ADVANCED MATERIALS
www.advmat.de

Preventing Vanishing Gradient Problem of Hardware Neuromorphic System by Implementing Imidazole-Based Memristive ReLU Activation Neuron

Jungyeop Oh, Sungkyu Kim, Changhyeon Lee, Jun-Hwe Cha, Sang Yoon Yang, Sung Gap Im, Cheolmin Park, Byung Chul Jang, and Sung-Yool Choi**

With advances in artificial intelligent services, brain-inspired neuromorphic systems with synaptic devices are recently attracting significant interest to circumvent the von Neumann bottleneck. However, the increasing trend of deep neural network parameters causes huge power consumption and large area overhead of a nonlinear neuron electronic circuit, and it incurs a vanishing gradient problem. Here, a memristor-based compact and energy-efficient neuron device is presented to implement a rectifying linear unit (ReLU) activation function. To emulate the volatile and gradual switching of the ReLU function, a copolymer memristor with a hybrid structure is proposed using a copolymer/inorganic bilayer. The functional copolymer film developed by introducing imidazole functional groups enables the formation of nanocluster-type pseudo-conductive filaments by boosting the nucleation of Cu nanoclusters, causing gradual switching. The ReLU neuron device is successfully demonstrated by integrating the memristor with amorphous InGaZnO thin-film transistors, and achieves 0.5 pJ of energy consumption based on sub-10 µA operation current and high-speed switching of 650 ns. Furthermore, device-to-system-level simulation using neuron devices on the MNIST dataset demonstrates that the vanishing gradient problem is effectively resolved by five-layer deep neural networks. The proposed neuron device will enable the implementation of high-density and energy-efficient hardware neuromorphic systems.

1. Introduction

Recent advances in artificial intelligence (AI) have revolutionized existing electronic industries by creating high-tech products that perform intelligent tasks. In particular, the convergence of AI and Internet-of-Things (IoT) technology can provide smart IoT edge devices that can easily offer AI services to the general public. However, the current growing trend of neural network parameters for advanced AI services requires energy-hungry data transfer between the processor and off-chip memory in the conventional von Neumann system.[1–3] To eliminate the energy-consuming data transfer of synaptic weights, in-memory computing with non-volatile memories has been extensively explored for vector-matrix multiplication (VMM) operation, and on-chip synaptic weight storage for multi-layer perceptrons with few hidden layers.[4–6] However, modern deep neural networks (DNNs) with delicate decision boundaries have been developed with several layers, such as 152 layers of ResNET,[7] where the output nodes of each layer generate outputs via the nonlinear activation function on the weighted sum. Most of the approaches for implementing the activation function utilize general processors to compute and propagate activation functions, which eventually causes energy-hungry data transfer in and out of memory, and requires a large-area analog-to-digital converter (ADC) at the end of the synapse array.[6,8] In addition to inefficient energy consumption and a large footprint, the use of logistic sigmoid or tanh activation functions in the hidden layer

J. Oh, J.-H. Cha, S. Y. Yang, C. Park, S.-Y. Choi
School of Electrical Engineering
Graphene/2D Materials Research Center
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
E-mail: sungyool.choi@kaist.ac.kr

S. Kim
Department of Nanotechnology and Advanced Materials Engineering
Sejong University
209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea

C. Lee, S. G. Im
Department of Chemical and Biomolecular Engineering
Graphene/2D Materials Research Center
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

B. C. Jang
School of Electronics and Electrical Engineering
Kyungpook National University
41566, 80 Daehakro, BukguDaegu, Republic of Korea
E-mail: bc.jang@knu.ac.kr

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adma.202300023

DOI: 10.1002/adma.202300023

causes a vanishing gradient problem when DNNs are trained using the backpropagation learning rule.[9] To overcome this problem, the Rectified Linear Unit (ReLU) activation function with a derivative size of one is widely utilized for DNN and convolutional neural networks (CNN).[10] Therefore, the implementation of ReLU activation functions using energy- and area-efficient hardware is essential for smart IoT edge devices.

In early efforts, conventional metal-oxide-semiconductor (CMOS) circuits [11] and ADC with reconfigurable function mapping[12] were investigated to implement nonlinear activation functions. However, their energy consumption is comparable to the energy consumed by an entire synapse array during the VMM operation. [6,13] Considering that recent DNNs require an increasing number of activation functions, the implementation of an energy- and area-efficient activation function that can be integrated with the synapse array is inevitable. Thus, the challenges of these circuit-based approaches have turned the research community's attention to emerging nanoelectronic devices for low-power, high-density neuron-device applications. Phase-change memory, magnetic tunnel junctions, leaky ferroelectric field-effect transistors, and single latch-up transistors have been investigated.[14–17] While successfully mimicking a leaky integrate-and-fire model for the spiking neuron, none of these devices implemented the ReLU activation function, which requires volatile and gradual resistance change. A recent study reported a ReLU activation function using a volatile Mott device based on VO_2 with a resistor heater.[13] However, the four-terminal Mott device is unsuitable for developing highly scaled neuron devices, and requires a large operation current (~mA) for heater operation and a small on/off ratio.[18] Furthermore, the VO_2 active material has poor thermal stability owing to its well-known metal-to-insulator transition occurring at a low temperature of 340 K.[19] Consequently, NbO_2 has been primarily studied for frequency ReLU neurons,[20] temporal coding LIF neurons,[21] and selector device[22] applications. In contrast, volatile diffusive memristors have low operating currents and high on/off ratios with fast switching speeds.[23,24] Despite these advantages, diffusive memristors suffer from abrupt resistive switching and internal ionic dynamics that feature digital output and temporal signal processing,[23] making it challenging to demonstrate a continuous activation function for implementing DNN. Therefore, it is necessary to engineer both the material and device structures to implement ReLU activation function with diffusive memristor.

In this study, we propose a compact and energy-efficient memristor-based ReLU, called mReLU, for neuromorphic hardware systems, which can be utilized to negate vanishing gradient problem in DNNs. The mReLU neuron device features an analog voltage output with a current input, as well as gradual and volatile switching, both of which are essential for ReLU activation function. First, we designed a memristor with an organic/inorganic hybrid switching structure to implement a volatile and gradual resistive switching. Imidazole-based copolymer switching layer is formed via solvent-free initiated chemical vapor deposition (iCVD), which preserves inherent material properties during copolymerization. The imidazole functional group, which has two favorable bonding sites with Cu, boosts the nucleation of Cu clusters, enabling the formation of cluster-type conductive filaments. For volatile switching, a hybrid device structure with different film densities was adopted, with an atomic layer deposition (ALD)-grown Al_2O_3 layer serving as a high-density switching layer. The mReLU neuron device was demonstrated by integrating the memristor with amorphous InGaZnO (a-IGZO) thin-film transistors (TFTs), and features sub-10 uA driving current with a fast switching of 650 ns. The energy consumption of the memristor-based ReLU neuron was calculated as 0.5 pJ, which is two orders of magnitude smaller than that of a recently reported device.[13] We performed a device-to-system-level simulation with mReLU neurons to investigate the training performance of DNNs according to the activation function. It was confirmed that mReLU could effectively reduce the gradient vanishing problem in the five-layer DNN for classifying the MNIST dataset. Finally, we simulated edge detection on a real-world image using mReLU to demonstrate the functional operation of the proposed device. Our results revealed that compact and energy-efficient mReLU neurons could enable highly integrated hardware neural networks by connecting adjacent layers with reduced performance degradation.

