

One memristor-one electrolyte-gated transistor-based high energy-efficient dropout neuronal units

Yalin Li(李亚霖), Kailu Shi(时凯璐), Yixin Zhu(朱一新), Xiao Fang(方晓), Hangyuan Cui(崔航源), Qing Wan(万青), and Changjin Wan(万昌锦)

Citation: Chin. Phys. B, 2024, 33 (**6**): 068401. DOI: 10.1088/1674-1056/ad39d6
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Memristor's characteristics: From non-ideal to ideal

Fan Sun(孙帆), Jing Su(粟静), Jie Li(李杰), Shukai Duan(段书凯), and Xiaofang Hu(胡小方)
Chin. Phys. B, 2023, 32 (**2**): 028401. DOI: 10.1088/1674-1056/ac7548

A review on the design of ternary logic circuits

Xiao-Yuan Wang(王晓媛), Chuan-Tao Dong(董传涛), Zhi-Ru Wu(吴志茹), and Zhi-Qun Cheng(程知群)
Chin. Phys. B, 2021, 30 (**12**): 128402. DOI: 10.1088/1674-1056/ac248b

Artificial synaptic behavior of the SBT-memristor

Gang Dou(窦刚), Ming-Long Dou(窦明龙), Ren-Yuan Liu(刘任远), and Mei Guo(郭梅)
Chin. Phys. B, 2021, 30 (**7**): 078401. DOI: 10.1088/1674-1056/ac05a3

SBT-memristor-based crossbar memory circuit

Mei Guo(郭梅), Ren-Yuan Liu(刘任远), Ming-Long Dou(窦明龙), and Gang Dou(窦刚)
Chin. Phys. B, 2021, 30 (**6**): 068402. DOI: 10.1088/1674-1056/abd7dc

Suppression of ferroresonance using passive memristor emulator

S Poornima
Chin. Phys. B, 2021, 30 (**6**): 068401. DOI: 10.1088/1674-1056/abd7e0

One memristor–one electrolyte-gated transistor-based high energy-efficient dropout neuronal units

Yalin Li(李亚霖)[1,†], Kailu Shi(时凯璐)[1,†], Yixin Zhu(朱一新)[1], Xiao Fang(方晓)[1], Hangyuan Cui(崔航源)[1], Qing Wan(万青)[2,‡], and Changjin Wan(万昌锦)[1,§]

[1] *School of Electronic Science & Engineering, and Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing* 210023, *China*

[2] *Yongjiang Laboratory (Y-LAB), Ningbo* 315202, *China*

(Received 8 March 2024; revised manuscript received 19 March 2024; accepted manuscript online 3 April 2024)

Artificial neural networks (ANN) have been extensively researched due to their significant energy-saving benefits. Hardware implementations of ANN with dropout function would be able to avoid the overfitting problem. This letter reports a dropout neuronal unit (1R1T-DNU) based on one memristor–one electrolyte-gated transistor with an ultralow energy consumption of 25 pJ/spike. A dropout neural network is constructed based on such a device and has been verified by MNIST dataset, demonstrating high recognition accuracies ($>90\%$) within a large range of dropout probabilities up to 40%. The running time can be reduced by increasing dropout probability without a significant loss in accuracy. Our results indicate the great potential of introducing such 1R1T-DNUs in full-hardware neural networks to enhance energy efficiency and to solve the overfitting problem.

Keywords: dropout neuronal unit, synaptic transistors, memristor, artificial neural network

PACS: 84.35.+i, 73.40.Mr, 77.55.–g, 81.05.Gc **DOI:** 10.1088/1674-1056/ad39d6

1. Introduction

Recent years have witnessed rapid development on hardware implementations of artificial neural network (ANN), which is thought to offer ultrahigh energy efficiency with massive throughput of data.[1,2] One fundamental building block for a hardware-based ANN is a synaptic device with updatable weights. A broad spectrum of devices such as memristors, electrolyte-gated transistors, electrochemical cells, and spintronics, have been proposed with multilevel characteristics and synapse-like properties.[3–11] The multilevel characteristics can be exploited for implementing *in situ* vector–matrix multiplication (VMM) operations which are dominant the computational load for artificial intelligence (AI) algorithms.[12] Some of these devices have been demonstrated with extreme high energy-efficiency (e.g., larger than tera operations per second per watt, TOPS/W), revealing the great potential for building an energy-efficient platform for machine learning.[13]

