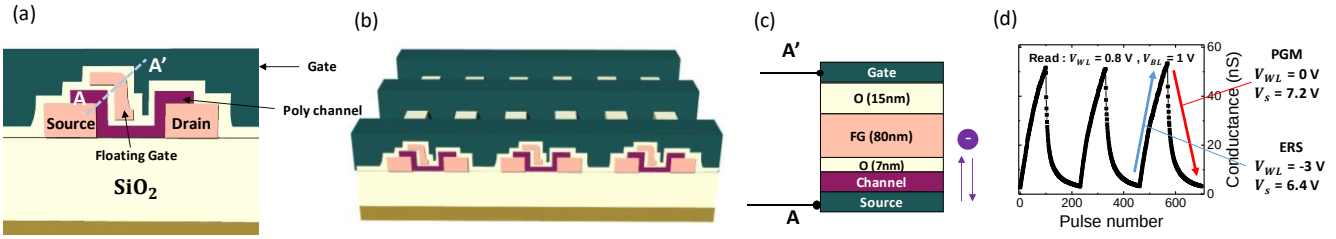

Fig. 1. Cross-sectional views of a TFT-type NOR flash memory cell cut in the (a) WL direction, and (b) bird's eye view of the NOR flash array. (c) Gate stack of the synaptic device. (d) Measured LTP and LTD characteristics of the device with the number of applied pulses. Program and erase pulses are applied between gate and source electrode.

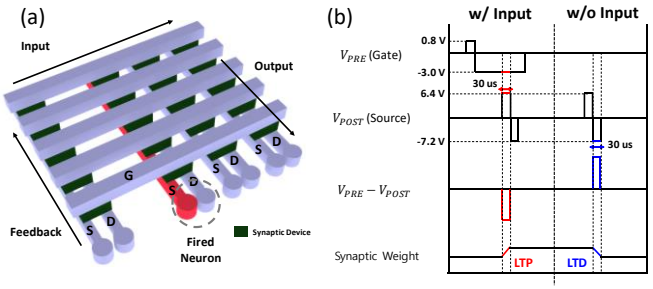

Fig. 2. (a) 3D schematic view to illustrate training in neural network. (b) Pulse scheme that enables pattern training by using simplified STDP rule.

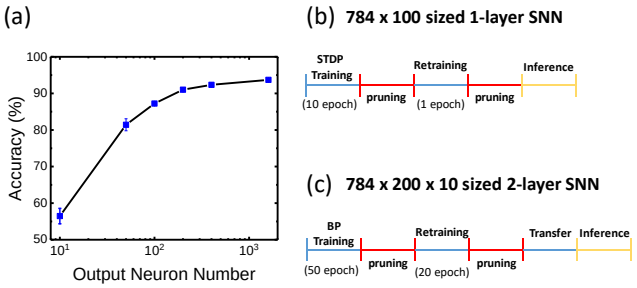

Fig. 3. (a) Recognition accuracy of the network versus output neuron number. Process for pruning in SNN trained by (b) STDP and (c) BP algorithms.

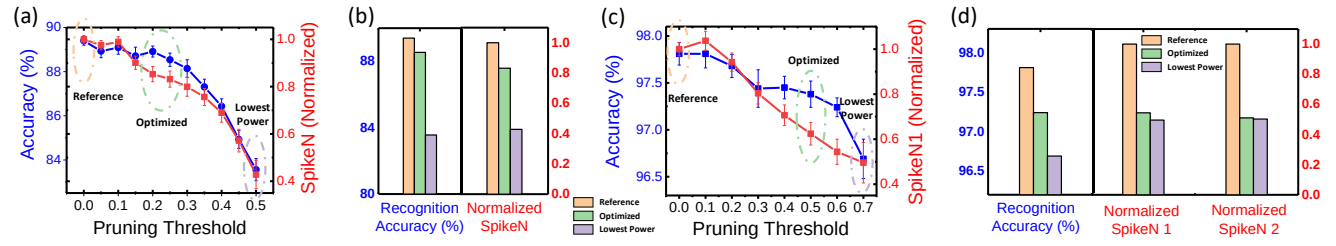

Fig. 4. The recognition accuracy and the required number of spikes in inferencing process in SNN trained by STDP rule (a), (b) and BP algorithm (c), (d). SpikeN1 and SpikeN2 represent the number of spikes in the hidden and output layers, respectively.

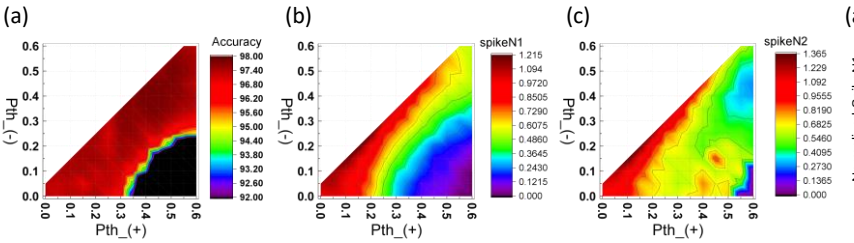

Fig. 5. (a) Accuracy and (b), (c) the number of spikes in each layer based on asymmetric pruning.

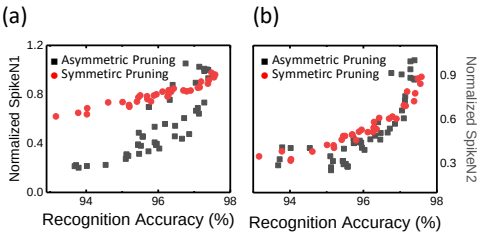

Fig. 6. The spike number in each layer vs recognition accuracy of the network under two pruning conditions.