Multi-level Memristors based on Two-dimensional Electron Gases in Oxide Heterostructures for High Precision Neuromorphic Computing

Sunwoo Lee
University of Southern California
Jaeyoung Jeon
Ajou University
Kitae Eom
University of Wisconsin-Madison https://orcid.org/0000-0002-8114-609X
Chaehwa Jeong
Korea Advanced Institute of Science and Technology https://orcid.org/0000-0001-9153-710X
Yongsoo Yang
Korea Advanced Institute of Science and Technology (KAIST) https://orcid.org/0000-0001-8654-302X
Ji-Yong Park
Ajou University https://orcid.org/0000-0001-5117-3532
Chang-Beom Eom
University of Wisconsin-Madison
Hyungwoo Lee (hyungwoo@ajou.ac.kr)
Ajou University

Article

Keywords: oxygen vacancies, tunneling conductance, variance-aware weight quantization

Posted Date: November 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1019162/v1

Abstract

Memristors are essential elements for hardware implementation of artificial neural networks. The key functionality of the memristors is to realize multiple non-volatile conductance states with high precision. However, the variation of device conductance limits the number of allowed states. Since actual data for neural network training inherently have a non-uniform distribution, the insufficient number of conductance states and the resultant inaccurate weight quantization may generate significant errors in the memristor-based computation. Herein, we demonstrate a multi-level memristor based on two-dimensional electron gas in a $Pt/LaAlO_3/SrTiO_3$ heterostructure. By redistributing oxygen vacancies, we precisely controlled the tunneling conductance of the device, achieving multiple conductance states (more than 27). The multi-level switching capability and the high retention performance allow us to implement a variance-aware weight quantization (VAQ), designed for improved computing accuracy. We verify that the VAQ provides greater accuracy in image classification process, as compared to conventional uniform quantization. These results provide valuable insight into developing high-precision multi-bit memristors for practical neuromorphic processors.

Introduction

The development of artificial neural networks has been spurred by attempts to understand and mimic the highly efficient synaptic connections in brain[1,2]. The essence of the biologically-inspired computing technology is the development of synaptic devices that are capable of responding and adapting to external stimuli, *i.e.*, the learning. The learning of synaptic devices can be implemented by modulating their synaptic weights, which is normally the conductance states of the device. When a synaptic device experiences external stimuli such as voltage or current pulses, it changes its synaptic weight and memorizes it in a non-volatile way. Such modulation of the individual synaptic weights between neurons enables the energy-efficient parallel computations based on neural network architectures[3].

Memristors are one of the leading candidates for such synaptic devices[4-7]. In particular, research effort has been focused on their capability of multi-level non-volatile switching, aiming for a high-level in-memory computation[8-14]. The ultimate goal of the memristors is an analog operation with a nearly infinite number of conductance states, imitating the analog operation of biological synapses. Considering the mechanism of resistive switching, in principle, most of conventional memristors should be able to stabilize a tremendous number of conductance states with a separation of a conductance quantum $\frac{2e^2}{h}$, where e and h represents the unit charge and Plank's constant, respectively[15-17]. However, because the uncontrolled resistive switching mechanisms and the defect-induced charge trapping phenomena inevitably cause the random fluctuation of output current, the weight values of the memristors always show non-negligible variation[17,18]. Thus, the conductance states of the memristors used to be quantized so that the output signals with variation can be rounded to the nearest state[19]. In this way, although the number of allowed states (*i.e.,* the number of the representable weight values) severely decreases, the variation issue can be resolved. At this stage, it is important to increase the on/off conductance ratio of

the devices to maximize the number of the conductance states. The minimization of current fluctuation is also essential to fully take advantage of the limited conductance range.

In addition, it is worth considering how to define the conductance states of the memristors. In most of the previous research, the multiple conductance states of memristors are defined by uniformly dividing the available conductance range. That is probably because the uniform states can be simply programmed by linearly-incremental forming voltages. However, based on the knowledge of neural network training, the actual weight values as well as intermediate data during training, such as activations, commonly have a normal-like distribution[20,21]. This implies that the small-weight-regime can contribute to the result of the training more dominantly than the high-weight-regime. Therefore, the conventional uniform configuration of the conductance states may not be optimal for achieving high accuracy of the memristor-based computations. The non-uniform conductance states, configured considering the data characteristics, will better represent the distribution of the weights and the activations. Therefore, to improve the fundamental performance of memristors, we focus on two issues: (1) to build an advanced memristive device that can realize a larger number of conductance states and (2) to adopt a variance-aware weight quantization technique for high accuracy.

Herein, we demonstrate a multi-level memristor based on the two-dimensional electron gas (2DEG) in a $Pt/LaAlO_3/SrTiO_3$ (Pt/LAO/STO) heterostructure. The voltage-driven migration of oxygen vacancies, which create defect states in the band gap of LAO, enables the precise control of the tunneling resistance in a non-volatile way. The 2DEG-based memristor achieved multiple conductance states, in excess of 27 states, with high reliability. Remarkably, the multi-level switching capability allowed us to implement the variance-aware weight quantization (VAQ), designed for higher computing accuracy. In the VAQ method, the conductance states of the 2DEG memristor are non-uniformly defined in a variance-aware manner, so that the overlap of conductance distributions between the neighboring states can be minimized. This enables the full use of the available conductance range of the memristor regardless of its noise characteristics. Furthermore, since the non-uniformly configured conductance states can better represent the normal-like distribution of the weights and the activations, the VAQ method reduces information loss effectively. We verify that the VAQ indeed provides greater accuracy in both matrix-matrix multiplication and convolution operation, as compared to the conventional uniform quantization. These results offer a significant step toward developing practical multi-bit memristors for neuromorphic applications.

Results

Multi-level memristors based on 2DEG in oxide heterostructures. The LAO/STO heterointerfaces have emerged as a new playground for exploring emergent electronic properties. The polarity discontinuity at the LaO^+/TiO_2^0 heterointerface generates an electric field pointing away from the bottom interface to the top surface in the LAO/STO heterostructure[22,23]. The built-in field is necessarily compensated by the formation of 2DEG at the bottom LAO/STO interface[24]. The oxide interface with the 2DEG was found to be highly conducting. Moreover, the 2DEG has shown many interesting physical properties distinct from

conventional semiconductor heterostructures[25,26], and thus offers possibilities for device applications[27-29]. We design a memristive device based on the 2DEG in the LAO/STO heterostructures. Fig. 1a shows the tunneling device configuration of a Pt/LAO/STO heterostructure. The highly-conducting 2DEG at the LAO/STO interface serves as a reliable bottom electrode in this device structure. When a positive bias voltage is applied to the top Pt electrode the 2DEG tunnels through the LAO layer, resulting in a tunneling current. We employ oxygen vacancy point defects to modulate the tunneling conductance. Since the oxygen vacancies in the LAO form intermediate energy levels within the bandgap[30], the distribution of the oxygen vacancies can determine the effective tunneling probability of the 2DEG (see the schematic band diagram of Fig. 1a). An as-grown LAO thin film has most of its oxygen vacancies at the top surface due to the internal built-in field[22,23,31]. By redistributing the surface oxygen vacancies, we can reversibly control the tunneling conductance of the device.