Many efforts have been paid on improving the precision of the synaptic devices, and the linearity and symmetry of the weight updating properties, which can further improve the performance of the hardware-based ANN. Furthermore, the hardware-based ANN can also be optimized by introducing advanced algorithms to reduce the computation and time costs. For example, a dropout neuronal unit (DNU) was proposed for realizing the dropout function which regularizes the model parameters by randomly dropping the hidden nodes of the neural network during the training process to avoid co-adaptation of these nodes.[14] An Ag/Ta_2O_5:Ag/Pt memristor was proposed with dropout functions based on its stochastic and volatile switching characteristics.[15] The ANN built with such DNU and a WO_3-based synapse remained a high accuracy of $>85\%$ with a dropout probability of 50%. During the training process, the memristor-based DNU is randomly switched between activated and dropout states. However, the activated state is involved in the formation of conductive pathway through the device, inevitably augmenting the energy consumption, and the memristor-based DNUs have been reported with an energy consumption of more than nanoJoule.

Here, we report a one memristor–one electrolyte-gated transistor-based dropout neuronal unit (1R1T-DNU) with a minimized energy consumption of 25 picoJoule per spike. The dropout function is achieved based on the stochastic dynamics of an Ag/IGZO/Al_2O_3/ITO memristor, which is capacitively coupled to the channel of an electrolyte-gated IGZO transistor. Such capacitively coupling would induce a current response through the IGZO channel, which is much lower than that of a memristor-based DNU. A dropout neural network is constructed based on such 1R1T-DNUs, which is verified by the MNIST dataset. High recognition accuracies of over 90% can be achieved even at a large range of dropout probabilities up to 40%. In addition, the accuracy can be approximately

[†]These authors contributed equally to this work.

[‡]Corresponding author. E-mail: wanqing@nju.edu.cn

[§]Corresponding author. E-mail: cjwan@nju.edu.cn

© 2024 Chinese Physical Society and IOP Publishing Ltd

maintained with a decreased running time. Our design offers a possibility for a high energy-efficient full-hardware neural networks with minimal effect of overfitting.

2. Experimental details

The structural diagram of the 1R1T is shown in Fig. 1(a), presented in the form of a bottom gate-integrated memristor. The SiO_2 films of electrolyte-gated transistors were deposited on ITO glass by plasma-enhanced chemical vapor deposition (PECVD) using 100-sccm SiH_4/N_2 mixture (5% SiH_4/95% N_2) and 450-sccm N_2O as reaction gases. The deposition time and the deposition temperature are 450 s and 100 °C, respectively. The 30-nm IGZO films were sputtered using magnetron sputtering in a pure argon atmosphere with a vacuum of 0.3 Pa. Al source and drain electrodes were deposited by thermal evaporation through a shadow mask. The channel width and length were 1000 μm and 80 μm, respectively. During the process of manufacturing transistors, a customized nickel mask is used to mask the other half of the substrate. Once the transistor is complete, it is then protected with tin foil.

The 5-nm Al_2O_3 films were deposited on the other half of the clean ITO glass by atomic layer deposition (ALD) at 200 °C using trimethylaluminum and water as reaction materials. The 150-nm IGZO films were sputtered using magnetron sputtering at an argon to oxygen ratio of 25:15 with a sputtering vacuum of 0.3 Pa. Finally, 120-μm diameter Ag top electrodes were deposited by thermal evaporation through a shadow mask. The electrical characteristics of the memristors and transistors as well as the circuits of the dropout neuronal unit constructed from them were investigated by the platform design automation (PDA) FS380 semiconductor parameter analyzer, the semiconductor parameter analyzer (Keithley 2636B) and the fiber coupled laser module (Changchun New Industry (CNI) Laser PGL-FC-405 nm) at room temperature with relative humidity of 50%.

