Kernel Application of the Stacked Crossbar Array Composed of Self-Rectifying Resistive Switching Memory for Convolutional Neural Networks

Yumin Kim, Jihun Kim, Seung Soo Kim, Young Jae Kwon, Gil Seop Kim, Jeong Woo Jeon, Dae Eun Kwon, Jung Ho Yoon, and Cheol Seong Hwang*

Herein, a feasible method is provided for circuit implementation of the convolutional neural network (CNN) in neuromorphic hardware using the multiple layers-stacked resistance switching random access memory (ReRAM). The specific ReRAM is accompanied by self-rectification functionality. The single-input multiple-output (SIMO) scheme is an optimum method in the extraction of the features of a letter with a versatile selection of the intended features, whereas the multiple-input single-output (MISO) scheme provides a highly efficient method to extract the features from the color image, which is composed of several component color images. The Pt/HfO2–x/TiN-based self-rectification ReRAM that is integrated into the sidewalls of the two-layer structure provides a sound framework for the circuit implementation of the SIMO and MISO schemes. The appropriate selection of the kernels for image compression and feature extraction greatly facilitates the CNN in neuromorphic hardware.

1. Introduction

In modern artificial intelligence systems, image recognition technology is a crucial part of the systems from simple characters\(^1\)–\(^3\) and picture\(^3\)–\(^5\) classifications to more complicated tasks, such as image captioning\(^6\)–\(^7\), speech recognition\(^8\)–\(^9\), language processing\(^10\), and even operation of self-driving vehicles\(^11\). The system is usually based on an artificial neural network (ANN) structure.

To make ANN recognize images effectively, convolutional neural network (CNN) is widely used\(^12\)–\(^22\). In CNN, the raw image is filtered to acquire a series of appropriately subsampled feature maps. Then, the filtered feature maps are fed into the subsequent fully connect network (FCN) to classify and identify the original image. One of the problems of CNN is that CNN is a vast memory usage for inputting the image. Another issue is about the shallow feature extraction layer, especially for large color images. For example, the memory size for the input image is proportional to the product of the side lengths of the image (or the number of pixels) for hand-written letters. For color images, tripled memory use is required to input the three component color images into the red, green, and blue channels. A large amount of memory increases the load for arithmetic calculation in CNN.

The performance of ANN has been enormously improved mainly by the development of the back-propagation method in the deep neural network. Such a deep layer structure requires a huge consumption of energy to perform so many vector-matrix multiplications (VMMs). Moreover, the repeated weight update requires frequent movement of data between the memory and processors, which hugely increases the time and energy consumption\(^13\)–\(^15\).

In this regard, the crossbar array (CBA) using the nonvolatile memory (NVM) is the core element for a physical calculation system, where the NVM could be the resistive switching random access memory (ReRAM)\(^15\)–\(^19\), or phase-change RAM (PcRAM)\(^20\)–\(^21\), instead of the software-based VMM. Despite great improvements in these NVM-based neuromorphic systems, two major problems remain. These are lack of high-performance selection devices and device variability (operation to operation and device to device). Transistor-based systems have been mainly developed to address the first problem\(^15\)–\(^16\),\(^18\)–\(^20\)\(^21\), which actually undermine the major merits of the CBA (3D stackability)\(^22\). However, introducing a high-functionality embedded selector, which is referred to as the self-rectifying ReRAM\(^23\)–\(^26\), greatly reduces the feature size and makes the stacking of the CBA highly feasible. For the second problem, off-chip learning followed by write-and-verify scheme is mostly favored during training, which greatly mitigates the problems related to the
variations. This strategy is also adaptable for the “mostly inference” system, of which the values are not being frequently updated. Alternatively, modifying the readout circuit structure\cite{27, 28} or still improving device fabrication parameters can be pursued to solve these problems. This work adopted a self-rectifying ReRAM, which embeds a high-functionality selector.

Such self-rectifying ReRAM greatly simplifies the system architecture, making the stacking of the CBA highly feasible and flexible. Due to such flexibility, a specific structure of the stacked CBA (S-CBA), whose fabrication method is described in the Experimental Section in detail, could be fabricated. The S-CBA is highly compatible with the CNN functionality because its various layers can simultaneously correspond to the multiple input or output vectors of the neuromorphic system.

This work provides two critical improvements, first, the methods for implementing the CNN convolution function using the self-rectifying ReRAM-based S-CBA, which is especially efficient in treating multiple input or output vectors, and second, fabrication of S-CBA using the self-rectifying ReRAM (Pt/HfO$_2$-x/TiN memory) and experimental demonstration of the CNN functionality for simple image processing. The adopted self-rectifying Pt/HfO$_2$-x/TiN ReRAM cells were rather variation prone, but the suggested application scheme demonstrated the immunity of the system performance over variation.

