Solid State Electronics 165 (2020) 107742

Contents lists available at ScienceDirect

Solid State Electronics

journal homepage: www.elsevier.com/locate/sse

Initial synaptic weight distribution for fast learning speed and high recognition rate in STDP-based spiking neural network

Jangsaeng Kim[a], Chul-Heung Kim[a], Sung Yun Woo[a], Won-Mook Kang[a], Young-Tak Seo[a], Soochang Lee[a], Seongbin Oh[a], Jong-Ho Bae[b], Byung-Gook Park[a], Jong-Ho Lee[a,*]

[a] Department of Electrical and Computer Engineering and the Inter-University Semiconductor Research Center (ISRC), Seoul National University, Seoul 151-742, Republic of Korea
[b] Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720, USA

ARTICLE INFO

The review of this paper was arranged by Jung-Hee Lee

Keywords:
Initial synaptic weight distribution
Spike-timing-dependent plasticity (STDP)
Homeostasis functionality
NOR flash memory
Spiking neural networks

ABSTRACT

We analyze that the initial synaptic weight distribution affects the performance, such as the learning speed, recognition rate and the power consumption in the spiking neural networks (SNNs) based on spike-timing-dependent plasticity (STDP) learning rule. A thin-film transistor (TFT)-type NOR flash memory is used as a synaptic device. In this fully connected two-layer neuromorphic system using the proposed pulse scheme, the results with and without the homeostasis functionality were analyzed separately. In addition, power consumption of the network in various initial synaptic weight distributions, and recognition rate that varies with the number of output neurons are also investigated. In pattern recognition for 28 × 28 MNIST handwritten patterns, higher performance is achieved in various aspects when the initial synaptic weights are distributed near the maximum value.

1. Introduction

In recent years, neuromorphic computing system has attracted a great deal of attention as an alternative of the von Neumann architecture and is being rapidly developed with the exponential data growth [1]. Especially, in the field of software, deep neural networks (DNNs) based on back-propagation algorithms show high recognition rate in pattern recognition [2,3]. Moreover, various efforts have been made to implement a hardware neural network (HNN) using electronic synaptic devices for low power consumption and improved speed [4,5]. As another approach to implementing HNN, spiking neural networks (SNNs) based on spike-timing-dependent plasticity (STDP) learning rule are being studied, and inference performance has been investigated in simulations [6–10]. In order to improve the performance of the network, various attempts have been made in both neural networks. In the case of DNNs, various synaptic weight initialization methods have been studied to improve the performance of the network [11–13]. However, no studies have been reported on the effect of the initial synaptic weight distribution in SNNs.

Previously, we proposed a TFT-type NOR flash memory array as a synaptic device [14] and the unsupervised online learning based on two-layer fully connected SNN was performed [15]. In this work, the impact of the initial synaptic weight distribution in proposed neural network is investigated in various aspects such as learning speed, recognition rate, and power consumption. Furthermore, the results with and without the homeostasis are analyzed respectively. In addition, training and inference are carried out on the various number of output neurons.

2. Device characteristics and network design

Fig. 1(a) shows the schematic 3-D array view of a TFT-type NOR flash memory cells as a synaptic device. A half-covered n^+ poly-Si floating gate (FG) is formed as a charge storage layer between the cross-point of the word line (WL) and the source line (SL). The thicknesses of the poly-Si active layer, tunneling SiO_2 layer, blocking SiO_2 are 20 nm, 7 nm, and 15 nm, respectively. The width of the control gate is 2 μm and the length between the source and drain is 0.5 μm. One synaptic device can be scaled down to $8F^2$ if the width of the control gate is scaled to the minimum feature size (F). The crossbar arrangement of synaptic devices is advantageous in terms of scaling of synaptic devices. Fig. 1(b) shows the measured conductance change of the synaptic device as the number of pulses applied to the WL and SL. 50 repeated erase pulses ($V_G = -3$ V, $V_S = 5$ V) and 300 repeated program pulses

* Corresponding author.
E-mail address: jhl@snu.ac.kr (J.-H. Lee).