3. Results and discussion

The *I*–*V* characterization of the Ag/IGZO/Al_2O_3/ITO device with ten repeated scans is shown in Fig. 1(b). The resistance of the memristor is tuned from its high resistance state (HRS) to low resistance state (LRS) with the applied voltage increases to its threshold voltage (V_{TH}). When the applied voltage decreases to its holding voltage (V_{HOLD}), the memristor spontaneously decays to the HRS, indicating the threshold switching property. Stochastic threshold switching behavior was obtained in the device, with the threshold voltage fluctuating around 1.0 V, suggesting that the threshold switching behavior of the memristor exhibits some degree of randomness, even when the same voltage pulse is applied repeatedly. Switching mechanism of the Ag-based threshold switching memristor can be explained by the formation of the Ag filaments between the top and bottom electrodes, driven by the electric field, and its rupture when self-diffuses to the minimum dielectric interface energy.[16,17] The thin and dense Al_2O_3 layer plays a crucial role on achieving a sharp switching from HRS to LRS, by hindering the stochastic switching under a low electric field.

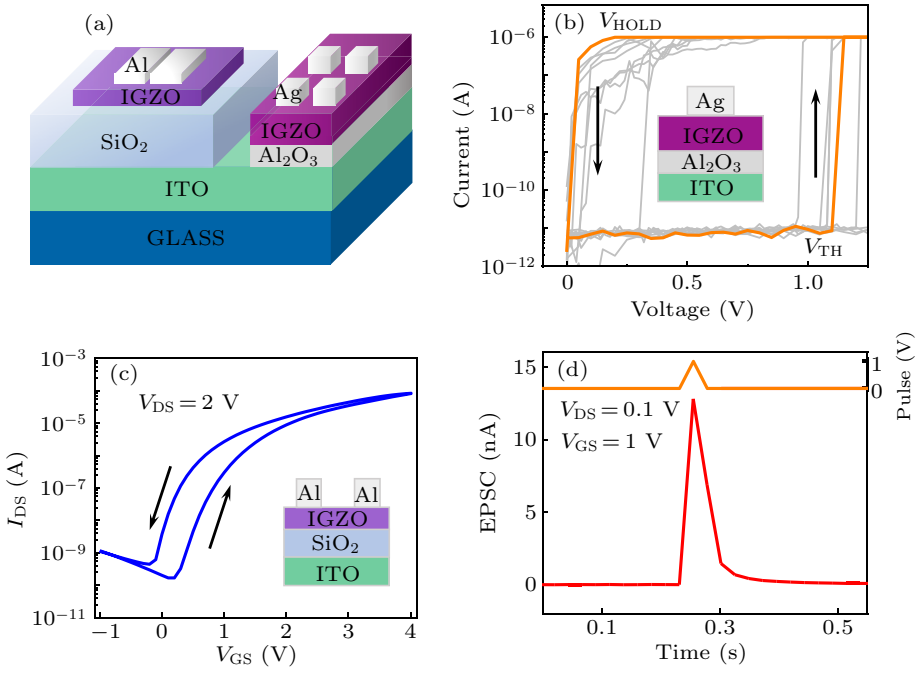


Fig. 1. (a) Schematic diagram of the device structure. (b) The *I*–*V* curve of the IGZO-based memristor in ten sweep loops. Inset: the cross-section schematic diagram of the memristor. (c) The transfer characterization of the transistor. Inset: the cross-section schematic diagram of the transistor. (d) A typical EPSC triggered by a gate voltage pulse (1.0 V, 25 ms) with $V_{DS} = 0.1$ V.

The transfer characterization of the electrolyte-gated IGZO transistor is shown in Fig. 1(c). An anticlockwise hysteresis window can be observed, which can be explained by the electrolyte double layer effect.[18,19] The SiO_2 thin film, prepared using low-temperature and low-vacuum technology with a large amount of reaction gas introduced in a short period of time, is sparse and porous. This makes it susceptible to absorbing moisture from the air. When subjected to an electric field, it ionizes ions and exhibits the characteristics of an electrolyte film. Under the influence of an electric field, ions migrate at a slower rate than electrons. This results in a significant hysteresis window in the transfer curve during scanning. A typical excitatory postsynaptic current (EPSC) response by a gate voltage pulse of 1.0 V (width = 25 ms) with $V_{DS} = 0.1$ V is shown in Fig. 1(d). The EPSC increases rapidly from 9.77 pA to 12.8 nA during the gate pulse and gradually decays back to the initial current after the pulse.