2. Results and Discussion

A DNN, which consists of numerous neuron layers, is essential for implementing advanced AI-based tasks. **Figure 1**a shows the structure of a fully connected DNN, in which several neurons are connected with synaptic weights. The input data applied to the DNN, such as images with dense features, are multiplied by the corresponding synaptic weights, and the multiplication results are accumulated in the neurons. Because the synapses connecting the neuron layers are only responsible for the linear VMM, it is essential to use the nonlinear activation function in the neuron layer to construct a complex decision boundary. Figure 1b shows two representative activation functions. To achieve the minimal loss function, which reflects the precision of the network, an optimal activation function should be adopted because the saturation speed of the chain-rule-based backpropagation algorithm is strongly dependent on the derivative of the activation function. The sigmoid function is utilized to consider its smooth curve with derivatives of less than one. However, deeper DNN training of the synapses far from the output layer is slowed down because of the vanishing gradient problem, which results in the partial derivative of the loss function being close to zero by the successively multiplied derivative of the activation function. However, because the ReLU activation function has derivatives of one in all positive outputs, the network with the ReLU activation function can quickly achieve the minimal loss function. In contrast to the temporal signal processing neurons used in spiking neural network, after the transition point of $x = 0$, the ReLU activation function produces a linear output that depends only on the current input, regardless of any previous inputs (Figure 1b). To implement this ReLU activation function, an artificial neuron device must have volatile, linear, and gradual switching characteristics. We developed a volatile memristor-based ReLU activation device, called mReLU, with these characteristics and low power consumption in compact devices for energy-efficient hardware neuromorphic systems. In the neuromorphic system harnessing the crossbar synaptic array, the voltage signal applied to all rows is multiplied by the corresponding conductance equivalent to the synaptic weights, and the results are accumulated at the column connected to the neuron in the form of a current

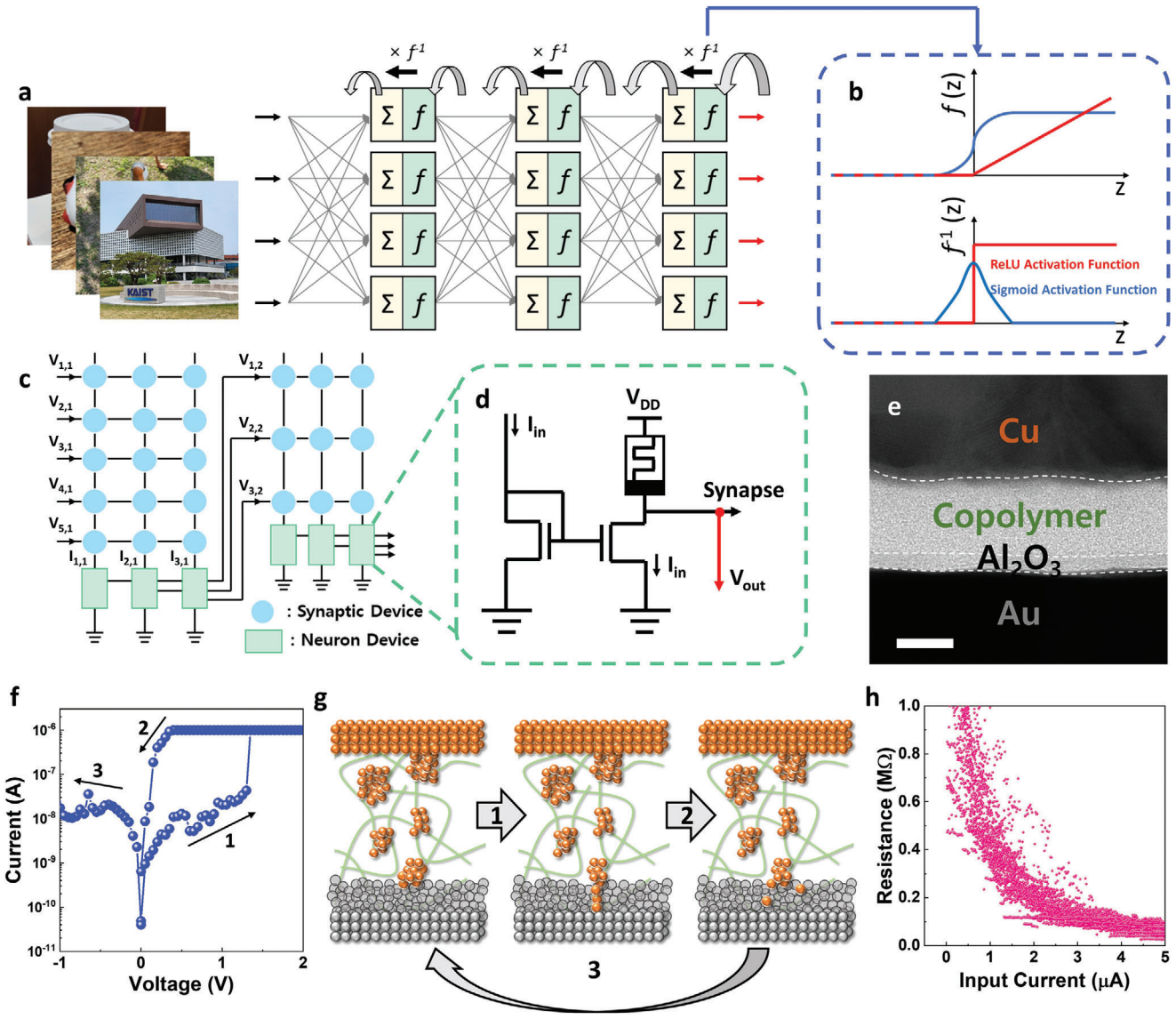

Figure 1. Activation functions in DNN and ReLU neuron implementation in neuromorphic system. a) Schematic illustration feed forward propagation and backpropagation of DNN. b) Two representative activation functions and their derivatives. c) Conceptual circuit diagram of fully analog hardware neuromorphic system. d) Schematic diagram of proposed mReLU neuron circuit. e) Cross-sectional TEM image of the hybrid memristor. f) The measured quasi-static current–voltage (I–V) switching characteristics with I_{CC} of 1 μA. g) Device schematics with CF morphology illustration for each resistance state. h) Resistance of the hybrid memristor when the I_{CC} is swept from 0 to 5 μA.

(Figure 1c). The proposed mReLU receives the weighted sum current from the synaptic array via a current mirror, and outputs the linear voltage generated by the voltage divider to be transferred to the next layer (Figure 1d). A volatile memristor with gradual resistance modulation by the input current is essential to mimic the ReLU activation function, which is the linear increase in the output voltage with respect to the input current. To implement these key features, we developed an mReLU device using an organic/inorganic double-layer memristor with a hybrid structure of Cu/copolymer (15 nm)/Al_2O_3 (3 nm)/Au, in which the copolymer was designed for essential gradual switching, as will be discussed in more detail subsequently (see Figure 1e). Figure 1f shows the I–V characteristics of the volatile switching of our mReLU device. It is noteworthy that the volatile switching of the memristor is essential for DNN neuron devices, since the neuron outputs should only depend on the current input signal, regardless of previous inputs. The mReLU showed 1) an abrupt conductance increases when the voltage exceeded the threshold during a positive bias forward sweep DC sweep with default delay required to reach the steady state owing to the fast operation speed of the memristor, and then 2) relaxed back to an insulating state during the positive backward sweep. The memristors exhibited gradual relaxation at each driving current (see Figure S3, Supporting Information), and this gradual relaxation shows a narrow distribution of relaxation speed.[25] Asymmetric hysteresis loops were observed because of the asymmetric

