electronics

MDPI

Article

An On-Chip Learning Method for Neuromorphic Systems Based on Non-Ideal Synapse Devices

Jae-Eun Lee [1], Chuljun Lee [2], Dong-Wook Kim [1], Daeseok Lee [1] and Young-Ho Seo [1,*]

[1] Department of Electronic Materials Engineering, Kwangwoon University, 20, Gwangun-ro, Nowon-Gu, Seoul 01897, Korea; wodms6364@naver.com (J.-E.L.); dwkim@kw.ac.kr (D.-W.K.); leeds@kw.ac.kr (D.L.)

[2] Department of Materials Science & Engineering, Pohang University of Science and technology (POSTECH), 77 Cheongam-Ro, Nam-Gu, Pohang, Gyeongbuk 37673, Korea; leecj@postech.ac.kr

* Correspondence: yhseo@kw.ac.kr

Received: 13 October 2020; Accepted: 16 November 2020; Published: 18 November 2020

Abstract: In this paper, we propose an on-chip learning method that can overcome the poor characteristics of pre-developed practical synaptic devices, thereby increasing the accuracy of the neural network based on the neuromorphic system. The fabricated synaptic devices, based on $Pr_{1-x}Ca_xMnO_3$, $LiCoO_2$, and TiO_x, inherently suffer from undesirable characteristics, such as nonlinearity, discontinuities, and asymmetric conductance responses, which degrade the neuromorphic system performance. To address these limitations, we have proposed a conductance-based linear weighted quantization method, which controls conductance changes, and trained a neural network to predict the handwritten digits from the standard database MNIST. Furthermore, we quantitatively considered the non-ideal case, to ensure reliability by limiting the conductance level to that which synaptic devices can practically accept. Based on this proposed learning method, we significantly improved the neuromorphic system, without any hardware modifications to the synaptic devices or neuromorphic systems. Thus, the results emphatically show that, even for devices with poor synaptic characteristics, the neuromorphic system performance can be improved.

Keywords: neuromorphic; neural network; synapse device; quantization; on-chip training

1. Introduction

Recently, deep learning has been implemented in systems central processing units (CPUs) and graphics processing units (GPUs), and has performed successfully in most fields that utilize an artificial neural network (ANN) [1,2]. However, bottlenecks are often caused by the strict hardware requirements, excessive memory access, and high power consumption [3,4]. Consequently, neuromorphic systems have attracted attention recently as an alternative to replace the Von-Neumann architecture [5–7]. Various types of synaptic devices have been researched for implementing these neuromorphic system [8–12]. Wong et al. studied in-memory computing based on Re-RAM after analyzing the structure of Von Neumann [13]. Qian et al. proposed a parallel convolution structure using memristor devices [14]. High energy efficiency was verified using this method, and the possibility as a next-generation computing was shown. Improving the undesirable synaptic characteristics remains a significant challenge, as nonlinear, discontinuous, and asymmetric conductance changes of potentiation and depression produce critical failures in on-chip learning performance [15,16].

Improving system reliability for non-ideal cases is also important. Generally, synaptic devices have a critical limitation on the levels of conductance they can access because the conductance has a certain

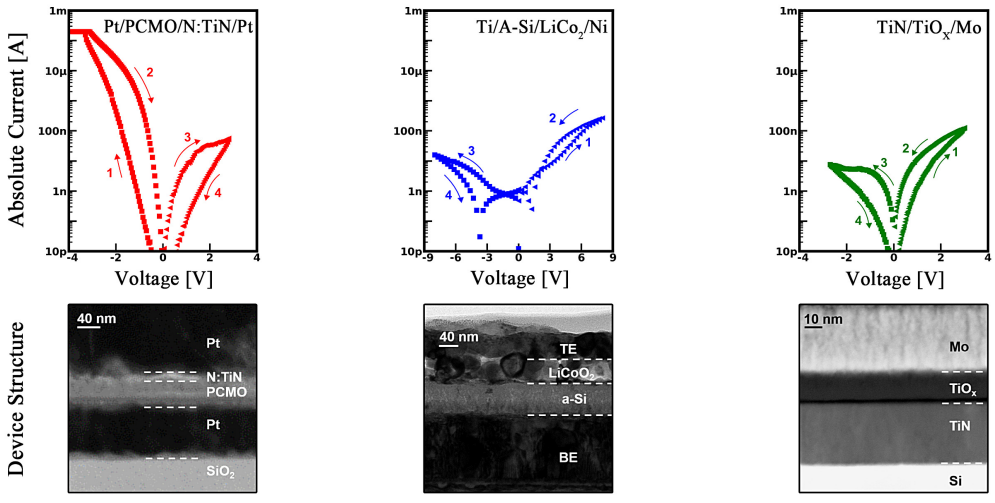
variation and uncertainty. Thus, many studies have been conducted to improve the performance of the overall neuromorphic system using the poor characteristics of such devices [17,18]. Kwon et al. [17] proposed an off-chip training method, such as changing the weight after training, so it is significantly different from this paper, which proposed an on-chip training method. This paper optimizes weight while reflecting the actual conductance of the device during on-chip training. In addition, Chang et al. [18] proposed learning techniques such as activation functions and threshold weight update scheme, and thus has a different perspective from the proposed method.

Considering these challenges, we have proposed a new method that can improve the performance of the entire neuromorphic system by quantizing the conductance used for learning in a synapse-based neural network (NN). Broadly, this provides a new on-chip training technique for neuromorphic systems. The method presented herein includes a conductance-based linear weighted quantization procedure and on-chip learning approach, intended for use in devices with non-ideal and undesirable characteristics. Using this method, we construct, train, and evaluate an NN that accurately and efficiently predicts the MNIST dataset.

