Complementary Graphene-Ferroelectric Transistors (C-GFTs) as Synapses with Modulatable Plasticity for Supervised Learning

Yue Zhou[1], Nuo Xu[2], Bin Gao[3], Yangyang Chen[1], Boyi Dong[1], Yi Li[1], Yuhui He[1] and Xiang Shui Miao[1]
[1]Wuhan National Laboratory for Optoelectronics, School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China
[2]Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA
[3]Institute of Microelectronics, Tsinghua University, Beijing, 100084 China
Email: heyuhui@hust.edu.cn miaoxs@hust.edu.cn

***Abstract*—**Novel complementary graphene-ferroelectric transistors (C-GFTs) based synapses are proposed and experimentally demonstrated for the first time. By exploiting the unique zero-bandgap property of graphene, GFT based synapses can be dynamically reconfigured between potentiative and depressive (PD) modes corresponding to hole- and electron dominated transport in graphene channels, respectively. Both modes demonstrate excellent linearity, small (2%) cycle-to-cycle variation and > 32 levels when used as synapses. By configurating the PD modes into a pair of C-GFTs, the hardware architecture of spiking neural networks (SNNs) can be substantially innovated, where the complicated circuitry previously required for supervised learning is now completely removed. With C-GFTs, a synapse footprint of 100 μm^2 and a power consumption of 8 pJ/per operation are demonstrated in the MNIST learning task.

I. Introduction

Remote Supervised Method (ReSuMe) is a well-accepted supervise learning (SL) algorithm for spiking neural networks (SNNs) for which a hardware implementation based on dual bipolar-RRAMs was reported with significantly enhanced performance on the MNIST benchmark test [1]. However, non-volatile transistor structures (as multi-terminal synaptic components) are more promising since they emulate the biological tripartite synapses more accurately (Fig.1), enabling concurrent learning [2]. In this work, the complementary graphene-ferroelectric transistors (C-GFTs) are proposed as synapses to implement ReSuMe. A pair of GFTs are pre-arranged through voltage-controlled ferroelectric polarization responding to the same excitation pulses in a complementary way (*i.e.* spike-timing dependent plasticity, STDP as well as anti-STDP). The primary advantage of using C-GFTs is that those complicated supervise circuits in the conventional SNNs based on single transistors (Fig.2) are now eliminated. Furthermore, the circuit complexity of pre- and post-neurons and those selection components in the dual-RRAM approach [1] are substantially simplified or even removed in our C-GFT designs (Fig.3). Excellent linearity, small cycle-to-cycle (c2c) variation, ultralow power and area consumptions are also demonstrated quantitatively in C-GFT based synapses.

II. Device Fabrication

Fig.4 shows schematically the structure of GFT, where a monolayer of graphene functions as the channel and organic P(VDF-TrFE) ferroelectric polymer as the gate dielectric. The top view of the fabricated device is shown in Fig. 5(a) with each unit characterized. The flowchart of the fabrication process is shown in Fig.5(b). Single layer graphene was wet transferred onto the pre-defined source and drain electrodes as the channel of a transistor. Next, PVDF was spin-coated on top of graphene followed by an annealing of 403 K. Finally, aluminum electrode was fabricated as the top gate. Note that the whole process was performed at temperature compatible with CMOS back-end-of-line (BEOL) process.

III. Results and Discussion

A. Electrical Characteristics of GFTs

The polarization versus electric field curves with different peak voltages are shown in Fig. 6 to demonstrate the ferroelectric properties of PVDF. The transfer characteristics G_{ds}-V_{gs} with several loops are measured as shown in Fig.7. In these curves, the left and right branches separated by the *Dirac* points represent hole- and electron-dominated transport, respectively, while the locations of *Dirac* points at positive or negative voltage denote the hole or electron doping (p- or n-type) of graphene. However, unlike the conventional graphene-channel transistors, here a pronounced hysteresis appears during the back-and-forth sweeping. The physical mechanisms are that the ferroelectric response is composed by a linear (paraelectric-, proportional to applied electric field) and a hysteretic (ferroelectric-, results in residual polarization even when the external field is removed) part. The latter results in non-volatile tuning of carrier types and densities as well as a shift of the graphene *Dirac* point. Two sets of hole and electron conduction branches during sweeping co-exist as shown in Fig.9. The initial conductance states of GFT (p-type in (a) and (b) while n-type in (c)) can be dynamically reconfigured through tuning the polarization in ferroelectric layer, when negative/positive gate voltage with amplitudes larger than the peak sweep voltages has reached. As a result, the GFT conductance will change in opposite directions under the same V_{gs} depending on different initial conduction types of the channel. This lays the physical foundation of complementary synapses.