2. Results and Discussion

A common method for implementing the convolution layer of the CNN is extracting the feature maps into smaller sizes through CBA bit lines (BLs), whereas a portion of the input image is fed into the CBA through word lines (WLs), which is commonly regarded as applying a kernel (or filter). In the neuromorphic hardware using the S-CBA, implementing such kernel functions can be readily understood from S1 Figure S1 and S2. Supporting Information, of which details are explained in S1-I (basic idea of applying the S-CBA to a kernel using an inverse mapping method). The novelty of the present work could be twofold: first, the entire image inputted to the WLs is processed through the BLs simultaneously, with multiple feature maps along the layer stacks, and a smaller processed image is outputted. Second, the large resistance variance problem in the Pt/HfO$_2$-x/TiN self-rectifying ReRAM does not affect the overall system performance, which is generally not the case for other systems based on the conventional paired conductance devices using Ohm’s law.

The first example showing the usefulness of the S-CBA is shown in Figure 1, where the four kernels are used to extract the feature of the hand-written number 5. First, the hand-written number 5, shown in a binarized form within a 14 × 14 (total 196) pixel (left panel in Figure 1a), is intended to be convoluted into the 13 × 13 compressed image with a right-edge feature, i.e., any subimage composed of 2 × 2 pixels, with the two right-hand pixels being white, will produce white pixels in the compressed image. This can be performed by constituting four kernels, as shown in the middle panel of Figure 1a. For the specific case of Figure 1a, the yellow squares in the kernels will correspond to the low-resistance state (LRS) in the hypothetical S-CBA, which will be described later. A $V_r$ of 6.3 V will be applied to the WLs connecting the cells when the corresponding pixels in the sub-image (red box in the left panel of Figure 1a) are white, whereas the other WLs corresponding to the black pixels become floated. A $V_r$ of 6.3 V was chosen to make an actual cell read voltage of 6 V, taking into consideration voltage drops in the peripheral circuit. Here, the actual cell read voltage (6 V) was determined by the current-voltage ($I$–$V$) characteristics of the self-rectifying ReRAM, which will be discussed in the next section. In this case, there will be three WLs at $V_r$, whereas the remaining one WL will be floated, as shown in the middle panel of Figure 1a.

The next important thing is to select the appropriate BLs to extract the feature of the right edge. This can be achieved by selecting BLs representing the kernels #1 and #3 and connecting them to the two inputs of an AND gate, as shown in the lower-right portion of Figure 1a. This is because the two kernels represent the up and down pixels of the right-edge feature of the subimage, and only when both pixels have high currents, the case can represent the right-edge feature in the compressed image. As the example shown in Figure 1a can fulfill this criterion, the corresponding pixels in the compressed image (upper-right panel of Figure 1a) become white, suggesting that these pixels represent the right-edge feature in the convoluted (or compressed) image. A similar idea can be applied to the other two different configurations of the four pixels with the right two pixels being commonly white, which were identified by the blue and green boxes in the left panel of Figure 1a. As these subimages have their two right pixels as white, they commonly result in the white pixels in the convoluted image, as indicated by the small blue and green boxes in the right panel of Figure 1a. If left-, lower-, or upper-edge features are intended to be extracted, only changing the BL connections to the AND gate can accomplish this goal. For example, the BLs corresponding to kernels #1 and #2 should be connected to the inputs of the AND gate to extract the lower-edge feature.