The composition of this paper is as follows: Section 2 briefly introduces the proposed system and the overall experimental process. Section 3 describes three devices fabricated to demonstrate the generality of the proposed method, proposes a quantization system, and compares and analyzes the experimental results. Finally, Section 4 concludes and suggests future research directions.

2. Analysis of Synapse Device

Three types of synaptic devices, based on PCMO, $LiCoO_2$, and TiO_x, were fabricated as shown in Figure 1 and characterized to evaluate and confirm the performance improvements provided by the proposed method. The previously listed synaptic devices had two-terminal structures of Pt/PCMO/N:TiN/Pt, Ti/a-Si/$LiCoO_2$/Ni, and TiN/TiO_x/Mo, respectively, with a detailed description of the fabrication methods provided in our previous works [11,19,20]. Using our conductance-based quantization method, we converted the analyzed conductance values into weights. We then constructed an NN composed of fully-connected layers, which could classify MNIST with the quantized weight. Finally, we demonstrated the effectiveness of the proposed method via training and evaluation.



(**a**) (**b**) (**c**)

Figure 1. Cross-sectional TEM images and I–V characteristics for (**a**) PCMO synaptic devices (Pt/PCMO/N: TiN/Pt); (**b**) $LiCoO_2$ synaptic devices (TE/$Li_{1-x}CoO_2$ /a-Si/BE); and (**c**) TiO_x synaptic devices.

Figure 1 shows the transmission electron microscopy (TEM) images and current-voltage (I–V) characteristics of the applied synaptic devices, which all exhibit conductance changes in response to applied biases. This conductance change in PCMO (Figure 1a) is attributed to the formation, or elimination, of the metal-oxide layer between the PCMO and N:TiN. When a negative (positive) bias is applied, oxygen in the PCMO (TiNOx) migrates to the N:TiN (PCMO) layer, and the sub-oxide layer of TiON is formed (dissolved), causing a change in the device conductance [19].

The $LiCoO_2$ conductance also changes under sequential positive and negative bias application (Figure 1b). $LiCoO_2$ is known as conductance-tunable material because the conductance can be altered by modifying its Li-concentration [11]. In our fabricated Ti (BE)/a-Si/ $LiCoO_2$/Ni (TE) stacked synaptic devices, a positive (negative) bias was applied to the TE (BE). This caused Li-ions within the $LiCoO_2$ (a-Si) to migrate towards the a-Si ($LiCoO_2$), causing the device conductance to decrease (increase).

Additionally, the conductance of TiO_x (Figure 1c) can be modulated via the oxygen concentration. Specifically because oxygen vacancies contribute to the local conducting path, the material conductance is impacted by the amount of oxygen vacancies within the TiO_x layer, where oxygen serves the role of a dopant [20]. Therefore, applying positive (negative) bias to the Mo (TE) of the devices provides a mechanism for controlling the device conductance through the bi-directional migration of oxygen ions.

In addition to the specific devices used in this paper, similar demonstrations with varying conductance levels may be possible. However, for the practicality of focusing the experiments, this paper is limited to the aforementioned devices.

3. Proposed On-Chip Learning

3.1. On-Chip Training

The previous studies and the proposed method are explained in terms of neural networks as follows. In the learning process, Kwon et al. use a method of updating weights by applying quantization after normalizing weights [17]. Chang et al. uses a method of controlling the activation function of neurons and applying a threshold to the weight of each neuron [18]. Each method has a different synaptic device. Gated Schottky diode (GSD) was used in [17], and Ta/HfO_2/Al-doped TiO_2/TiN was used in [18]. In this paper, PCMO, Li, and TiO_x are used. All neural network architectures used a fully connected (FC) layer, and hard-sigmoid, sigmoid, and ReLU were used as activation functions, respectively. The method of [17] is for off-chip training, and the rest is for on-chip training. Table 1 summarizes the characteristics of each study.

Table 1. Comparison of the previous and proposed method.

Items	[17]	[18]	Proposed
Dataset	MNIST	MNIST	MNIST
Synaptic device	Gated schottky diodes	Ta/HfO_2/Al-doped TiO_2/TiN	PCMO, Li, TiO_x
Neural network	784-256-10 FC layer	784-300-10 FC laye	784-300-100-10 FC layer
Activation function	Hard-sigmoid	Sigmoid, ReLU	ReLU
Training	Off-chip	On-chip	On-chip

Training of neural networks refers to the process of finding optimized weights and biases of neurons using loss functions. After performing an operation that quantitatively compares the correct answer and the result of the neural network using the loss function, the weight and bias of the neuron are updated using the result. This process is repeated many times to optimize the neural network. The number of iterations to use all the data of the prepared dataset at once probabilistically is defined as one epoch. In other words, the training process consists of many epochs. In the process of learning (or training),

various experimental conditions for neural networks can be adjusted. In general, adjustable experimental conditions are called the hyperparameters. Various types of learning can be performed depending on the number of hyperparameters included in the learning and learning conditions, and the learning time can be greatly varied. When such learning is performed in a hardware-type neural network (neuromorphic system) chip composed of a manufactured synapse, this learning is called the on-chip training. In order to perform on-chip training, the neuromorphic system hardware should have almost all of the functions for learning. In addition, the deep learning algorithm that has completed learning can immediately infer the result using the learned neuromorphic system. On the other hand, the off-chip training is a method of performing learning outside of the neuromorphic system using software, etc. After external learning is completed, the weights are post-processed according to the neuromorphic system, or the neuromorphic system is fabricated using the post-processed weights.