B. Complementary GFT as Synapses

The GFT conductance modulation by a series of identical pulse train is measured, where various V_{gs} pulse heights are presented while the source-drain voltage V_{ds} is fixed at 0.1 V. Here opposite behaviors are observed in p- and n-type GFTs

(Fig. 10(a) and (b)). In order to attain high linearity for synaptic weight update, appropriate V_{gs} pulse heights have to be carefully selected. Too small amplitudes of pulse heights ($|V_{gs}| \leq 4V$) fail to reach the upper and lower limits of graphene channel conduction. On the contrary, although good ON/OFF ratios are gained, too large amplitudes of V_{gs} pulses ($|V_{gs}| \geq 10V$) lead to fast convergence to the lowest/highest conductance states, leaving insufficient number of intermedium conductance states and thus low precision of the synaptic weights. By using V_{gs} pulses with the heights ±8V and widths 10ms, the analog weight modulation of n- and p-channel GFT synapses are then measured and demonstrated in Fig.11. The benchmark used to measure the performance for level-based computing is shown in Table 1 with significantly reduced non-ideal factors in our C-GFT synapses [3]. Moreover, for the first time we report the functioning of synapses in potentiative and depressive modes (P- and D-mode): given positive/negative V_{gs} pulses the P-mode GFT synapse undergoes conductance increasing/decreasing while the D-mode counterpart takes decreasing/increasing. Fig.12(a) and (b) give a more comprehensive picture where the influence of initial channel conductance G_0 has been taken into account. Conductance modulation with different pulse widths (D-mode as an example) is shown in Fig.13. The demonstrated characteristics of both the pulse height and width dependence are then utilized in the design of STDP and anti-STDP. The voltage waveforms designed to realize STDP in P-mode and anti-STDP in D-mode are shown in the left of Fig.14. The corresponding synaptic weight tuning under various initial weights is then measured and demonstrated in the middle and right of the figure. Furthermore, by using completely antisymmetric waveforms, opposite results are obtained (Fig.15). Notably, synapses under this situation play the role of inhibitory synapses since the negative presynaptic spike results in postsynaptic potentials decreasing (w<0).

C. Implementation of ReSuMe using C-GFT Synapses

ReSuMe employs window function to drive the spike timings of output neuron to the desire ones as follows:

$$\frac{dw}{dt} = [S_d(t) - S_o(t)]\left[a + \int_0^\infty W(\tau) S_i(t-\tau) d\tau\right] \quad (1)$$

where w is the synaptic weight, $S_x(t) = \sum \delta(t - t_x)$ is the spike sequence of the desire, output or input neuron with subscript x=d, o or i and $W(\tau)$ is the window function to convolute with the input. The potentiative and depressive GFT synapses together as a complementary synapse pair are used to implement ReSuMe, as seen in Fig. 16. The basic idea is that the incremental and decremental weight changes required by the positive and negative signs of S_d(t) and S_o(t) are implemented respectively through the potentiative and depressive properties of the complementary synapses: given the positive timing difference between the input and desired spikes ($t_d - t_i > 0$), the weight of P-mode GFT synapse is increased (Δw_p>0) according to STDP; on the other hand, timing difference between the input and output ($t_o - t_i > 0$) results in decreased weight of D-mode synapse due to the anti-STDP (Δw_d<0); The relative timings of output and desired spikes then determine whether the overall synaptic strengths are enhanced or reduced. In Fig.17 the time chart of signals shows an example where the first-round output fires earlier than the desired ($t_o < t_d$) when the C-GFT are excitatory synapses. t_o approaches t_d round-by-round during the learning process. Note that the C-GFT used as inhibitory synapses have also been designed as seen in Fig.16(b). Several prominent advantages would be gained with them. As shown in Fig.19, the weight tuning of inhibitory synapses helps facilitate the training convergence, reducing the total on-line learning time. Furthermore, the introduction of negative weights mitigates the limit of small ON/OFF ratio of the graphene channel, leading to larger range of spike timing modulation. Hence it is essential to include inhibitory synapses in the design to guarantee the high accuracy of ReSuMe training. Also note that a novel refresh mechanism in which the waveforms for excitatory and inhibitory synapses are periodically exchanged is developed to circumvent the weight saturation phenomenon, as shown by the evolution of weight updates in Fig.18.