The next step is to implement such functionality into the S-CBA. For this purpose, a hypothetical S-CBA is composed of four stacked layers, of which the number corresponds to the four kernels in Figure 2a, where each layer has 196 (= 14 × 14) WLs and 169 (= (14 − 2 + 1) × (14 − 2 + 1)) BLs, as shown in Figure 1b. In this case, the key idea is how to program the four kernels to the large-sized S-CBA so as to achieve simultaneous outputs from the BLs, which contain the convoluted information described earlier. For this purpose, the pixels in the original input image were numbered, as shown in Figure 1c, and this matrix-like pattern is vectorized, as shown in Figure 1d. Inputting this vectorized image to the S-CBA can be accomplished by applying the 6.3 V of $V_r$ to all the WLs corresponding to the white pixels in the input vector. The four wires within a column (of which the total number is 169) in Figure 1b constitute one BL, which
Figure 1. The SIMO simulation that compresses the binarized hand-written number into the feature map. a) The convolution results in the simulation. The binarized hand-written number 5 of 14 × 14 pixels (left panel) is converted to the 13 × 13 compressed image (right panel). The four kernels shown in the middle panel were used in the simulation to extract the right-edge feature, i.e., any subimage composed of 2 × 2 pixels with the two right-hand pixels being white. The yellow and white squares and V marks in the kernels indicate the LRS and HRS cells, and the $V_{\text{r}}$ (≈ 6.3 V) is applied to the WLs, respectively. b) The schematic of the S-CBA used in the simulation. c) The numbered original input image. d) The LRS mapping status of 196 × 169 × 4 S-CBA in the simulation. The vectorized input pattern was fed into the WLs.
will eventually produce the current signal that determines white or black in the corresponding cells in the convoluted image. The convoluted image can be achieved by converting the output vector composed of BL signals into the $13/C_2^{13}$ matrix, as shown in the right panel of Figure 1a. The ReRAM cells belonging to one BL, ca., BL1, are displayed, as shown in the first panel of Figure 1d. As there are four layers in each BL, there are four strings of ReRAM cells, which correspond to the kernels #1–#4. It should be noted that the BL1 corresponds to the first $2/C_2^2$ subinput pattern indicated by the red square, as shown in Figure 1c. The next step is to program the four strings in a given BL according to the configurations of the kernels. For example, kernel #1 in Figure 1a has an LRS status at the lower right corner, which corresponds to pixel number 16 when kernel #1 is placed at the red square location in Figure 1c. Therefore, the corresponding ReRAM cell in string 1, which belongs to the first layer in S-CBA, 16th from the top, will be set to LRS, whereas all other cells in the same string remained in high-resistance state (HRS), as shown in the first column of the first panel of Figure 1d. For representing kernel #2, where the LRS status is assigned to the pixel number 15, the second string will have the 15th ReRAM cell from the top in the LRS, whereas all other...
cells remain in the HRS. For kernels #3 and #4, similar operations can be performed for strings 3 and 4 with their second and first ReRAM cells being in the LRS while all others are in the HRS. When all the WLs are biased, as discussed earlier, all the four strings in BL1 will show low currents because none of the LRS cells in the BL1 correspond to \( V_r \). The first pixel in the convoluted image becomes black. The movement of the selected subimage by one-pixel distance toward the right-hand direction in the original image can be represented by shifting the selected LRS cell locations in the BL2. As shown in the second panel of Figure 1d, the 17th, 16th, 3rd, and 2nd ReRAM cells in the strings 1–4, respectively, are set to LRS while all other cells remained in HRS. When the WLs are biased as discussed earlier, none of these strings in BL2 will show high current; as for the BL1, so the second pixel in the convoluted image will be black too. For one more shifted subimage, the cells in the four strings of the BL3 can be programmed, as shown in the third panel of Figure 1d, which will also produce a black pixel in the convoluted image. Such an arrangement of the LRS cells in all the strings of all BLs can be repeated up to the last BL, i.e., BL169 (the fourth panel of Figure 1d). Such a mapping scheme is denoted as “Orthogonal Mapping” because each state combination per stack does not overlap with each other. When the selected subimage reaches the location indicated by a large pink box in the original image of Figure 1a, the first white pixel is recorded in the convoluted image (small pink box in the right panel of Figure 1a). For the previous step, i.e., when the subimage located one pixel left of the large pink box, the output must also be black because the top-right pixel among the four selected pixels was black. It should be noted that in the cases where two right pixels within the four selected pixels are white, white pixels will be produced in the convoluted image. Although it was not explicitly mentioned earlier, selecting the right-edge feature can be readily achieved by connecting the strings 1 and 3 to the two inputs of AND gates, which are located at the end of each BL (actually composed of four strings from each layer). Other features can also be easily selected by connecting different strings to the AND gates.

One of the most critical features of the CNN hardware using the specific S-CBA structure is the simultaneous determination of all the pixels in the convoluted image by simultaneously reading the output currents of all AND gates at each BL for a given input image (one input of the vectorized image). If another feature extraction, such as an upper edge, for the given number, is necessary, an identical S-CBA with only different connections of the strings to the AND gates can provide the desired functionality. This demonstrates the high efficiency and flexibility of the suggested circuit configuration. Although the kernel configurations in this example were chosen based on the authors’ intention, they (or even with a larger size) can be configured using the software-trained kernel weights, which might be useful for the “mostly inference” system.

Because the circuit configuration produces multiple outputs upon a single input, this method can be termed as a single-input multiple-output (SIMO) scheme. Such a SIMO scheme might be appropriate for cases where the input can be represented by a single-input vector, as shown in Figure 1. In contrast, multiple-input single-output (MISO) scheme might be more appropriate for cases where there are several simultaneous input vectors, such as the feature extraction of a color image, which is described in detail, as shown in Figure 2. In this case, the same S-CBA can be feasibly utilized where the inputs and outputs are reversed compared with Figure 1.