A synaptic array is produced by connecting synaptic devices, and an additional circuit is added to this to create a neuromorphic system. This neuromorphic system corresponds to hardware capable of executing deep learning algorithms. Because synaptic devices have limitations on conductance values that can be stably expressed, they cannot have high performance when learning and inferring with the precision of the original deep learning algorithm. A technique that selects conductances that can show the best performance of the neuromorphic system among the conductance values that the synaptic device can express and trains the synaptic device using the selected conductances is a quantization algorithm. A deep learning model using such a quantization algorithm is a quantized neural network. The quantized neural network is a neuromorphic system to which the proposed quantization algorithm is applied.

3.2. Proposed Method

The synaptic characteristics, including long-term potentiation and depression, was modeled and measured, as shown in Figure 2a–d, respectively. Modeling was performed with Equation (1) [21] to apply these characteristics to the NN simulation:

$$G = \begin{cases} ((G_{LRS}^{\alpha} - G_{HRS}^{\alpha}) \times w + G_{HRS}^{\alpha})^{1/\alpha} & \text{if } \alpha \neq 0 \\ G_{HRS} \times (G_{LRS}/G_{HRS})^{w} & \text{if } \alpha = 0. \end{cases} \quad (1)$$

G_{LRS} and G_{HRS} are the low resistance state (LRS) and high resistance state (HRS) conductance, α is a parameter representing the device conductance characteristics, and w is an internal variable. During the learning process, w increases and decreases when potentiation and depression pulses, respectively, are applied to each synaptic device. The modeling function in Equation (1) is well suited for analyzing both linearity and symmetry. Specifically, if $\alpha = 1$, it has the highest possible degree of linearity, whereas $\alpha > 1$ or $\alpha < 1$ indicate that it is concave or convex, respectively. Figure 2a shows the model function with respect to device behavior, where α values change, and α_p and α_d denote parameters for potentiation and depression characteristics, respectively. Since fabricated synaptic devices cannot be perfect, there is disturbance in the distribution of measured conductance. The disturbance of this distribution may be due to the imperfections of the synaptic device, but may also be caused by the incompleteness of the experimental environment that occurred during the measurement. Outliers deviating from the mean distribution appear particularly in Figure 2b. Outliers are approximated to quantized values by a function estimated based on identical pulses after applying the proposed quantization. Typical synapse devices show nonlinear conductance property, which should be conducted by pulse numbers with the same width and amplitude. Thus, we proposed conductance level-based quantization method which utilize a representative conductance for specific conductance range. It means that, under the identical pulses, more linear conductance changes can be achieved.

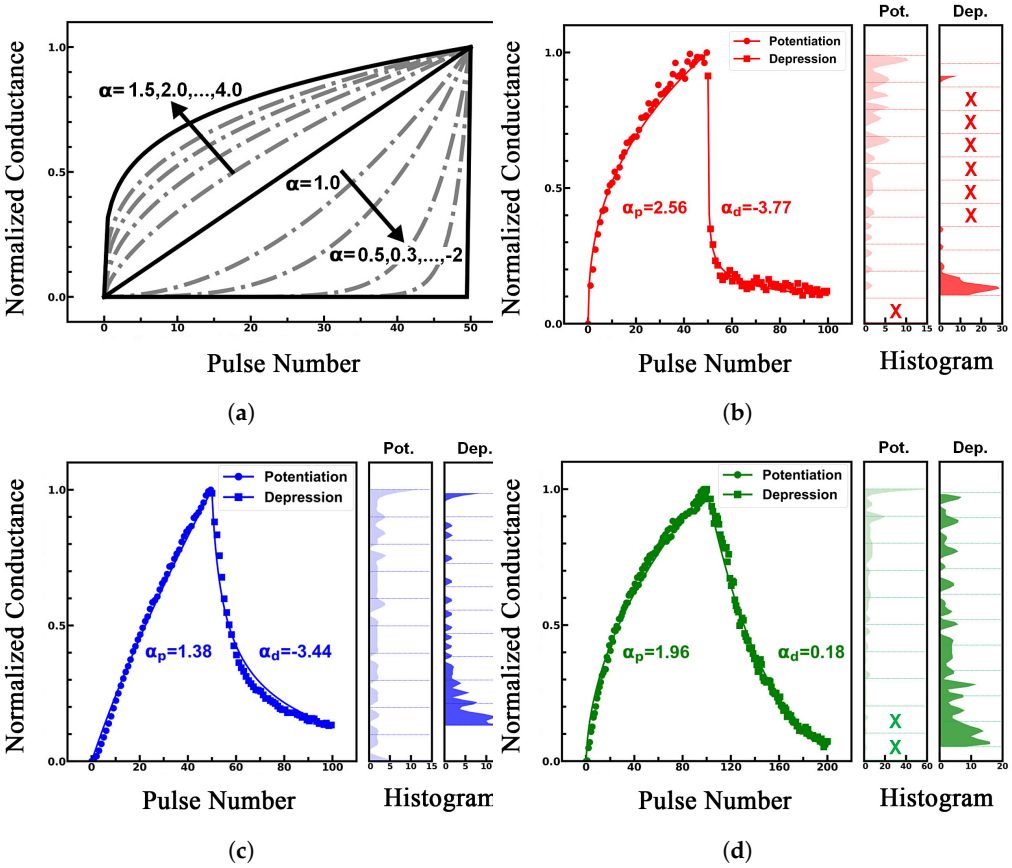


Figure 2. The synaptic characteristics including the potentiation, depression characteristics of the synaptic devices; (**a**) modeling function, (**b**) PCMO, (**c**) $LiCoO_2$, and (**d**) TiO_x devices for identical pulses and the histogram of measured conductance values on 10 level. The points correspond to measured conductance values and the solid line represents the modeling function.