electrode structure of the mReLU device. During the negative bias sweep, 3) no abrupt increase in conductance was observed, even at voltages over the threshold. Figure 1g shows a schematic of the operational principle of asymmetric volatile switching. In the copolymer layer, the Cu injected during the initial electroforming process forms a local conduction path in the shape of a cluster, whereby the origin is presented in **Figure 2**. The ALD-grown Al_2O_3 layer with a relatively high density compared to the copolymer layer enables volatile switching because a small amount of Cu penetrates into the Al_2O_3 layer owing to the high film density and the penetrated Cu nanoclusters (NCs) are dissolved by the high mechanical stress gradient.[26,27] Generally, a local increase in impurities changes the physical, chemical, transport, and stress gradient properties. The local stress σ can be estimated by $\sigma = \kappa \frac{\Delta n_D}{n}$, where Δn_D [cm^{-3}] denotes the local change of n_D due to cation migration, n denotes the background atomic density of the electrolyte, and $K = E/3(1-2\nu)$ symbolizes the bulk modulus, where E denotes the Young's modulus and ν denotes the Poisson coefficient.[27] When compared to pV3D3 material, Al_2O_3 material has a higher density, Young's modular, and Poisson coefficient (Table S1, Supporting Information). When Cu cations penetrate the Al_2O_3 layer, this creates a high mechanical stress gradient, which makes Cu cations move back toward the top electrode to recover the stress field, enabling volatile switching of mReLU device. When a positive voltage was applied to the top electrode (TE), the minimal amount of Cu injected into the Al_2O_3 near the bottom electrode (BE) increased the conductance by reducing the tunneling gap (1 in Figure 1f). However, as the applied electric field weakened, the infiltrated Cu atoms decomposed or returned to the adjacent Cu clusters by surface diffusion to minimize the surface energy, thereby reverting to the insulating state (2 in Figure 1f). Owing to the asymmetric electrode structure, when a negative voltage was applied to the TE, the tunneling gap between the BE and Cu cluster near the Al_2O_3 layer did not decrease, and the resistance state remained unchanged (3 in Figure 1f). Our mReLU device shows successfully controllable conductance tuning by modulation of the tunneling gap in the low-current input range (~μA), as shown in Figure 1h. Therefore, conductance control with a low input current indicates that our mReLU device is suitable for the implementation of a low-power operating ReLU activation function.

The main driving principle of the mReLU device with structure of Cu/copolymer/Al_2O_3/Al to realize the ReLU function is the formation of a pseudo-CF composed of metal NCs with a tunneling gap. The tunneling gaps between NCs in the pseudo CF can be controlled by the electrochemical reactions and surface diffusion of active metal NCs, such as Cu or Ag, considered as bipolar electrodes,[28–30] which are electrochemically influenced by the electric current flowing through the CF. Note that the morphology of the CF determines how the conductance changes in response to the flowing current. Figure 2a compares the conductance modulation according to the continuity of the conduction path in the switching medium: continuous CF and pseudo CF, which exhibited a quantum jump in conductance as the applied input current increased. As the tunneling gaps are filled with infiltrated Cu, atomically thin metal filaments are formed between the TE and BE (Figure 2b). The resulting conductance abruptly increases to the quantum conductance ($G_0 = 2e^2/h \sim 77.5$ μS) owing to the switched conduction mechanism from electron tunneling to ballistic transport by atomic point contact formation.[31,32] However, gradual conductance evolution of CF is achieved with metal NC-based pseudo CF as its dominant conduction mechanism is electron tunneling between metal NCs.[33] As illustrated in Figure 2c, NCs are discretely connected to the adjacent metal NCs through the tunneling gap, even when the tunneling gap between specific NCs disappears by merging. Therefore, in pseudo CF, the dominant conduction mechanism remains electron tunneling, which results in gradual conductance modulation.

To form the pseudo-CF essential for gradual conductance switching, the morphology of the CF should be controlled by engineering the switching-medium material. The morphology of CF is determined by the ion mobility and redox reactions in the switching medium.[29] Copolymers synthesized using various monomers with different ion mobilities and electrochemical potentials are promising candidates as switching medium to modulate the CF morphology as well as a useful platform for scrutinizing the physics of CF morphology. To develop an mReLU device, we controlled the formation of metal NCs by copolymerizing vinylimidazole (VI) with two favorable bonding sites with Cu in the V3D3 polymer matrix with low ion mobility.[34] VI binds chemically to Cu ions, and also offers a reduction site for Cu ions in the medium owing to its electron-donating nature,[35,36] as shown in Figure 2d. Therefore, the uniformly distributed VI enhances the reduction rate of the Cu ion, serves as a nucleation site for Cu NCs, and contributes to the formation of pseudo CF based on Cu NCs. To synthesize a uniformly distributed functional material, an initiated chemical vapor deposition (iCVD) process was utilized (see Figure 2e). The iCVD process is a novel solvent-free technique for forming a polymer thin film via initiator radicals generated by thermal filaments, and is suitable for the formation of memristor materials owing to its excellent film uniformity, thickness control, low-temperature processing, and copolymerization of monomers with different polarities.[37] The 1,3,5-trimethyl-1,3,5-trivinyl cyclotrisiloxane (V3D3), utilized as a copolymer backbone, has a symmetric structure with three cross-link sites, enabling high crosslinking density, which facilitates excellent thermal/chemical stability (Figure 2f).[38] However, the V3D3 material with high ring strain owing to the small ring size of the symmetrical cyclosiloxane makes it difficult to interact with Cu;[39] thus, the pV3D3 switching layer has low ion mobility and forms a continuous thick conical CF owing to the lack of reduction sites in the medium, as demonstrated via high-resolution transmission electron microscopy (TEM) in our previous work.[31] Therefore, by copolymerizing VI containing imidazole functional groups with high polarity on the stable V3D3 polymer backbone, V3D3-VI copolymer materials have been developed to enhance the nucleation of Cu NCs and induce Cu NCs-based pseudo CF.

For a detailed investigation of the copolymer films, a composition analysis of the copolymer was conducted via X-ray photoelectron spectroscopy (XPS). Five different polymeric thin films were synthesized at different monomer flow rates. In the XPS survey shown in Figure 2g, as the VI flow rate increased, it was observed that the N1s peak increased while the Si2p peak decreased. Because V3D3 contains a silicon atom and VI contains a nitrogen atom, it enables the estimation of the contents of VI and V3D3 by investigating the ratio between silicon and nitrogen, which is used as a process-monitoring index for material optimization