The left panel of Figure 2a shows a simplified color image, indicating a sailing boat on the sea and a tree at the seashore region, shown in a digital image composed of 14 × 14 pixels. As all the colors in the digital image are composed of three component colors: red (R), green (G), and blue (B), the image can be decomposed into three component images, as shown in the middle panel of Figure 2a, with identical pixel sizes. (In fact, if the image was taken by a digital camera, there must be already three component color images.) As shown in Figure 2, the purpose of the convolution is to delineate the border between the water (blue color in the original image) and the other region, as shown in the upper-right panel of Figure 2a. This can be achieved by applying kernels to the component images, as shown in the middle panel of the same figure. For instance, under this circumstance, the first subimage of the R component image (indicated by the white square in the corresponding R component image) will generate a high current at the lower right corner of the corresponding kernel as that location has both \( V_r \) (−5.3 V) input and LRS cell in the kernel. The negative bias voltage is necessary because of the unidirectional current flow (from WL to BL) due to the self-rectifying functionality of the ReRAM cell (see the next section for details). All the other locations in the same kernel will not generate a high current.

In contrast, the first subimage of the G component image under the application of the corresponding kernel with a similar configuration will not produce any high current because there are no cases where the \( V_r \) and LRS cells match. Nonetheless, if these two outputs are connected via an OR gate, as shown in the lower right panel of Figure 2a, the output could be high (white square) in the convoluted image, as shown in the right panel of Figure 2a. Given the kernel configuration for the B component image (all cells are in HRS), no high current signal can be achieved from this kernel despite how the pixels are distributed in the image. As a result, the convoluted image will have white pixels for the subimages in the original image containing either red or green pixels at their lower right corner. In contrast, all the subimages containing only blue pixels will be compressed into the black pixels in the convoluted image, which will show the boundary between the sea and other portions, as shown in the right panel of Figure 2a. Therefore, it can be anticipated that such type of convolution can be readily achieved when the three vectorized inputs of the three color component images are simultaneously inputted to the BLs of the S-CBA and outputs, which is produced from the WLs, as shown in Figure 2b. In this case, the number of BLs and WLs must be 196 and 169, respectively, because of the inverted input and outputs compared with the SIMO case, as shown in Figure 1. In such a parallel configuration of the connection between the strings in each BL and corresponding WL, the OR functionality could be automatically achieved without the adoption of an actual OR gate, as the current flowing into the WL could be high for the cases where any of the connected strings generate high currents. As for the case of Figure 2, the shift of the subimages in each component color image can be implemented to the S-CBA by programming the LRS cells, as shown in Figure 2c. In this case, layers (strings) 1, 2, and 3 were assumed to coincide with the R, G, and B.
components of the images. By this arrangement of the input vectors to the strings in each layer of the BLs, the simultaneous inputs of the three color components can be made.

Moreover, the convolution of the input color image to the feature map, as shown in Figure 2a, can be accomplished in a single step by reading the current at all WLs and converting the current values to white or black pixels in the output vector, which then will be converted again to the matrix format with the 13 × 13 size. It can be understood that the nonselection of a certain color during convolution can be achieved by letting all the four squares in the corresponding kernels be white. This means that all the ReRAM cells in the corresponding layer remain in the HRS, not by selecting the inputs of the hypothetical OR gate. (The OR gate in Figure 3a was added to show the concept but not to indicate that such OR gates are actually necessary.) Therefore, this type of circuit configuration can be a typical example of using the S-CBA in MISO scheme.

Such MISO can also be used to relay one part of the CNN circuit to the next one in a large neural or neuromorphic network because one CNN generally produces several output vectors, as can be shown in Figure 1, which can then be fed into the next CNN layer. Therefore, under such a circumstance, the sequential adoption of the S-CBA in the SIMO and MISO schemes will be a highly useful method to expedite the complicated inputs as well as saving the energy. For such an application of the MISO to the sequential CNN, the output current must be converted to a voltage signal in an active mode. This can be accomplished by adopting a modified circuit of the usual current mirror (CM), as shown in Figure S3a, Supporting Information.

Figure 3a–d shows the output voltage distribution according to the BL number from the first to the fourth layers, respectively, when a $V_r$ of 6.3 V was simultaneously applied to all the WLs in the circuit, as shown in Figure 1b (SIMO scheme). In fact, the direct outputs of the S-CBA at each BL must be current, which could be converted to the voltage signal ($V_{out}$ in Figure 3a–d) using another CM circuit, as shown in Figure S3b, Supporting Information. The compressed image in the right panel of Figure 1a was the outcome of the combined results of Figure 3a,c via the AND gate, of which the results are included in the inset of Figure 3a. That is, the white pixels in that image correspond to the cases where both BLs in layers 1 and 3 have low $V_{out} \approx 0$ V, i.e., the compressed image in Figure 1a was just a converted (into the matrix) image of the voltage vector in the inset figure in Figure 3a.

