ARTICLES
https://doi.org/10.1038/s41565-021-00874-8
nature nanotechnology

Energy-efficient Mott activation neuron for full-hardware implementation of neural networks

Sangheon Oh[1], Yuhan Shi[1], Javier del Valle[2], Pavel Salev[2], Yichen Lu[1], Zhisheng Huang[1], Yoav Kalcheim[2], Ivan K. Schuller[2] and Duygu Kuzum[1]✉

To circumvent the von Neumann bottleneck, substantial progress has been made towards in-memory computing with synaptic devices. However, compact nanodevices implementing non-linear activation functions are required for efficient full-hardware implementation of deep neural networks. Here, we present an energy-efficient and compact Mott activation neuron based on vanadium dioxide and its successful integration with a conductive bridge random access memory (CBRAM) crossbar array in hardware. The Mott activation neuron implements the rectified linear unit function in the analogue domain. The neuron devices consume substantially less energy and occupy two orders of magnitude smaller area than those of analogue complementary metal–oxide semiconductor implementations. The LeNet-5 network with Mott activation neurons achieves 98.38% accuracy on the MNIST dataset, close to the ideal software accuracy. We perform large-scale image edge detection using the Mott activation neurons integrated with a CBRAM crossbar array. Our findings provide a solution towards large-scale, highly parallel and energy-efficient in-memory computing systems for neural networks.

As the amount of data for computing exponentially increases, data transfer between memory and processor turns into a major bottleneck dominating the system-level energy consumption. In-memory computing has been proposed to circumvent this bottleneck, which arises from von Neumann architecture, by minimizing or eliminating the energy-consuming data transfer between memory and processor[1,2]. In-memory computing with emerging non-volatile memories (eNVMs)[3–6] has shown promising results for on-chip storage of weights and computation of multiply–accumulate (MAC) operations for a single layer[7–9]. However, modern deep neural networks (DNNs) consist of hundreds of layers (for example, ResNet has 152 layers[10]) such that the outputs of each layer are individually connected to artificial neurons applying non-linear activation functions on weighted sums. Most in-memory computing approaches using eNVMs still rely on general processors to compute and propagate activation functions of each layer. However, activations that move in and out of the memory can dominate the energy consumption of in-memory computing-based accelerators[8,11–13]. Moreover, computation of one element of activation using analogue-to-digital converters (ADCs) consumes energy comparable to the energy consumed by a whole synaptic array for a MAC operation[13]. Since DNNs need to have a very large number of activations to achieve high accuracy[13], it is critical to develop energy- and area-efficient implementations of activation functions, which can be integrated on the periphery of the synaptic arrays. Recent works have investigated analogue complementary metal–oxide semiconductor (CMOS) circuits[14] and ADCs with reconfigurable function mapping[15] for the implementation of non-linear activation functions. However, a compact and energy-efficient nanodevice implementing the non-linear activation functions has yet to be demonstrated.

Here we propose a volatile four-terminal Mott activation neuron device based on vanadium dioxide (VO_2) for compact and energy-efficient implementation of activation functions. The Mott activation neuron consists of a nanowire heater for precise control of the temperature of the VO_2 film. First, we experimentally demonstrate that the resistance of the Mott activation neuron can be switched linearly and gradually to emulate a rectified linear unit (ReLU) activation function, which is the most widely used activation function. The Mott activation neuron can generate an output voltage, which follows the ReLU activation function for a given weighted sum current. Then, we study the energy efficiency of the Mott activation neuron in comparison to activation function circuits with an analogue CMOS[14] or reconfigurable digital ADC[15]. We investigate the performance of hardware neural networks implemented with the Mott activation neurons in terms of energy, latency, peripheral neuron/circuit area and classification accuracy. Lastly, we fabricate CBRAM crossbar arrays and Mott activation neuron arrays to demonstrate edge detection using convolutional neural networks in hardware. Our results show that the small size and energy efficiency of the Mott activation neuron enable direct stacking of synaptic layers in neural networks and achieve substantial gains in energy efficiency and area while providing high accuracy.

Mott activation neuron

Neural networks consist of a set of neurons organized in layers, connected with synaptic weights (Fig. 1a). The inputs applied to the networks are multiplied by the corresponding weights and the multiplication results are accumulated in neurons. Then, the output of a neuron is calculated by passing the MAC results through a non-linear activation function. In-memory computing architectures map these neural network operations onto the arrays of eNVM devices. The weights are stored in arrays of eNVM devices, and the weighted sum is calculated using Kirchhoff's current law[16]. While in-memory computing allows the local storage of the weights in compact and energy-efficient synaptic devices, the activation function calculations are still implemented with general processors or large and complex neuron peripheral circuits (Fig. 1b). This substantially degrades energy and area efficiency at the system level. The activation function we target is the ReLU, which is the most widely used

[1]Electrical and Computer Engineering Department, University of California San Diego, La Jolla, CA, USA. [2]Department of Physics, University of California San Diego, La Jolla, CA, USA. ✉e-mail: dkuzum@eng.ucsd.edu