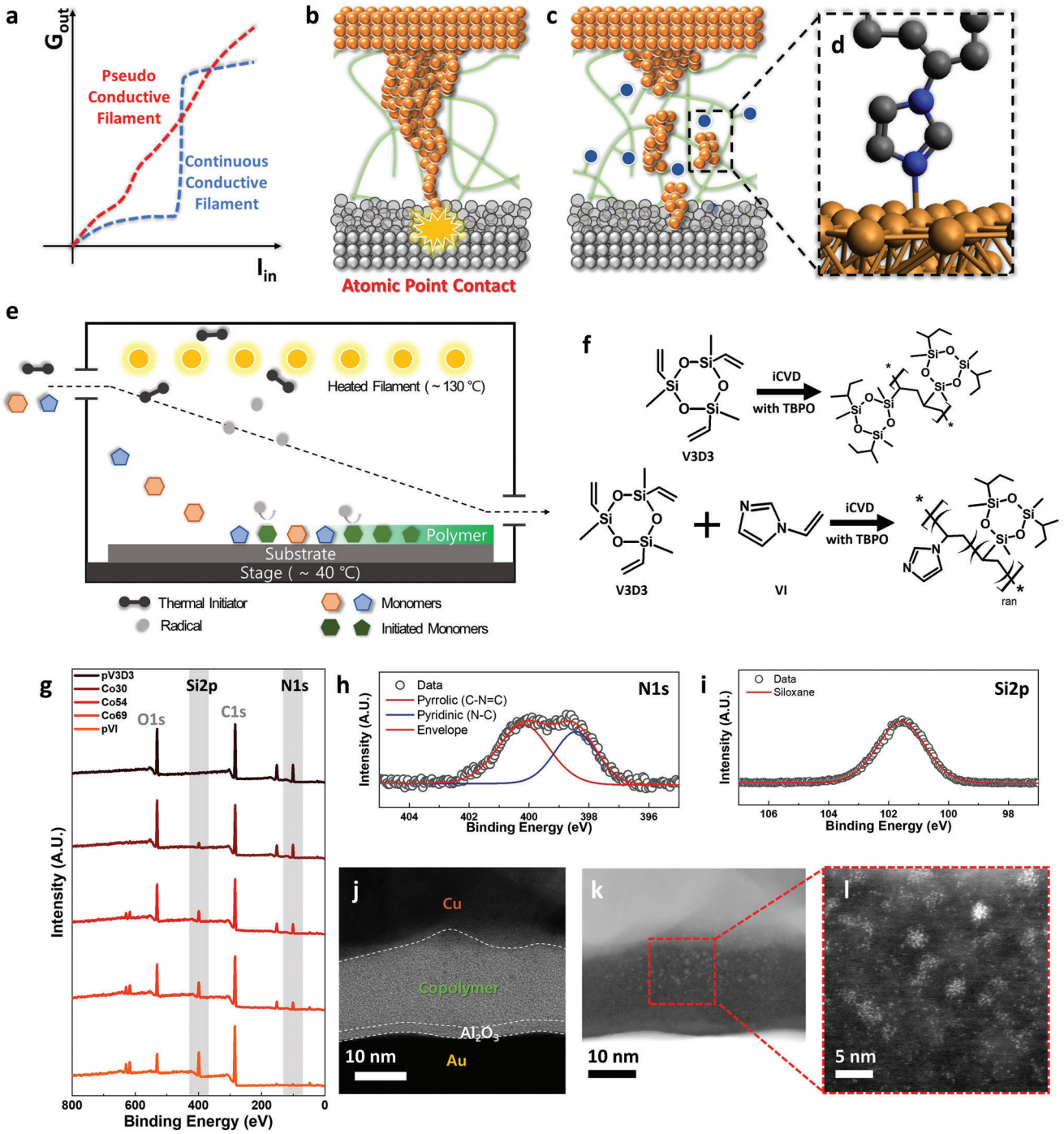

Figure 2. CF morphology modulation with copolymer switching layer. a) Comparison of change in conductance with respect to the input current by the CF morphology. Schematic illustration of CF morphology: b) continuous CF and c) pseudo CF. d) Chemical bond geometric of imidazole functional group on Cu (111) surface. e) Schematic illustration of iCVD process. Decomposed thermal initiators by heated filaments change into radicals, and react with monomers. Monomers react with radicals, change into initiated monomers, and are polymerized onto the substrate. f) Synthetic scheme of the pV3D3 and the p(V3D3-*co*-VI) film. XPS spectra: g) survey scans of copolymer films and high-resolution scan of h) N1s and i) Si2p of the copolymer film with 30% of VI content. Cross-sectional STEM analysis: j) BF-STEM image and k) HAADF-STEM image of the copolymer memristor after forming. l) Atomic-resolution HAADF-STEM images of the CF region.

(See Figure S1, Supporting Information). With this process monitoring index, it was confirmed that the flow rate of monomers enables precise control of the VI content in the copolymer film, which is related to important kinetic factors, such as metal mobility and redox rate. Among various copolymers, a copolymer film with 30% of VI content was adopted as the optimized switching medium for the generation of sub-10 nm Cu NCs. In the optimized copolymer, the deconvolution results of N1s peak consist of 400.3 and 398.4 eV peaks, which are identified as the imidazole ring containing pyridinic (N–C3) and pyrrolic (C–N=C) bonds, respectively (Figure 2h).[40] Other characteristic peaks corresponding to the siloxane (Si–O) bond were found in the Si2p spectra at 101.5 eV, which is reported to be the cyclosiloxane peak of V3D3 (Figure 2i).[34] Notably, the imidazole and cyclosiloxane groups were not damaged by the initiator or thermal filament of the iCVD process during deposition.

The morphology of the CF in the copolymer switching layer was investigated via cross-sectional scanning TEM (STEM) analysis. In the bright-field (BF)-STEM image, it was observed that Cu metal penetrated into the copolymer switching layer in the form of NCs in the local region of sub-50 nm (Figure 2j). Electron energy loss spectroscopy (EELS) identified the penetrated Cu NCs by revealing peaks at 935 and 956 eV, which correspond to the L2 and L3 edges of Cu (See Figure S2, Supporting Information). The Cu NCs were clearly described by high-angle annular dark field (HADDF)-STEM, as shown in Figure 2k,l. Therefore, nonvolatile Cu NCs with a size of sub-5 nm were successfully formed in the copolymer switching medium, which resulted from many reduction sites and a high reduction rate, limiting further growth of NC by the distributed VI.

To demonstrate the functionality of the copolymer-based mReLU device in implementing the ReLU activation function, we investigated its switching characteristics. The mReLU device returned to a high resistance state in the backward sweep during 100 consecutive DC cycles with an external compliance current (I_{CC}) of 1 μA, indicating reliable volatile switching (**Figure 3**a). Note that this volatile characteristic is inevitable for the implementation of ReLU neurons because the ReLU neuron output must be determined only by the input at any particular time, regardless of the history of the input and output. The volatile nature of the mReLU device results from the high stress gradient formed by the mechanical stress of the high-density Al_2O_3 layer compared to that of the pV3D3 layer, resulting in the surface diffusion or dissolution of the Cu NCs.[27] The stress gradient between Al_2O_3 and pV3D3 layers generates the additional force for Cu cation migration toward the top electrode. The cumulative distribution of the operational set and hold voltages showed a narrow distribution without overlap (Figure 3b). Contrary to diffusive memristors based on dielectric doping,[23] our ReLU device requires metal injection from the active electrode, enabling asymmetrical switching with rectifying features. In addition to volatile switching, the gradual conductance modulation necessary to implement the ReLU function was experimentally demonstrated in our ReLU device (Figure 3c). Because the Cu supply for forming Cu NCs is sufficient from the active TE, the tunneling gap can be adjusted in response to the flowing current. As the ReLU neuron circuit based on our mReLU device receives a summation current as input, current pulses with different heights (I_{CC}s) of 1 μA and 100 nA were applied. Figure 3d exhibits the reliable volatile switching of our mReLU device via a consecutive current pulse. The resistance of the device was well controlled with respect to the I_{CC} and exhibited volatile characteristics. Even when operated consecutively, the resistance state of the mReLU does not drift, and the output depends only on the input, regardless of the resistance state caused by prior electrical stimulation. Moreover, to further analyze the gradual switching and energy consumption, pulse voltages were applied to our mReLU device. A series resistance of 1 MΩ was attached to the mReLU device to prevent a hard breakdown. The resistances of the mReLU device were well modulated in response to 1ms-wide voltage pulses with various amplitudes, as shown in Figure 3e. The resistance decreases as the amplitude of the voltage pulse increases. This is attributed to the decreased tunneling gap between the Cu NCs owing to the pulse voltage, which is consistent with the results shown in Figures 1h and 3c. The latency of the mReLU was 650 ns, which is defined as the time interval between the pulse and saturation point of the output pulse (Figure 3f). Considering 650 ns-wide pulse voltage, the energy consumption of the mReLU neuron is calculated as 0.5 pJ, which is approximately 400 times smaller than the energy consumption of the recently reported Mott ReLU device (199.5 pJ) (**Table 1**). This ultralow energy consumption is attributed to the tunneling conduction mechanism of our mReLU device. Moreover, the proposed mReLU can be used as a replacement of area and energy-intensive CMOS activation neurons, which most general computing purpose neuromorphic hardware relies on.

To demonstrate the feasibility of our mReLU device for emulating the ReLU neuron, we integrated an mReLU and a-IGZO transistor (Figure 3g). An integrated a-IGZO transistor is necessary to map the gradual switching of mReLU onto the output voltage (V_{out}) for the ReLU function by forming a voltage divider circuit. The a-IGZO transistor has been adopted to implement the ReLU function, which generates low output in range of the negative input voltage, primarily because the wide bandgap (≈3 eV) of a-IGZO material can suppress the off current and the gate-induced drain leakage current. This is the reason why many semiconductor industry and research groups have studied the a-IGZO channel material for stackable static random access memory (SRAM) and 3D dynamic random access memory (DRAM) Cell transistor.[41,42] An optical microscopy (OM) image of the integrated device is shown in Figure 3h. The a-IGZO transistor was fabricated with a W/L ratio of 100 μm/10 μm, and the mReLU was patterned at 50 μm × 50 μm on the drain region of the a-IGZO transistor (Figure S4, Supporting Information). The mReLU, capable of a low-temperature process below 150 °C, is suitable for integration with a-IGZO transistors. The weighted sum current from the synaptic array flows into the a-IGZO transistor through the current mirror circuit, which results in resistance modulation of the mReLU. Considering that the supply voltage (V_{DD}) is divided between the a-IGZO transistor and mReLU, V_{out} increases linearly as a result of the decreasing resistance of our mReLU device in response to the increasing sum current. To explore the feasibility of a large-scale neuron array, we investigated the device-to-device variation of a-IGZO transistors and the bilayer memristors that constitute the proposed neuron circuit (Figures S5 and S6, Supporting Information). It was found that 20 a-IGZO transistors exhibited a narrow threshold voltage distribution of 0.038 V, and the uniform characteristics for the volatile and incremental