To build the 2DEG-based memristor, we synthesized a LAO thin film on a TiO_2-terminated (001) STO substrate by pulsed laser deposition (PLD) with *in-situ* monitoring of reflection high-energy electron diffraction (RHEED). Fig. 1b shows the oscillation and the patterns of RHEED, indicating the layer-by-layer growth of the single-crystalline LAO thin film. After the growth of the film, the LAO/STO heterostructure was slowly cooled down to room temperature without oxygen gas injection or post-annealing, so that the oxygen vacancies are not fully removed. The Pt electrodes were subsequently fabricated on the top surface of the LAO thin film through a conventional lift-off process. Further details of the sample fabrication are found in the “Methods” section.

Fig. 1c and 1d show the atomic force microscopy (AFM) images measured on the surface of a thermally-treated STO (001) substrate and an as-grown LAO thin film, respectively. The surface of the as-grown LAO film is atomically flat and smooth, indicating the high quality of the film. The step-and-terrace structure on the LAO surface, which is almost identical to that on the STO substrate, implies that the layer-by-layer growth mode is well preserved throughout the deposition process. The high-angle annular dark field scanning transmission electron microscopy (HAADF-STEM) image taken from the same sample (Fig. 1e) also indicates the high quality of the LAO/STO heterostructure. The line profile along (001) from the STEM image (Fig. 1f) confirms that the thickness of the LAO thin film is exactly 12 unit-cells, as we designed, and the atomic intermixing at the interface is minimal. Fig. 1g shows the out-of-plane θ-2θ X-ray diffraction (XRD) pattern around (002) STO peak. Only a single peak representing the (002) reflection of the LAO is found, ensuring the epitaxial nature of the single-crystalline LAO thin film. All of these structural analyses confirm the high crystallinity and the well-defined heterointerface of the LAO/STO heterostructure, regardless of its oxygen deficiency. Considering that the switching performance of a memristor is strongly influenced by the roughness of the switching interface, this atomically sharp and flat interface is an important advantage of our device based on the LAO/STO heterostructure.

We examine the electrical switching performance of the tunneling device based on the Pt/LAO/STO heterostructure. Fig. 2a shows the 50 successively measured *I-V* curves of the device. In this study, unless

otherwise stated, the bias voltage is applied to the bottom contact for the 2DEG (Supplementary Fig. 1). As indicated by the pinched hysteresis loops, the 2DEG-tunneling device exhibits a repeatable bipolar resistance switching behavior. The asymmetry of the hysteresis is attributed to different energy band offsets at the top Pt/LAO and the bottom LAO/STO interfaces. The *I-V* characteristics show that the positive voltage results in the Off-switching (*i.e.,* increasing the tunneling resistance), while the negative voltage results in the On-switching. This switching polarity provides a clue to the underlying mechanism of the resistive switching in the device. Note that the as-grown LAO film has most of its oxygen vacancies at near the top surface due to the inherent built-in field in the LAO. When a negative voltage is applied to the bottom 2DEG interface, the electropositive oxygen vacancies migrate from the top surface toward the bottom 2DEG interface. This migration of the oxygen vacancies reduces the effective height and width of the tunneling barrier and, hence, results in the increase of the tunneling current. On the other hand, when a high positive voltage is applied, the oxygen vacancies move back to the top surface, resulting in the off-switching (see Supplementary Fig. 2 for more details).

The gradual change of I_v at the positive voltage regime in Fig. 2a gives us a hint that we can implement multiple conductance states rather than binary states. Fig. 2b shows the multiple conductance states of the same device, programmed by different writing voltage V_{write}. We first fully turned off the device by applying the V_{write} of +9V and then gradually turned the device on by applying incremental V_{write}. For each conductance state, I_v is measured at a reading voltage V_{read} of +0.5 V. Remarkably, we could implement 27 discrete conductance states with the On/Off ratio of % at the V_{write} ranged from -3.00 V to -4.25 V. In that specific range of V_{write}, the switching of the conductance states is quite reliable and reproducible. We have not observed degradation or undesired state change during the reading process, indicating good retention in the V_{write} regime. In fact, more conductance states and a higher On/Off ratio could be achieved with a broader range of V_{write}. However, when the V_{write} was smaller than -3 V, we observed that the change of the conductance was nonlinear. As the V_{write} increased over -4.25 V, the output current was further increased but became unstable and noisy. Thus, we consider only this linear and reliable regime for the following study. The representative conductance value and the standard deviation for each conductance state are given in Fig. 2c. The representative values are obtained by averaging 200 conductance values measured at each state. Besides the outliers (marked by red triangles), the overall change of the conductance value is desirably linear. In an artificial neural network, considering the required linear relationship between the input signals and the weight change for the network training, the linear dependence of conductance on V_{write} enables more efficient and accurate training (Supplementary Note 3). Therefore, the high linearity of the conductance change, without additional doping[32] or multilayer stacking[33], makes this 2DEG memristor a promising and attractive candidate for synaptic applications.

On the other hand, it should be noted that the variance of the output current at each conductance state is not ideally small and has a clear dependence on the conductance value itself. As shown by the error bars in Fig. 2c, the standard deviation of the output current at each state increases with the conductance of the device. In particular, the conductance distributions of higher states (*i.e.,* from state 23 to state 27) overlap to each other due to their large variations. This large current variance (*i.e.,* the electrical noise of

the output current) is inevitable at the high-conductance regime, as discussed in Supplementary Fig. 3. To further clarify the noise characteristics, we measured the current power spectral density (PSD) $S_I(f)$ of the 2DEG memristor. Fig. 2d shows the PSD spectra at 4 distinct conductance states. After setting each conductance state, the PSD spectrum of the I_V was measured at +1 V. All the PSD spectra show a typical $1/f$ behavior, indicating the presence of charge traps with a wide range of time constants. It is also clearly seen that the fluctuation of I_V becomes stronger as the device conductance increases. This current noise and the resultant state-overlap issue must be confronted because the inaccurate state assignment results in basic and fatal errors in interpreting weight values of the memristor during computations. In the following section, we demonstrate that the state-overlap issue can be effectively circumvented by customizing the conductance state configuration based on the variance of the output current at each state.

Variance-aware quantization. In principle, the overlap issue can be simply resolved if we selectively use only the conductance states whose distributions do not overlap with each other at all. Thus, in the case of previous memristors, the weight values are assigned to uniformly- and coarsely-defined conductance states. However, due to the large variance of the output at the higher states, as shown by Fig. 2c, the conventional method allows only a small number of conductance states, degrading the performance of the memristors. If we ignore the variance and just use more finely-defined conductance states, the state-overlap issue will become severe, and it will severely degrade the computing accuracy of the memristor. Therefore, we propose a VAQ method that effectively resolves the overlap issue without compromising the performance of the memristors. The multi-level switching capability and the high retention performance of our 2DEG memristor allows us to implement the VAQ. Fig. 3a schematically depicts the state configuration for the uniform quantization and the VAQ. Note that the conductance states are non-uniformly defined for the VAQ, such that the conductance distributions hardly overlap across any two neighboring states.