Fig.20 shows the crossbar synapse array architecture to execute ReSuMe at a hardware level. The MNIST task has been used as an example in which the earliest spiking of one neuron in the output layer indicates the type of input pattern (Fig.21). Fig.22 shows the excitatory (a) and inhibitory (b) synaptic weights respectively. The learning accuracy by checking with 10^4 test digits is shown in Fig.23. It demonstrates that by using the inhibitory synapses and the designed refresh mechanisms the recognition accuracy is significantly improved from 40% to 86%. Table 2 indicates the significant predominance in area, energy and latency by using C-GFT.

IV. CONCLUSION

C-GFTs have been proposed and demonstrated for the first time as synapses for an SNN. The complementary pair of devices exhibits potentiative and depressive conductance behaviors depending on the type of channels. A novel SNN architecture design and training method are thus proposed for ReSuMe, a supervised learning algorithm. Calibrated simulations have shown that a recognition accuracy of 86% for the MNIST learning task is achieved at the cost of power consumption of 8 pJ/per synaptic operation (total consumption about 63 mJ for 1M images) and a cell area of 100 μm^2.

Acknowledgment

The work was supported by the National Natural Science Foundation of China under Grant Nos. 61841404, 51732003.

REFERENCES

[1] C.-C. Chang, *et al.*, IEDM pp. 15.5.1-15.5.4, 2018.
[2] N. J. Allen, *et al.*, Nature vol. 457, p. 675, 02/04/online 2009.
[3] P. Chen, *et al.*, IEDM pp. 6 1.1-6.1.4, 2017.
[4] M. Jerry *et al.*, IEDM pp. 6.2.1-6.2.4, 2017.
[5] M. Seo *et al.*, *IEEE EDL,* vol. 39, no. 9, pp. 1445-1448, 2018.
[6] J. Woo *et al.*, *IEEE EDL,* vol. 37, no. 8, pp. 994-997, 2016.
[7] S. H. Jo et al , *Nano Letters,* vol. 10, no. 4, pp. 1297-1301, 2010.

Motivation

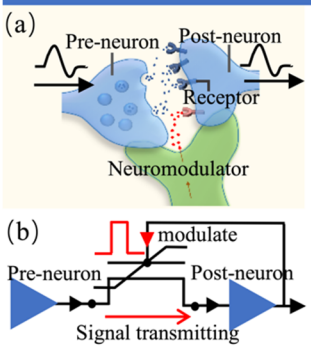


Fig. 1. (a) Biological tripartite synapse. (b) The transitive synapse.

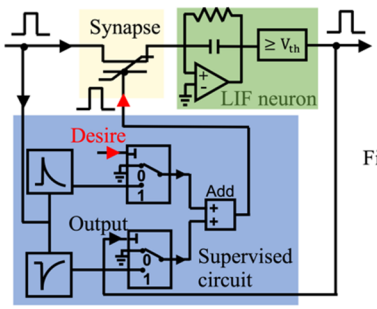


Fig. 2. Implementation of hardware SNN based on conventional single nonvolatile transistors

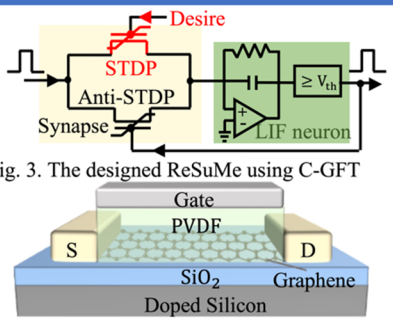


Fig. 3. The designed ReSuMe using C-GFT

Fig. 4. Graphene-ferroelectric transistor structure where the graphene channel is 5 μm long and 10 μm wide, and PVDF is 100 nm thick.