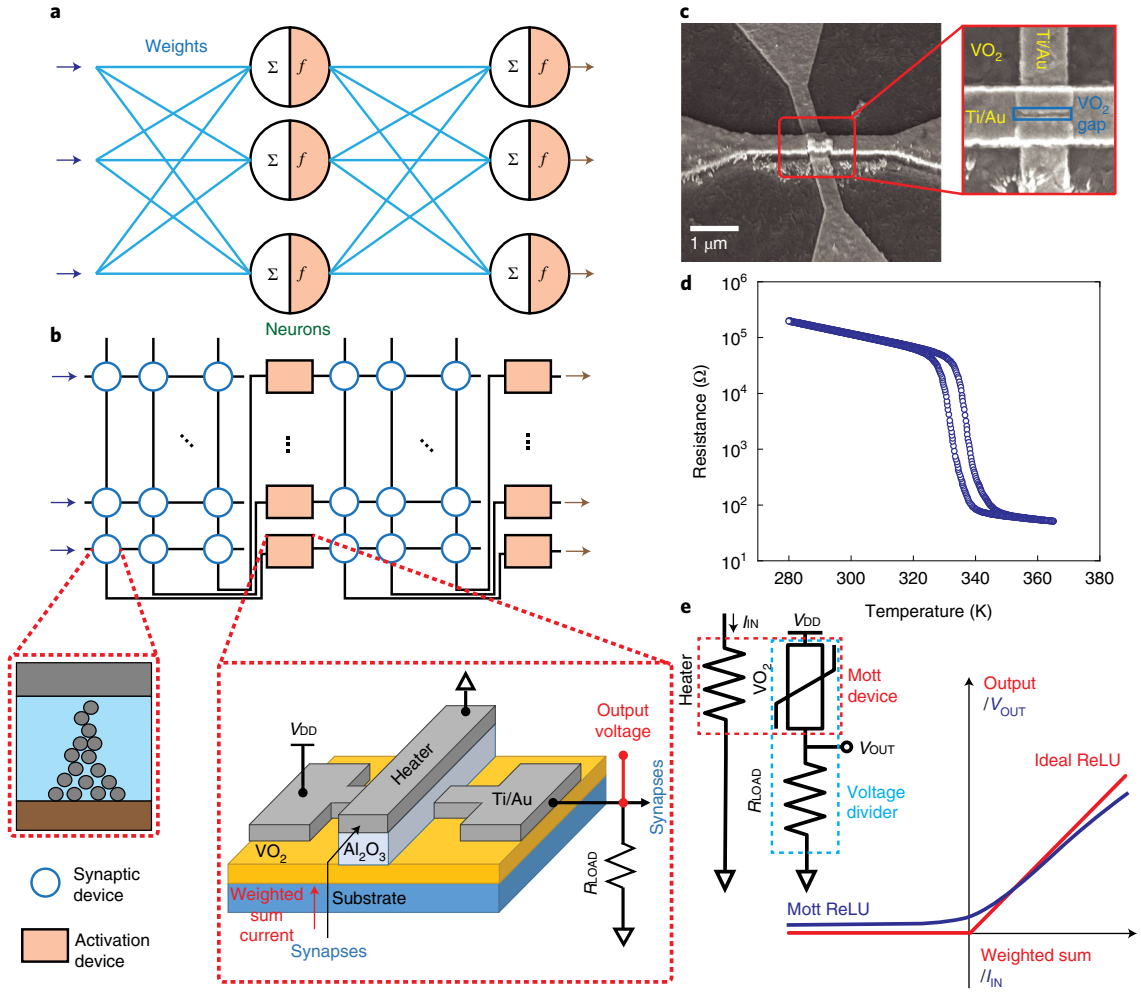

Fig. 1 | The Mott ReLU device for the hardware implementation of a neural network. a,**b**, An illustration shows a neural network (**a**) and hardware implementation (**b**) of the neural network with synaptic and activation (or neuron) devices. Σ represents a weighted sum while *f* represents the activation function. In **b**, the inset on the left shows a schematic of a resistive memory cell. The inset on the right shows a schematic of the Mott device with a nanowire heater. Mott activation devices allow direct stacking of multiple eNVM arrays for DNNs. The heater is connected to a column of presynaptic arrays and gets a weighted sum current. Then, one pad of the VO_2 gap is connected to V_{DD} and the other pad is connected to the next synaptic array. The pad connected to the next synaptic array is also connected to a load resistor. Weights are stored in eNVM devices, and weighted sum currents from each column are fed into the Mott ReLU. Then, the output of the Mott ReLU is applied as the input to the next layer. **c**, A scanning electron microscope image of the Mott device (scale bar, 1 μm). The inset shows the nanowire heater on the top of the 50 nm VO_2 gap. **d**, Resistance of the VO_2 gap when the temperature is swept from 280 K to 365 K. **e**, An illustration shows how a Mott device will be used as a ReLU activation function. The output of the ReLU activation function will be represented by V_{OUT} of the Mott ReLU device while the weighted sum input to ReLU will be represented by the input current (I_{IN}) to the Mott ReLU device.

activation function. The output of the ReLU activation function (that is, $f(x) = \max(0, x)$) depends only on current input regardless of previous inputs and resistance states. In addition, the output of the ReLU function is linear after the transition point (that is, $x = 0$). In order to emulate the ReLU activation function, the device should exhibit volatile, linear and gradual resistive switching. We developed a four-terminal VO_2-based activation device (illustrated in the inset of Fig. 1b on the bottom) that exploits a thermal-driven Mott transition of VO_2 to embody these characteristics in a single nanodevice. The Mott ReLU device uses a nanowire heater (that is, Ti (20 nm)/Au (30 nm)) to control the resistive switching of a lateral, 50 nm VO_2 gap beneath it. The heater and the VO_2 gap are electrically insulated by a 70 nm Al_2O_3 layer. A scanning electron microscope image of a fabricated device is shown in Fig. 1c, and detailed fabrication procedures are discussed in the Methods. The heater generates heat through Joule heating, depending on the magnitude of the weighted sum current generated by each column of the eNVM array. Then, the generated heat is transferred to the VO_2 film through the electrical insulator (that is, the Al_2O_3 layer) and induces the phase transition from the insulating states to the metallic states, which results in a resistivity drop. The temperature-dependent resistance of the VO_2 gap is shown in Fig. 1d. To map the gradual resistivity changes of the VO_2 gap onto the output voltage (V_{OUT}), a voltage divider circuit is implemented as illustrated in the inset of Fig. 1e. The supply voltage (V_{DD}) is divided into the voltage drop across the VO_2 gap and the load resistor, depending on the resistance ratio of the VO_2 gap and the load resistor. As the resistance of the VO_2 gap decreases, the voltage drop across the VO_2 gap decreases, which results in the increment of the output voltage (or the voltage drop across the load resistor). As a result, the resistive switching of the VO_2 gap allows

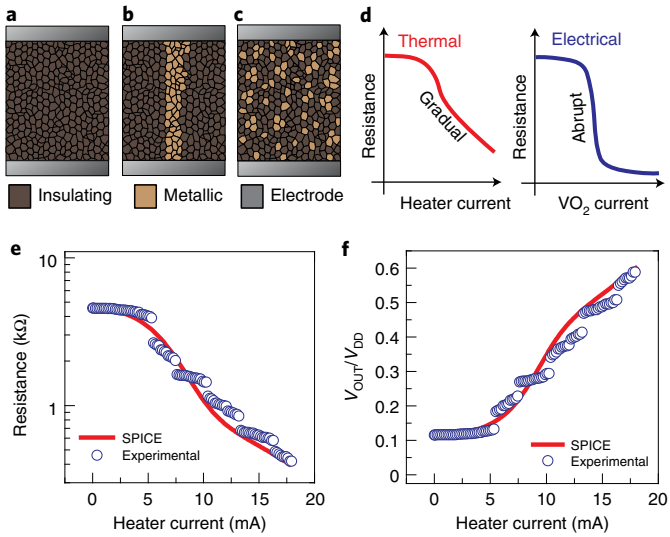

Fig. 2 | Switching mechanisms of VO_2 gap. a–c, Schematics show a VO_2 gap with no bias (**a**), filamentary switching (**b**) and thermal-driven switching (**c**). **d**, As compared to thermal-driven domain-wise switching, electrical filamentary switching shows an abrupt change in resistance. **e**, Resistance of the VO_2 device as a function of heater current showing ~77 levels. It shows gradual and linear resistive switching when the input current is larger than 5 mA. **f**, Voltage ratio between the output (V_{OUT}) and the supply voltage (V_{DD}) of the device as a function of heater current with a 1,900-Ω-load resistor as a representative example. Symbols are experimental data and lines are SPICE simulation results. The *V*–*I* characteristic is similar to the ReLU function shown in Fig. 1e.

the output voltage to emulate the ReLU activation function as illustrated in Fig. 1e. Since the output of the Mott ReLU device is voltage, it can be directly applied to the next layer as an input voltage. Therefore, multiple synaptic layers can be directly stacked on each other for driving the next layers by eliminating complex digital circuits and ADCs between the layers. Moreover, the small size of the Mott ReLU device allows the integration of the device for each column of the synaptic array, which eliminates the need for time multiplexing and hence, enables fully parallel operations.

The main operating principle of the Mott ReLU device is the Mott transition (or insulator-to-metal transition) of the VO_2 gap. The Mott transition of the VO_2 gap can be induced by either electrical filamentary switching or thermal-driven domain-wise switching[17,18]. When a voltage bias above the threshold is applied across the VO_2 gap, Joule heating due to the bias induces filament formation, and the filament is widened as the voltage increases (Fig. 2a,b). Since the filament formation is a cascading avalanche effect, the resistance switching is abrupt[19]. By contrast, when the transition is driven by temperature, only the domains whose critical temperature is below the device temperature transit to the metallic phase (Fig. 2a,c). Since the transition temperature of each domain exhibits variations[20], the number of domains switched to metallic phases gradually increases as the temperature increases. As a result, the resistance of VO_2 gradually decreases as the temperature increases (Fig. 2d). This gradual switching behaviour of VO_2 was previously confirmed by scanning microwave microscopy imaging of the VO_2 film[20]. The Mott ReLU device is engineered to exploit this thermal-driven linear resistive switching for emulating the linear increment of the ReLU activation function, as shown in Fig. 2e. Then, this linear resistive switching of the VO_2 gap is projected to the output voltage. The ratio between V_{OUT} and V_{DD} of the Mott ReLU device with a 1,900-Ω-load resistor is demonstrated in Fig. 2f as a representative example. Two potential practical issues regarding the Mott transition are discussed in Supplementary Note 1.