As a proof of concept, we selected the uniformly-separated (Fig. 3b) and the non-uniformly-separated (Fig. 3d) 12 conductance states out of the total 27 states, implemented by our 2DEG memristor (Supplementary Fig. 4). To estimate the quantization errors depending on the state configuration, we measured 80 conductance values at each state and assigned them to the nearest conductance state. The heatmaps of the incorrectly quantized weight values when using the uniform states and the non-uniform states are given in Fig. 3c and 3e, respectively. The horizontal and the vertical axis of the heatmaps represent the intended weight values and the actually-quantized weight values, respectively. These heatmaps clearly show that the VAQ with the non-uniform states decreases the quantization errors effectively. Because the conductance states for the VAQ are configured to have minimal overlaps between the data distributions, the measured weight values are less likely assigned to the incorrect neighboring states, achieving lower quantization errors. The difference between the two quantization schemes is particularly significant at states 9 to 11, where the uniformly-separated states have largely overlapped

distributions. This result empirically proves that the conductance states defined in a variance-aware manner exploit the available conductance range without suffering from the distribution overlap issue.

Evaluation of the VAQ: matrix-matrix multiplication. One of the major advantages of the VAQ is that the conductance states configured for the VAQ can accurately represent the non-uniform data distribution. In general, the weight values as well as the intermediate data, generated in neuromorphic network training, have normal-like distributions (Supplementary Fig. 5). The conventional uniformly-configured conductance states can well represent the large weight values, that take up only a small portion of the entire model parameters, while having significant quantization errors for the many small weight values (Supplementary Fig. 6). On the contrary, the non-uniform conductance states configured for the VAQ is suitable to represent such normal-like data distributions. Because a larger number of states is assigned to the small weight values (appearing more frequently) than to the large weight values (appearing less frequently), the non-uniform state configuration is advantageous for representing the normal-like data distribution.

To verify this advantage of the VAQ, we perform matrix-matrix multiplication, which is one of the most fundamental arithmetic operations for neural network training, using the conventional uniform quantization and the VAQ, and then compare their outputs. Fig. 4a shows the schematic illustration of an image classification process with convolutional neural network (CNN)[39]. Modern CNNs typically have a fully connected layer at the end of the model to perform the classification. We first calculate the output activation matrix by multiplying the original floating-point input activation matrix by the weight matrix collected from ResNet20[35] output layer during its training on CIFAR-10 dataset[34]. Note that we use the input activations and the weight values collected from the actual neural network training so that the experimental results demonstrate the superior performance of VAQ under the realistic settings (Supplementary Note 8). The histogram of the collected input activations and the weight values are given by Fig. 4b and 4c, respectively. Then, we quantize the original output matrix and consider the quantized output matrix as ground-truth. When evaluating the performance of a quantization scheme, we calculate the output activations using the quantized input activations and weights and compare them to the ground-truth output activations.

Fig. 4d and 4e show the number of the correctly- and wrongly-quantized elements of the output matrix using the uniform quantization and the VAQ, respectively. Apparently, the uniformly-configured states do not appropriately represent the original data. Most of the quantized elements are assigned to the incorrect neighbor states. Even though this error rate is somewhat exaggerated due to the small number of allowed states (*i.e.,* 12), it is clear that the quantization error is severe in this conventional quantization scheme. On the other hand, the VAQ method remarkably improves the accuracy. This improvement can be explained by the following two reasons. First, the VAQ represents the small data using finer-grained quantization states and, hence, achieves a lower quantization error of the small activations as compared to the uniform quantization. Since the activations at each layer have a normal-like distribution, improving

the accuracy of the small activations naturally results in reducing the total quantization error. Second, the more accurate small input activations can reduce the quantization error of the large output activations, because each output activation is computed by summing up the products of multiple input activations and weights. This performance gain can recover the quantization errors increased by the VAQ's coarser-grained states.

Fig 4f shows the mean squared error (MSE) for each state, calculated using the different state configurations. While the uniform quantization yielded high errors especially in the small-value regime, the VAQ method markedly reduced the errors. For example, the MSE for the state 5 is reduced by ~82.4% by adopting the VAQ. Although the non-uniform state configuration effectively reduced the errors in the small-value regime, the errors in the large-value regime were slightly increased. This is not surprising because, in the VAQ method, we set a relatively smaller number of states for the large-value regime. Thus, the errors in the large-value regime can be weakly increased, but they are sufficiently small to be ignored. This result confirms that the conductance states configured for the VAQ can accurately represent the realistic weight values, particularly in the small-value regime. We performed another matrix-matrix multiplication experiment using natural language processing data and obtained a similar result (Supplementary Fig. 7).

Evaluation of the VAQ: a heavily re-used convolution filter. When multi-level memristors are used for performing convolution operations, which is another key operation in neural network training, the accurate quantization is a crucial prerequisite for successful training. This is because the convolution operation repeatedly uses the same weight matrix to calculate all individual output activations. Depending on the stride size, each input activation can also be repeatedly used to calculate multiple output activations. Thus, the quantization errors at the input side can snowball into greater errors as the training progresses. In this section, we demonstrate that the VAQ can effectively reduce the quantization errors during the convolution operation.

Fig. 5a shows a training image, sample #1888 from Fashion-MNIST dataset[36]. The inset schematically shows the calculation of the output activation matrix from the input training image. The detailed process is given in Supplementary Note 10. The top panel of Fig. 5b shows the numerical input values for each pixel of the image. Note that most of the normalized data are lied between 0.2 and 0.6. This implies that the realistic data are not uniformly distributed over the entire range. To simulate the convolution operation based on the 2DEG memristors, we quantized the non-uniformly distributed input data to 12 states before applying the convolution filter. The middle panel of Fig. 5b shows the quantized input data using the conventional uniform quantization method. It is clearly shown that the uniform quantization leads to loss of a significant amount of information in the input data. A majority of the input activations are quantized into smaller state values than their actual values due to the limited number of states for the small data values. The bottom panel of Fig. 5b shows the quantized input data using the VAQ method. Unlike the uniform quantization, the finer-grained states for the small data values represent the input data

more accurately. Particularly, while the uniform quantization zeroes out many small input data, the VAQ keeps them in non-zero states mitigating the loss of information.