We considered the alpha value as a variable, so we set its range to vary from −10 to 10, with 0.01 increments, to find the alpha value that would produce the smallest error in the modeling function and measured the corresponding conductance. Using Equation (1), α_p and α_d are extracted by calculating the difference between the predicted conductance (G) and the measured conductance of the fabricated device. First, the conductance is predicted using Equation (1), increasing a from –10 to 10 in increments of 0.01. Next, the difference between the predicted value and the measured value is calculated using MSE (mean square error), and a with the smallest difference is found. In this model, PCMO, $LiCoO_2$, and TiO_x have (α_p, α_d) of (2.56, −3.77), (1.38, −3.44), and (1.96, 0.18), respectively. Clearly, their nonlinear and asymmetric properties lessen in severity along the order of PCMO, $LiCoO_2$, and TiO_x.

The measured conductance change characteristics of the three devices are shown in Figure 2b–d. Each section contains a plot of the conductance of potentiation and depression, according to identical pulses, and histograms for each of these. The measured conductance histogram is concentrated in a specific range, and there is a range for which no data exist, indicated by 'X'. In the case of using 20 conductance values for each device, a linear device such as Li does not have a nonexistent conductance value, and TiO_x and PCMO have two and seven nonexistent conductance values, respectively. In other words, it can be seen that the number of nonexistent conductance values increases as the device has a large nonlinear property. Generally, the widespread use of electrical conductance properties has been limited by problems such as repeated measurement displacement and non-uniform distributions of measured conductance.

Considering this, we trained and evaluated our method using only the conductance levels available to actual, practical synaptic devices.

In this work, the proposed quantization method was applied to the neuromorphic system shown in Figure 3. The weight was adjusted with pulses in the synaptic cell array to obtain the loss value as a result of the calculation. The weight which can be implemented by a pair of synaptic cells was adjusted with pulses in the synaptic cell array to obtain the loss value as a result of the calculation. The gradient is calculated based on the calculated loss value, and then this is used to update the weight. The weight is quantized using our proposed quantizer described below, and on-chip learning is performed by inserting a pulse with a corresponding weight into a synaptic cell array.

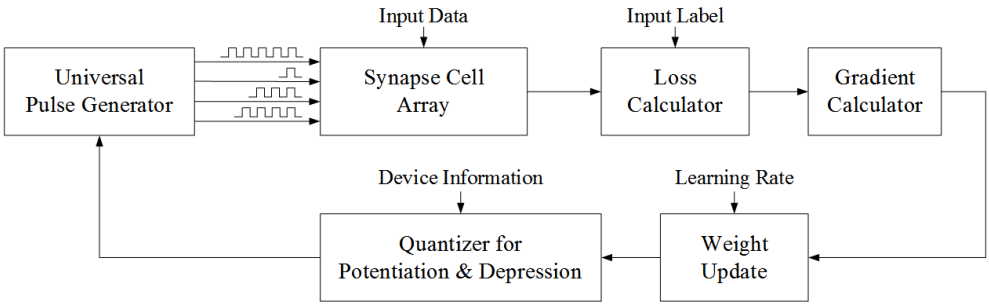


Figure 3. Proposed quantization system.

The proposed linear weighted quantizer is described in detail in Figure 4. After precisely analyzing the characteristics of the various devices using Equation (1), the conductance values are obtained based on identical pulses. Then, normalization is performed to easily convert the conductance value into the weight of the neural network. Normalizing all conductance by finding the maximum and minimum values of the conductance, the conductance has a discrete value between 0 and 1. Among the normalized discontinuous conductances, we extract the N-level weights that most closely match the uniform function. At this time, the potentiation and depression characteristics of the device are separately applied based on the weight magnitude change. This process maximizes the linearity of the device's inherent conductance. In general, if the synaptic device is more linear, learning of a neuromorphic-based deep learning system using the synaptic device is well performed [17,18,22]. However, it is very difficult for a typical synaptic device to have an electrically linear operation characteristic, and it has a nonlinear operation characteristic as shown in Figure 2b–d. However, even if a synaptic device has a nonlinear operation characteristic, it can perform a similar operation to a linear one if a quantization algorithm with high linearity is applied based on identical pulses. That is, if a quantization algorithm capable of performing a linear operation is applied to a synaptic device having a nonlinear operation characteristic, the learning efficiency of a neuromorphic system can be improved.

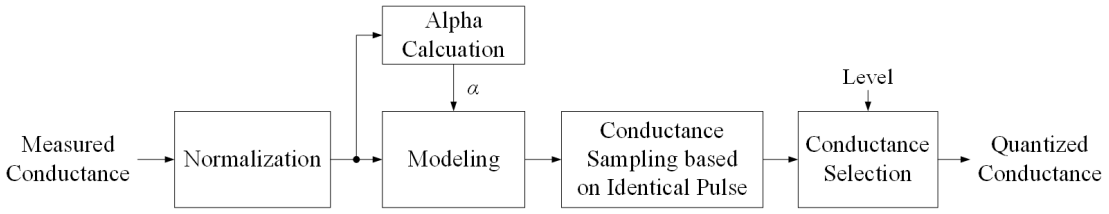


Figure 4. The proposed process of converting the measured conductance into the quantized conductance; after applying the modeling function to the normalized conductance, the conductance is sampled based on the identical pulse. Next, the conductance that maximizes linearity according to the level is selected, and the quantized conductance is extracted.

Figure 5 schematically shows the process of converting the measured conductance of the fabricated synapse device into quantized conductance using the proposed conductance quantization algorithm. The top three graphs in Figure 5 are for potentiation, and the bottom three graphs are for depression. As shown in Figure 5a, the measured conductance from the synapse device has a somewhat disordered form. Using Equation (1), we model the function closest to the trend of unaligned measured conductance. α used to model the function can be calculated using a variety of methods. Through repeated various experiments, the optimal α for each synapse device is obtained in advance. The modeled function is shown in Figure 5b. The conductance that can be expressed by the actual synapse device is sampled among conductance values that can be expressed by the modeled function. This process is shown in Figure 5c. Eventually, the neuromorphic system composed of synapse devices is learned using the conductance of Figure 5c.