To further assess the compatibility of the Mott ReLU device for implementing the ReLU activation function, we extensively characterized its switching characteristics. In addition to gradual switching, the resistive switching should be volatile to implement ReLU function in synaptic arrays. That is because the output of the ReLU activation function should depend only on the input at that moment, regardless of previous inputs and resistance states. The volatile switching of the Mott ReLU device is experimentally verified in Fig. 3a. When 1-ms-wide current pulses with various amplitudes are applied to the heater, the resistance of the device is switched and maintained only when the current pulse is high. Furthermore, the output voltage for a given input current should not exhibit a high level of variation, which could degrade neural network performance. Figure 3b demonstrates that each resistance state of the device shows only ~4% or less variation when the resistance states are iteratively measured. The impact of this small variation on the neural network performance is studied in the section, Neural network implementations. Lastly, the endurance of the device should be high to allow a large number of weighted sum operations in hardware. For the inference with the MNIST dataset[21], each Mott ReLU device should generate its output for 10,000 times per epoch (or a whole testing set). Hence, the device should endure this large number of cycling operations. Figure 3c experimentally demonstrates that the Mott ReLU device shows no sign of ON/OFF ratio degradation up to 5,000 cycles. It has been shown that an endurance larger than 10^{10} cycles can be easily achieved with VO_2 devices[22]. Furthermore, we performed pulse measurements to investigate the power consumption and the latency of the Mott ReLU device. Figure 3d shows the total power consumption as a function of heater current, as well as the power consumed by the heater and the VO_2 gap separately. The total power consumption of the Mott ReLU device is dominated by the heater. The latency of the Mott ReLU is 61.4 ns, measured as the time difference between the first saturation point of the input and output pulse (Fig. 3e). The energy consumption of the Mott ReLU is 199.5 pJ for a 65 ns pulse width.

The Mott ReLU device can replace complex peripheral circuits for activation function calculation. Therefore, it is important to compare the performance of the Mott ReLU device against other implementations of activation functions (that is, analogue CMOS[14] and digital ADC[15] circuits discussed in the Methods). The performance benchmarking (Supplementary Note 2) results of the Mott ReLU device against the analogue CMOS circuit[14] and the digital ADC implementation[15] are summarized in Table 1. The energy consumption of the Mott ReLU can be further reduced by optimizing the device to have more heat confinement on the VO_2 gap. As the heat generated by the heater is more confined to the VO_2 gap, the device requires less heater current to achieve the same temperature on the VO_2 gap[23]. Therefore, by replacing the heater material with a higher thermal resistance material (for example, Ti has a thermal resistance ~10 times higher than that of Au), the energy consumption of the device can be lowered. To determine the energy consumption of an optimized device, we developed an empirical thermal model of our device (Supplementary Fig. 1a,b) as discussed in Supplementary Note 3, which shows good agreement with experimental data as shown in Fig. 2e,f. The power consumption of the Mott ReLU can be reduced by ~25 times, down to 128 μW, by increasing the thermal resistance of the nanowire heater (Supplementary Fig. 2a). Moreover, the latency can be reduced to ~3.8 ns (Table 1) by minimizing the parasitic capacitance of the Mott ReLU below 10^{-11} F (Supplementary Fig. 2b), which would result in a total reduction of ~300 times in energy consumption down to 0.638 pJ (Table 1). Our experimental results show that the Mott ReLU device achieves a 450–1,500 times improvement in area and 1.5–3 times improvement in latency while achieving low