As a reference, the original output data of the convolution operation without any quantization process is given in Fig. 5c. Since we apply a convolution filter on the image with a stride of , the input image matrix of size provides an output activation matrix of size . Fig. 5d and 5e present the quantized convolution outputs computed based on the uniform quantization and the VAQ method, respectively. The uniform quantization fails to precisely represent the small output values making the overall image speckled and noisy, while the VAQ provides a comparable output image to the original output. It should be noted that we do not consider any specific activation function in this simulation. If an activation function was used, such as sigmoid or hyperbolic tangent, the output activations would be rescaled to a range of 0 to 1. That is, the data distribution can be shifted to the small value regime, and the uniform quantization likely loses more information compared to the VAQ. Likewise, if rectified linear unit (ReLU)[37] was used, the magnitude of the overall data flow would be significantly reduced since all the negative values are zeroed out. Regardless of the type of activation function, therefore, the VAQ is expected to result in smaller quantization errors than those of the uniform quantization. Throughout these studies, we show the promise of the multi-level 2DEG memristors for neuromorphic computing applications. Particularly, the VAQ implemented by the 2DEG memristors offers an attractive way to reduce the inevitable and critical quantization errors and thereby enables the practical use of the prototypical multi-level memristors.

Discussion

Lastly, we address the importance of the quantization performance of multi-level memristors. Given the overlapped weight distributions across neighboring states, increasing the number of states does not always reduce the total quantization error. Thus, reducing the distribution overlap is essential to fully utilize the given conductance range. The VAQ tackles the overlap issue through the quantization states defined in a variance-aware manner. This approach effectively improves the quantization accuracy of all individual devices, and the higher device-level accuracy ends up improving the whole network performance. The quantization error at each layer strongly affects the classification or regression performance of the whole network. Since the data propagate through all the layers forward and backward, the quantization error accumulates across all the layers. Moreover, the training is usually performed using stochastic gradient descent (SGD)[38] that is an iterative optimization algorithm. Thus, the quantization errors are accumulated even across multiple training iterations. Given such characteristics of neural network training, reducing the quantization error at each layer is critical to achieve good classification or regression performance. As discussed above, the non-uniformly configured states for the VAQ can better represent the real data distribution and, thereby, reduce the quantization errors at each network layer effectively. Therefore, the VAQ will allow the 2DEG memristors to achieve the higher classification and regression performances.

In conclusion, we have demonstrated a novel memristive device based on the 2DEG in the Pt/LAO/STO heterostructure. The 2DEG memristor could realize multiple conductance states, over 27, with high

reliability. Through applying different writing voltages, we could specifically set the conductance state to desired values. Remarkably, the multi-level switching capability of the 2DEG memristor allowed us to explore an advanced weight quantization strategy, the VAQ method, aiming to minimize the quantization errors. In this variance-aware manner, we could effectively reduce the state-overlap issue and, thereby, enhance the computing accuracy. We demonstrated that the VAQ method resulted in significantly improving the computing accuracy for the image classification process, as compared to the conventional uniform quantization method. These results indicate the high potential of the multi-level 2DEG memristors for neuromorphic computing applications. The precisely-configurable conductance states of the 2DEG memristors can serve as the weight values of the artificial synaptic interconnections in neural networks. Moreover, the VAQ implemented by the 2DEG memristors suggests an efficient and powerful strategy to achieve greater accuracy in computation. Therefore, we believe that our results will provide a stepping stone for full hardware implementation of convolutional neural networks and relevant neuromorphic processors.

Methods

Device Fabrication. LAO thin films were epitaxially grown on TiO_2-terminated STO (001) substrates using PLD. To obtain the TiO_2-terminated STO substrates, as-received STO substrates were etched using buffered-HF for 1 min and thermally annealed at 1000°C for 6 hours. To grow LAO films, a single-crystalline LAO target was ablated using a KrF (248 nm) excimer laser at a repetition rate of 3 Hz and with a fluence of 2.0 J/cm^2. During the growth, the temperature of PLD chamber was kept as 550°C. The oxygen partial pressure for growing LAO was 10^{-3} mbar. After growing the LAO films, the samples were slowly cooled down to room temperature without changing oxygen partial pressure. Subsequently, the 50-nm-thick Pt electrodes were fabricated by a lift-off process.

STEM Measurements. A cross-section specimen for TEM analysis was fabricated by using a focused ion beam (FIB) machine (Helios G4, FEI), and the sample thickness of sub 30 nm was achieved via FIB milling. The LAO/STO interface structure was measured at the atomic resolution using a Titan Double Cs corrected TEM (Titan cubed G2 60-300, FEI) in high-angle annular dark field scanning TEM (HAADF-STEM) mode. The microscope was operated at 200 kV accelerating voltage with the beam convergence semi-angle of 17.9 mrad. The inner and outer angles of the HADDF detector were chosen as 40 and 200 mrad, respectively. An 2048 × 2048 image of the LAO/STO interface was measured with 1 μs dwell time and 8.1 pm pixel size. The total electron dose was about 1.91 × 10^5 electrons Å^{-2}.

XRD Measurements. The lattice structure of the LAO thin film was determined by XRD. A D8 Discover (Bruker AXS) high-resolution X-ray diffractometer with a Cu K_α source (λ = 1.5405 Å) was used for the XRD measurements. The out-of-plane θ–2θ XRD pattern of the as-grown LAO/STO heterostructure was measured at around (002) STO peak.

Electrical Characterizations. The *I-V* measurements and the multi-level switching experiments were conducted using a semiconductor analyzer, Keithley 4200. For all the electrical characterizations, an input

voltage (*i.e.,* V_{write} or V_{read}) is directly applied to the bottom 2DEG interface and the output current is measured through the top Pt contact. To minimize the contact resistance, we widely deposited a commercial silver paste onto the side wall of the sample and contacted the silver surface using a commercial metal prove tip. As for the *I-V* curve measurements, the input voltage was swept from -9 V to +9 V and *vice versa*. Totally, 50 cycles of *I-V* curves are successively measured without changing the environment. To demonstrate the capability of multi-level switching, we initialized the 2DEG memristor by applying +9 V to the bottom contact. By this initialization process, the device was completely turned off. Subsequently, we gradually switched the device on by applying an incremental voltage from -3 V to -4.25 V. The conductance states were identified by measuring the V_{read} of +0.5 V.

Noise Characterizations. The electrical noise spectra of the device were measured using a dynamic signal analyzer, SR785 (Stanford Research Systems) along with a low-noise current amplifier, SR 570 (Stanford Research Systems). After programming the conductance state by applying different V_{write}, the current power spectral density was measured at a constant voltage of +1 V. The maximum frequency was set as 102.4 kHz.

Declarations

Data Availability

The data that support the findings of this study are available from the corresponding author on reasonable request.

Acknowledgments

This work is supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2021R1C1C1011219). Work at the University of Wisconsin-Madison is funded in part by the Gordon and Betty Moore Foundation's EPiQS Initiative, grant GBMF9065 to C.-B.E. and Vannevar Bush Faculty Fellowship (ONR N00014-20-1-2844). Transport measurements and analysis at the University of Wisconsin–Madison was supported by the US Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences (BES), under award number DE-FG02-06ER46327 (C.B.E.). J.-Y. P. acknowledges the support from NRF grant funded by the Korea government (MSIT) (No. 2019R1A2C1007913). The work at the KAIST is supported by the NRF grants funded by Korea government (MSIT) (No. 2020R1C1C100623911) and KAIST singularity professor program. C. J. was also partially supported by the KAIST-funded Global Singularity Research Program (M3I3) for 2021. The STEM experiment was conducted using a double Cs corrected Titan cubed G2 60-300 (FEI) equipment at KAIST Analysis Center for Research Advancement (KARA).