The EPSC can be accurately modeled using ion-related electric-double-layer modulation and a stretched-exponential decay function. When a voltage pulse is applied on the gate electrode, protons are attracted to the channel, resulting in an increase in channel current through electrostatic coupling. This process is equivalent to a charge–discharge process. Subsequently, the ions spontaneously diffuse to the equilibrium state at the end of the pulse.[20]

The operating mechanism is schematically illustrated in Fig. 2(a). As shown in Fig. 2(a), the increase of voltage (V_M) applied on the top electrode of the memristor would induce the formation of Ag filaments between the top and bottom electrodes. Then, an abrupt switching from the HRS to the LRS would result in an abrupt increase of voltage on the gate terminal (V_G) of the electrolytic-gated transistor component in the 1R1T, which in turn causes a surge in the channel current (I_{DS}, shown by the blue line).[21] Figure 2(b) shows the transfer characterization of the 1R1T. A clear current surge can be observed at a V_M close to 2.0 V, which is consistent with our proposed mechanism. A low subthreshold slope of 85 mV/dec can be achieved at this region. The voltage pulses of 1.0 V (width = 25 ms) are applied to the 1R1T with $V_{DS} = 0.1$ V. As shown in Fig. 2(c), the EPSC can be either activated (red line) and inactivated (blue line), which is randomly occurred by applying the same voltage pulses. The memristor can be either switched from HRS to LRS or not, with a certain probability which is dependent on the amplitude of the voltage pulse (V_M). In this case, if the memristor can be switched from HRS to LRS, the memristor is functioned as a conducting wire, and the voltage pulse could effectively trigger an EPSC. On the other hand, if the memristor cannot be switched from HRS to LRS, the memristor is functioned as a resistor with a large resistance, and the voltage pulse could not effectively trigger an EPSC. In other words, such random EPSC responses are originated from the stochastic switching of the memristor.

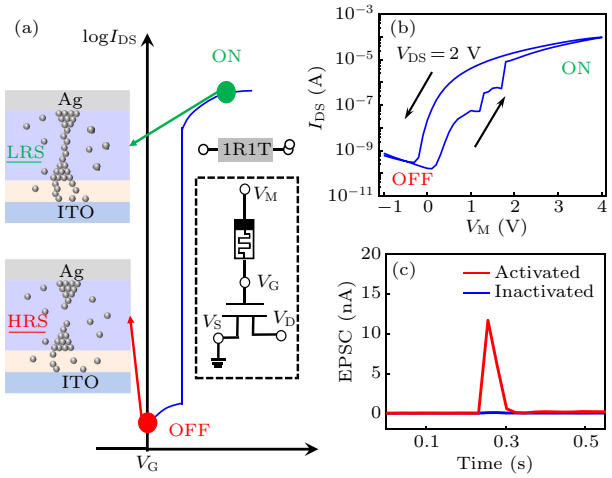


Fig. 2. (a) Operation principle of the proposed 1R1T. Inset: the equivalent circuit of the 1R1T. (b) The transfer characterization of the 1R1T. (c) Different EPSCs of the 1R1T.

The EPSCs of the 1R1T in response to varied V_M are shown in Fig. 3(a), amplitudes of V_M were set from 0.2 V to 2.0 V in 0.2-V steps and $V_{DS} = 0.1$ V. The test was repeated 25 times under different V_M. As shown in the inset, the EPSCs can be either triggered by the V_M of 1.2 V (noted as: Activated) or inactivated (noted as: Inactivated). If the EPSC value exceeds 1 nA, it is defined activated; otherwise, it is inactivated. The current values for activated and inactive states differ by more than 10 times. The activated probability (P_A) is defined as

$$P_A = \frac{n_{\text{Activated}}}{n_{\text{Total}}},$$

where $n_{\text{Activated}}$ are the numbers of the activated EPSC, n_{Total} are the numbers of total EPSCs.