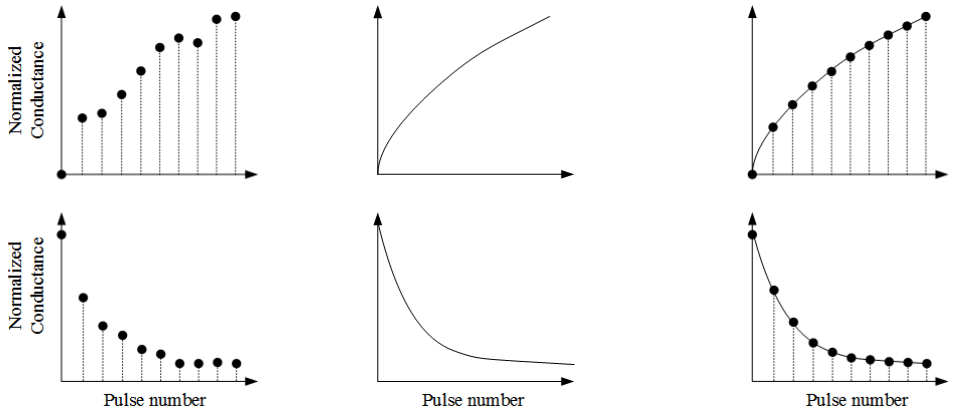


Figure 5. Generation process of the quantized conductance; (**a**) normalization to use the measured conductance as a weight; (**b**) after extracting the value of alpha using normalized conductance, modeling the synaptic device; (**c**) sampling based on identical pulses to apply the device characteristics to a function modeling.

A typical method is the most basic method of quantizing synaptic devices. This method samples and quantizes the conductance of synaptic devices at regular intervals. That is, when this method is used, the nonlinearity of the device can be reflected as it is. The typical conventional method [22] and proposed linear weighted quantization method are compared directly in Figure 6a,b. For each method, the conductance representative values of 5 are depicted in dots. In the typical method, quantized values are sampled by setting the stride to 2, resulting in nonlinearity because the device characteristics are

applied as-is. However, in our proposed method, it is possible for the sampled stride to be applied flexibly, enabling even a nonlinear device to operate linearly.

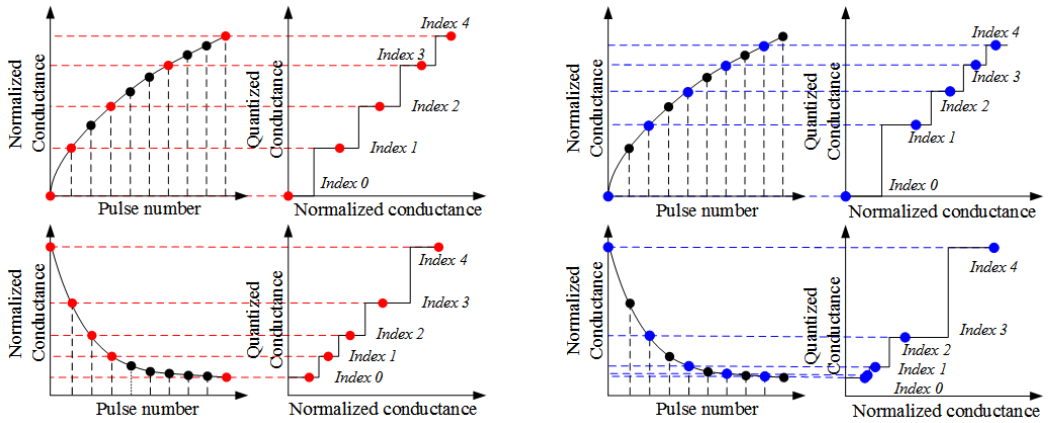


Figure 6. Conductance quantization (**a**) proposed and (**b**) typical methods. In (**a**,**b**), the top is the potentiation graph and the bottom is the depression graph. In addition, the left is the normalized conductance for the identical pulse, and the right is the quantized conductance for the normalized conductance. The index *N* means the Nth quantization index.

4. Experiment and Results

To thoroughly evaluate our proposed method, back-propagation-based on-chip learning was performed by classifying the handwritten images of the MNIST dataset with 28 × 28 sizes. The training and test data had sizes of 60,000 and 10,000, respectively. Additionally, the NN structure is shown in Figure 7, which consists of two hidden layers with 300 and 100 neurons, respectively, and an output layer with 10 neurons. If we use the structure of convolutional neurons that can extract spatial characteristics more efficiently than FC neurons, batch normalization that leads to efficient learning results by controlling the distribution of results by layer, dropout that randomly removes neurons, and learning techniques such as weight decay, we would have better learning and inferring performance. However, we fundamentally experimented with on-chip learning and used the most basic deep learning model based on FC layer to show the original effect. After performing softmax on the output layer, the number drawn on the input image was predicted based on the most probably output location. We used ReLU (Rectified Linear Unit) and CEE (Cross-Entropy Error) as activation and loss functions, respectively. We extracted random input data in 100 mini-batch units and performed forward through NN. The back-propagation was then performed based on the loss function. After training and inferring the NN with both methods using the same training method, the learning results of each method were compared and analyzed. In the neural network respect, from input layer to the first hidden layer, input data are multiplied by synaptic weights, and summated to derive intermediate results. Then, the intermediate results are calculated by the activation function to lead results of the first hidden layer. The synaptic weights, in synaptic cell respect, can be realized by conductance of the synaptic cells, and the input data are implemented by applied voltages on the memristor.

