The future of Neuromorphic Engineering——Memristor: Characteristics and Applications

Yuan Lei & Wenrui Yu

Abstract—**Despite the fact that neuromorphic computing architectures such as Neural Networks are widely employed in computer vision and data science applications, the development of neuromorphic computing has long been losing its connection with bionics and neuroscience. In this case, the robustness and power consumption of the present neuromorphic computing is not even close to the physical limits of semiconductors. Due to its scalability, tunability and resistive switching characteristics, the memristor is found to be the promising solution to a more compact and energy-efficient computing architecture. The object of this article is to find why memristor prevails in neuromorphic computing by illustrating its basic characteristics and evaluating its performance in real applications. The data in this review is drawn from papers describing the development in memristor-based computing architecture. In this paper, we first conclude the structure and electrical properties of memristor by reviewing previous publications regarding the modeling, fabrication and instrumentation of memristor. Then assessing the in-situ method neuromorphic computing architecture, by combining information of several papers related to the developments of carbon nanotube applications and evaluating the error rate and energy efficiency of recently constructed networks. As a result, we highlight the applications' high performance in traditional classification issues, owing to their low energy consumption and small size. However, device offset and process variability affection are the main difficulties that still needed to be solved, which lead future research a long way to go.**

Index Terms—**Neuromorphic engineering, memristor, neuromorphic computing, crossbar array, neural network**

I. INTRODUCTION

Neuromorphic engineering, also known as neuromorphic computing, is a subject to simulate the neuron activity in the nervous system with Very Large Scale Integration (VLSI) electronic circuits [1]. As the demand for high-speed and large-scale computing increases rapidly, the speed of conventional computing architectures is limited by the shared bus between the program memory and data memory, namely the 'von Neumann bottleneck', thus neuromorphic computing receives more attention nowadays.

The hardware-level implementation of neuromorphic computing architectures can be realized by memristors. The memristor is a type of non-linear two-terminal electronic device that is capable of nonvolatile resistive switching [2]. Compared to traditional electronic components, which are known as resistors, capacitors and inductors, memristors have drawn a lot of attention from both industry and academia. The memristor has lower leakage, better scalability, lower power consumption [3]. Furthermore, it has a biomimetic function, showing that it has the potential to simulate and construct artificial synapses. With these characteristics, organizing memristors into large arrays leads to a promising future of high-speed computing, as the intermediate computing matrix can be stored as memristor conductance [4]. This non-von Neumann architecture, which simulates the neuron connections, may have the opportunity to cause iteration in the computing system.

This review article aims to assess the in-situ method to implement neuromorphic computing architectures using memristor-arrays by evaluating the error rate/accuracy and energy efficiency. With analysis, this article may give an inspiration of architecture of computing system in the future. Based on this target, the article will first focus on the resistive switching characteristics of memristor by reviewing and concluding the latest progress on the research of memristor material, structure and fabrications. Then paying more attention to discuss its three latest applications: perceptron [5], LSTM [6] and CNN [7].

The article will first illustrate the physical structure and electrical properties of memristors in Section II. In Section III, its recent applications in neuromorphic computing will be explored. Then, the article will point out the existing problems of memristor fabrication and application in Section IV. Finally, in Section V, a conclusion will be drawn on the performance of these applications.

II. THE CHARACTERISTICS OF MEMRISTORS

This section illustrates the prerequisites for exploring the application of memristor in neuromorphic computing. In the first part of this section, the basic model of memristor will be presented. Then, the resistive switching behavior will be discussed. The tunability of the memristor will be further illustrated in the second part of this section.