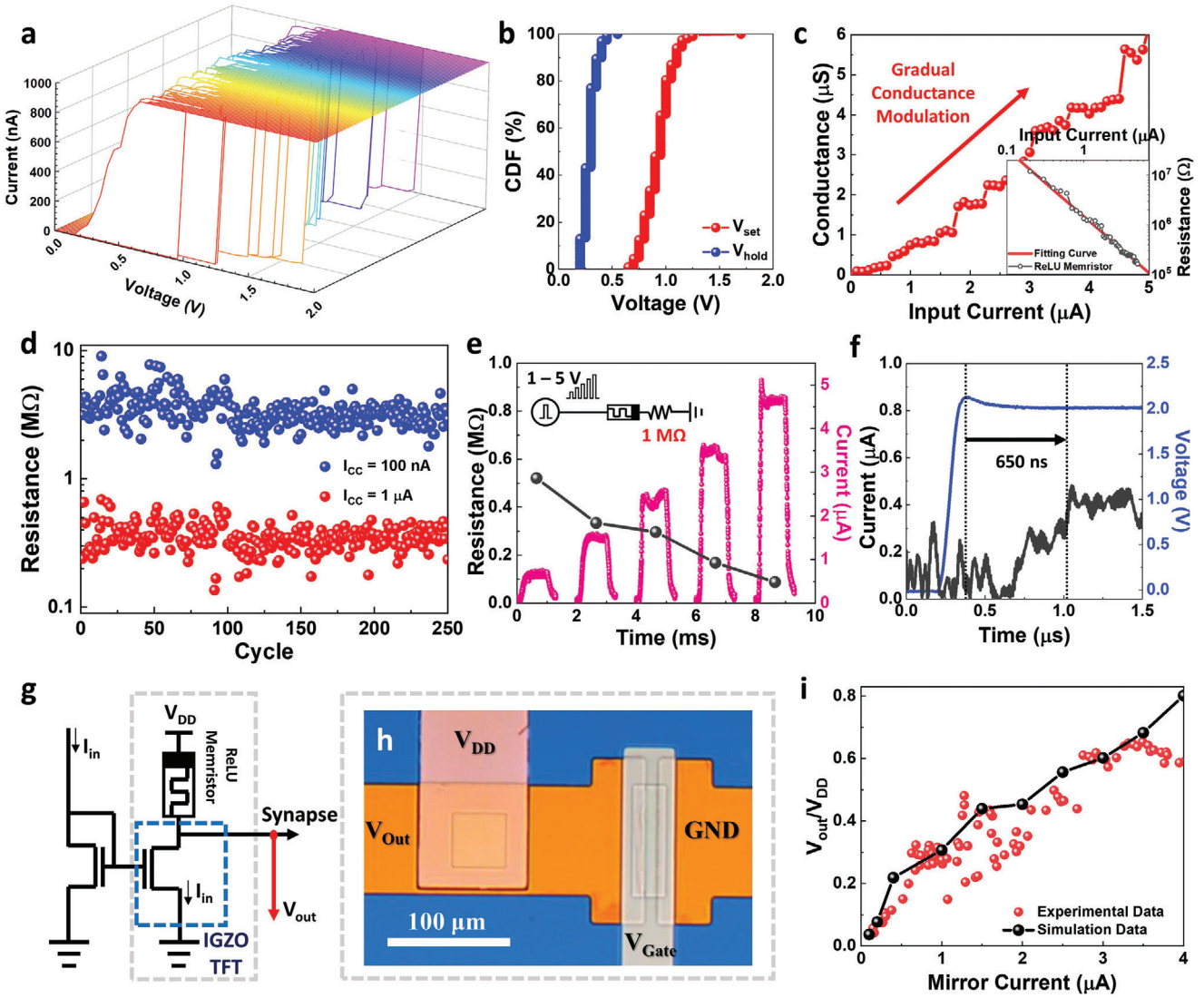

Figure 3. The electrical characteristics of the hybrid memristor and mReLU. a) Consecutive 100 DC sweep curves of the device. b) Cumulative curves of set voltage and hold voltage with consecutive 200 DC sweep. c) Conductance switching curves with the current sweep from 0 to 5 μA of the device. Inset shows resistance fitting as the function of the current. Pulse switching of the hybrid memristor. d) Resistance state change of the device by flowing 100 nA and 1 μA current pulse with 5 ms pulse with and 10 ms period, alternatively. e) Resistance change of the device with 1 MΩ of series resistance by stimulating the incremental voltage pulse. f) Voltage pulse applied to the device and the current flow as a function of time. g) Schematic diagram of integrated mReLU neuron. h) OM image of integrated mReLU neuron with 1 IGZO TFT-1 hybrid memristor. i) Voltage ratio between V_{OUT} and V_{DD} of the device as a function of mirror current. Red circles are experimental data and black circles are SPICE simulation results.

conductance update characteristics of 25 bilayer memristors were verified. The threshold voltage and transconductance of the a-IGZO transistor and the current-conductance characteristics of the mReLU were used to model the integrated device (Figure S7a, Supporting Information). The input current was successfully copied to the output transistor through the current mirror (Figure S7b, Supporting Information). The relationship between the current flowing through the output transistor and V_{out} is illustrated in Figure 3i. The result of SPICE, well known as a general-purpose circuit simulator, shows a complete correlation between the mirrored current and V_{out}, indicating that a function for the input current and output voltage has a ReLU functionality. The experimentally measured results exhibit the same trend as the simulation results; thus, demonstrating the feasibility of using mReLU neurons for implementing ReLU neurons. The output of our mReLU device was the voltage. There is no need for large-area analog-digital circuitry because this output voltage can be directly applied to the following neuron layer as the input voltage. Additionally, the small footprint of our mReLU device allows the integration of each column of the synaptic array, enabling fast parallel operation without time multiplexing.

To confirm the applicability of the proposed mReLU neurons to the DNN, a device-to-system-level simulation for evaluating the vanishing gradient problem was performed. The role of neurons in DNN inference includes the summation of weighted currents from synapses, imparting nonlinearity to the

Table 1. Summary of analog artificial neurons for the hardware-based neuromorphic system.

Neuron Materials	Switching Mechanism	Structure Complexity	Neuron Model	Energy Consumption	Neural Network Model	Learning Rule
$Pt/SiO_xN_y:Ag/Pt$[44]	Diffusive Dynamics	1M1C	LIF	–	SNN	STDP
$GeSb_2Te_5$[45]	Phase Change	1M	IF	5 pJ	–	–
$TiN/NbO_x/TiN$[21]	Metal-Insulator Transition	1M	LIF	0.5 pJ	Temporal Coding SNN	Temporal BP
Silicon NpN junction[46]	Baristor	1T	LIF	6 pJ	–	–
MJT Material[15]	Magnetic Tunneling Junction	1T	LIF	7.1 fJ	SNN	STDP
$Hf_{0.5}Zr_{0.5}O_2$[47]	Ferroelectric	1F-1T1R	LIF	–	SNN	–
CMOS[11]	Analog CMOS	> 3 OPAMPs	Sigmoid	3410 pJ	DNN	BP
$TiN/NbO_x/a\text{-}Si$[20]	Metal-Insulator Transition	1T-1M	Frequency-ReLU	–	SNN	Stochastic Gradient Descent
Cu/p(V3D3-*co*-VI)/Al_2O_3/Al [This Work]	**Diffusive Dynamics**	**2T-1M**	**ReLU**	**0.5 pJ**	**DNN**	**BP**