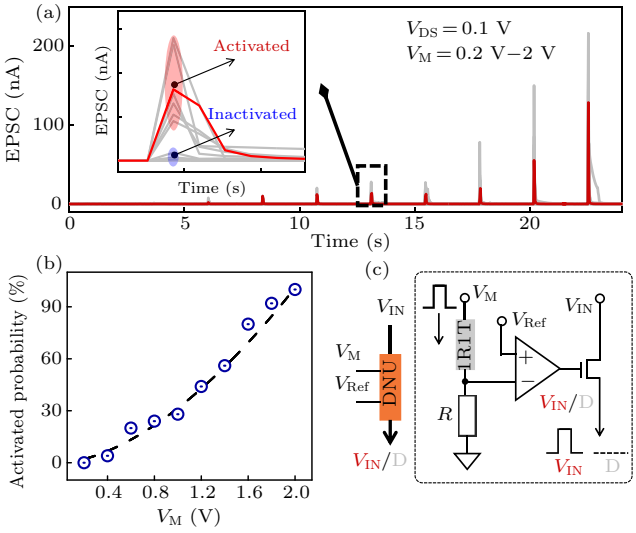


Fig. 3. (a) The EPSCs of the 1R1T in response to varied V_M. Inset: the EPSCs partially triggered by the V_M of 1.2 V. (b) The activated probability of the 1R1T. (c) Hardware implementation of the 1R1T-DNU.

The activated probability of the 1R1T can be tuned by V_M, as shown in Fig. 3(b). The P_A increases from ≈ 0% to ≈ 100%, when the V_M increases from 0.2 V to 2 V, as the

larger $V_{\rm M}$ is more prone to induce an HRS to LRS transition of memristor.[22] The minimal activated power consumption of the 1R1T is estimated to 25 pJ per pulse at the $V_{\rm M} = 0.4$ V by considering both the leakage current and the channel current. The average power consumption of a single pulse ($P_{\rm Pulse}$) is calculated using the following formula:

$$P_{\rm pulse} = \frac{\sum_{j=1}^{n_{\rm Activated}} (V_{\rm DSj} \cdot I_{\rm DSj} + V_{{\rm M}j} \cdot I_{{\rm GS}j}) \cdot T_{\rm Pulse})}{n_{\rm Activated}} \times \{V_{\rm M} \in [0\ {\rm V}, 2\ {\rm V}]\},$$

where $n_{\rm Activated}$, $V_{{\rm DS}j}$, $I_{{\rm DS}j}$, $V_{{\rm M}j}$, $I_{{\rm GS}j}$, and $T_{\rm Pulse}$ are the numbers of activated EPSCs, the source–drain voltage, source–drain current, voltage applied on the top electrode of the memristor, leakage current of transistor, and pulse width of the $V_{\rm M}$ when EPSCs are activated, respectively. A DNU based on such activated probability tunable 1R1T devices was shown in Fig. 3(c). The 1R1T device is connected to a resistor (R), and the voltage over R is set as the inverting input of the comparator. The reference voltage ($V_{\rm Ref}$) for the noninverting input is calculated as 1 nA multiplied by R. The switch connected to the output end of the comparator determines whether to dropout $V_{\rm IN}$. For instance, if the $V_{\rm M}$ is 1.2 V, and the EPSC value of 1R1T is greater than 1 nA, the voltage over R exceeds $V_{\rm Ref}$, resulting in a low output voltage from the comparator. This voltage is insufficient to activate the switch, and the $V_{\rm IN}$ is consequently dropout (noted as: D). Otherwise, the switch is turned on and $V_{\rm IN}$ can transmit to the next stage as usual (noted as: $V_{\rm IN}$). As shown by the equivalent circuit, the input voltage ($V_{\rm IN}$) can transmit to the output through a transistor which switches at a certain level of probability, and such a switching probability is dependent on the value of $V_{\rm M}$. This switching probability is equivalent to the previously mentioned $P_{\rm A}$, which is defined as the dropout probability ($P_{\rm D}$).

A three-layer ($784 \times 256 \times 10$) dropout neural network was constructed based on such 1R1T-DNU for classifying MNIST handwritten digits, as shown in Fig. 4(a). The DNUs at the first layer can be activated or dropout (labeled by "×"), and the activated probability is dependent on $V_{\rm M}$.[23] The connected synapses between each layer were constructed based on our previous IGZO nanofiber photoelectric neuromorphic transistors.[24] The updating of conductance (served as synaptic weight) based on such transistors as shown in Fig. 4(b), the structure of the transistor is shown in the inset. The device exhibits a precision of 5 bits (32 states) through excitatory stimulation (light pulse, denoted by '+') and inhibitory stimulation (electrical pulse, denoted by '−'). A readout circuits of one full connected layer is shown in Fig. 4(c). The recognition task was implemented through simulation,[25] and the recognition accuracies and running times under different $P_{\rm D}$ were calculated as shown in Fig. 4(d).