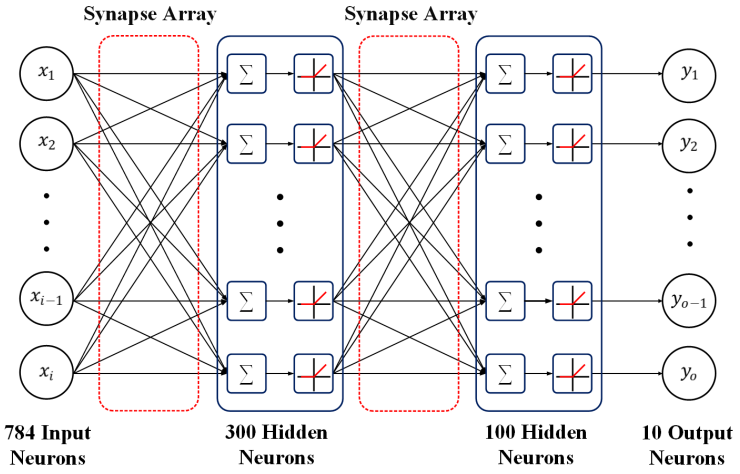
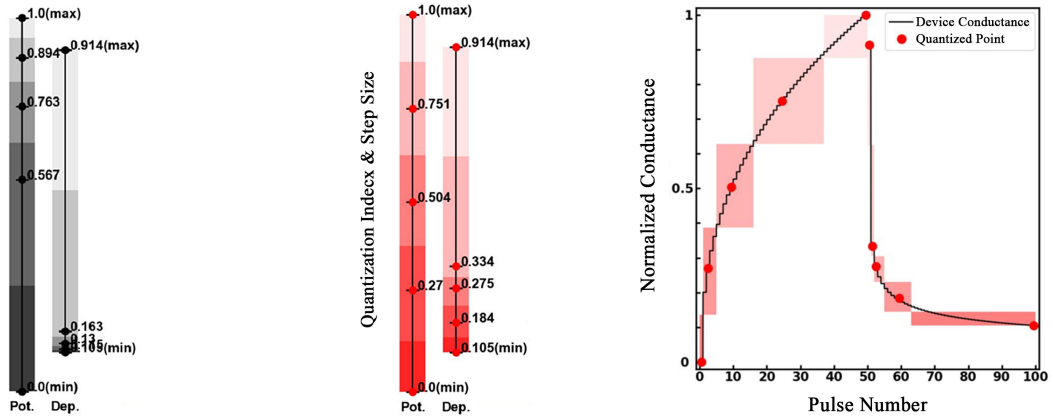


Figure 7. The architecture of the neural network used in the experiment; The network consisted of two hidden layers with 300 and 100 hidden neurons and one output layer with 10 output neurons. ReLU is used as an activation function.

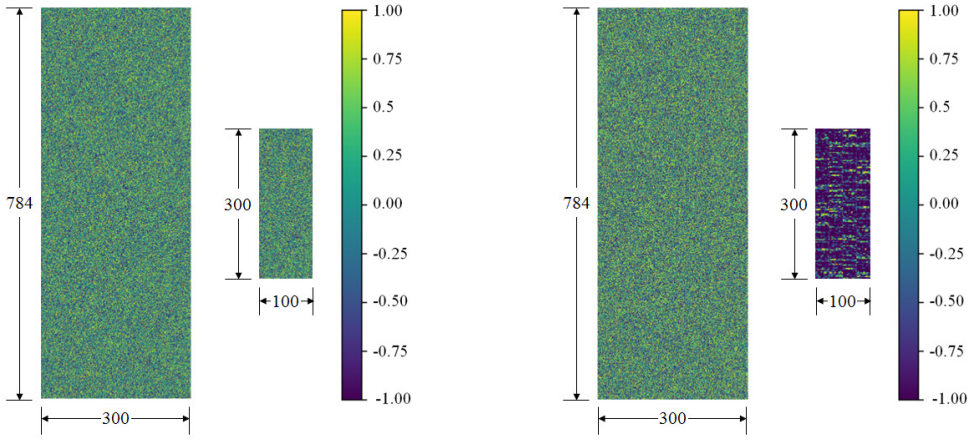
The proposed linear weighted quantization technique and the typical method [22] are compared in Figure 8a,b. For conductance ranges, conductance values representing specific ranges are expressed as dots. In Figure 8c, the nonlinear conductance of the device based on the identical pulse is represented by the black line, and the conductance value corresponding to the specific pulse range is represented by the point using the proposed quantization method. The value of the point in these is the same as the value of the point in Figure 8a. In both cases, potentiation and depression of PCMO synaptic devices showed quantized conductance at five levels after normalization. The typical method is very nonlinear because it applies the characteristics of the device as it is. However, the proposed linear-weighted quantization technique employs a conductance quantization method that maximizes linearity within the device's inherent characteristics. That is, even if the characteristics of the device are not linear, by quantization conductance that improves linearity, it is possible to show learning characteristics such as an ideal device has linearity.



(**a**) (**b**) (**c**)

Figure 8. Quantization index (Quantized Point) and step size of (**a**) typical and (**b**) proposed method, and (**c**) quantized potentiation and depression over the pulse number.