https://doi.org/10.1016/j.sse.2019.107742

Available online 12 December 2019

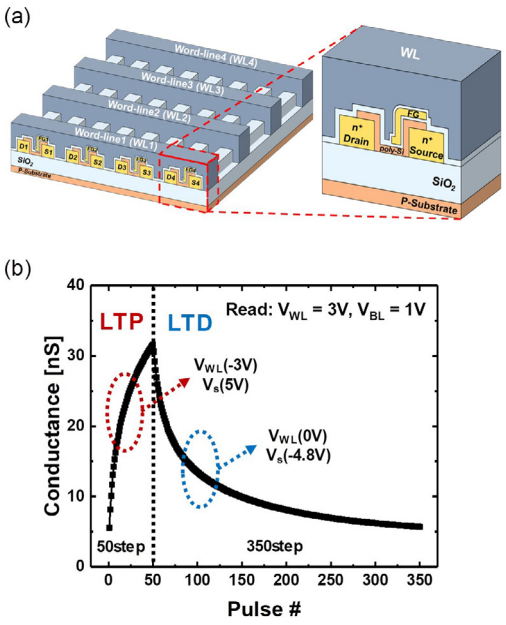

Fig. 1. (a) Crossbar array of TFT-type NOR flash memory cells used as synaptic devices. (b) Measured LTP/LTD characteristics of proposed device.

(V_G = 0 V, V_S = −4.8 V) are applied. The conductance change represents the measured long-term potentiation (LTP)/long-term depression (LTD) characteristics of the proposed synaptic device.

The pulse scheme of presynaptic (PRE) and postsynaptic (POST) neurons for selective STDP-based weight update depending on the overlapped voltage between PRE (X_{pre} = V_G) and POST (X_{post} = V_S) neurons is represented in Fig. 2(a). When a POST neuron fires, a POST feedback pulse is applied to all connected synaptic devices through a common SL. The weights of synaptic devices contributing to the spike of the POST neuron are potentiated under the erase condition (X_{pre} = −3 V, X_{post} = 5 V). The weights of the other synapses are depressed under the program condition (X_{pre} = 0 V, X_{post} = −4.8 V) since only POST feedback pulse is applied. The weights of synapses connected to the not fired POST neurons do not change. Under these conditions, the simplified STDP learning rule used for updating synaptic weights can be obtained as illustrated in Fig. 2(b). In this figure, r_{LTP} and r_{LTD} represent the increase and decrease of the weight magnitudes, respectively.

Fig. 3(a) shows the schematic illustration of a fully connected two-layer system with a 28 × 28 input neurons and multiple output neurons. Output neurons exploited leaky integrate-and-fire (LIF) neuron model and lateral inhibition functionality. The system level pattern learning simulation was performed by the software Python using the measured LTP/LTD characteristics of a TFT-type NOR flash memory cell in Fig. 1(b) and simplified STDP learning rule in Fig. 2(b). The neural circuit and operation mechanism used in this simulation are the same as those used in the previous work [16]. In this simulation, full binary MNIST datasets consisting of 60,000 training data and 10,000 test data is used. In all cases, the learning is done in 3 epochs since the recognition rate saturates after 3 epochs. The uniform random distribution of the initial synaptic weight is divided into four parts between the minimum weight (W_{min}) and the maximum weight (W_{max}) as shown in Fig. 3(b).

3. Results and discussion

3.1. Learning speed and recognition rate

Fig. 4(a) shows the recognition rate of the proposed fully connected SNN with 500 output neurons as a parameter of the initial synaptic