Figure 3e shows the results from the MISO scheme, as shown in Figure 2. In this case, all the BLs in Figure 2b are simultaneously biased with a $V_r$ of −5.3 V, which was determined to match an actual reading voltage of 6 V, as in the SIMO case. The output voltages of the 169 WLs through the voltage conversion circuit are displayed. As discussed earlier, this is actually the outcome of the OR-gated layers 1 and 2. When the criterion for the high or low value of $V_{out}$ was settled at 0.5 V, there was no error in discriminating the results; all pixels with $V_{out} >$ or $< 0.5$ V coincided with the white or black color. Therefore, the compressed image in Figure 2a was just a converted (into the matrix) image.

![Figure 3](image.png)

**Figure 3.** The output voltage distribution according to the BL number for a) layer 1, b) layer 2, c) layer 3, and d) layer 4, respectively, when a $V_r$ of 6.3 V was simultaneously applied to the WLs, as shown in Figure 1d (SIMO scheme). The inset of (a) shows the combined output voltages of layers 1 and 3 via the AND gate. e) The output voltage distribution according to the WL number when a $V_r$ of −5.3 V was simultaneously applied to the BLs, as shown in Figure 2c (MISO scheme).
of the voltage vector in Figure 3e. While the $V_{out}$ results in Figure 3a–d were close to either 1.2 or 0 V, the data in Figure 3e also contained somewhat scattered values. This is due to the random variation in ReRAM resistance. In the MISO scheme, due to multiple inputs, a larger number of ReRAM cells affect the output current of each WL than the SIMO scheme, which causes the output voltage fluctuation, as shown in Figure 3e. Even with such variation, the discrimination was successful, as shown in Figure 3e.

Finally, the extendability of such a CNN scheme based on the fabricated self-rectifying Pt/HfO$_2$/TiN ReRAM was examined by simulation examining the reading margin (RM) and writing margin (WM) of the S-CBA using the HSPICE and physical model developed by the authors. The fundamental assumptions that were used for the simulations, as shown in Figure S4, Supporting Information, are that the I–V characteristics of each cell are not influenced by the presence of the neighboring memory cells. This can be accomplished only when the sneak current is sufficiently suppressed. This could be the case due to the high functionality of the embedded self-rectification. The sneak current influences not only the RM but also the WM, of which the details have been reported elsewhere.$^{[31,32]}$ The simulations shown in Figure S4, Supporting Information, confirmed that it is possible to provide the system with an integration density of more than 1 Mb with sufficient RM and WM, when an appropriate voltage scheme is adopted. As the required density of the S-CBA was $\approx 0.13$ Mb in Figure 1 and 2, this value confirms that the simulation results shown in Figure 3 are reliable. Nevertheless, there can be other CNN circuits that may require a higher integration density of the CBA. For these cases, using a lower wire resistance material, such as $W$, which is widely used as an electrode in the semiconductor industry, is helpful to decrease the parasitic resistance effect that is more detrimental in determining the WM.$^{[32]}$ Layer stacking is generally favorable in achieving the lower overall wire resistance due to the generally shorter length of the WLs and BLs.$^{[22]}$ The high efficiency of the suggested SIMO and MISO schemes could be identified from the low power consumption when the inputs in Figure 1a and 2a are assumed. The power consumption was estimated by the HSPICE simulation, of which details are shown in S1-VI, Table S1, and Figure S7, Supporting Information. The consumed power was 0.46 and 1.17 $\mu$W for the SIMO and MISO cases, respectively.

3. Conclusions

In conclusion, the two-layered S-CBA was demonstrated which was composed of self-rectifying Pt/HfO$_2$/TiN ReRAM cells with sufficiently high-memory functionality. This array could be a flexible platform for performing the VMM based on Kirchhoff’s law. The multiple-layer CBA could be configured as having one WL contacts several layers in a single biline, which could be used for either SIMO or MISO schemes. Both schemes can greatly facilitate the simultaneous acquisitions of the multiple input and output vectors, which then eventually improves the system performance. The various types of kernels can be easily adapted based on the S-CBA circuit geometry, which also contributed to the high performance of the hardware-based CNN circuit. Although the experimentally proven integration density was only $4 \times 2 \times 2$ (16 bits) due to the limited university research facilities, it could be confirmed by the HSPICE simulation that the provided ReRAM cell performance can offer an approximately 1 Mb density without involving the sneak current issue.