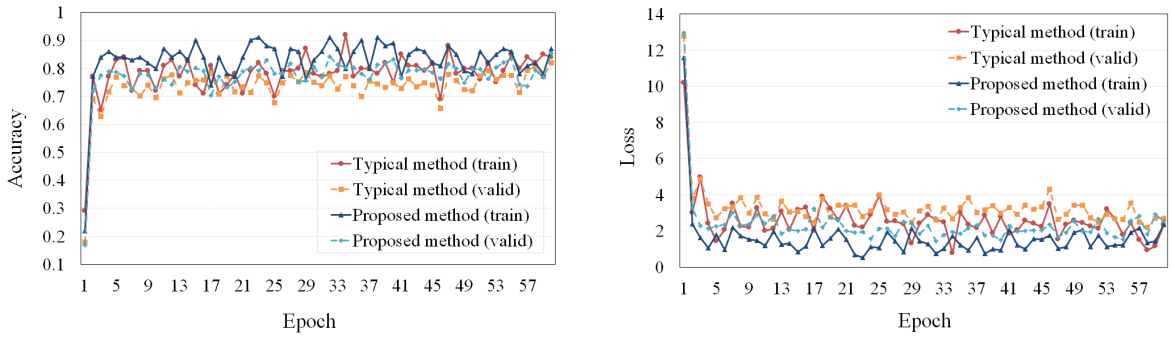
When initializing the weights, the conductance values of the neural network are determined to have a uniform distribution over the maximum and minimum range of conductance of each synapse device. As training progresses, these initial weights are quantized using the proposed method and converged to limited conductance values. Figure 9 shows the results of training about weights with the conductance values of five levels using the quantization technique proposed in a neural network composed of Li devices. The initialized weight map is shown in Figure 9a, and the weight map that has been trained is shown in Figure 9b. Compared to Figure 9a, the weight map in Figure 9b is quantized and optimized with several specific values.



(**a**) (**b**)

Figure 9. The weight map (**a**) before and (**b**) after training using the proposed method with five levels. These maps are made using the result of $LiCoO_2$. (**a**,**b**) are composed of three sub-figures. The first and second are weight maps of 784 × 300 and 300 × 100, respectively. Since two weight maps with different dimensions are expressed at the same height, the size of the pixels representing a weight in the two sub-figures are different. The third is information of weight size.

Figure 10 shows the process of training a neuromorphic system composed of Li devices using the proposed method. Figure 10a is a graph that evaluates the performance of the network during the training process, and the accuracy of the result inferred by the network and the correct answer is used as the evaluation method. Figure 10b shows the error between the inferred result and the correct answer. In Figure 10, "valid" represents the validation result, and "train" represents the training result. The "valid" result is a result of inferring using data that is not used for training, and can better show the general performance of the network. As shown in Figure 10, it shows very low performance initially, but it can be seen that the network performance improves rapidly. That is, it can be confirmed that the proposed method is suitable for training a network. In addition, in almost all results, it can be seen that the proposed method shows better performance than the typical method. When learning the three synapse devices, the learning rate was selected as 0.5 for PCMO, 0.3 for Li, and 0.5 for TiO_x through an experimental optimization process. In network training, the selection of the learning rate is very important. Because the characteristics of each device are very different, the network designer must experiment to find the learning rate that best optimizes the network.



(**a**) (**b**)

Figure 10. Performance comparison of typical and proposed method using training and validation results (**a**) accuracy, (**b**) loss. Since MNIST consists of 60,000 training data and mini-batch is set to 100, 1 epoch is 600 iterations in probability.

Using our proposed technique, we improved the performance of a neuromorphic system very simply, without using complicated calculations, additional circuits, or additional processing steps, and using only realistic device conductance values. Furthermore, by implementing only these limited device conductance values, the possibility of on-chip learning for practical synaptic devices is effectively demonstrated. Clearly, maximizing the linearity of synaptic devices significantly improves the learning capabilities of inherently nonlinear synaptic devices.

In Figure 11, we directly observe the merits of the proposed method, which improves the NN accuracy. Compared with the typical method, the accuracy is improved by a remarkable 37.7% when the PCMO device has five levels of conductance. The analytical results for the three devices show that a more nonlinear device has a correspondingly higher rate of accuracy improvement.

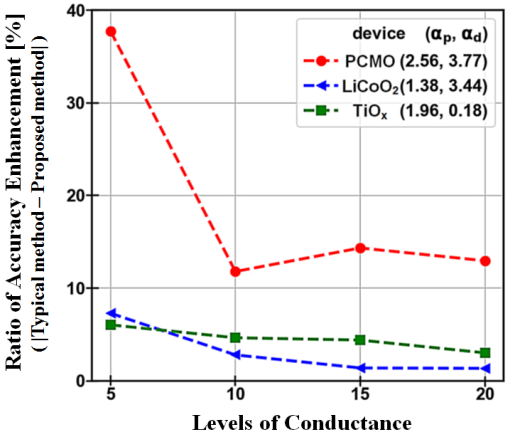


Figure 11. Ratio of accuracy enhancement between the proposed and typical method [22].

5. Conclusions

In this paper, we analyzed the conductance variation in three synaptic devices for on-chip learning and proposed a conductance-based linear weighted quantization method and on-chip learning method. Using on-chip learning for non-ideal synaptic device, we proved the generalizability of the proposed method by successfully training and evaluating the NN, which exhibited an accuracy improvement rate of 37.7% for the PCMO device. This demonstrates that, although the device had nonlinear, discontinuous, and asymmetrical properties, it can still achieve high accuracy. Furthermore, our proposed method offers

practicality and strong applicability, as it does not require additional circuits, adjustment of identical pulses, or any advances engineering approaches for materials and devices. Thus, even for inherently nonlinear, discontinuous, and asymmetrical devices, high neuromorphic system performance is possible. In the future, we intend to proceed with a study that considers conductance variations.