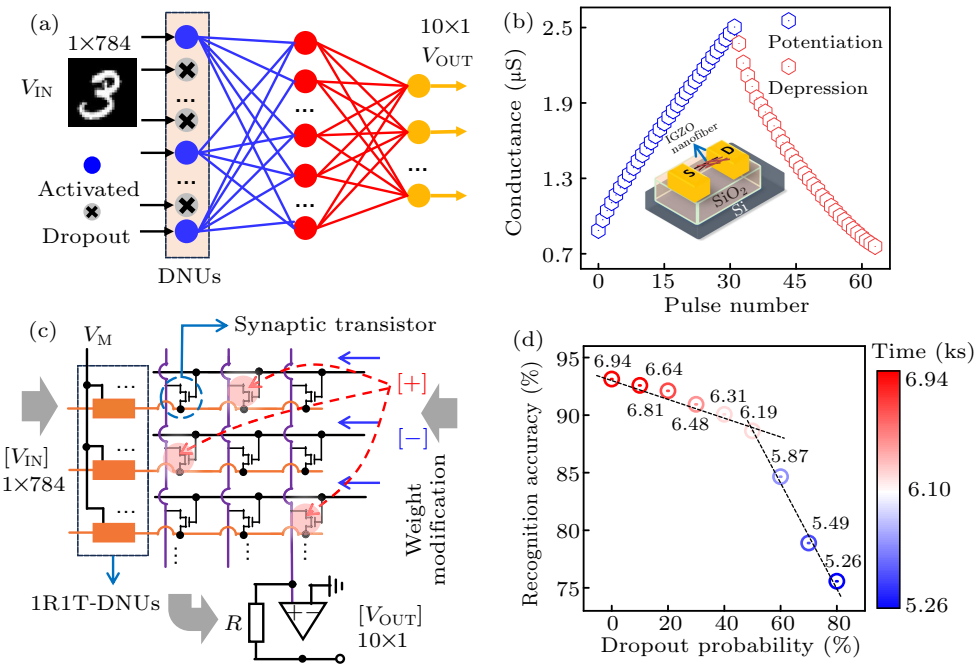


Fig. 4. (a) Schematic illustration of the dropout neural network. (b) Conductance updating by light and electric pulses for potentiation and depression, respectively. (c) The readout circuit of one fully-connected layer introducing the 1R1T-DNUs. (d) Recognition accuracies and running times with varied dropout probabilities.

The simulation uses the gradient descent algorithm with analog weight update based on the conductivity parameters of our synaptic devices. The 784 inputs correspond to 28×28 grayscale pixels of an MNIST digit can be randomly dropout with a certain probability. The output reveals the classification result of the hand-written digits from 0 to 9. During each epoch, the ANN was trained on a randomly selected set of 1×10^4 patterns from a training data set of 6×10^4 images.

The recognition accuracy was then tested on a separate set of 1×10^4 images from the testing data. The pattern recognition accuracies under various dropout probabilities (P_D) within 50 training epochs are shown in Fig. 4(d). The significant decrease in recognition rate, as depicted in Fig. 4(d), can be attributed to a high dropout probability of 50%. This elevated dropout probability induces over-regularization, where a significant number of feature values are randomly eliminated during each training iteration. Consequently, the neural network is unable to extract sufficient information from the training data, leading to a significant decrease in recognition accuracy. The recognition accuracy decays slowly to $\sim 90\%$ with P_D from 0 to 40%, and decays dramatically with a larger P_D. In addition, the running time can be reduced from 6.94 ks to 5.26 ks with the P_D increases from 0 to 80%. As a certain level of dropout probability can avoid the overfitting problem, our results indicate the feasibility of such dropout neural network which can maintain a high recognition accuracy ($> 90\%$) at a wide range of dropout probabilities ($\geq 40\%$).

4. Conclusion and perspectives

In summary, one memristor–one electrolyte-gated transistor-based dropout neuronal unit was built with a low activated power consumption of 25 pJ/spike. A dropout neural network was constructed based on such 1R1T-DNUs, which is capable of classifying handwritten digit with an accuracy of over 90% at a large range of dropout probability up to 40%. What is more, the high accuracy can be approximately maintained with a reduced running time by introducing the such DNUs. These results reveal the great potential of such 1R1T-DNU for constructing full-hardware artificial neural networks with high energy-efficiency and the ability to solve the overfitting problem.