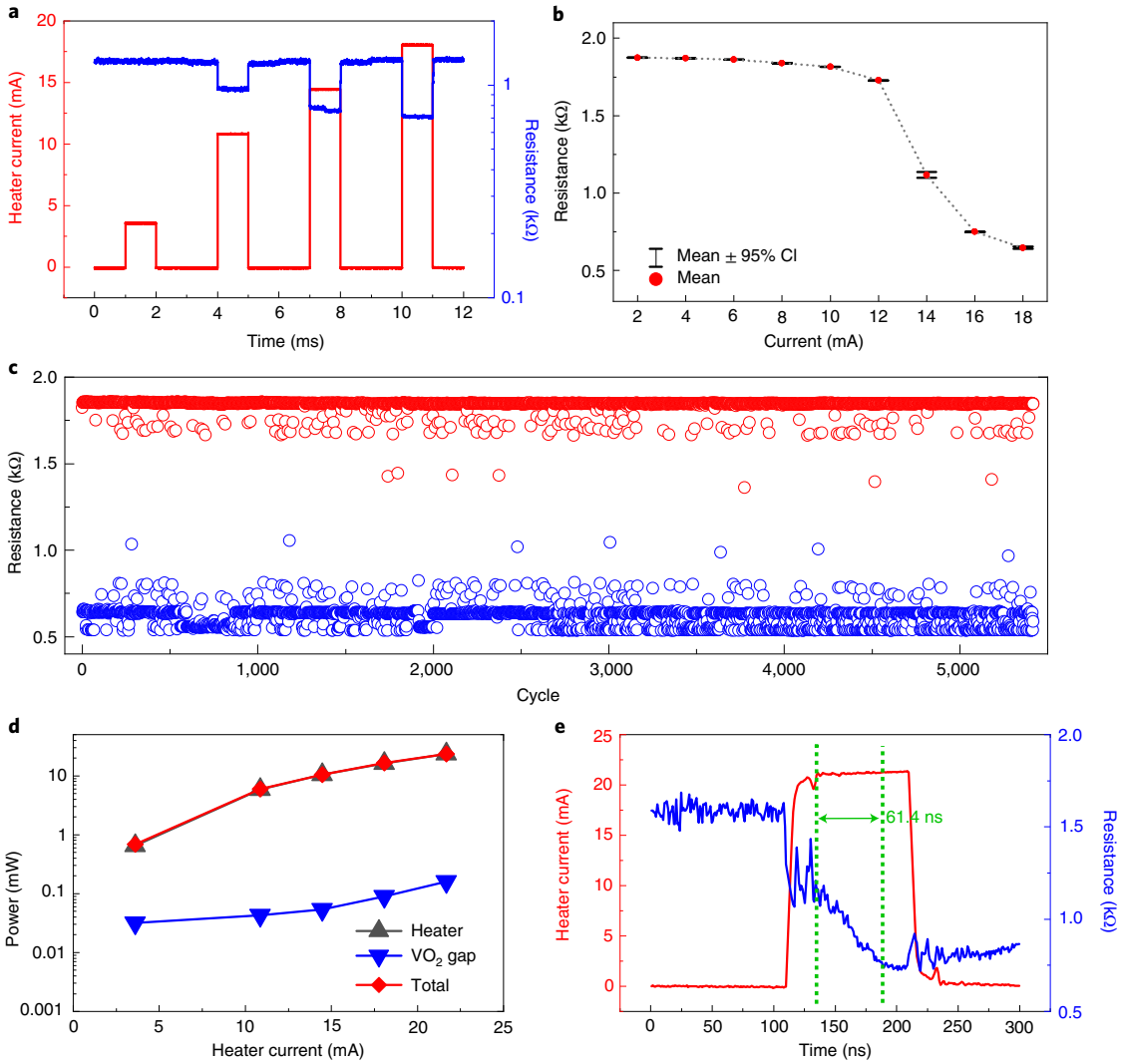

Fig. 3 | Electrical characteristics of the Mott ReLU device. a, Resistance of the VO_2 gap when a current pulse is applied to the heater. The resistance stays at a low resistance state only when the bias is applied. **b**, Cycle-to-cycle (or intra-device) variation of each resistance state of the Mott ReLU device. For each data point, the heater cools down to set the resistance of the VO_2 gap back to the no-bias case before applying another bias to the heater. The circle symbol represents the mean value while the error bars represent a 95% critical interval (CI). **c**, Endurance of the Mott ReLU device. The state of the device is alternately switched between the highest (red symbols) and lowest resistance states (blue symbols) by flowing 0 mA and 18 mA current through the heater, respectively. **d**, Power consumption of each component of the Mott device (that is, the heater and the VO_2 gap) with various heater currents. The power consumption of the Mott device is dominated by the heater. **e**, Heater current applied to the device and the resistance of the VO_2 gap as a function of time. 61.4 ns after the input to the Mott device is stabilized, the output of the Mott device becomes stable, as indicated by the green dashed lines.

energy consumption. Moreover, the optimization of the Mott ReLU device can further reduce the energy consumption and improve the latency, offering substantial gains in area, latency and energy efficiency as a replacement to the analogue CMOS[14] and digital[15] ADC circuits.

Neural network implementations

We have demonstrated that the Mott ReLU neurons can provide smaller area and better energy efficiency as compared to the other circuit implementations. It is also critical to evaluate the network-level performance using the Mott ReLU devices for hardware implementation of DNNs. To compute the accuracy of neural network implementations with the Mott ReLU device, we simulated multilayer perception (MLP; Fig. 4a) and LeNet-5 (ref. [21]; Fig. 4b; the details on the configurations of the networks are discussed in the Methods). The schematic and transmission electron

Table 1 | The performance of the activation device or circuit

	Mott	Analogue CMOS[14]	Digital ADC[15]
Energy (experimental/optimal, pJ)	199.5/0.638[a]	3,410	19.4
Latency (experimental/optimal, ns)	61.4/3.8[a]	91.91	207
Area (μm²)	0.64	951.06	289[b]
Leakage (μW)	27.0	11,060	-

[a]Shows projected optimal energy and latency when the thermal resistance of the heater is increased by ×10 and the parasitic capacitance of a Mott ReLU is <10^{-11} F. [b]This area is only the area per neuron circuit. The digital ADC implementation needs a shared circuit, which occupies 0.086 mm² of area. Comparison of Mott ReLU, analogue CMOS ReLU[14] and digital ADC with reconfigurable function mapping[15] at the single ReLU level. The energy, latency and leakage power are evaluated from the experimental measurement results shown in Supplementary Fig. 2a,b. For the energy estimation, we used a 65 ns pulse for the Mott ReLU case.

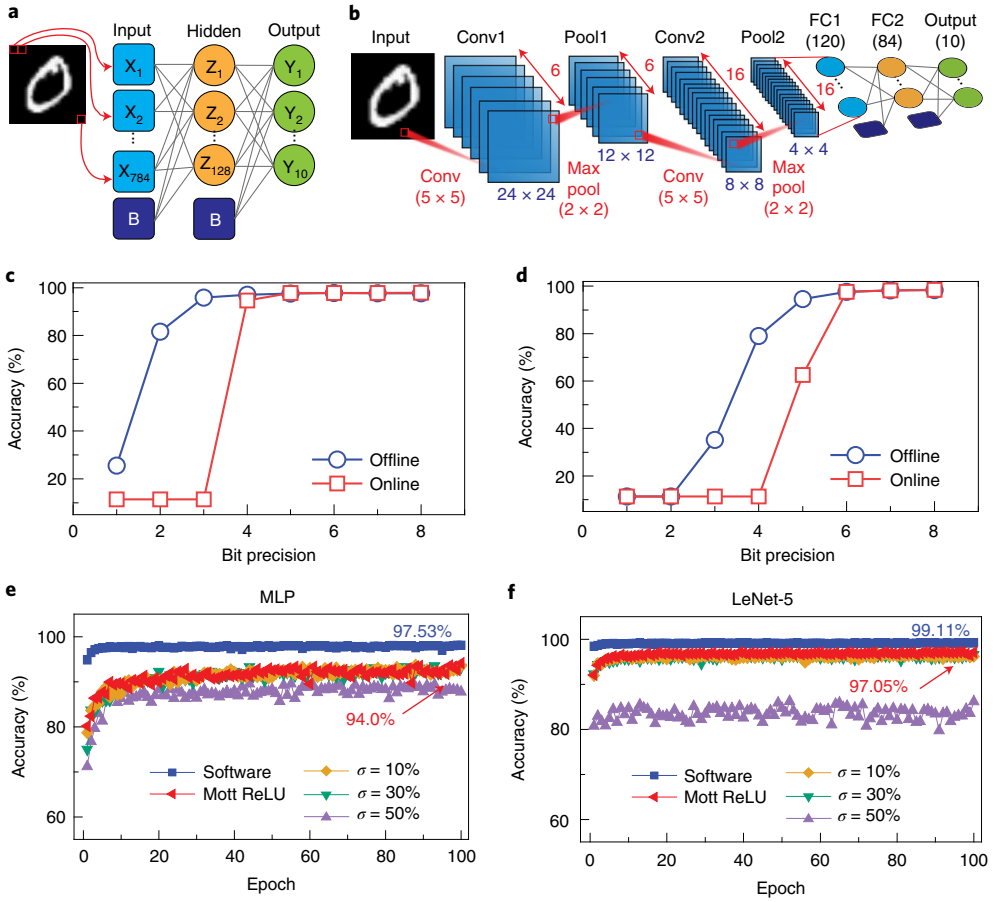

Fig. 4 | Network-level implementations. a,**b**, A schematic of MLP (**a**) and LeNet-5 (**b**) networks used for simulations with the Mott ReLU. MLP consists of an input layer (X), a hidden layer (Z) and an output layer (Y) with bias (B) for the input and hidden layers. MLP has one ReLU layer and LeNet-5 has four ReLU layers after convolutional (Conv) and fully connected (FC) layers. **c**,**d**, Accuracy of MLP (**c**) and LeNet-5 (**d**) with the ReLU activation function for offline classification (blue circle symbol) and online learning (red square symbol). The ReLU activation function is quantized to have 1- to 8-bit precision. MLP needs 5-bit precision while LeNet-5 requires 6-bit precision to prevent the accuracy drop. **e**,**f**, The network simulation results for MLP (**e**) and LeNet-5 (**f**) for the whole MNIST set for each epoch (60,000 images). Experimental measurement results from Fig. 2e,f are used for these simulations. The Mott ReLU achieves an accuracy comparable to the ideal ReLU implemented in software (blue square symbol) unless the cycle-to-cycle (or intra-device) variation of the Mott ReLU device (σ) is higher than 50%. Red triangle, yellow diamond, green triangle and purple triangle symbols represent results for the no variation, $\sigma = 10\%$, $\sigma = 30\%$ and $\sigma = 50\%$ cases, respectively.