Author Contributions: Conceptualization and methodology, D.L. and Y.-H.S.; software and validation, J.-E.L. and C.L.; writing–review and editing, D.-W.K.; project administration and funding acquisition, Y.-H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07043220). The present Research has been conducted by the Research Grant of Kwangwoon University in 2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
2. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
3. Yu, S. Neuro-inspired computing with emerging nonvolatile memorys. *Proc. IEEE* **2018**, *106*, 260–285. [CrossRef]
4. Prezioso, M.; Merrikh-Bayat, F.; Hoskins, B.D.; Adam, G.C.; Likharev, K.K.; Strukov, D.B. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nat. Cell Biol.* **2015**, *521*, 61–64. [CrossRef] [PubMed]
5. Thomas, A. Memristor-based neural networks. *J. Phys. D Appl. Phys.* **2013**, *46*, 9. [CrossRef]
6. Kuzum, D.; Yu, S.; Wong, H.-S.P. Synaptic electronics: materials, devices and applications. *Nanotechnology* **2013**, *24*, 382001. [CrossRef] [PubMed]
7. Kandel, E.R.; Schwartz, J.H. *Principles of Neural Science*, 2nd ed.; Elsevier: New York, NY, USA, 1985.
8. Chanthbouala, A.; Garcia, V.; Cherifi, R.O.; Bouzehouane, K.; Fusil, S.; Moya, X.; Xavier, S.; Yamada, H.; Deranlot, C.; Mathur, N.D.; et al. A ferroelectric memristor. *Nat. Mater.* **2012**, *11*, 860–864. [CrossRef] [PubMed]
9. Suri, M.; Bichler, O.; Querlioz, D.; Traoré, B.; Cueto, O.; Perniola, L.; Sousa, V.; Vuillaume, D.; Gamrat, C.; DeSalvo, B. Physical aspects of low power synapses based on phase change memory devices. *J. Appl. Phys.* **2012**, *112*, 054904. [CrossRef]
10. Vincent, A.F.; Larroque, J.; Locatelli, N.; Ben Romdhane, N.; Bichler, O.; Gamrat, C.; Zhao, W.S.; Klein, J.-O.; Galdin-Retailleau, S.; Querlioz, D. Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse for Neuromorphic Systems. *IEEE Trans. Biomed. Circuits Syst.* **2015**, *9*, 166–174. [CrossRef] [PubMed]
11. Lee, C.; Lee, J.; Kim, M.; Woo, J.; Koo, S.-M.; Oh, J.-M.; Lee, D. Two-Terminal Structured Synaptic Device Using Ionic Electrochemical Reaction Mechanism for Neuromorphic System. *IEEE Electron Device Lett.* **2019**, *40*, 546–549. [CrossRef]
12. Govoreanu, B.; Kar, G.; Chen, Y.-Y.; Paraschiv, V.; Kubicek, S.; Fantini, A.; Radu, I.; Goux, L.; Clima, S.; Degraeve, R.; et al. 10×10 nm^2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation. In Proceedings of the 2011 International Electron Devices Meeting, Washington, DC, USA, 5–7 December 2011; pp. 729–732.
13. Ielmini, D.; Wong, H.-S.P. In-memory computing with resistive switching devices. *Nat. Electron.* **2018**, *1*, 333–343. [CrossRef]
14. Yao, P.; Wu, H.; Gao, B.; Tang, J.; Zhang, Q.; Zhang, W.; Yang, J.J.; Qian, H. Fully hardware-implemented memristor convolutional neural network. *Nat. Cell Biol.* **2020**, *577*, 641–646. [CrossRef] [PubMed]
15. Burr, G.W.; Shelby, R.M.; Sidler, S.; Di Nolfo, C.; Jang, J.; Boybat, I.; Shenoy, R.S.; Narayanan, P.; Virwani, K.; Giacometti, E.U.; et al. Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element. *IEEE Trans. Electron Devices* **2015**, *62*, 3498–3507. [CrossRef]

16. Lee, D.; Park, J.; Moon, K.; Jang, J.; Park, S.; Chu, M.; Kim, J.; Noh, J.; Jeon, M.; Lee, B.H.; et al. Oxide based nanoscale analog synapse device for neural signal recognition system. In Proceedings of the IEEE International Electron Devices Meeting (IEDM), Washington, DC, USA, 4–7 December 2015.
17. Kwon, D.; Lim, S.; Bae, J.-H.; Lee, S.-T.; Kim, H.; Kim, C.-H.; Park, B.-G.; Lee, J.-H. Adaptive Weight Quantization Method for Nonlinear Synaptic Devices. *IEEE Trans. Electron Devices* **2018**, *66*, 395–401. [CrossRef]
18. Chang, C.-C.; Chen, P.-C.; Chou, T.; Wang, I.-T.; Hudec, B.; Chang, C.-C.; Tsai, C.-M.; Chang, T.-S.; Hou, T.-H. Mitigating Asymmetric Nonlinear Weight Update Effects in Hardware Neural Network Based on Analog Resistive Synapse. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2018**, *8*, 116–124. [CrossRef]
19. Lee, C.; Koo, S.-M.; Oh, J.-M.; Lee, D. Compensated Synaptic Device for Improved Recognition Accuracy of Neuromorphic System. *IEEE J. Electron Devices Soc.* **2018**, *6*, 403–407. [CrossRef]
20. Park, J.; Lee, C.; Kwak, M.; Chekol, S.A.; Lim, S.; Kim, M.; Woo, J.; Hwang, H.; Lee, D. Microstructural engineering in interface-type synapse device for enhancing linear and symmetric conductance changes. *Nanotechnology* **2019**, *30*, 305202. [CrossRef] [PubMed]
21. Jang, J.; Park, S.; Burr, G.W.; Hwang, H.; Jeong, Y. Optimization of Conductance Change in $Pr_{1-x}Ca_xMnO_3$-Based Synaptic Devices for Neuromorphic Systems. *IEEE Electron Device Lett.* **2015**, *36*, 457–459. [CrossRef]
22. Dundar, G.; Rose, K. The effects of quantization on multilayer neural networks. *IEEE Trans. Neural Netw.* **1995**, *6*, 1446–1451. [CrossRef] [PubMed]