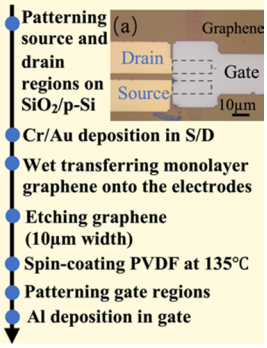


Fig. 5. (a) Top view and (b) Key process steps

Electric Characteristics of Complementary Graphene-ferroelectric Transistors (C-GFT)

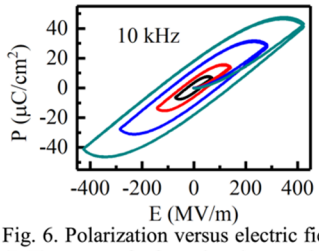


Fig. 6. Polarization versus electric field (P-E) curves of a ferroelectric capacitor

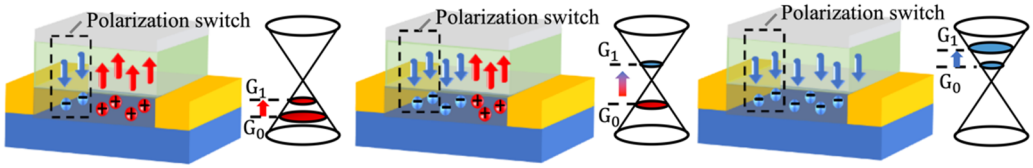


Fig. 8. Left sub-figures show corresponding polarization states in ferroelectric layer at G_1 states in Fig. 9. while right sub-figures show the shifting of Fermi level from G_0 to G_1.

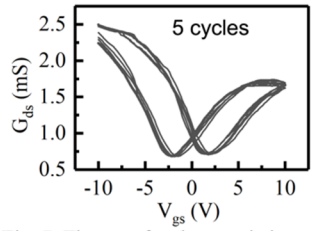


Fig. 7. The transfer characteristic curves

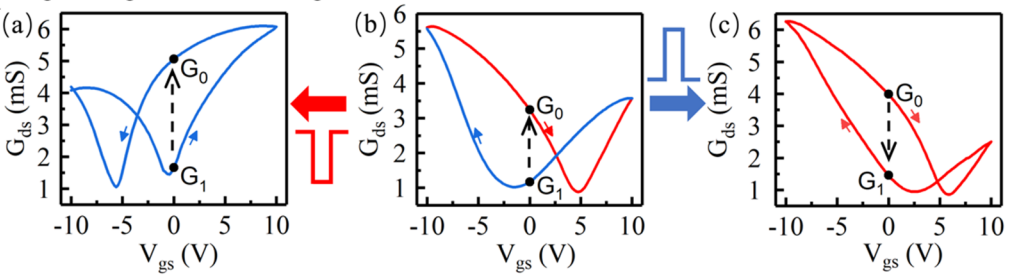


Fig. 9. Gate voltage sweeping under different initial polarization states, which are dynamically reconfigurable.

Graphene-ferroelectric Transistors as Complementary Synapses

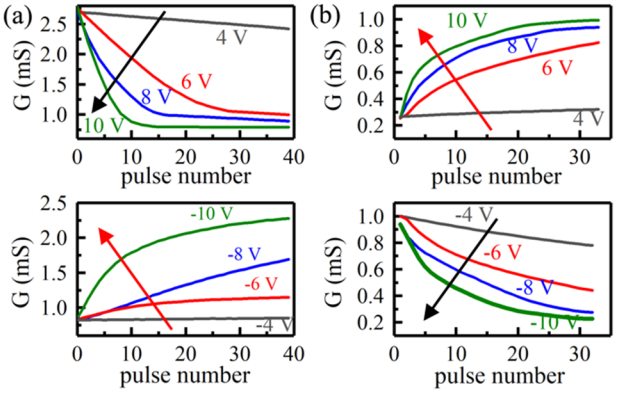


Fig. 10. (a) p-channel and (b) n-channel conductance modulation given a series of identical RESET/SET V_{gs} pulses with various amplitudes.

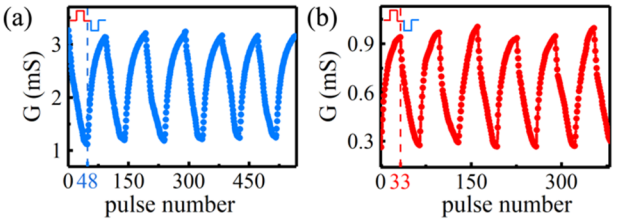


Fig. 11. (a)(b) Analog weight update in depressive/potentiative modes. Both are with pulse amplitudes ±8V and widths 10ms.