microscopy image of the CBRAM cell (Supplementary Fig. 3a) are shown in Supplementary Fig. 3b,c, respectively. Table 2 summarizes the accuracy results of the ideal (that is, software ReLU) and Mott ReLU cases for both MLP and LeNet-5. We investigated both online learning (that is, training is done on the hardware) and offline classification cases (that is, only inference is done on the hardware). When the ReLU activation functions of MLP (Fig. 4c) or LeNet-5 (Fig. 4d) are quantized, the accuracy degradation is not significant unless the precision is ~6 bit or higher. Since the precision of the Mott ReLU device is high enough (~6 bit), the accuracy degradation due to the Mott ReLU is negligible as compared to the accuracy degradation due to the synaptic devices (~10% for MLP and ~3% for LeNet-5). This is mainly because of the limited precision (~5 bit) of the CBRAM devices[24]. The neural networks with variations (cycle-to-cycle in Fig. 4e,f and device-to-device in Supplementary Fig. 4a,b) on the Mott ReLU are also investigated, and it is verified that no significant accuracy degradation due to the variations occurs (Supplementary Note 4). Since the Mott ReLU achieves accuracies close to the ideal software, the accuracy will not be a limiting factor for implementing activation functions using the Mott ReLU device.

System-level performance benchmarking

To evaluate the performance of the hardware system for neural networks with the Mott ReLU device, we performed system-level performance benchmarking for offline classification using the NeuroSim platform[25]. NeuroSim is a C++-based circuit-level macro-model for neuro-inspired architectures. We modified NeuroSim to integrate Mott ReLU peripherals with CBRAM syn-

Table 2 | Network simulation results

	Online learning		Offline classification	
	MLP	LeNet-5	MLP	LeNet-5
Software (64 bit)	97.53%	99.11%	97.53%	99.11%
Mott ReLU (~6 bit)	94.0%	97.05%	94.42%	98.38%
CBRAM (~5 bit) with Mott ReLU (~6 bit)	84.2%	94.21%	89.97%	98.35%

The accuracy results of MLP and LeNet-5 for ideal software (64 bit), 64-bit weights with Mott ReLU (~6 bit) and CBRAM (~5-bit weights) with Mott ReLU (~6 bit). The results show that the Mott ReLU can achieve accuracy comparable to the ideal software ReLU.

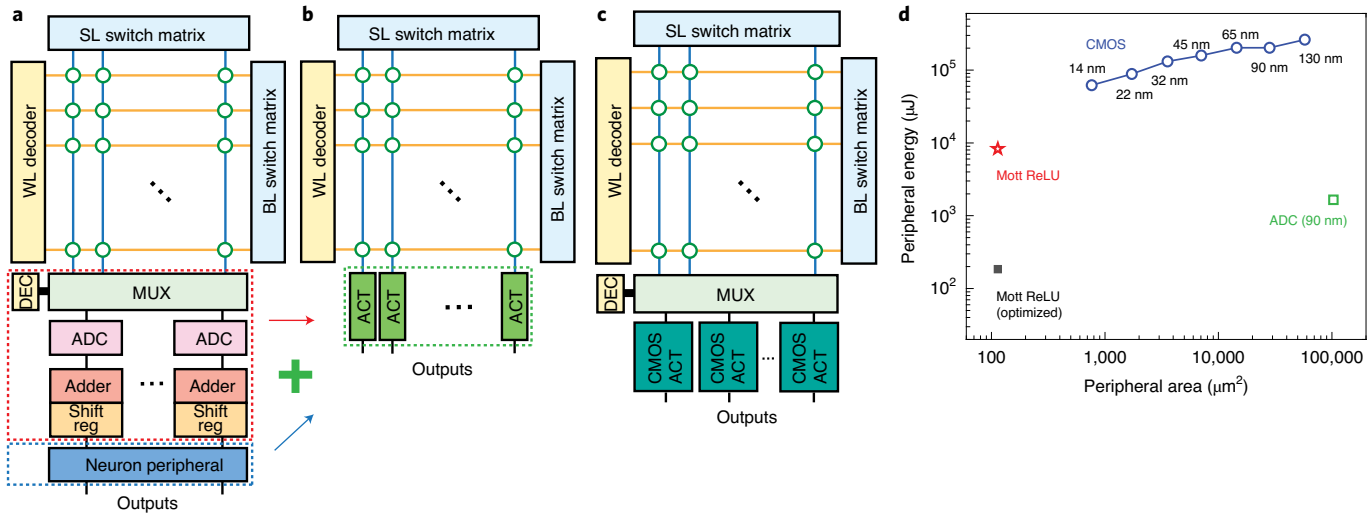

Fig. 5 | System-level benchmarking results. a–c, An illustration of a synaptic core and neuron peripheral circuits implemented with conventional digital circuits (**a**), Mott ReLU circuits (**b**) and analogue CMOS ReLU circuits (**c**). The Mott ReLU device can replace the ADCs, adder, shift register and neuron peripheral circuits. The CBRAM synaptic core has a wordline (WL) decoder (DEC), bitline (BL) switch matrix and source line (SL) switch matrix. In contrast to the CMOS analogue activation circuit (ACT), the Mott ReLU device can be integrated for each column due to its small size. **d**, Peripheral energy versus peripheral area in different technology nodes for CBRAM synaptic core with CMOS ReLU peripheral for LeNet-5 implementation. A CBRAM synaptic core with digital ADC peripherals (green square symbol) and Mott ReLU is also presented as a reference. The parameters for different technology nodes of CMOS circuits are adopted from the predictive technology model[27,28]. The Mott ReLU continues to provide substantial gains in energy and area even though the CMOS is scaled down to a 14 nm node. The star symbol shows performance results using experimentally measured Mott ReLU characteristics, while the black square symbol shows projected performance results using an optimized Mott ReLU device (that is, the thermal resistance of the heater is increased by ×10 and the parasitic capacitance is below 10^{-11} F). The system-level energy consumption using the optimized Mott ReLU can be further reduced by ~50 times.

aptic cores. We compared the synaptic cores with the Mott ReLU peripheral against the ones with peripheral circuits implemented by analogue[14] and digital[15] CMOS ReLU circuits. For the Mott ReLU peripheral, the experimental results on energy and latency (Fig. 3d,e) are integrated into the NeuroSim platform[25]. The peripheral circuits of analogue CMOS ReLU circuits for the NeuroSim platform[25] are developed based on the SPICE simulations. The dynamic energy, leakage power and latency of the Mott ReLU and CMOS ReLU activation circuits shown in Table 1 are integrated into the circuit modules.

The architecture of the hardware systems with conventional digital peripheral circuits, the Mott ReLU device and analogue CMOS circuits are illustrated in Fig. 5a–c, respectively. In contrast to the conventional analogue one-transistor one-resistor (1T1R) architecture with digital neuron peripheral (Fig. 5a)[25], the Mott ReLU device allows a simpler synaptic core design (Fig. 5b) by avoiding multiplexer (MUX) sharing (Supplementary Note 5) and replacing complex circuits and ADCs. Before system-level benchmarking, we first investigated whether the Mott ReLU device can drive the inputs to the next synaptic array without additional circuits by performing circuit simulation with Simulation Program with Integrated Circuit Emphasis (SPICE; Supplementary Note 6). This result (Supplementary Fig. 5a,b) clearly demonstrates that the Mott ReLU device can generate stable output to drive the next synaptic layer without additional circuits. The system-level performance benchmarking results are summarized in Supplementary Table 1. The architecture with the Mott ReLU (65 ns input pulse) provides substantial gains over the architectures with analogue CMOS and digital ADC implementations (Supplementary Note 7). Lastly, we compared the performance of synaptic cores with the Mott ReLU and analogue CMOS circuits considering technology scaling (130 nm to 14 nm) as discussed in the Methods. The results in Fig. 5d demonstrate that the experimentally measured Mott ReLU provides ~10 times the energy gain regardless of the CMOS technology node. Moreover, the system-level gain in energy can be further improved up to ~100 times using the optimized Mott ReLU in comparison to the analogue CMOS ReLU. More importantly, the Mott ReLU achieves an orders of magnitude smaller peripheral circuit area in comparison to both the digital ADC and analogue CMOS implementations of the activation function. The system-level performance results show that the Mott ReLU device offers a promising approach to replace power-hungry and large-area activation function circuits in the neuron periphery.