Author Contributions

HL and SL conceived the project. KE and CBE fabricated and characterized thin film samples. JYJ carried out electrical transport measurements. SL performed simulations and calculations. JYP supervised electrical characterizations. CJ and YY performed STEM experiments. SL, JYJ, KE, CBE, and HL prepared the manuscript. HL directed the overall research.

Competition Interests

The authors declare no competing financial interests.

References

1. Williams, R. S. What's Next?[The end of Moore's law]. *Computing in Science & Engineering* **19**, 7-13 (2017).

2. Misra, J. Saha, I. Artificial neural networks in hardware: A survey of two decades of progress. *Neurocomputing* **74**, 239-255 (2010)

3. Ielmini, D. Ambrogio, S. Emerging neuromorphic devices. *Nanotechnology* **31**, 092001 (2019).

4. Chua, L. Memristor-the missing circuit element. *IEEE Trans. Circuit Theory* **18**, 507-519 (1971).

5. Strukov, D. B., Snider, G. S., Stewart, D. R. Williams, R. S. The missing memristor found. *Nature* **453**, 80-83 (2008).

6. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61-64 (2015).

7. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641-646 (2020).

8. Ielmini, D. Wong, H. S. P. In-memory computing with resistive switching devices. *Nature Electronics* **1**, 333-343 (2018).

9. Nandakumar, S. R. et al. A 250 mV Cu/SiO2/W memristor with half-integer quantum conductance states. *Nano letters* **16**, 1602-1608 (2016).

10. Lee, M. J. et al. A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta 2 O 5− x/TaO 2− x bilayer structures. *Nature materials* **10**, 625-630 (2011).

11. Kim, K. M. et al. Self-limited switching in Ta2O5/TaOx memristors exhibiting uniform multilevel changes in resistance. *Advanced Functional Materials* **25**, 1527-1534 (2015).

12. Chanthbouala, A. et al. A ferroelectric memristor. *Nature materials* **11**, 860-864 (2012).

13. Schranghamer, T. F. Oberoi, A. Das, S. Graphene memristive synapses for high precision neuromorphic computing. *Nature communications* **11**, 1-11 (2020).

14. Zhu, X. Li, D. Liang, X. Lu, W. D. Ionic modulation and ionic coupling effects in MoS 2 devices for neuromorphic computing. *Nature materials* **18**, 141-148 (2019).

15. Terabe, K. Hasegawa, T. Nakayama, T. Aono, M. Quantized conductance atomic switch. *Nature* **433**, 47-50 (2005).

16. Xue, W. et al. Controllable and stable quantized conductance states in a Pt/HfOx/ITO memristor. *Advanced Electronic Materials* **6**, 1901055 (2020).

17. Yi, W. et al. Quantized conductance coincides with state instability and excess noise in tantalum oxide memristors. *Nature communications* **7**, 1-6 (2016).

18. Yu, S. et al. Characterization of low-frequency noise in the resistive switching of transition metal oxide HfO 2. *Physical Review B* **85**, 045324 (2012).

19. Jacob, B. et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2018).

20. Franchi, G. Bursuc, A. Aldea, E. Dubuisson, S. Bloch, I. TRADI: Tracking deep neural network weight distributions. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer International Publishing, (2020).

21. Bellido, I. Fiesler, E. Do backpropagation trained neural networks have normal weight distributions? *International Conference on Artificial Neural Networks*. Springer, London, (1993)

22. Nakagawa, N. Hwang, H. Y. Muller, D. A. "Why some interfaces cannot be sharp. *Nature materials* **5**, 204-209 (2006).

23. Lee, H. et al. Direct observation of a two-dimensional hole gas at oxide interfaces. *Nature materials* **17**, 231-236 (2018).

24. Ohtomo, A. Hwang, H. Y. A high-mobility electron gas at the LaAlO 3/SrTiO 3 heterointerface. *Nature* **427**, 423-426 (2004).

25. Brinkman, A. et al. Magnetic effects at the interface between non-magnetic oxides. *Nature materials*. **6**, 493–496 (2007).

26. Reyren, N. et al. Superconducting interfaces between insulating oxides. *Science* **317**, 1196–1199 (2007).

27. Mannhart, J. Schlom, D. G. Oxide interfaces—an opportunity for electronics. *Science* **327**, 1607–1611 (2010).

28. Cheng, G. et al. Sketched oxide single-electron transistor. *Nature Nanotechnology*. **6**, 343–347 (2011).

29. Wu, S. et al. Nonvolatile Resistive Switching in Pt/LaAlO3/SrTiO3 Heterostructures *PHYSICAL REVIEW X* **3**, 041027 (2013).

30. Mitra, C. Lin, C. Robertson, J. Demkov, A. A. Electronic structure of oxygen vacancies in SrTiO 3 and LaAlO 3. *Physical Review B* **86**, 155105 (2012).

31. Zhong, Z. Xu, P. X. Kelly, P. J. Polarity-induced oxygen vacancies at LaAlO 3/ SrTiO 3 interfaces. *Physical Review B* **82**, 165127 (2010).

32. Chandrasekaran, S. Simanjuntak, F. M. Saminathan, R. Panda, D. Tseng, T. Y. Improving linearity by introducing Al in HfO2 as a memristor synapse device. *Nanotechnology* **30**, 445205 (2019).

33. Jiang, Y. et al. Linearity improvement of HfOx-based memristor with multilayer structure. *Materials Science in Semiconductor Processing* **136**, 106131 (2021).

34. Krizhevsky, A. Hinton, G. Learning multiple layers of features from tiny images. **7** (2009).

35. He, K. Zhang, X. Ren, S. Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016).

36. Xiao, H. Rasul, K. Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:* **1708.07747** (2017).

37. Nair, V. Hinton, G. E. Rectified linear units improve restricted boltzmann machines. *Icml*. (2010).

38. Robbins, H. Monro, S. A stochastic approximation method. *The annals of mathematical statistics* **400-407** (1951).

39. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural computation* **1**, 541-551 (1989).