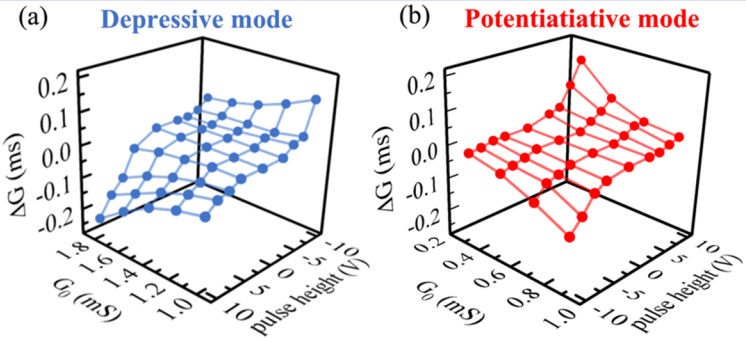
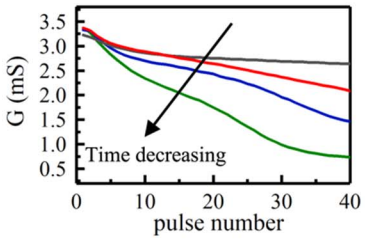


Fig. 12. (a)(b): the conductance change ΔG versus pulse height and the initial conductance G_0 (pulse widths Δt = 10 ms) in depressive/potentiative modes.

	[4]	[5]	[6]	This work
Device	Si-FeFET	FinFET	ReRAM	C-GFT
States	20	50	40	52
Nonlinearity	5.54/−8.08	1.58/−7.57	1.94/−0.61	1.3/−1.1
Asymmetry	13.62	9.15	2.55	2.4
C2C variation	<0.5%	N/A	5%	2%

Table 1. Parameters of various types of synapse devices



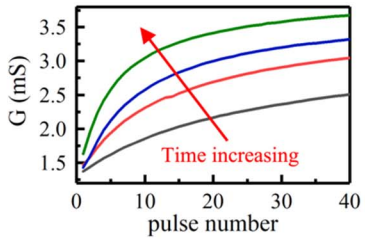
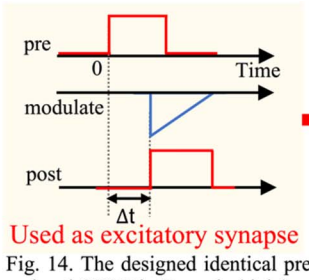


Fig. 13. p-channel conductance modulation by V_{gs} various widths from 5ms to 20ms.



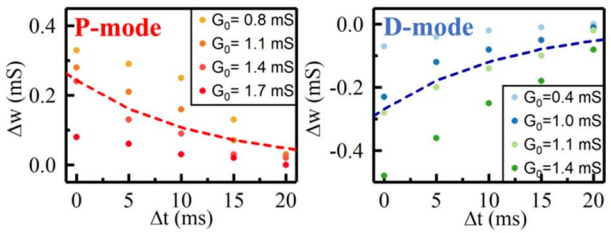
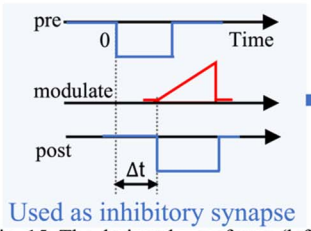


Fig. 14. The designed identical pre-, modulated and post waveforms (left) to realize STDP in P-mode (middle) and anti-STDP in D-mode (right), and the measured results where different initial weights G_0 are considered.



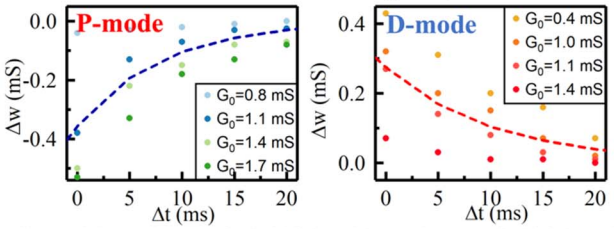


Fig. 15. The designed waveforms (left) to realize anti-STDP in P-mode (middle) and STDP in D-mode (right), and the measured results. Note that here GFT is used as inhibitory synapse.