Acknowledgements

Project supported by the National Key Research and Development Program of China (Grant Nos. 2021YFA1202600 and 2023YFE0208600) and in part by the National Natural Science Foundation of China (Grant Nos. 62174082, 92364106, 61921005, 92364204, and 62074075).

References

[1] López C 2023 *Advanced Materials* **35** 2208683
[2] Prezioso M, Merrikh-Bayat F, Hoskins B D, Adam G C, Likharev K K and Strukov D B 2015 *Nature* **521** 61
[3] Ahn E C 2020 *npj 2D Materials and Applications* **4** 17
[4] Sun K, Chen J and Yan X 2021 *Adv. Funct. Mater.* **31** 2006773
[5] Zhu L Q, Wan C J, Guo L Q, Shi Y and Wan Q 2014 *Nat. Commun.* **5** 3158
[6] Zhu Y, He Y, Chen C, Zhu L, Mao H, Zhu Y, Wang X, Yang Y, Wan C and Wan Q 2022 *Appl. Phys. Lett.* **120** 113504
[7] Zhu Y, Peng B, Zhu L, Chen C, Wang X, Mao H, Zhu Y, Fu C, Ke S, Wan C and Wan Q 2022 *Appl. Phys. Lett.* **121** 133502
[8] Lin X, Long H, Ke S, Wang Y, Zhu Y, Chen C, Wan C and Wan Q 2022 *Chin. Phys. Lett.* **39** 068501
[9] Wu X, Zhang Y, Zhang T and Zhang B 2023 *Chin. Phys. B* **32** 100506
[10] Xiang T, Chen F, Li X, Wang X, Yan Y and Wang L 2023 *Chin. Phys. B* **32** 117301
[11] Yang Y, Fu C, Ke S, Cui H, Fang X, Wan C and Wan Q 2023 *Chin. Phys. B* **32** 118101
[12] Amirsoleimani A, Alibart F, Yon V, Xu J, Pazhouhandeh M R, Ecoffey S, Beilliard Y, Genov R and Drouin D 2020 *Advanced Intelligent Systems* **2** 2000115
[13] Demasius K U, Kirschen A and Parkin S 2021 *Nat. Electron.* **4** 748
[14] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 *Journal of Machine Learning Research* **15** 1929
[15] Huang H M, Xiao Y, Yang R, Yu Y T, He H K, Wang Z and Guo X 2020 *Adv. Sci.* **7** 2001842
[16] Wang Z, Joshi S, Savel'ev S E, Jiang H, Midya R, Lin P, Hu M, Ge N, Strachan J P, Li Z, Wu Q, Barnell M, Li G L, Xin H L, Williams R S, Xia Q and Yang J J 2017 *Nat. Mater.* **16** 101
[17] Wang Z, Wu H, Burr G W, Hwang C S, Wang K L, Xia Q and Yang J J 2020 *Nat. Rev. Mater.* **5** 173
[18] Wan C, Zhou J, Shi Y and Wan Q 2014 *IEEE Electron Dev. Lett.* **35** 414
[19] Wan X, Feng P, Wu G D, Shi Y and Wan Q 2015 *IEEE Electron Dev. Lett.* **36** 204
[20] Shao F, Feng P, Wan C, Wan X, Yang Y, Shi Y and Wan Q 2017 *Adv. Electron. Mater.* **3** 1600509
[21] Huang Q, Huang R, Pan Y, Tan S and Wang Y 2014 *IEEE Electron Dev. Lett.* **35** 877
[22] Mazady A and Anwar M 2014 *IEEE Trans. Electron Dev.* **61** 1054
[23] Sanjar K, Rehman A, Paul A and JeongHong K 2020 8*th International Conference on Orange Technology (ICOT)*, 18–21 December, 2020, pp. 1–4
[24] Zhu Y, Peng B, Zhu L, Chen C, Wang X, Mao H, Zhu Y, Fu C, Ke S, Wan C and Wan Q 2022 *Appl. Phys. Lett.* **121** 133502
[25] Mendenhall J and Meiler J 2016 *Journal of Computer-Aided Molecular Design* **30** 177