Figures

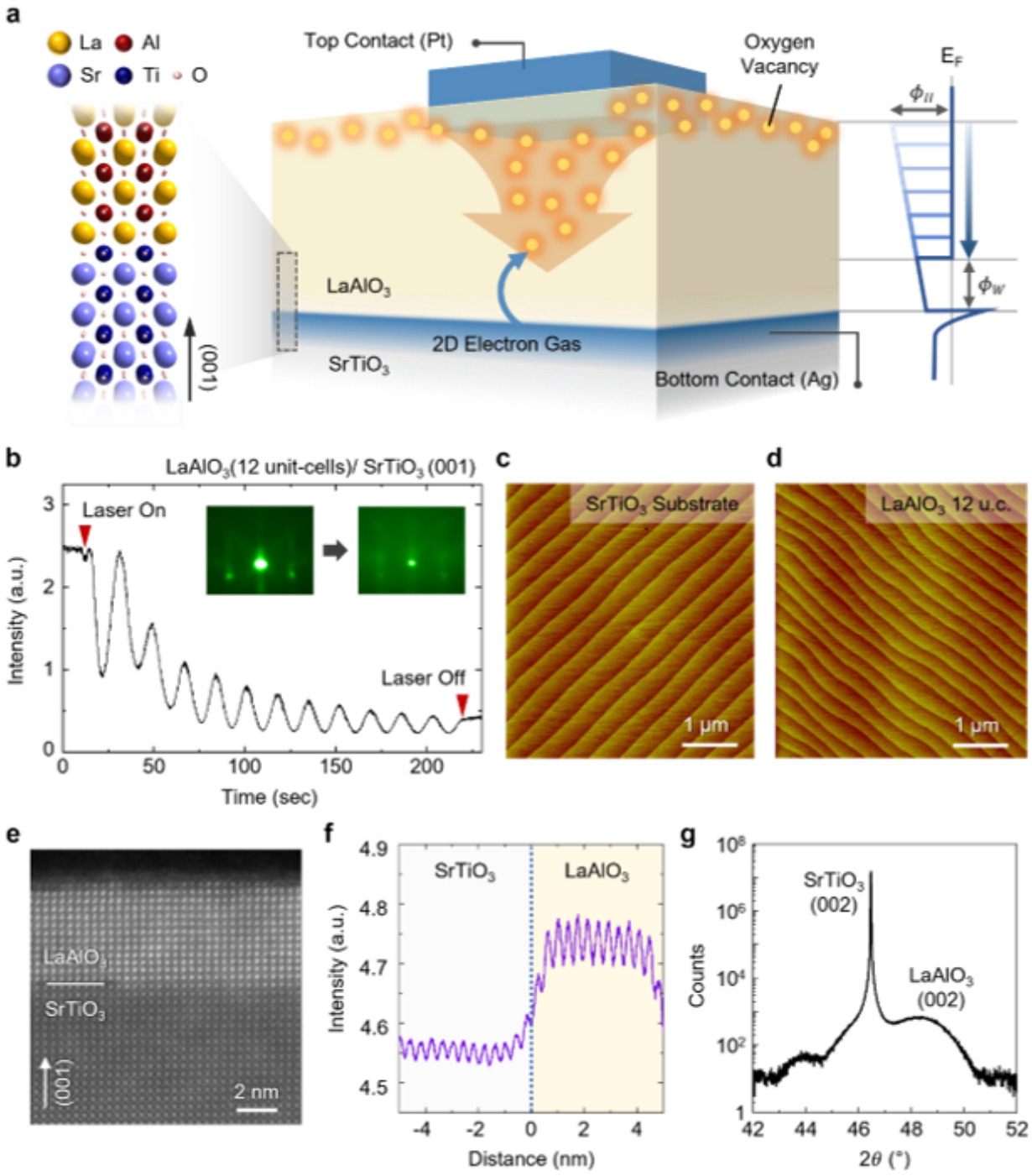

Figure 1

Memristive device based on a Pt/LAO/STO heterostructure. a Schematic depicting the mechanism for the resistive switching in the oxide heterostructure. The spatial distribution of oxygen vacancies determines the tunneling probability of the 2DEG between the LAO/STO interface and the top Pt electrode. b Thickness-dependent evolution of the in-situ RHEED intensity oscillation during the PLD deposition of LAO thin films. The insets show the RHEED patterns before and after the film growth. c AFM topography image measured on the surface of a thermally-treated STO (001) substrate. d AFM topography image measured on the surface of an as-grown LAO thin film. e HAADF-STEM image of the LAO/STO

heterostructure. f Intensity of a line profile along (001) obtained from the STEM image. g XRD θ-2θ scan of the LAO/STO heterostructure.