summation currents, and sending out the outputs to the next neuron layer (**Figure 4**a), which correspond to Kirchhoff's law in the synaptic crossbar array, rectifying function of mReLU, and voltage output, respectively. Furthermore, in addition to calculating the derivative value of the neuron output for error-backpropagation, the neuron for training the DNN sends the output to the next neuron layer. Because the mReLU device has a derivative value of zero or one on its output, the computational overhead for calculating the derivative of the activation function can be minimized. The performance of the proposed mReLU neuron was evaluated using a five-layer DNN with a 784-100-100-100-100-10 structure (Figure 4b). MNIST handwritten digits with 60 000 training samples and 10 000 test samples with 28 × 28 pixels were utilized. The performance of the DNN was evaluated based on the mReLU and sigmoid activation functions used in the hidden neuron layer, and the gradient vanishing problem with a minibatch size of 500 was analyzed through the cross-entropy loss function using the Softmax activation function. The recognition rate of the neural network using the mReLU and sigmoid functions achieved recognition rates of 95.4% and 91.2%, respectively (Figure 4c). This difference in the recognition rate is attributed to the learning speed of the neural network, which is supported by the rapid convergence of the loss function by the fast training of the neutral network (Figure 4d). The difference in the convergence speed of the loss function is interpreted as gradient loss occurring in the deep layer, which may be caused by the distribution of the output. The distribution of the output is significantly affected by the number of input neurons and initial synapse weight distribution, and was evaluated using the He normal distribution weight initialization (Figure 4e,f).[43] The sigmoid function exhibits a wider output distribution in the initial neuron layer, which has more input neurons than the other layers (Figure 4e); therefore, the derivatives of the neuron tend to be smaller in the initial neuron layer, indicating the vanishing gradient problem. In contrast, the mReLU function features a uniform output distribution in all neuron layers with an average value close to zero, indicating that half of the neurons have a derivative of one (Figure 4f). Therefore, although the sigmoid function showed a smaller weight update in the first synaptic layer than in the fifth synaptic layer (Figure 4g), the mReLU function exhibited a uniform weight update in both the first and fifth synaptic layers (Figure 4h). To verify the neural network performance degradation due to the non-ideal effect of mReLU neuron, we performed device-to-system level simulations of five-layer DNN with mReLU neuron for MNIST handwritten digit recognition (Figure 4i). All neuron circuits, except for the software neuron, return a maximum output voltage of 2 V, which corresponds to the driving voltage of the mReLU circuit. mReLU neurons with different temporal variations were evaluated, while the ideal mReLU neuron represents a neuron without temporal variation. The mReLU neuron, with a maximum value of 2 V and 9.9% of temporal variation, achieved a recognition rate of 95.4%, while software ReLU exhibited a 97% recognition rate.

To demonstrate the functional operation of our mReLU neuron, we simulated the edge detection operation using convolutional filters on a real-world image (**Figure 5**a). We recognized images by combining low-level edges, and this function is implemented through a convolution. During convolution, a filter is applied to each pixel and the local neighborhood of the image, producing an output from the weighted sum between the filter weights and the input pixel values. Simple vertical and lateral filters were used to evaluate the performance of the mReLU neurons (Figure S8, Supporting Information). From four representative 10 × 10 input patches (Figure 5b), the summation currents obtained through the convolution operation with vertical and horizontal filters were activated using mReLU neurons. Figure 5c shows the summation current convolved through the vertical and horizontal filters. It was observed that the magnitude and sign of the current change depended on the pixel position. The output voltage from the mReLU is shown in Figure 5d, and the maximum voltage is bounded to 2 V as V_{DD} and rectified as a positive number. Figure 5e shows the output voltage for the entire image. Utilizing the mReLU activation function, the lateral and vertical edges were detected well. Therefore, the mReLU neuron device has great potential as a compact low-power neuron device for hardware-based neuromorphic systems.

3. Conclusion

We developed a compact and energy-efficient memristor-based ReLU activation neuron device that can solve the chronic

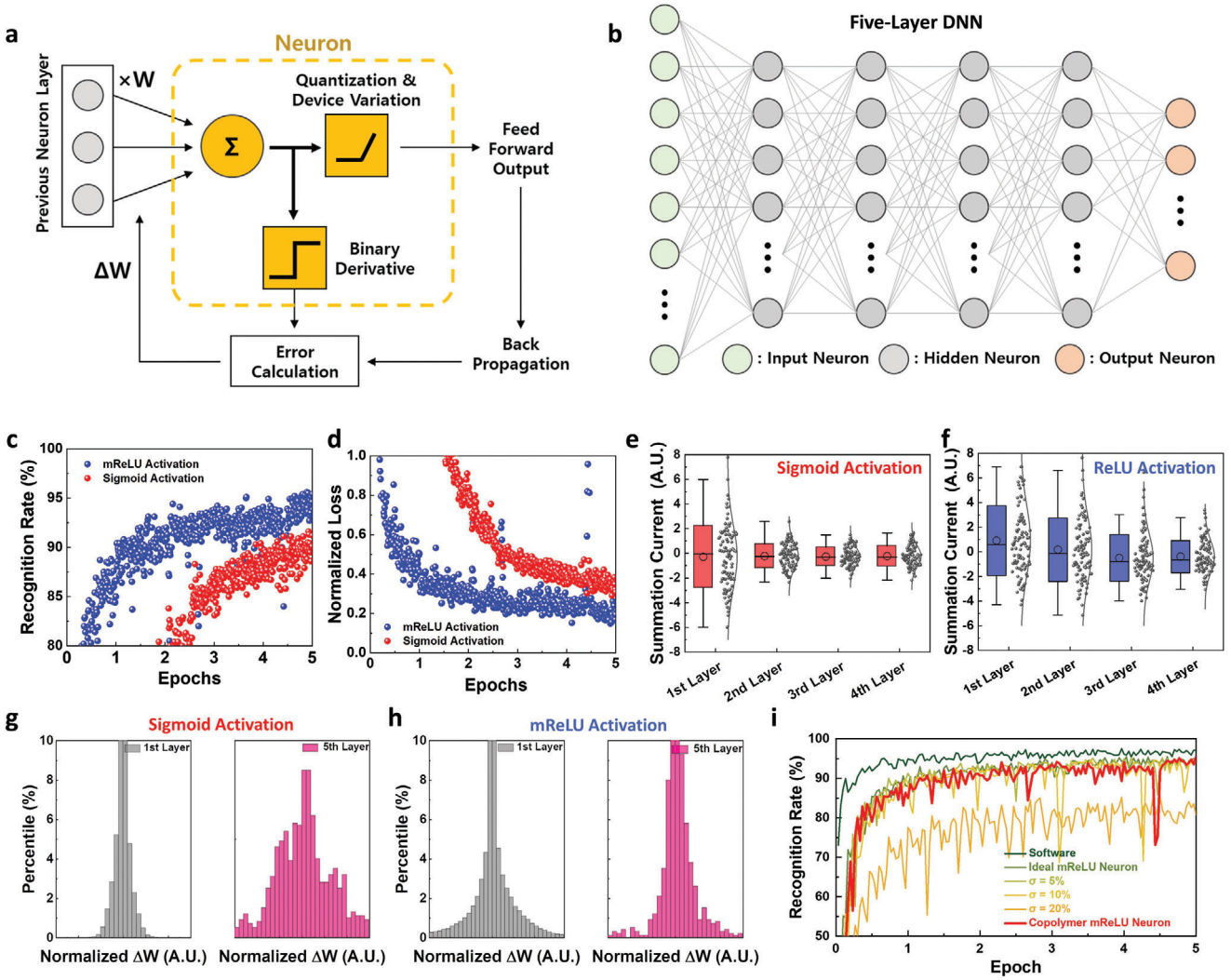

Figure 4. Vanishing gradient simulation for DNN a) Conceptual diagram of the neuron role in DNN. b) A five-layer DNN used to classify handwritten digits in the MNIST database. c) Accuracy and d) normalized loss with the sigmoid and the ReLU activation function. Summation current output of the trained DNN with e) the sigmoid and the ReLU activation function. Probability distribution of the weight update using g) the sigmoid and h) the ReLU activation function. i) The network simulation results for DNN with mReLU activation neuron.

vanishing gradient problem in advanced DNN. The mReLU neuron device has current input and voltage output, which allows direct connection with the adjacent synapse layer. To emulate the volatile and linear resistance changes of the memristor, a hybrid switching structure was introduced. By copolymerizing the monomers, the CF morphology was modulated to cluster-based CF with gradual conductance modulation, and a volatile nature was achieved by utilizing a high-density Al_2O_3 layer. The fabricated device achieves low power consumption based on sub-10 μA operation and fast switching thanks to cluster-based CF. We performed a device-to-system-level simulation to investigate hardware neuromorphic systems based on mReLU, demonstrating that our mReLU neuron device can effectively solve the vanishing gradient problem. Furthermore, we demonstrated that the mReLU neuron device can be used as an activation function for edge detection and neurons in DNNs. We believe that the proposed mReLU can reduce the energy and area overload on the peripheral circuit for implementing the activation function, and solve the chronic vanishing gradient problem in DNN, thereby providing a solution for energy-efficient hardware neuromorphic systems.