Integration of Mott ReLU devices with crossbar arrays

To demonstrate the integration of Mott ReLU devices with synaptic arrays in hardware, we fabricated CBRAM crossbar arrays (Fig. 6a) and a Mott ReLU device array (Fig. 6b) as explained in the Methods. We designed a custom printed circuit board (PCB; Fig. 6c) to interface and integrate the CBRAM and the Mott ReLU chips in hardware. Each column of the crossbar array is directly connected to Mott ReLU devices (Fig. 6d) to investigate how the weighted sum current generated by the array controls the output voltage of the Mott ReLU devices. First, we varied the input voltage to the crossbar array (−250 to 250 mV) while programming the weights of ~2/3 of the synaptic devices on a column to the low resistance state, and the rest to the high resistance state. Figure 6e shows that the output voltage exhibits ReLU characteristics as the input voltage to the CBRAM devices is increased from −250 mV to 250 mV. Then, we varied the synaptic weights in the column while the input voltage was fixed at 130 mV. As the ratio of devices programmed to the low resistance state increases, the output voltage exhibits ReLU characteristics (Fig. 6f). These experimental results demonstrate that the weighted sum current that depends on the input voltage and the weights (resistance) of the synaptic devices can successfully drive the Mott ReLU neurons to implement ReLU activation function.

For a large-scale hardware demonstration, we implemented a convolutional edge detection operation[26] with filters (Supplementary

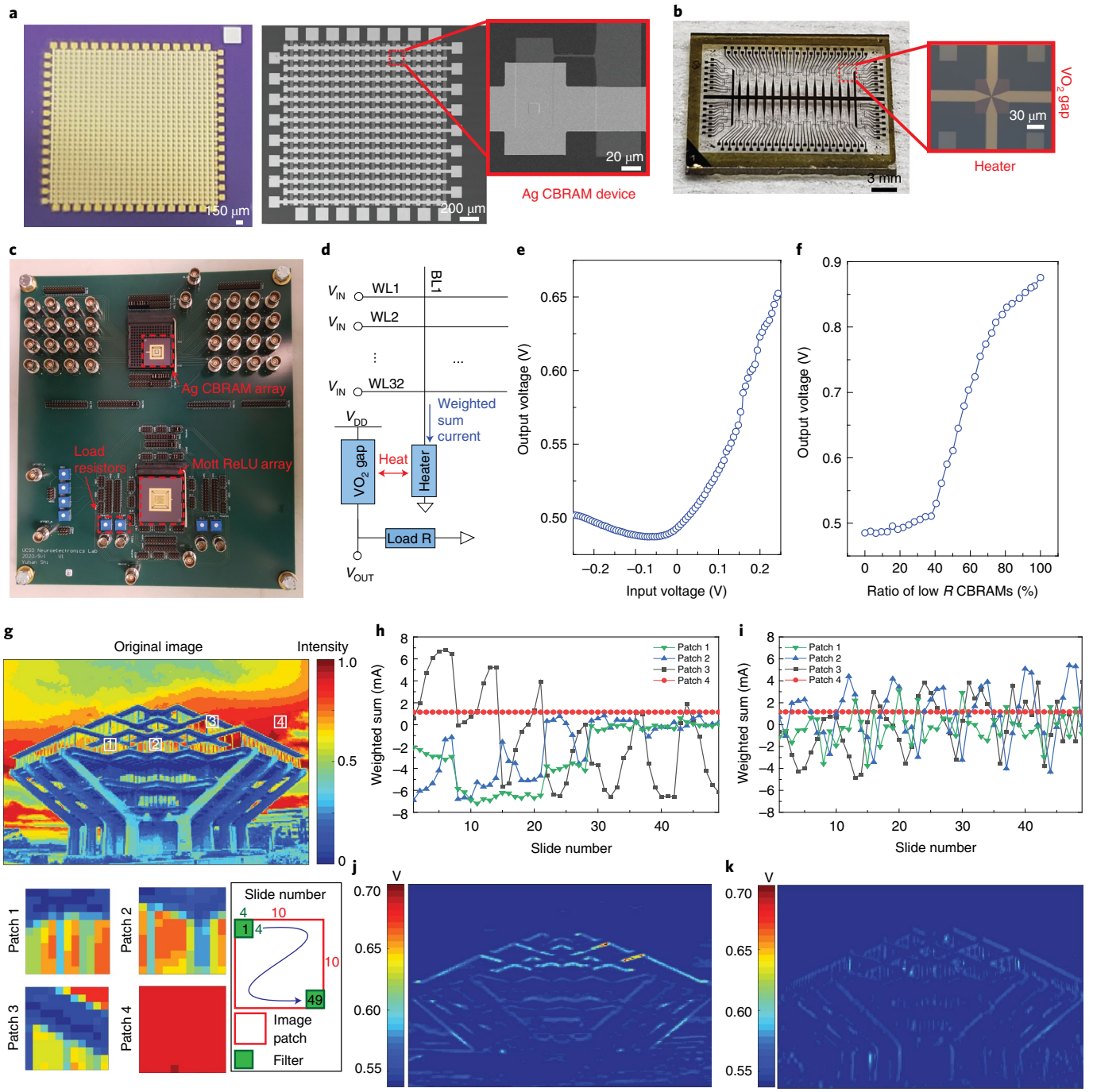

Fig. 6 | Hardware demonstration of the integration of Mott ReLU devices and a synaptic array. a, An optical image (scale bar, 150 μm) of a CBRAM crossbar array (32 × 32) and the scanning electron microscope image of a 16 × 16 CBRAM array (scale bar, 200 μm). We use a 16 × 16 array for the following hardware implementation. **b**, A Mott ReLU device array that contains 44 devices (scale bar, 3 mm). The insets in **a** and **b** (scale bars, 20 μm and 30 μm, respectively) show single devices, the CBRAM and Mott ReLU device, respectively. **c**, Image of the custom PCB board with the Mott ReLU and the CBRAM arrays wire bonded onto it to demonstrate neural network operation. **d**, An illustration explains how a Mott ReLU device is connected to a column of the CBRAM array with a load resistor (Load R) in hardware. **e,** Output voltage of the Mott ReLU device as the input voltage (V_{IN}) to the CBRAM array is swept from −250 mV to 250 mV when ~2/3 of devices on a column of the CBRAM array are set to a low resistance state while the others are set to a high resistance state. For the Mott ReLU device, 1.1 V is applied as V_{DD} to the VO_2 gap with a 3.3-kΩ-load resistor connected in series, and 7 mA of offset current is applied to the heater. **f**, Measured output voltage of a Mott ReLU device when the percentage of CBRAM devices at the low resistance state is varied from 0% to 100%. **g**, A 180 × 270 image used for edge detection. Colour bar represents the pixel intensity of the image. Four representative 10 × 10 patches and a schematic of the convolution operation are shown below. The schematic illustrates that the convolution operation is done by sliding the 4 × 4 filters on the image patches 49 times. **h**,**i**, For a lateral filter (**h**) and vertical filter (**i**), the experimentally measured weighted sum currents of the CBRAM array during the convolution operations for these four patches are shown. The weighted sum current produced by the CBRAM array during the convolution operation is fed to the Mott ReLU array to perform ReLU operation. **j**,**k**, Panels **j** and **k** show the output voltage of the Mott ReLU devices for the whole image during the convolution and ReLU operations for the lateral and vertical filters, respectively. Colour bar represents the output voltage of the Mott ReLU.