4. Experimental Section

Experimental Procedure for S-CBA Fabrication and Device Test: For the fabrication of the planar-type Pt/HfO$_2$/TiN device, a 50 nm-thick TiN film was deposited on a SiO$_2$/Si substrate using a sputtering system (Endura, Applied Materials). The TiN film was patterned by photolithography followed by a photoresist lift-off process as the bottom electrode (BE). For the film, a Ti target, with a diameter of 12 inches, was used with 15 standard cubic centimeter per minute (sccm) of Ar gas and 85 sccm of N$_2$ gas. DC power and substrate temperature were kept at 5000 W and 200 °C, respectively. The base and operating pressures were $\approx 10^{-8}$ torr and 4 mtorr, respectively. Afterward, 10 nm-thick HfO$_2$ films were deposited through atomic layer deposition (ALD) at a substrate temperature of 230 °C using $\text{H}_2$[($\text{NH}_3$)]$_2$ and O$_2$ as the H precursor and oxygen source, respectively. After that, a plasma treatment was performed in a showerhead-type ALD chamber using 1000 sccm of Ar gas and 100 sccm of O$_2$ gas (see the next subsection for the performance). Then, a 50 nm-thick Pt top electrode (TE) was deposited by the electron beam evaporator (ZZS550-2/D, Maestech) and patterned by photolithography followed by photoresist lift-off process for the crossbar structure.

For the fabrication of the two-layered vertical-type Pt/HfO$_2$/TiN S-CBA, SiO$_2$ (130 nm)/TiN (50 nm)/Si$_3$N$_4$ (50 nm)/TiN (50 nm) layers were sequentially deposited on a SiO$_2$/Si substrate. For the TiN BE deposition, a different sputtering system (SRN120, SORONA) was used for vertical device fabrication. In the system, a Ti target with a diameter of 4 in. was used with 20 sccm of Ar gas and 3 sccm of N$_2$ gas. RF power and substrate temperature were kept at 500 W and room temperature. The base and operating pressures were $\approx 10^{-7}$ torr and 1 mtorr, respectively. The plasma-enhanced chemical vapor deposition system (PlasmaPro System100, Oxford Instruments) was used for the SiO$_2$ and Si$_3$N$_4$ films, which were used as isolation layers in the S-CBA device. Then, the multilayers were patterned and etched into a line shape with a width of 10 μm. For the SiO$_2$ layer etching, a reactive-ion etching system (RIE 80 plus, Oxford Instruments) with 40 sccm of CF$_4$ gas and 4 sccm of O$_2$ gas was used with an RF power of 200 W. In this etching condition, the SiO$_2$ layer was well etched, whereas that of the Si$_3$N$_4$ and TiN films were suppressed. For the TiN/Si$_3$N$_4$/TiN layers etching, another inductively coupled plasma (ICP)-type etching system was used (PlasmaPro System100 Cobra, Oxford Instruments) with 33 sccm of BCl$_3$ gas and 50 sccm of Cl$_2$ gas. The ICP and RF powers were set to 1000 W and 250 W, respectively. In this etching condition, both nitride films were well etched, whereas that of the SiO$_2$ layer was limited. After the etching process, a 10 nm-thick HfO$_2$ film was deposited and the plasma treated with the same equipment and conditions as the planar-type device. Then, an 80 nm-thick Pt TE was deposited by an electron beam evaporator (ZZS550-2/D, Maestech) and patterned by photolithography followed by the photoresist lift-off process for the crossbar structure (line width of 10 μm). The schematics for the overall fabrication process of the vertical device are shown in Figure S5, Supporting Information. Due to the difficulty in patterning the Pt TE at the side-wall areas of higher height-line density of the vertical-type device. The energy-dispersive spectroscopy (EDS)
mapping and cross-sectional transmission electron microscope images of the vertical-type device were observed using field-emission-transmission electron microscopy (FE-TEM) (JEOL, JEM-F200). The I–V characteristics and 4 × 2 × 2 S-CBA kernel application results were measured using a Hewlett-Packard 4145B semiconductor parameter analyzer. During the measurement, the Pt TE was biased, and TiN BE was grounded. All the simulations were performed using HSPICE.

**Self-Rectifying ReRAM and Fabrication of S-CBA:** The aforementioned devices were tested as a charge trap-based ReRAM cell. While the HfO₂/TiN interface constituted a quasi-Ohmic contact that provided a fluent carrier (electron) migration between the abundant trap sites in the HfO₂ film and TiN electrode upon a positive bias application, the Pt/HfO₂ interface constituted a Schottky-type interfacial contact, suppressing the electron injection from the Pt to HfO₂ upon a negative bias application to the Pt TE. The details of such electron trapping or detrapping-based ReRAM are reported in previous studies. In short, the ReRAM was set (switching from HRS to LRS) upon the sufficiently high positive bias application to the Pt, which trapped the electrons within the HfO₂ layer, whereas the device was reset (switching from LRS to HRS), upon the application of a sufficiently negative bias, which detrapped the electrons within the HfO₂ layer. Moreover, the high Schottky barrier at the Pt/HfO₂ interface always suppresses the current flow upon a negative bias application. Therefore, the devices are provided with the self-rectification functionality.