4. Experimental Section

Polymeric Film Deposition via iCVD: Five different polymeric layers were deposited into the iCVD reactor (Daeki Hi-Tech Co, Ltd.) with the vaporized monomers and initiator. V3D3 (95%, Gelest, USA) and VI (99%, Aldrich, USA) were used as the monomers, and *tert*-butyl peroxide (TBPO, 98%, Aldrich, USA) was used as the initiator.

Device Fabrication: The hybrid memristors were fabricated with a-IGZO transistors on a thermal grown 90 nm SiO_2 substrate. Thermal evaporation and lift-off process were used for the Au (50 nm)/Cr (5 nm) of source, drain and gate electrode. 30 nm thick a-IGZO channel was deposited through RF sputtering, and was patterned with wet-etch process.

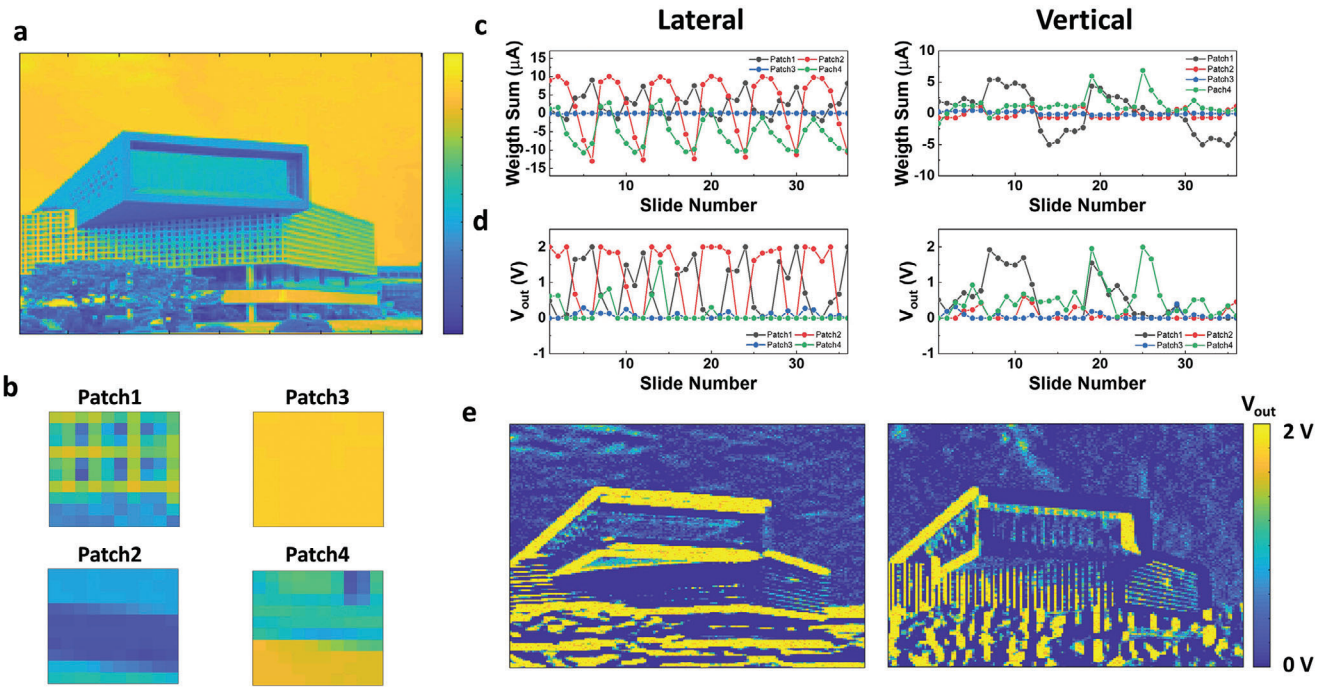

Figure 5. Edge detection demonstration. a) A real-world 160 × 160 image for edge detection. Color map represent the pixel intensity. b) Four representative 10 × 10 patches. c) Summation current outputs for the lateral filter (left) and the vertical filter (right) during the convolution operation. d) Output voltages of mReLU neuron fed from the lateral filter (left) and the vertical filter (right). e) Output voltage of the mReLU neuron for the whole image during the convolution for the lateral filter (left) and the vertical filter (right).

Thermal ALD grown 30 nm thick Al_2O_3 layer was used as gate dielectric. Active layer of the memristor was patterned on the gate dielectric with wet etch process. 3 nm thick ALD grown Al_2O_3 layer and 15 nm thick p(V3D3-*co*-VI) film deposited with iCVD process were used as memristor switching dielectric. For the top active electrode, 50 nm thick Cu film was deposited with thermal evaporation, and was patterned with wet-etch process.

Device Characterization: Keithley 4200 semiconductor parameter analyzer was used to investigate the electric characteristics of the fabricated devices. Electrical measurements were performed under ambient air condition at room temperature. Keithley 4200 and Keithley 4225-PMU (pulse generator), and 4225-RPM (remote amplifier/switch) were used to perform DC sweep measurements and voltage pulse measurement, respectively. A cross-sectional TEM sample was prepared using a focused ion beam (FIB, FEI Helios Nano Lab 450 HP). STEM images were obtained using a JEOL ARM 200F instrument at an accelerating voltage of 200 kV and a FEI Titan G^2 60–300 with a dual spherical aberration Cs-corrector at an accelerating voltage of 300 kV. EELS spectra were obtained using a Gatan Imaging Filter.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

J.O. and S.K. contributed equally to this work. This work was supported by the Basic Science Research Program through the NRF, funded by the Ministry of Education (NRF Award No. NRF-2022R1A2B5B02002189, NRF-2020M3F3A2A01082618, NRF-2022R1C1C1006557 and NRF-2021R1C1C1008949).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

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

Keywords

deep neural network, initiated chemical vapor deposition, neuromorphic computing, ReLU activation neuron, vanishing gradient problem

Received: January 2, 2023
Revised: March 5, 2023
Published online: April 27, 2023

[1] J. Hasler, B. Marr, *Front Neurosci* **2013**, *7*, 118.
[2] M. A. Zidan, J. P. Strachan, W. D. Lu, *Nat. Electron.* **2018**, *1*, 22.
[3] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, Y. Shi, *Nat. Electron.* **2018**, *1*, 216.
[4] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, *Adv. Mater.* **2018**, *30*, 1705914.
[5] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, H. Qian, *Nature* **2020**, *577*, 641.
[6] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, *Nature* **2022**, *608*, 504.