A. Scalability & Restive Switching

As shown in fig.1, a simplified equivalent circuit of the memristor is presented in sub-graph a, where V is the voltmeter, A stands for the ammeter, w for the length of the doped region of the device, D for the total length of the device. The doped region, which has low resistance R_{ON}, and the undoped (or low dopant) region, which behaves much higher resistivity R_{OFF}, are connected in series. The boundary of the two regions can be moved by the applied external voltage V_{ex}, which equals to $v_0 sin(\omega_0 t)$, since it causes the drift of charged dopants [8]. If assuming that the device has an average ion mobility μ, the memresistance of such a device can be formalized as:

$$M(q) = R_{ON,OFF}\left(1 - \frac{\mu R_{ON}}{D^2} q(t)\right). \qquad (1)$$

$R_{ON,OFF}$ is the effective resistance of the high-resistive and low-resistive state of memristor. The term q(t) represents the total electron flowing through the device with respect to time. q(t) would become larger in absolute value for higher dopant mobility μ and smaller semiconductor film thickness D. Due to the factor $1/D^2$, memristance could become more significant when scaling down the device scale.

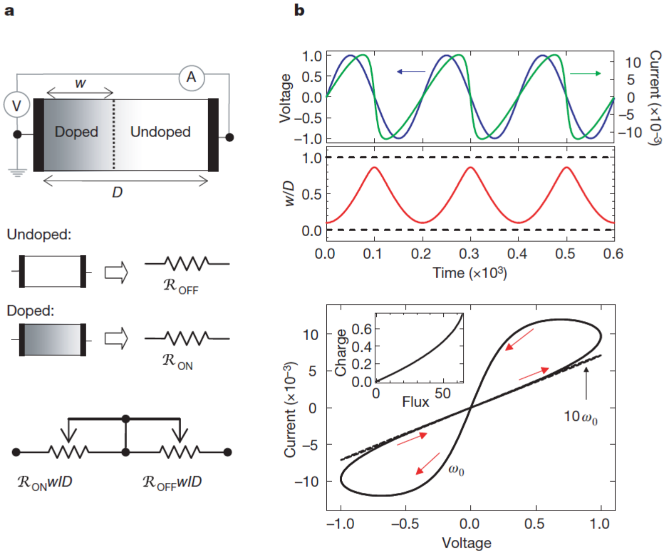

Fig. 1. (a) The coupled variable-resistor model for a memristor, (b) the transient response (the applied voltage in blue, the resulting current in green, the width of doped area in red) and the I-V characteristics of the memristor model [3, pp. 2].

Fig.1 b shows the I-V characteristics of the memristor and the graph of the applied voltage and resulting current as a function of time t. The applied voltage is $v_0 sin(\omega_0 t)$ and the resistance ratio is R_{OFF}/R_{ON}=160, where v_0 is the magnitude of the applied voltage and ω_0 is the frequency. We can conclude from the graph that the I-V curve of the memristor does not have the same linear property as that of resistors, nor does it have the integral or differential behavior as that of capacitors and inductors. Instead, the equivalent resistance (memristance) changes with respect to the direction of the voltage variation. In other words, the memristor has the capability to "memorize" the previous state, which is called resistive switching. The resistive switching of the memristor is similar to the behavior of synapsis in the human brain, hence, making a great advance towards neuromorphic computing and biomimetic engineering.

B. *Tunability*

Different from the definition of resistance and capacitance, the memristance is less physically determined, which means there are more variables that can tune the performance of the memristor. As expressed in equation (2), resistance is defined as the conductivity ρ times the ratio of the length L and the cross-section area S of the resistor, while in equation (3) capacitance can be expressed as the permittivity ϵ times the area of capacitor S divided by 4π times the product of electrostatic force constant k and the distance d between the plates of the capacitor. However, this is not the case for memristor.

$$Resistance = \rho\frac{L}{S} \qquad (2)$$

$$Capacitance = \frac{\epsilon S}{4\pi k d} \qquad (3)$$

The memristance can be tuned by many factors, such as doping [9], structured optimization [10], electrode interface engineering [11], as well as grain boundary [12]. Thus, some relevant mechanisms have been proposed to modify the performance of memristor. A graphene/MoS2−xOx/graphene (GMG) memristor [13] is reported by Wang. MoS2−xOx is obtained by partial oxidation of 2D MoS2 nanosheets. The switching of the device can be tuned by the migration of S and O atoms caused by the Joule heating-induced thermo-electrophoresis effect. In addition, Sangwan reported a memristor-based on single-layer MoS2 by utilizing the grain boundaries (GBs) [12]. The grain boundary can be repeatedly modulated by the bias voltages, leading to the transition between the high resistive state and low resistive state with an on/off ratio up to 1×10^3.