As the pristine HfO₂ film in the Pt/HfO₂/TiN structure had very low trap density to induce a sufficient memory window, the pristine Pt/HfO₂/TiN device showed only threshold switching performance (see Figure S6, Supporting Information). The HfO₂ film was treated under the Ar plasma ambient to induce a sufficiently high density of defects, mostly oxygen vacancy. Details of such a process, experimental results, and plasma conditions are shown in this section and Figure S6, Supporting Information.

A planar CBA with a single-layer configuration was fabricated in 4 × 4 density with a line or space dimension of 10 μm, as shown in Figure 4a. This was to prove the functionality of the self-rectification of the ReRAM cells. Figure 4b shows the switching I–V curves of all the 16 ReRAM cells.
where all the cells work effectively according to the intention. Memory switching was observed in the positive bias region, whereas the current in the negative bias region was always suppressed. Figure 4c shows the high functionality of the self-rectifying ReRAM cells. In this case, cell #6, shown in red font in Figure 4a, was programmed to HRS, whereas all the other cells were programmed to LRS, which corresponds to the worst case in CBA in reading the selected HRS cell (cell #6 in this case). Despite the unfavorable distribution of the memory cell states, the HRS of cell #6 can be well determined, suggesting a sufficient suppression of the sneak current through the neighboring LRS cells due to the self-rectifying functionality of the involved ReRAM cells.

Similar functionality can be confirmed from the two-layer stacked S-CBA with \(2 \times 2\) configuration, as shown in Figure 5a, which shows the schematic diagram of the fabricated S-CBA. Figure 5b,c shows the cross-section TEM image and composition mapping results using EDS in TEM. The devices were well fabricated according to the intended structure and stacking configuration. In this device configuration, the HfO\(_2\)-x film on the side-wall region sandwiched between the Pt TE and TiN BE was the functional ReRAM layer. The stripe-shaped Pt TE and TiN BE served as the WL and BL, as shown in Figure 1 and 2. Figure 5d,e shows the I–V test results of all the eight ReRAM cells corresponding to the results in Figure 4b,c, respectively, demonstrating that the ReRAM cells fabricated on the side-wall region of the S-CBA function quite well according to the purpose. The cross-talk effect was also sufficiently suppressed in the S-CBA structure due to the high functionality of the self-rectifying cells on the side walls (Figure 5e). Although it was improbable to fabricate the four-layer stacked S-CBA with \(196 \times 169\) cells in each layer to experimentally demonstrate the SIMO scheme in Figure 1 due to the limited capability of the university fabrication facilities, the data in Figure 4 were sufficient to extract the device features, which are necessary to simulate the CNN function, as shown in Figure 1 and 2. Moreover, the feasibility of suppressing the cross-talk effect in such a larger array size (\(4 \times 196 \times 169 = 132496 \times 0.13\) Mb) was confirmed by simulation via using both the HSPICE and physical model developed by the authors, based on the I–V curves in Figure 5d. The simulation methods are shortly described at the end of this section, and detailed information about them is available in SI-II and Figure S3, Supporting Information.

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

Acknowledgements
This work was supported by Future Semiconductor Device Technology Development Program (20003655) through the Ministry of Trade, Industry, & Energy (MOTIE, Korea) (project name: Development of self-rectifying resistance switching materials and devices for selectorless crossbar application).

Conflict of Interest
The authors declare no conflict of interest.

Author Contributions
Y.K. and J.K. contributed equally to this work. Y.K. designed and fabricated the devices, analyzed electrical data, and performed the chemical and material analyses. He also devised the analytical model to estimate the achievable array size of the device. J.K. conceived the kernel application of the device and performed the HSPICE simulations to demonstrate the kernel application in the CNN system and identify the achievable array size. S.S.K., Y.J.K., and G.S.K. contributed to the device design and fabrication. J.W.J. and D.E.K. supported data analysis. J.H.Y. guided the ideas of the kernel application of the device. C.S.H. directed the whole work and prepared the manuscript.

Keywords
convolutional neural networks, hafnium oxide, kernels, self-rectifying resistive switching random access memory, stacked crossbar arrays

Received: October 6, 2019
Revised: October 27, 2019
Published online: December 9, 2019