Fig. 6a,b) followed by a ReLU operation on a real-world image with the CBRAM crossbar and the Mott ReLU array using the custom PCB as discussed in the Methods. The weighted sum current resulting from the convolution operation from four representative 10×10 input patches (Fig. 6g) with both the lateral and vertical edge detection filters (Supplementary Fig. 6a,b) mapped using a differential pair scheme (Supplementary Fig. 6c) are shown in Fig. 6h,i, respectively. The weighted sum current generated during the convolution operation is fed to the Mott ReLU devices to perform the ReLU operation on the weighted sum. The output voltages of the Mott ReLU devices as a result of the weighted sum with the lateral and vertical filters for the whole input image are shown in Fig. 6j,k, respectively. These results show that lateral and vertical edges of the image are detected by implementing corresponding filters using the Mott ReLU devices integrated with the CBRAM crossbar array in hardware. The successful edge detection using the Mott ReLU devices integrated with the CBRAM crossbar array proves the feasibility of using Mott ReLU neurons as activation units for in-memory computing systems.

Conclusions

We introduced a nanoscale, Mott-transition-based device for the ReLU activation function. The device exhibits volatile, linear and gradual resistive switching of a VO_2 film controlled by the metal nanowire heater on top of it. The Mott ReLU device shows minimal cycle-to-cycle variation and long endurance, which are important for hardware implementation of neural networks. We have shown that the Mott ReLU devices generate an output voltage, which follows the ReLU activation function, with the given input current. This allows the Mott ReLU device to drive the synaptic devices on the next layer directly. We performed system-level simulations for a hardware implementation of neural networks with the Mott ReLU devices. Moreover, we experimentally demonstrated that the Mott ReLU devices can be integrated with CBRAM crossbar arrays to perform filtering operations of convolutional neural networks. Our findings suggest that the device with Mott-transition-based activation can achieve substantial gains in energy, latency and area compared to the digital or analogue circuit implementations of the activation function, while maintaining high accuracy. The small size and high energy efficiency of the Mott device provide a solution towards large-scale, highly parallel and energy-efficient in-memory computing systems for DNNs.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41565-021-00874-8.

Received: 12 February 2020; Accepted: 2 February 2021;
Published online: 18 March 2021

References

1. Wong, H. S. P. et al. Metal–oxide RRAM. *Proc. IEEE* **100**, 1951–1970 (2012).
2. Zidan, M. A., Strachan, J. P. & Lu, W. D. The future of electronics based on memristive systems. *Nat. Electron.* **1**, 22–29 (2018).
3. Kang, D.-H. et al. A neuromorphic device implemented on a salmon-DNA electrolyte and its application to artificial neural networks. *Adv. Sci.* **6**, 1901265 (2019).
4. Ge, R. et al. Atomristor: nonvolatile resistance switching in atomic sheets of transition metal dichalcogenides. *Nano Lett.* **18**, 434–441 (2018).
5. van de Burgt, Y. et al. A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nat. Mater.* **16**, 414–418 (2017).
6. Zhao, X. et al. Confining cation injection to enhance CBRAM performance by nanopore graphene layer. *Small* **13**, 1603948 (2017).
7. Chakrabarti, B. et al. A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit. *Sci. Rep.* **7**, 42429 (2017).
8. Kim, S. et al. Binarized neural network with silicon nanosheet synaptic transistors for supervised pattern classification. *Sci. Rep.* **9**, 11705 (2019).
9. Oh, S., Huang, Z., Shi, Y. & Kuzum, D. The impact of resistance drift of phase change memory (PCM) synaptic devices on artificial neural network performance. *IEEE Electron Device Lett.* **40**, 1325–1328 (2019).
10. He, K., Zhang, X., Ren, S. & Sun J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
11. Kataeva, I. et al. Towards the development of analog neuromorphic chip prototype with 2.4M integrated memristors. In *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)* 255–259 (IEEE, 2019).
12. Gao, B. et al. Ultra-low-energy three-dimensional oxide-based electronic synapses for implementation of robust high-accuracy neuromorphic computation systems. *ACS Nano* **8**, 6998–7004 (2014).
13. Yang, T.-J. & Sze, V. Design considerations for efficient deep neural networks on processing-in-memory accelerators. In *Proc. IEEE International Electron Devices Meeting (IEDM)* 514–517 (IEEE, 2019).
14. Krestinskaya, O., Salama, K. N. & James, A. P. Learning in memristive neural network architectures using analog backpropagation circuits. *IEEE Trans. Circuits Syst. I* **66**, 719–732 (2019).
15. Giordano, M. et al. Analog-to-digital conversion with reconfigurable function mapping for neural networks activation function acceleration. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **9**, 367–376 (2019).
16. Ambrogio, S. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60–67 (2018).
17. Stefanovich, G., Pergament, A. & Stefanovich, D. Electrical switching and Mott transition in VO_2. *J. Phys. Condens. Matter* **12**, 8837–8845 (2000).
18. Qazilbash, M. M. et al. Mott transition in VO_2 revealed by infrared spectroscopy and nano-imaging. *Science* **318**, 1750–1753 (2007).
19. del Valle, J. et al. Subthreshold firing in Mott nanodevices. *Nature* **569**, 388–392 (2019).
20. Madan, H., Jerry, M., Pogrebnyakov, A., Mayer, T. & Datta, S. Quantitative mapping of phase coexistence in Mott-Peierls insulator during electronic and thermally driven phase transition. *ACS Nano* **9**, 2009–2017 (2015).
21. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
22. Radu, I. P. et al. Switching mechanism in two-terminal vanadium dioxide devices. *Nanotechnology* **26**, 165202 (2015).
23. del Valle, J., Salev, P., Kalcheim, Y. & Schuller, I. K. A caloritronics-based Mott neuristor. *Sci. Rep.* **10**, 4292 (2020).
24. Shi, Y. et al. Neuroinspired unsupervised learning and pruning with subquantum CBRAM arrays. *Nat. Commun.* **9**, 5312 (2018).
25. Chen, P.-Y., Peng, X. & Yu, S. NeuroSim: a circuit-level macro model for benchmarking neuro-inspired architectures in online learning. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **37**, 3067–3080 (2018).
26. Shrivakshan, G. & Chandrasekar, C. A comparison of various edge detection techniques used in image processing. *Int. J. Comput. Sci. Issues* **9**, 269–276 (2012).
27. Zhao, W. & Cao, Y. Predictive technology model for nano-CMOS design exploration. *ACM J. Emerg. Technol. Comput. Syst.* **3**, https://doi.org/10.1145/1229175.1229176 (2007).
28. Zhao, W. & Cao, Y. New generation of predictive technology model for sub-45 nm early design exploration. *IEEE Trans. Electron Devices* **53**, 2816–2823 (2006).