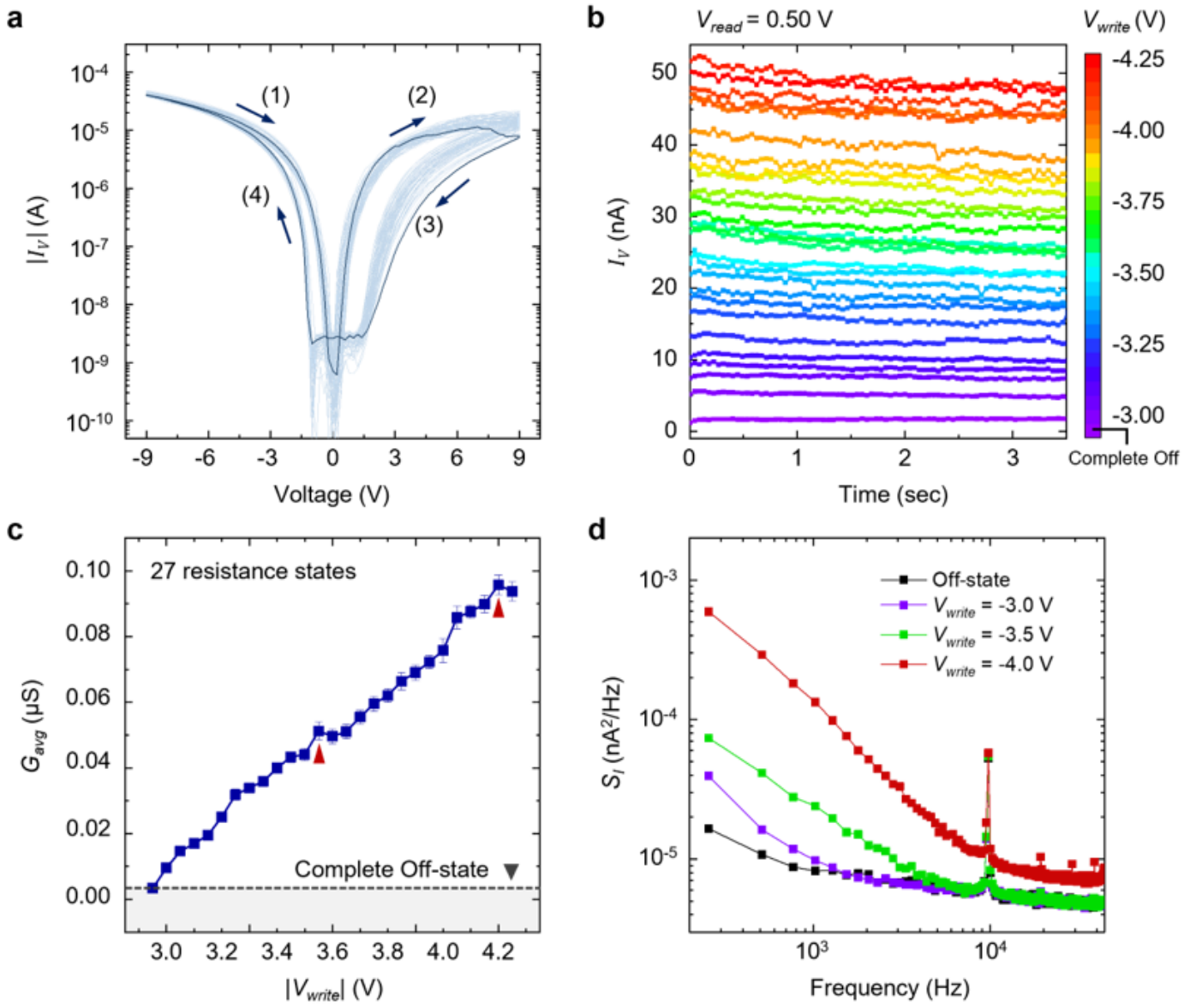

Figure 2

Multi-level switching behavior of the 2DEG-based memristor. a I-V characteristics of the 2DEG memristor. The arrows represent the directions of the voltage sweep. Totally, 50 sets of I-V curves are represented. The first cycle is highlighted by the deep-blue color. b Sequentially programmed conductance states, showing the representative 27 states. c The averaged conductance values for each state. The error bars represent the standard deviation of the mean values. The red triangles indicate the outliers. d Current power spectral density of output currents at the representative 3 different states and the off-state.

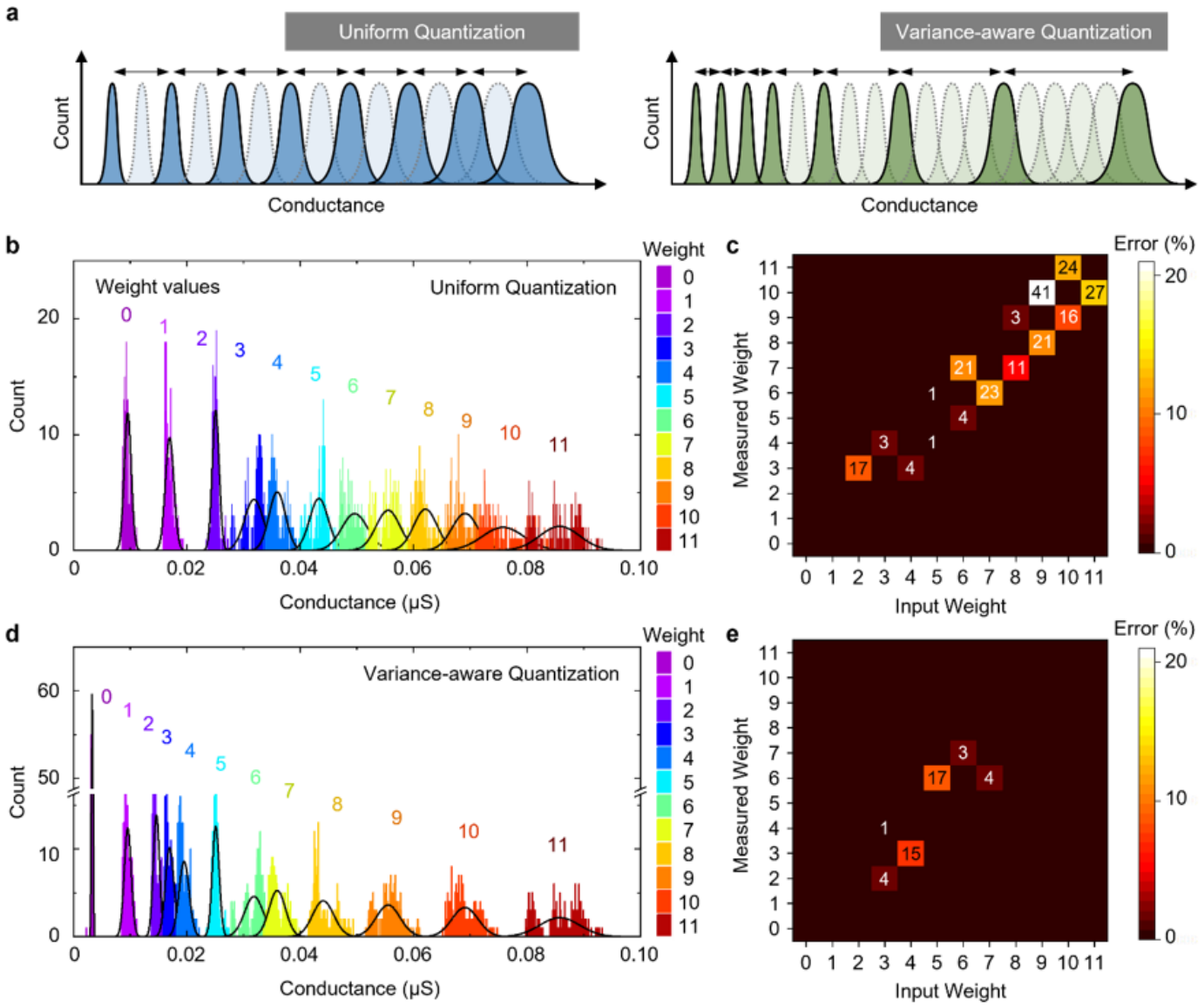

Figure 3

Variance-aware weight quantization for the 2DEG memristor. a Schematics showing the conventional uniform quantization and the variance-aware quantization methods. b Conductance histogram of uniformly-separated 12 conductance states. The black lines represent the normal distribution fitting curves. c Simulated heatmap of measurement error, calculated based on the uniformly-separated states. d Conductance histogram of the nonuniformly-separated 12 conductance states. The separation between each state is set small in the low current regime, while it is expanded as the current increases. e Simulated heatmap of measurement error, calculated based on the nonuniformly-separated states. Note that the error is significantly reduced by the variance-aware quantization method.