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffin, Proc. of the IEEE 1998, 86, 2278.
[2] D. Cireșan, U. Meier, J. Schmidhuber, arXiv preprint arXiv 2012, 1202.2745.
[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, in Advances in Neural Information Processing Systems 2012, pp. 1097–1105.
[4] K. Simonyan, A. Zisserman, arXiv preprint arXiv 2014, 1409.1556.
[5] K. He, X. Zhang, S. Ren, J. Sun, in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2016, pp. 770–778.
[6] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2015, pp. 3156–3164.
[7] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2016, pp. 4651–4659.
[8] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu, IEEE/ACM Trans. on Audio Speech and Lang. Process. 2014, 22, 1533.
[9] T. N. Sainath, A. Mohamed, B. Kingsbury, B. Ramabhadran, in 2013 IEEE int. Conf. on Acoustics, Speech and Signal Processing 2013, pp. 8614–8618.
[10] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, J. Mach. Learn. Res. 2011, 12, 2493.
[11] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Montiort, U. Muller, J. Zhang, Z. Zhang, J. Zhao, K. Zieba, arXiv preprint arXiv 2016, 1604.07316.
[12] Y. LeCun, Y. Bengio, G. Hinton, Nature 2015, 521, 436.
[13] J. Backus, ACM 2007, 21, 1977.
[14] S. Yu, Proc. of the IEEE 2018, 106, 260.
[15] Z. Wang, C. Li, W. Song, M. Rao, D. Belkin, Y. Li, P. Yan, H. Jiang, P. Lin, M. Hu, J. P. Strachan, N. Ge, M. Barnell, Q. Wu, A. G. Barto, Q. Qiu, R. S. Williams, Q. Xia, J. J. Yang, Nat. Electron. 2019, 2, 115.
[16] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, R. S. Williams, in Proc. of the 53rd Annual Design Automation Conf. 19, ACM 2016, p. 19.
[17] M. Prezioso, F. M. Bayat, B. Hoskins, G. Adam, K. K. Likharev, D. B. Strukov, Nature 2015, 521, 61.
[18] Z. Wang, S. Joshi, S. Savelev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo, H. Jiang, P. Lin, C. Li, J. J. Hoon, N. K. Upadhyay, J. Zhang, M. Hu, J. P. Strachan, M. Barnell, Q. Wu, H. Wu, R. S. Williams, Q. Xia, J. J. Yang, Nat. Electron. 2018, 1, 137.
[19] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, W. D. Lu, Nat. Electron. 2019, 1, 290.
[20] G. W. Burr, P. Narayanan, R. M. Shelby, S. Sidler, I. Boybat, C. di Nolfo, Y. Leblebici, in 2015 IEEE Int. Electron Devices Meeting (IEDM) 2015, p. 4.
[21] H. Tsai, S. Ambrogio, C. Mackin, P. Narayanan, R. M. Shelby, K. Rocki, A. Chen, G. W. Burr, in 2019 Symp. on VLSI Technology 2019, pp. T82–T83.
[22] J. Y. Seok, S. J. Song, J. H. Yoon, K. J. Yoon, T. H. Park, D. E. Kwon, H. Lim, G. H. Kim, D. S. Jeong, C. S. Hwang, Adv. Funct. Mater. 2014, 24, 5316.
[23] J. H. Yoon, S. J. Song, I. H. Yoo, J. Y. Seok, K. J. Yoon, D. E. Kwon, T. H. Park, C. S. Hwang, Adv. Funct. Mater. 2014, 24, 5086.
[24] J. H. Yoon, K. M. Kim, S. J. Song, J. Y. Seok, K. J. Yoon, D. E. Kwon, T. H. Park, Y. J. Kwon, X. Shao, C. S. Hwang, Adv. Mater. 2015, 27, 3811.
[25] J. H. Yoon, S. Yoo, S. J. Song, K. J. Yoon, D. E. Kwon, Y. J. Kwon, T. H. Park, H. J. Kim, X. L. Shao, Y. Kim, C. S. Hwang, ACS Appl. Mater. Interfaces 2016, 8, 18215.
[26] J. H. Yoon, D. E. Kwon, Y. Kim, Y. J. Kwon, K. J. Yoon, T. H. Park, X. L. Shao, C. S. Hwang, Nanoscale 2017, 9, 11920.
[27] W. Bae, K. J. Yoon, C. S. Hwang, D. K. Jeong, Nanotechnology 2016, 27.48, 485201.
[28] W. Bae, K. J. Yoon, C. S. Hwang, D. K. Jeong, IEEE Trans. Electron Devices 2017, 64.4, 1591.
[29] L. Gao, P. Y. Chen, S. Yu, IEEE Electron Device Lett. 2016, 37, 870.
[30] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, Q. Xia, Nat. Electron. 2017, 1, 52.
[31] K. J. Yoon, W. Bae, D. K. Jeong, C. S. Hwang, Adv. Electron. Mater. 2016, 2, 1600326.
[32] Y. Kim, Y. J. Kwon, J. Kim, C. H. An, T. Park, D. E. Kwon, H. C. Woo, H. J. Kim, J. H. Yoon, C. S. Hwang, Adv. Electron. Mater. 2019, 5, 1800806.
[33] Y. Kim, Y. J. Kwon, D. E. Kwon, K. J. Yoon, J. H. Yoon, S. Yoo, H. J. Kim, T. H. Park, J. W. Han, K. M. Kim, C. S. Hwang, Adv. Mater. 2018, 30, 1704320.