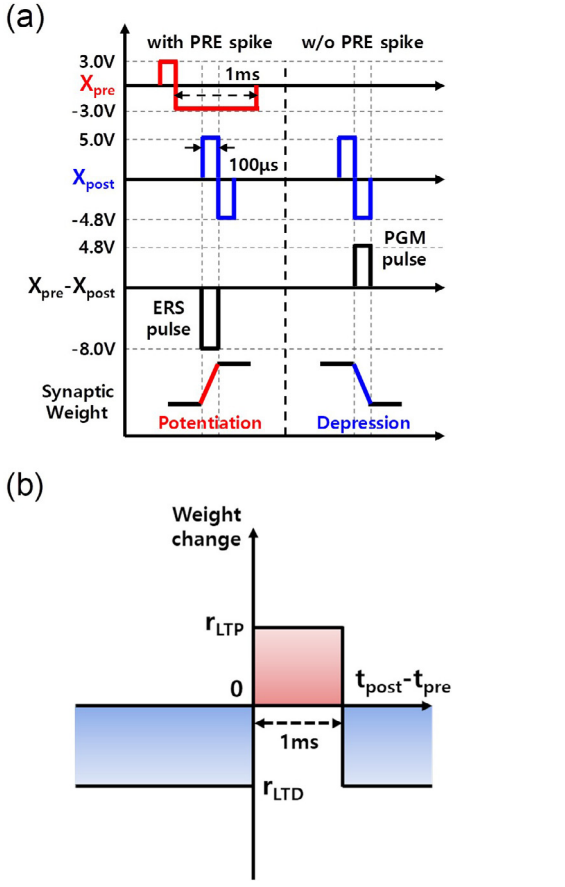

Fig. 2. (a) Pulse scheme of PRE and POST neurons for STDP learning rule. (b) Simplified STDP learning rule.

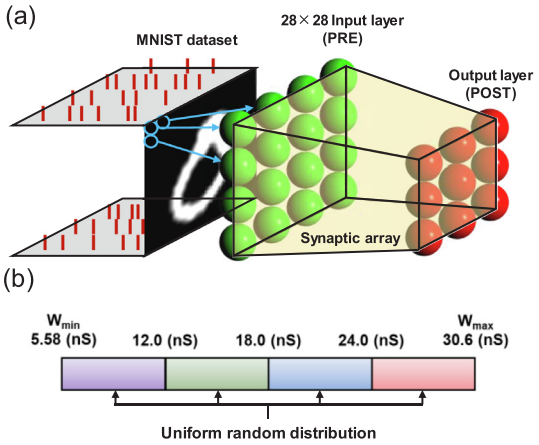

Fig. 3. (a) Schematic illustration of a fully connected two-layer neuromorphic system. (b) Initial synaptic weight distributions.

weight distribution in Fig. 3(b). The homeostasis functionality [17] is not used in this simulation. It is clearly seen that significantly higher recognition rate (~92%) and faster learning speed are achieved when the initial synaptic weights are distributed near the maximum value. Higher initial synaptic weights lead to a rapid increase in membrane potential, resulting in less time consumption for learning. The learning time of the red line in Fig. 4(a), where the initial synaptic weight distribution is highest, is about 7 and 23 times less than those of the blue and black lines in Fig. 4(a) for the lower weight distribution. Here, the learning time is defined as the time taken to reach 90% of the saturation recognition rate. On the other hand, when the initial synaptic weights

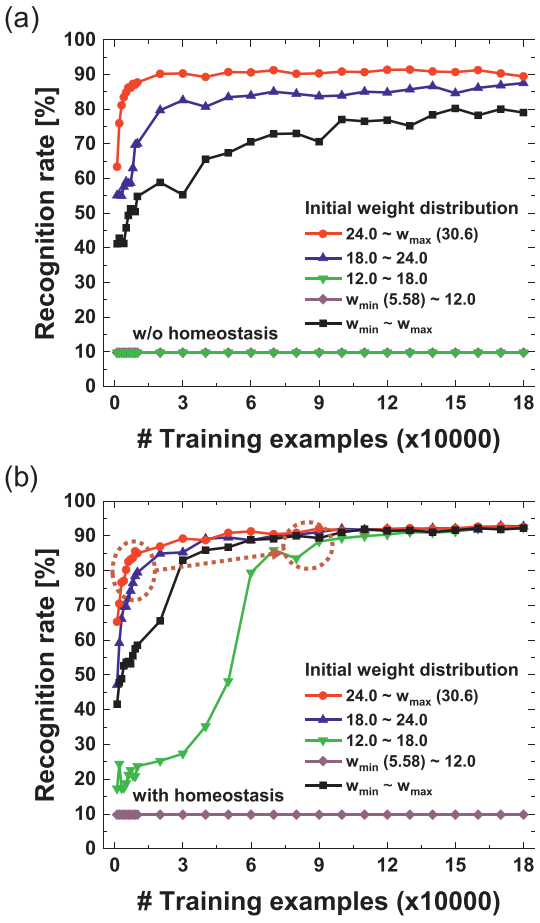

Fig. 4. (a) Recognition rate without the homeostasis and (b) with the homeostasis as a parameter of the initial synaptic weight distribution.

are distributed near the minimum value (green and purple lines in Fig. 4(a)), the output neurons are not fired and eventually learning does not work properly.