To summarize, the memristor is capable of "memorizing" the previous state, and its scalability and tunability make it possible to enhance the performance while lowering the power consumption of very large-scale integration.

III. Applications in Neuromorphic Computing

Using memristors to build the crossbar arrays is becoming a popular choice in neuromorphic computing with its unique features, especially in the construction of neural networks. Through the timeline we can observe, the complexity and functional completeness of the structure are gradually developing.

To present the recent applications in neuromorphic computing, this section is divided into three parts. In each part, one application is explained and evaluated. The memristor-based fully-connected (FC) network is covered in the first part, the recurrent deep neural network (RNN) in the second part and the convolutional neural network (CNN) in the third part.

A. *Two-layer perceptron*

Perceptron is the basis of neural network and support vector machine. In 2018, *Can Li* [5] partitions 128×64 one-transistor one-memristor crossbar array (1T1R) to construct a perceptron with two fully connected layers. The multiple-layer perceptron is a simple neural network constructed by fully connected layers. The network contains 64 input neurons, 54 hidden neurons and 10 output neurons.

With self-adaptive in-situ learning, figure 2 presents the accuracy curve of both experiment and detect-free simulation. The classification accuracy reaches 91.71% on 10,000 test samples. The gap between the experiment with 11% devices has hardware imperfections and software simulation is approximately 2%-4%. Meanwhile, with simulation, the accuracy can reach 97% for a larger (1024 × 512) memristor array.

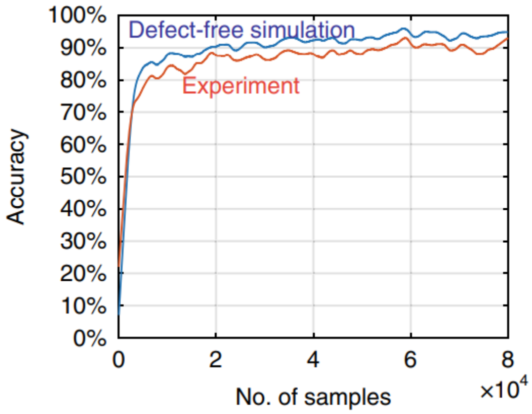

Fig. 2. Minibatch Classification accuracy of experiment and detect-free simulation during online training [5, pp. 5].

The results show high accuracy in solving classification problems with a simple neural network. The self-adapting capability of hardware detection and size scalability gives the technique of memristor crossbar array a promising future of full hardware-implemented high-speed computing.

B. Recurrent deep neural networks with long short-term memory (LSTM) units

LSTM model is one type of recurrent neural network (RNN) that avoids log-term dependency problems and thus performs well in the prediction problems of time series, such as speech and video. However, the complex structure of LSTM networks leads to the speed limitation of computing on conventional hardware, as a great number of parameters need to be transferred between separate chips.

The memristor approach implemented in-situ LSTM analog matrix units with 128×64 1T1R [6] in 2019. The parameters are saved as the memristor conductance and dynamically adjusted to gain closer outputs during training.

Figure 3 indicates the classification accuracy of the testing set during 50 epochs. Grey lines are the 50-time repeated results of simulations that show 1.4% random conductance update error; the blue line is the experimental result that reaches a maximum of 79.1% accuracy in a certain classification problem of human gait identification.

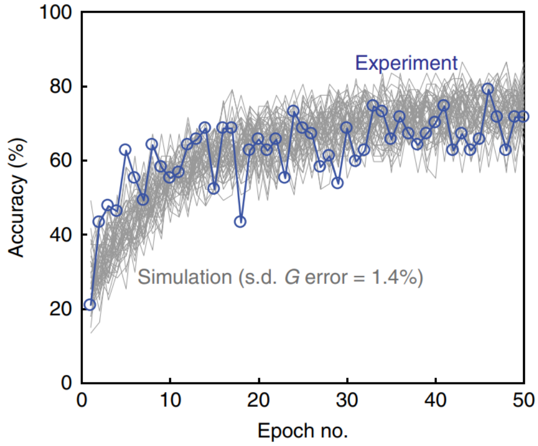

Fig. 3. The tendency of classification accuracy on testing set during 50 epochs [6, pp. 6].