Methods

Mott device fabrication. To fabricate the Mott-transition-based activation devices, a 70 nm VO_2 film is grown by reactive sputtering on top of an Al_2O_3 substrate in 4 mtorr Ar/O_2 (8% O_2) ambient at 520 °C. Then Ti (20 nm)/Au (30 nm) electrodes are patterned using electron-beam (e-beam) lithography and e-beam evaporation to define the 50 nm VO_2 gap. Next, 70 nm Al_2O_3 is deposited as the insulating layer. A Ti (20 nm)/Au (30 nm) nanowire heater is patterned on top of the Al_2O_3, aligning with the VO_2 gap using e-beam lithography and e-beam evaporation. To isolate each device, the VO_2 film outside the active area is etched with reactive ion etching. The resistance of the heater is ~30 Ω, while the resistance of the VO_2 gap without bias to the heater is ~10 kΩ.

Device measurement set-up. To measure the thermal gradual resistance switching of VO_2 while preventing electrical switching, we apply a 1 μA current source, using a Keithley 6221, to the VO_2 gap. The current is small enough not to initiate electrical switching. Then, we measure the voltage across the VO_2 gap using a Keithley 2182A. The resistive switching of the VO_2 gap is solely controlled by the heat generated by the heater on the top of the VO_2 gap. The heat generation is controlled by a voltage source connected to the heater. We measure the current flow through the heater to measure the heat generation using an oscilloscope. For the variability and endurance measurement, the Keithley 6221 is used to apply a current pulse train to the heater. Then the resistance of the VO_2 gap is extracted by measuring the voltage across the VO_2 gap using the Keithley 2182A while applying constant 1 μA current through the gap using another Keithley 6221. The ambient temperature is controlled by a Lake Shore TTPX Probe station for all the measurements.

CMOS ReLU implementation. The analogue CMOS circuit consists of three operational amplifiers, which amplify the input current and convert the input current to the output voltage, and an analogue switch that implements the rectifying function. The digital ADC circuit is implemented using ADC with reconfigurable function mapping. In order to evaluate the energy and latency of these three different ReLU implementations as an activation function, we assume that all implementations get an identical weighted sum result as an input to the Mott ReLU device or digital/analogue CMOS circuits. The area of each implementation is calculated from the layout of the device or circuits.

Neural network configuration. The MLP used for network simulations is composed of 785 input neurons (that is, 1 input neuron for bias and the other 784 neurons for MNIST dataset inputs), 128 hidden neurons and 10 output neurons. Each output neuron represents one of the digits (from 0 to 9). The hidden neurons have the ReLU activation function, while the output neurons have the soft-max activation function. LeNet-5 has six 5×5 convolutional filters for 28×28 MNIST input images. The outputs from the convolutional filters are fed to the ReLU activation function. Then, the outputs of ReLU activation functions are down-sampled using 2×2 max pooling. The second convolutional layer has sixteen 8×8 convolutional filters with 2×2 max pooling. The outputs from the last max-pooling layer are fed into the FC layers, which have 120 input neurons (FC1), 80 hidden neurons (FC2) and 10 output neurons (Output). The input neurons and hidden neurons of the FC layers have ReLU activation functions, while the output neurons have soft-max activation functions.

In the network simulations, the ReLU activation functions on the neuron layers (that is, the hidden layer of MLP and convolutional layers and FC layers of LeNet-5) are implemented with the Mott ReLU based on its experimental measurement results. A 1,900-Ω-load resistor is connected to the Mott ReLU, and 5 mA of offset current is applied to the Mott ReLU through an additional row on the synaptic array to shift the transition point to 0 mA. The weights are mapped onto the arrays of CBRAM devices by using the characteristics of CBRAM devices. The CBRAM cells used for the simulations exhibit ~40 conductance levels (~5 bit) and an ON/OFF ratio of 100. For the network simulation, the weights of the network ranging from −1 to 1 are mapped to the minimum (~1 μS) and maximum (~100 μS) conductance of CBRAM cells. Similarly, the outputs of the ReLU activations (0 to 785) are also linearly mapped to the output voltages of Mott ReLU devices (0 to 200 mV).

LeNet-5 requires a larger fanout for the FC1 layer. To address this, we incorporated a time multiplexing approach. By enabling a subset of columns of the synaptic array sequentially with the switch matrix, the number of devices connected to each Mott ReLU can be controlled. Since our architecture already has a switch matrix, this approach is directly implemented in performance benchmarking simulations with NeuroSim. It is important to note that larger-scale DNN models may require additional peripheral circuit blocks including buffers if they have many layers with large fanout. These blocks could be integrated with the synaptic arrays in the future and accounted for the performance benchmarking for different models.

Convolutional filtering with the Mott ReLU device integrated with CBRAM array. To implement convolutional filtering using the Mott ReLU and CBRAM array for image edge detection, the PCB is controlled by a semiconductor parameter analyser (Agilent 4155C) and a switch matrix (HP E5250A). Then, biasing and measurement are done by the semiconductor parameter analyser (Agilent 4155C). The 4×4 lateral and vertical filters are programmed into the columns of the crossbar array by unrolling the filters into 16×1 vectors on the CBRAM array. For each filter, the positive and negative weights are represented using two columns of the crossbar array to form a differential pair (that is, $G = G^+ - G^-$). The input image (180×270) is quantized (16 levels) and converted into a voltage pulse train of four binary pulses (250 mV for '1' and 0 mV for '0'). For the column representing negative weights, a negative voltage pulse train is applied as input to form a differential pair with the column representing positive weights (that is, $I = I^+ - I^-$). For the convolution operation, a filter slides over the input image and the weighted sum currents from the pair are combined and fed into a Mott ReLU device. For Mott ReLU devices, 1.1 V is applied to the VO_2 gap, load resistors are set to 3.3 kΩ and 7 mA of offset current is applied to the heater.

Data availability

The data that support the plots and other results of this paper are available from the corresponding author upon request.

Code availability

The software codes used for this study are available from the corresponding author upon request.

Acknowledgements

This work was supported by Office of Naval Research (N000142012405 and N00014162531), Samsung Electronics, the National Science Foundation (ECCS-1752241, ECCS-2024776 and ECCS-1734940), the National Institutes of Health (R21 EY029466, R21 EB026180 and DP2 EB030992) and Qualcomm Fellowship. The experimental aspects of this work were supported as part of the Quantum Materials for Energy Efficient Neuromorphic Computing (Q-MEEN-C) Energy Frontier Research Center (EFRC), funded by the US Department of Energy, Office of Science, Basic Energy Sciences under award #DE-SC0019273. The fabrication of the devices was performed at the San Diego Nanotechnology Infrastructure (SDNI) of the University of California San Diego, supported by the National Science Foundation (ECCS-1542148).

Author contributions

S.O., Y.S., I.K.S. and D.K. conceived the idea. S.O. developed the SPICE model and performed SPICE and network simulations. S.O., Y.S. and Z.H. performed the system-level benchmarking simulations. P.S., J.d.V. and Y.K. fabricated and measured the electrical characteristics of the Mott ReLU devices under the supervision of I.K.S.; S.O. and Y.L. fabricated the CBRAM array. Y.S. designed the PCB. S.O. and Y.S. performed the measurements for the CBRAM array and Mott ReLU device integration. All the authors discussed the results and contributed to the writing of the manuscript. I.K.S. and D.K. supervised the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41565-021-00874-8.

Correspondence and requests for materials should be addressed to D.K.

Peer review information *Nature Nanotechnology* thanks Jinfeng Kang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.