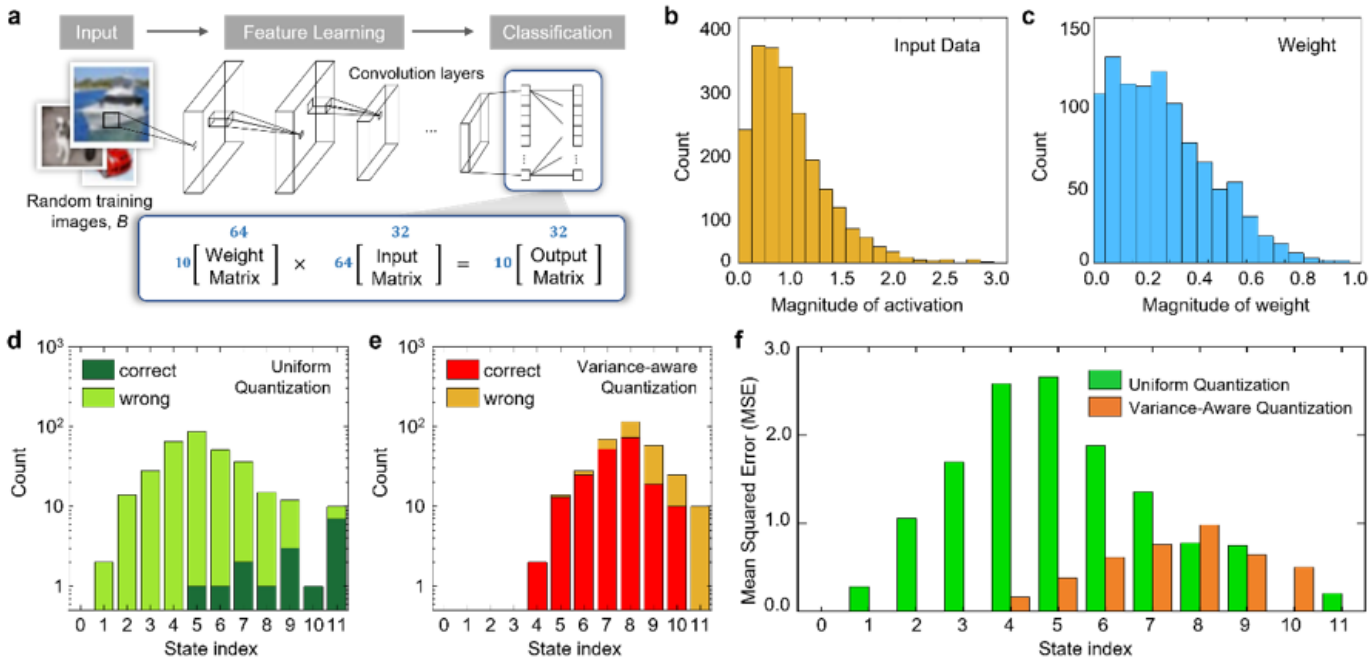

Figure 4

Variance-aware weight quantization for image classification problems. a An illustration of basic matrix operations for neural network training. The output activation matrix (10×32) is computed by multiplying the weight matrix (10×64) by the input activation matrix (64×32). We collected the data from a neural network designed for image classification tasks (ResNet20 output layer). b Histograms of the input data (i.e., activations from the previous layer) at the ResNet-20 output layer during training on CIFAR-10 dataset. c Histograms of the weight values at the same ResNet-20 output layer. d Histograms of the number of correct/wrong element-wise uniform quantization of the output matrix. We first get the ground-truth output matrix by quantizing the product of the two floating-point input matrices. We consider the quantization is correct if the quantized output element is the same as the corresponding ground-truth element. e Histograms of the number of correct/wrong element-wise variance-aware quantization of the output matrix. f Mean squared error of the output matrix. The error is calculated for 12 states separately.

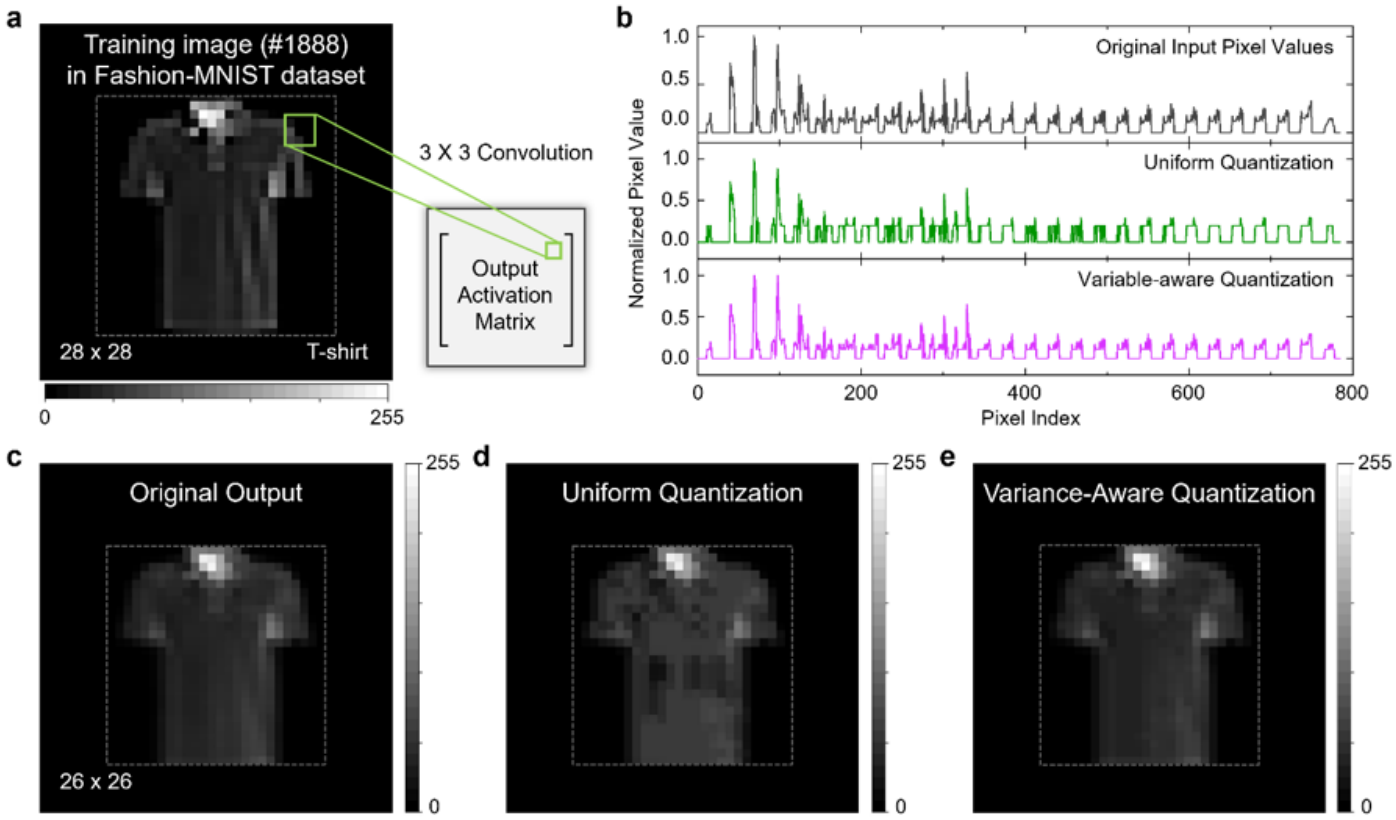

Figure 5

Variance-aware weight quantization for the convolution operation. a The training sample image (image #1888) of Fashion-MNIST dataset. The image consists of 28×28 gray-colored pixels. We apply a 3×3 convolution filter (stride of 1×1) to the input image and compare the output. b Normalized original input pixel values (top), the input pixel values quantized using the uniform states (middle), and the input pixel values quantized using the variance-aware states (bottom). c The original output of the convolution operations without quantization. d The output obtained by applying the uniform quantization. e The output obtained by applying the variance-aware quantization.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 2DEGMemristorSupplementary.docx