Fig. 4(b) shows the recognition rate of the same network as Fig. 4(a) as a parameter of the initial synaptic weight distribution. Unlike the previous case, the homeostasis functionality is used in this simulation. The homeostasis functionality adopts the method used in the previous research [17]. If the firing rate of a particular output neuron exceeds the target firing rate, the threshold voltage of the neuron is increased and in the opposite case, the threshold voltage of the neuron is decreased. When the homeostasis functionality is used in the output neurons, the recognition rates for all initial synaptic weight distributions saturate to a relatively higher value (~93% when the initial synaptic weights are distributed near the maximum value) than that of the network without the homeostasis functionality. However, learning the network requires much more time and more training data. The learning time of the red line in Fig. 4(b), where the initial synapse weight distribution is highest, is about 10 times less than that of the green line in Fig. 4(b) for the lower weight distribution. Even if the initial synaptic weight distribution consists of low values, the recognition rate can increase to a relatively higher value after several iterations (green line in Fig. 4(a) and (b)). However, when the initial synaptic weights are distributed near the minimum value (purple line in Fig. 4(b)), the output neurons are still not fired although homeostasis functionality is applied. This is because there is not enough weighted sum for the neuron to fire for the time the input comes in, and the membrane potential of the output neuron is reset to zero when the next input comes in.

In the proposed SNN, when the weights of the synaptic devices are

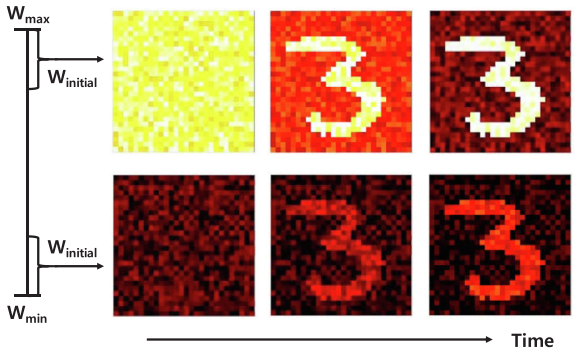

Fig. 5. Changes in weight map for the same time period under high and low initial synaptic weight distribution conditions.

updated by an output spike from a fired neuron, the synaptic devices to which the input signal was applied are potentiated and the others are depressed. LTD characteristic in LTP/LTD curve is more abrupt than LTP near the maximum value of synaptic weight. In addition, background region is relatively larger than pattern region. Moreover, high initial synaptic weight distribution facilitates fast integration in the membrane capacitor and therefore output neurons fire more often. Thus, trained synaptic weight maps are formed more quickly in learning certain patterns using a high initial synaptic weight distribution. These phenomena can be seen in Fig. 5. Besides, the SNN without the homeostasis functionality shows a similar recognition rate to that of the SNN with the homeostasis functionality when the initial synaptic weights are distributed near the maximum value as represented in Fig. 4(a) and (b). There is no depression in the weight of synapses connected with neurons that do not fire, so the likelihood of firing will still be high enough due to the fast integration in the membrane capacitor even if homeostasis is not used, resulting in a homeostasis functionality. This means that by setting the initial synaptic weight distribution near the maximum value, the circuitry needed to implement the homeostasis functionality can be saved. This can bring additional benefits in terms of power consumption and area occupation. In short, when the initial synaptic weights are distributed at a higher value, considerably high performance is achieved in various aspects such as the learning speed and recognition rate regardless of whether the homeostasis functionality is used or not.