The results show the efficiency of in-situ training on the memristor crossbar arrays for constructing LSTM units. Based on the fact of low power consumption and inference latency, the experiment presents the potentiality to integrate different functional arrays in one chip and achieve high-speed neuromorphic computing.

C. Five-layer convolutional neural network

CNN is the most extensively used neural network since it shows out-performed accuracy and computation speed when dealing with image signal problems. However, a huge number of convolutional arithmetic operations are involved in CNN. Thus, ultra-fast parallel multiply-accumulate hash rate and high-power consumption are in great demand for devices running CNN-related applications.

A five-layer CNN is first built with 2,048 memristor arrays to process an image recognition problem on a handwritten-digit database [7] in 2020. To avoid the possible non-ideal features of the hardware, a hybrid method is proposed by adjusting the last fully connected layer. The network is first trained ex-situ and gained the weights. Then the weights are transferred to the arrays and used to train the last FC network in situ.

Figure 4 shows the error rate of with and without hybrid training. In each group, with hybrid training, the error rate decreases 1.38%, 1.74% and 2.34% respectively and thus gets closer to the reference error rate (2.01%) trained ex-situ in Python with TensorFlow.

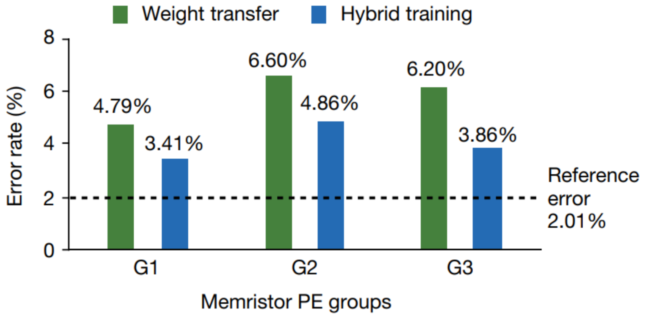

Fig. 4. The error rate of direct weight transfer and hybrid training method in three different PE groups on the test set [7, pp. 5].

The implemented structure presents superior hardware performance. With calculations of the detailed metrics, the energy efficiency is 11,014 GOP $s^{-1}W^{-1}$ and the performance density is 1,164 GOP $s^{-1}mm^{-2}$. Compared with Tesla V100 graphics processing units (GPU), this method is 110 times better in energy efficiency and 30 times better in performance density.

In summary, the memristor-based CNN computing system trades off training accuracy and energy cost. It achieves high accuracy of 96 percent in a certain image recognition problem with much lower energy consumption than GPUs.

IV. Discussions on Existing Problems

The neural network realized by using memristors shows low-power, small-size characteristics. In recent research [5-7], the memristor-based structure presents high-speed computing capability with acceptable accuracy currently. However, training with the memristor-based architectures mainly focuses on small in-situ units, thus achieving more complicated operations is still need the help of ex-situ methods. Based on this fact, there are some problems that need to be mentioned when the scale of neural networks keeps increasing.

The process of scaling down in memristor fabrication usually involves electron beam lithography [14]. As the scale of the memristor is down, then the cross-section area of the electrode decrease, which leads to an increase in the resistance. Normally the electrode is directly attached to the memristor in series so that the effective resistance R_{ON} and R_{OFF} in equation (1) will have a relatively large offset, which leads to a non-negligible offset in memristance. What's more crucial, the device offset can be more difficult to eliminate in large-scale integration. This in turn makes the problem critical.

The accuracy curve of defect-free simulation and experiment [5] in Section III has already indicated the hardware implementation could not achieve the same performance as software simulation. Though variable methods and architectures are proposed by researchers to get closed to the ideal accuracy, potential factors such as offsets and defects still bring negative effects.