[7] K. He, X. Zhang, S. Ren, J. Sun, presented at Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June **2016**.
[8] X. Peng, S. Huang, H. Jiang, A. Lu, S. Yu, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2020**, *40*, 2306.
[9] G. B. Goh, N. O. Hodas, A. Vishnu, *J. Comput. Chem.* **2017**, *38*, 1291.
[10] X. Glorot, A. Bordes, Y. Bengio, presented at Proc. of the 14th Int. Conf. on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, April **2011**.
[11] O. Krestinskaya, K. N. Salama, A. P. James, *IEEE Trans Circuits Syst I Regul Pap* **2018**, *66*, 719.
[12] M. Giordano, G. Cristiano, K. Ishibashi, S. Ambrogio, H. Tsai, G. W. Burr, P. Narayanan, *IEEE J Emerg Sel Top Circuits Syst* **2019**, *9*, 367.
[13] S. Oh, Y. Shi, J. Del Valle, P. Salev, Y. Lu, Z. Huang, Y. Kalcheim, I. K. Schuller, D. Kuzum, *Nat. Nanotechnol.* **2021**, *16*, 680.
[14] R. Yang, H. M. Huang, X. Guo, *Adv. Electron. Mater.* **2019**, *5*, 1900287.
[15] A. Sengupta, P. Panda, P. Wijesinghe, Y. Kim, K. Roy, *Sci. Rep.* **2016**, *6*, 30039.
[16] H. Mulaosmanovic, E. Chicca, M. Bertele, T. Mikolajick, S. Slesazeck, *Nanoscale* **2018**, *10*, 21755.
[17] J.-K. Han, M. Seo, J.-M. Yu, Y.-J. Suh, Y.-K. Choi, *IEEE Electron Device Lett.* **2020**, *41*, 1157.
[18] D. Ruzmetov, G. Gopalakrishnan, J. Deng, V. Narayanamurti, S. Ramanathan, *J. Appl. Phys.* **2009**, *106*, 083702.
[19] C. Wu, F. Feng, Y. Xie, *Chem. Soc. Rev.* **2013**, *42*, 5157.
[20] X. Zhang, Z. Wang, W. Song, R. Midya, Y. Zhuo, R. Wang, M. Rao, N. K. Upadhyay, Q. Xia, J. J. Yang, Q. Liu, M. Liu, presented at IEEE Tech. Dig. Int. Electron Devices Meet. , San Francisco, CA, USA **2019**.
[21] X. Zhang, Z. Wu, J. Lu, J. Wei, J. Lu, J. Zhu, J. Qiu, R. Wang, K. Lou, Y. Wang, T. Shi, C. Dou, D. Shang, Q. Liu, M. Liu, presented at IEEE Tech. Dig. Int. Electron Devices Meet. , San Francisco, CA, USA **2020**.
[22] E. Cha, J. Woo, D. Lee, S. Lee, J. Song, Y. Koo, J. Lee, C. G. Park, M. Y. Yang, K. Kamiya, K. Shiraishi, B. Magyari-Köpe, Y. Nishi, H. Hwang, presented at IEEE Tech. Dig. Int. Electron Devices Meet. , Washington, DC, USA **2013**.
[23] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, *Nat. Mater.* **2017**, *16*, 101.
[24] R. Midya, Z. Wang, S. Asapu, X. Zhang, M. Rao, W. Song, Y. Zhuo, N. Upadhyay, Q. Xia, J. J. Yang, *Adv. Intell. Syst.* **2019**, *1*, 1900084.
[25] F. Ye, F. Kiani, Y. Huang, Q. Xia, *Adv. Mater.* **2022**, 2204778.
[26] T. Jung, S. Jeon, *J. Mech. Sci. Technol.* **2022**, *40*, 042201.
[27] S. Ambrogio, S. Balatti, S. Choi, D. Ielmini, *Adv. Mater.* **2014**, *26*, 3885.
[28] J. Kang, T. Kim, S. Hu, J. Kim, J. Y. Kwak, J. Park, J. K. Park, I. Kim, S. Lee, S. Kim, *Nature Commun.* **2022**, *13*, 4040.
[29] Y. Yang, P. Gao, L. Li, X. Pan, S. Tappertzhofen, S. Choi, R. Waser, I. Valov, W. D. Lu, *Nat. Commun.* **2014**, *5*, 4232.
[30] W. Wang, M. Wang, E. Ambrosi, A. Bricalli, M. Laudato, Z. Sun, X. Chen, D. Ielmini, *Nat. Commun.* **2019**, *10*, 81.
[31] B. C. Jang, S. Kim, S. Y. Yang, J. Park, J.-H. Cha, J. Oh, J. Choi, S. G. Im, V. P. Dravid, S.-Y. Choi, *Nano Lett.* **2019**, *19*, 839.
[32] K. Terabe, T. Hasegawa, T. Nakayama, M. Aono, *Nature* **2005**, *433*, 47.
[33] J. H. Yoon, J. Zhang, P. Lin, N. Upadhyay, P. Yan, Y. Liu, Q. Xia, J. J. Yang, *Adv. Mater.* **2020**, *32*, 1904599.
[34] K. Pak, H. Seong, J. Choi, W. S. Hwang, S. G. Im, *Adv. Funct. Mater.* **2016**, *26*, 6574.
[35] N. Kovačević, I. Milošev, A. Kokalj, *Corros. Sci.* **2015**, *98*, 457.
[36] D. Kumar, V. Jain, B. Rai, *Corros. Sci.* **2020**, *171*, 108724.
[37] S. J. Yu, K. Pak, M. J. Kwak, M. Joo, B. J. Kim, M. S. Oh, J. Baek, H. Park, G. Choi, D. H. Kim, *Adv. Eng. Mater.* **2018**, *20*, 1700622.
[38] H. Moon, H. Seong, W. C. Shin, W.-T. Park, M. Kim, S. Lee, J. H. Bong, Y.-Y. Noh, B. J. Cho, S. Yoo, *Nat. Mater.* **2015**, *14*, 628.
[39] U. Das, G. Zhang, B. Hu, A. S. Hock, P. C. Redfern, J. T. Miller, L. A. Curtiss, *ACS Catal.* **2015**, *5*, 7177.
[40] B. Wang, H.-J. Liu, Y. Chen, *RSC Adv.* **2016**, *6*, 2141.
[41] S.-W. Chang, T.-H. Lu, C.-Y. Yang, C.-J. Yeh, M.-K. Huang, C.-F. Meng, P.-J. Chen, T.-H. Chang, Y.-S. Chang, J.-W. Jhu, T.-Z. Hong, C.-C. Ke, X.-R. Yu, W.-H. Lu, M. A. Baig, T.-C. Cho, P.-J. Sung, C.-J. Su, F.-K. Hsueh, B.-Y. Chen, H.-H. Hu, C.-T. Wu, K.-L. Lin, W. C.-Y. Ma, D.-D. Lu, K.-H. Kao, Y.-J. Lee, C.-L. Lin, K.-P. Huang, K.-M. Chen, et al., presented at Tech. Dig. Int. Electron Devices Meet. , San Francisco, CA, USA **2021**.
[42] S. Subhechha, N. Rassoul, A. Belmonte, H. Hody, H. Dekkers, M. J. van Setten, A. Chasin, S. H. Sharifi, S. Sutar, L. Magnarin, U. Celano, H. Puliyalil, S. Kundu, M. Pak, L. Teugels, D. Tsvetanova, N. Bazzazian, K. Vandersmissen, C. Biasotto, D. Batuk, J. Geypen, J. Heijlen, R. Delhougne, G. S. Kar, presented at IEEE Symp. on VLSI Technology & Circuits Digest of Technical Papers, Honolulu, HI, USA **2022**.
[43] K. He, X. Zhang, S. Ren, J. Sun, presented at Proc. IEEE Int. Conf. on Computer Vision, Santiago, Chile, Dec **2015**.
[44] Z. Wang, S. Joshi, S. Savel'ev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo, *Nat. Electron.* **2018**, *1*, 137.
[45] T. Tuma, A. Pantazi, M. L. Gallo, A. Sebastian, E. Eleftheriou, *Nat. Nanotechnol.* **2016**, *11*, 693.
[46] J.-W. Han, M. Meyyappan, *IEEE Electron Device Lett.* **2018**, *39*, 1457.
[47] J. Luo, L. Yu, T. Liu, M. Yang, Z. Fu, Z. Liang, L. Chen, C. Chen, S. Liu, S. Wu, presented at Tech. Dig. Int. Electron Devices Meet, San Francisco, CA, USA **2019**.