3.2. Power consumption

Fig. 6(a) and (b) represent the power consumption of the proposed fully connected SNN with homeostasis functionality until the recognition rate saturates as a parameter of the initial synaptic weight distribution in Fig. 3(b). The power consumption during the training the network is compared through the amount of charge charged and discharged in the membrane capacitor (Fig. 6(a)) and the number of times the synaptic weights have been updated (Fig. 6(b)). The amount of consumed power is normalized based on the smallest value (red bar in Fig. 6(a) and (b)) in this figure.

When the initial synaptic weights are highly distributed, the amount of charge charged and discharged in the membrane capacitor is up to 7 times smaller than that of the lowest initial synaptic weight distribution as shown in Fig. 6(a). Similarly, Fig. 6(b) shows that the number of weight updates for the highest synaptic weight distribution is reduced by 6.5 and 5.6 times for potentiation and depression, respectively, compared to those with the lowest synaptic weight distribution. This means that a higher initial synaptic weight distribution is beneficial in terms of power consumption and less training data is needed for network learning.

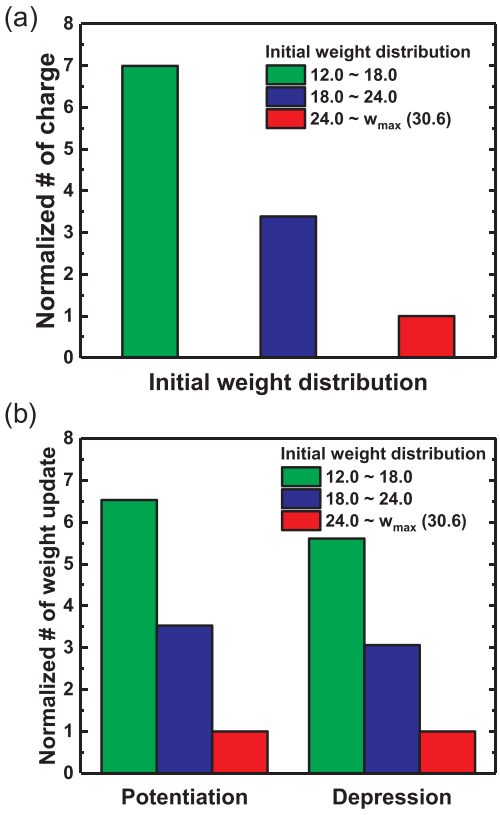

Fig. 6. Normalized power consumption of the network obtained by (a) the number of charge charged and discharged in the membrane capacitor and (b) the number of times the synaptic weights were updated until the recognition rate is saturated.

3.3. Change in the number of output neurons

The recognition rate versus the number of output neurons is also investigated as shown in Fig. 7(a), (b), and (c). The three figures are illustrated by different initial synaptic weight distributions. Solid lines represent the recognition rate with the homeostasis functionality, while dashed lines represent the recognition rate without the homeostasis functionality. As the number of output neurons increases, the recognition rate increases for all initial synaptic weight distributions similar to previous study [15]. As the initial synaptic weight distribution becomes lower, the recognition rate of the network without the homeostatic functionality becomes significantly lower than the recognition rate with the homeostatic functionality, as shown in Fig. 7(b) and (c). However, Fig. 7(a) shows that when the initial synaptic weights are distributed near the maximum value, it has a similar recognition rate regardless of whether the homeostasis functionality is used or not. The result shows the significance of the initial synaptic weight distribution clearly.

4. Conclusion

In this paper, we have investigated the impact of the initial synaptic weight distribution in various aspects and proposed a method to achieve high performances in 2-layer spiking neural networks. Simulation results show that higher initial synaptic weight distribution is beneficial in terms of recognition rate, learning speed and power consumption regardless of whether or not the homeostasis functionality is used. In particular, learning speed and power consumption were improved almost 10 times and 7 times, respectively. These results show that the weight initialization should be considered important in spiking neural networks.