In addition, the hardware integration of memristor-made synapses in neural networking will suffer more from process variability, thus, the properties of each hardware-implemented neuron can vary greatly [14]. Furthermore, the resolution of memristance states is limited compared to that of resistance and capacitance. And due to the variability that integration has induced, it is still unknown what is the correct trade-off between the size of the memristor and desired resolution.

V. Conclusions

The aim of this article was to assess the implementation of memristor crossbar array-based neuromorphic computing architectures. The characteristics of memristor, low leakage, conducive scalability, low power consumption and biomimetic function, show its potentiality to simulate artificial synapses and construct large-scale networks.

The latest applications not only show the trend in the construction development of neuromorphic computing architectures but also reveal the possibility to further practical high-speed computing applications. Despite the construction of the neural network is gradually getting more complex, the training results are still sustaining superior performance with the evaluation of accuracy. Extensive experiments present the potentiality of the memristor array in neuromorphic computing with its unique characteristics of high speed and low energy consumption.

With the fast-increasing demand for high-speed computing and large-scale network construction, conventional computing architectures are facing limitations. In the future, constructing large-scale high-speed computing integrated architectures with memristors is one possible method to break the 'von Neumann bottleneck'.

However, the complicated construction is limited by device offset and process variability currently. Thus, how to avoid the negative factors that may affect large-scale integration becomes a critical problem needed to be deeply researched.

References

[1] D. Monroe, "Neuromorphic computing gets ready for the (really) big time," *Communications of the ACM*, vol. 57 no. 6, pp. 13–15, 2014. doi:10.1145/2601069.

[2] L. Chua, "Memristor-the missing circuit element," *IEEE Transactions on circuit theory,* vol. 18, no. 5 pp. 507-519, 1971.

[3] D. Strukov, G. Snider, D. Stewart, et al., "The missing memristor found," *Nature*, vol. 453, pp. 80–83, 2008.

[4] Q. Xia and J. Joshua Yang. "Memristive crossbar arrays for brain-inspired computing," *Nature materials*, vol. 18, no. 4, pp. 309-323, 2019.

[5] Li, Can, et al. "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks." *Nature communications*, vol.9, no.1, pp. 1-8, 2018.

[6] C. Li, Z. Wang, M. Rao, et al., "Long short-term memory networks in memristor crossbar arrays," *Nature Machine Intelligence*, vol.1, no. 1, pp. 49-57, 2019.

[7] P. Yao, H. Wu, B. Gao, et al., "Fully hardware-implemented memristor convolutional neural network," *Nature,* vol. 577, no.7792, pp. 641-646, 2020.

[8] J. Blanc and D. Staebler, "Electrocoloration in SrTiO - vacancy drift and oxidation reduction of transition metals," *Phys. Rev.* vol. B no. 4, pp. 3548–3557, 1971.

[9] S. Gao, F. Zeng, M. Wang, et al., "Implementation of Complete Boolean Logic Functions in Single Complementary Resistive Switch," *Sci. Rep.* vol. 5, no. 15467, 2015.

[10] E. Linn, R. Rosezin, C. Kügeler, et al., "Complementary Resistive Switches for Passive Nanocrossbar Memories," *Nat. Mater.* vol. 9, no. 5, pp. 403-406, 2010.

[11] S. Kim, J. Kim, S, Choi, et al., "Direct Observation of Conducting Nanofilaments in Graghene-Oxide-Resistive Switching Memory," *Adv. Funct. Mater.,* vol. 25, no. 43, pp. 6710-6715, 2015.

[12] V. Sangwan, D. Jariwala, I. Kim, et al., "Gate-Tunable Memristive Phenomena Mediated by Grain Boundaries in Single-Layer MoS2," *Nat. Nanotechnol.*, vol. 10, no. 5, pp. 403-406, 2015.

[13] M. Wang, S. Cai, C. Pan, et al., "Robust Memristors Based on Layered Two-Dimensional Materials," *Nat. Electron.*, vol. 1, no. 2, pp. 130-136, 2018.

[14] Q. Xia, "Nanoscale resistive switches: devices, fabrication and integration," *Applied Physics*, vol. 102, no. 4, pp. 2909-2914, 2010.