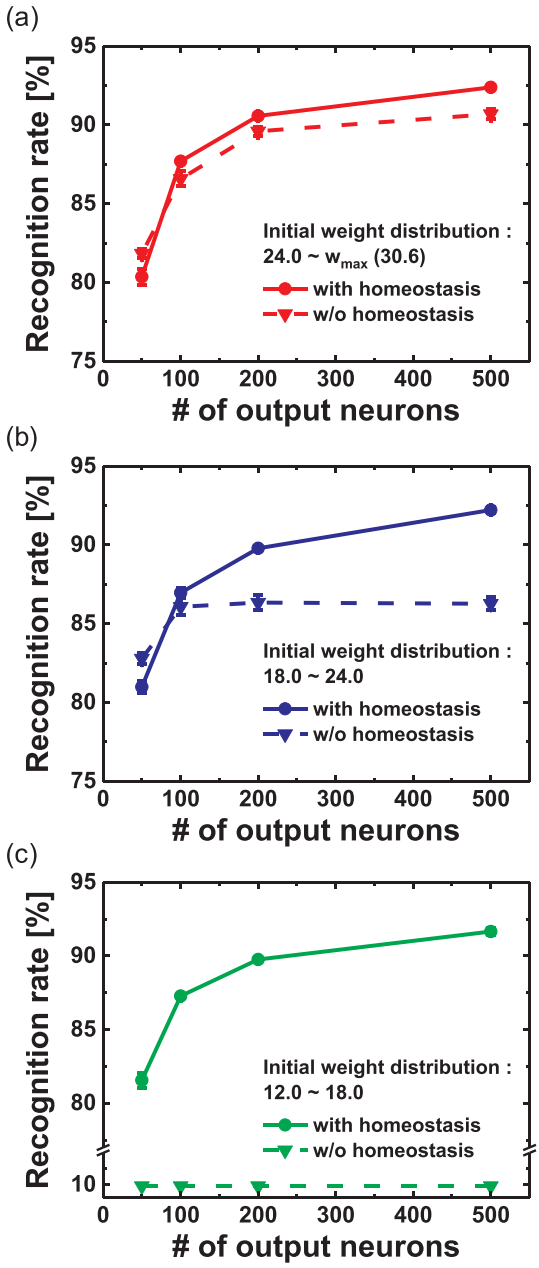

Fig. 7. Recognition rate versus the number of output neurons when the initial synaptic weights are distributed near the (a) maximum value, (b) medium value and (c) minimum value.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Brain Korea 21 Plus Project in 2019, and National Research Foundation of Korea (NRF-2016M3A7B4909604).

References

[1] Indiveri G, Liu S-C. Memory and information processing in neuromorphic systems. Proc IEEE Aug. 2015;103(8):1379–97.

[2] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–6.
[3] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015.;521:436–44.
[4] Burr GW, Shelby RM, Sidler S, di Nolfo C, Jang J, Boybat I, et al. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. IEEE Trans Electron Devices 2015;62(11):3498–507.
[5] Merolla PA, Arthur JV, Rodrigo A-I, Cassidy AS, Sawada J, Akopyan F, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science 2014;345(6197):668–73.
[6] Masquelier T, Thorpe SJ. Learning to recognize objects using waves of spikes and Spike Timing-Dependent Plasticity. Int Joint Conf Neural Networks IJCNN 2010:1–8.
[7] Suri M, Bichler O, Querlioz D, Cueto O, Perniola L, Sousa V, et al. Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction. IEDM Tech Dig 2011:79–82.
[8] Carlos ZR, Luis C-MA, Jose P-CA, Timothee M, Teresa S-G, Bernabe L-B. On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex. Front Neurosci 2011;5:26.
[9] Yu S, Gao B, Fang Z, Yu H, Kang J, Wong H-SP. A neuromorphic visual system using RRAM synaptic devices with Sub-pJ energy and tolerance to variability: experimental characterization and large-scale modeling. IEDM Tech Dig 2012:239–42.
[10] Sidler S, Pantazi A, Wozniak S, Leblebici Y, Eleftheriou E. Unsupervised learning using phase-change synapses and complementary patterns. Int Conf Artif Neural Networks (ICANN) 2017:281–8.
[11] Yam JYF, Chow TWS. A weight initialization method for improving training speed in feedforward neural network. Neurocomputing 2000;30(1–4):219–32.
[12] Mercedes F-R, Carlos H-E. Weight initialization methods for multilayer feedforward. Eur Symp Artif Neural Netw (ESANN) 2001:119–24.
[13] Mercedes F-R, Carlos H-E. A comparison among weight initialization methods for multilayer feedforward networks. Int Joint Conf Neural Netw (IJCNN) 2000:543–8.
[14] Kim C-H, Lee S, Woo SY, Kang W-M, Lim S, Bae J-H, et al. Demonstration of unsupervised learning with spike-timing-dependent plasticity using a TFT-type NOR flash memory array. IEEE Trans Electron Devices (TED) 2018;65(5):1774–80.
[15] Lee S, Kim C-H, Oh S, Park B-G, Lee J-H. Unsupervised online learning with multiple postsynaptic neurons based on spike-timing-dependent plasticity using a thin-film transistor-type NOR flash memory array. J Nanosci Nanotechnol (JNN) 2019;19(10):6050–4.
[16] Kang W-M, Kim C-H, Lee S, Woo SY, Bae J-H, Park B-G, et al. A spiking neural network with a global self-controller for unsupervised learning based on spike-timing-dependent plasticity using flash memory synaptic devices. Int Joint Conf Neural Netw (IJCNN) 2019:1–7.
[17] Querlioz D, Bichler O, Dollfus P, Gamrat C. Immunity to device variations in a spiking neural network with memristive nanodevices. IEEE Trans. Nanotechnol 2013;12(3):288–95.

Jangsaeng Kim received the B.S. degree in electrical engineering from Seoul National University (SNU), Seoul, Korea, in 2018, where he is currently pursuing the M.S. degree with the Department of Electrical and Computer Engineering. He is with the Inter-University Semiconductor Research Center, SNU. His current research interests include neuromorphic system and its application in computing.

Chul-Heung Kim received the B.S. degree in electrical engineering from Seoul National University (SNU), Seoul, Korea, in 2013, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. He is with the Inter-University Semiconductor Research Center, SNU. His current research interests include neuromorphic system and its application in computing.

Sung Yun Woo received the B.S. degree in electrical engineering from Kyungpook National University, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University (SNU), Seoul, Korea. He is with the Inter-University Semiconductor Research Center, SNU. His current research interests include neuromorphic system and neural networks.

Won-Mook Kang received the B.S. degree in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University (SNU), Seoul, Korea. His research interests include neuromorphic system.

Young-Tak Seo received the B.S. degree in electrical engineering from Seoul National University (SNU), Seoul, Korea, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. He is with the Inter-University Semiconductor Research Center, SNU. His current research interests include neuromorphic system and its application in computing.

Soochang Lee received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His current research interests include neuromorphic systems and its application in computing.

Seongbin Oh received the B.S. degree in electrical engineering from Seoul National University (SNU), Seoul, Korea, in 2018, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. He is with the Inter-University Semiconductor Research Center, SNU. His current research interests include neuromorphic system and its application in computing.

Jong-Ho Bae received the B.S. degree in Electrical Engineering from Pohang University of Science and Technology (POSTECH), in 2011 and the Ph.D. degree in Electrical and Computer Engineering from Seoul National University (SNU), Seoul, Republic of Korea. He is currently with EECS, University of California, Berkeley. His current research interests include NCFET, FeFET, Sensor, and Neuromorphic Computing.

Byung-Gook Park (M'90) received the B.S. and M.S. degrees in electronics engineering from Seoul National University (SNU), Seoul, South Korea, in 1982 and 1984, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1990. He joined as an assistant professor with the Department of Electrical and Computer Engineering, SNU, in 1994, where he is currently a professor.

Jong-Ho Lee (SM'01-F'16) received the Ph.D. degree from Seoul National University (SNU), Seoul, South Korea, in 1993, in electronic engineering. He was a Post-Doctoral Fellow with the Massachusetts Institute of Technology, Cambridge, MA, USA, from 1998 to 1999. He has been a professor with the School of Electrical and Computer Engineering, SNU, since 2009. He is a Lifetime Member of the Institute of Electronics Engineers of Korea (IEEK) and IEEE Fellow.