Complementary Synapses to Implement ReSuMe

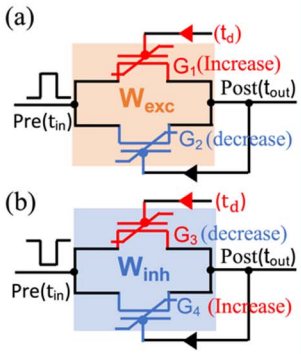


Fig. 16. The complementary GFT as (a) excitatory and (b) inhibitory synapses.

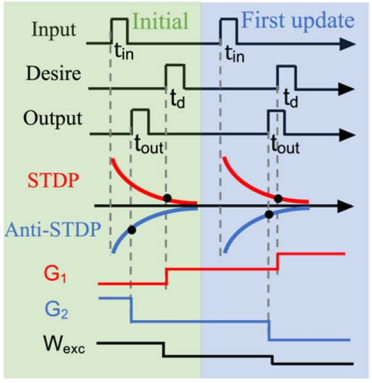


Fig. 17. Time chart of excitatory synapse training.

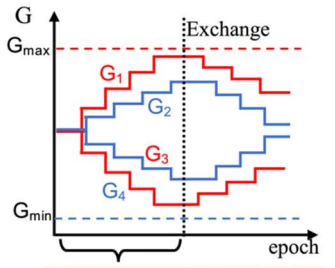


Fig. 18. Periodical refresh by exchanging the waveforms for excitatory and inhibitory synapses

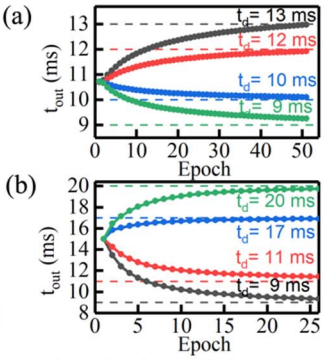


Fig. 19. The evolution of t_{out} with training epochs (a) without and (b) with inhibitory synapses.

Test with MNIST

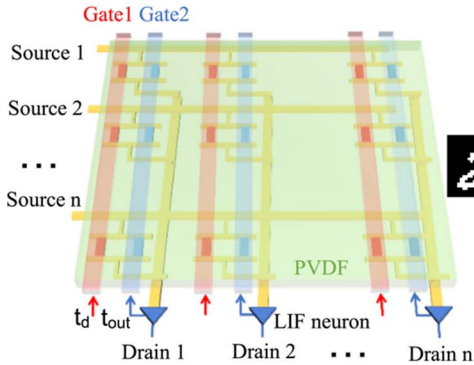


Fig. 20. The crossbar architecture to execute ReSuMe by using complementary GFT synapses

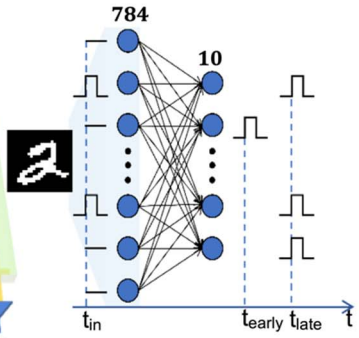


Fig. 21. MNIST classification with 784 input neuron and 10 output neurons

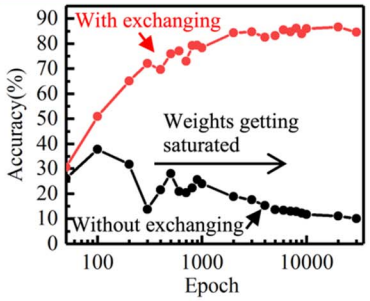


Fig. 23. The recognition accuracy with and without exchanging refresh.

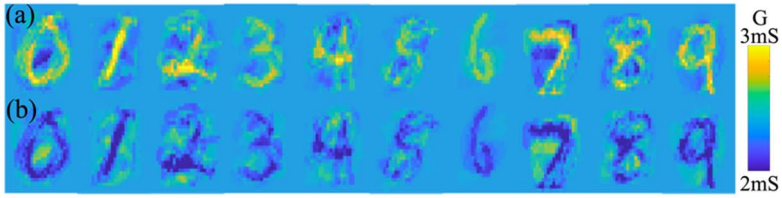


Fig. 22. The synaptic weights of (a) excitatory and (b) inhibitory synapses of 10 output neurons after training.

Features	This work	[4]	[5]	[7]
Supervised circuit	Not required	required	required	required
Accuracy	86%	90%	41%	73%
Energy	63 mJ	98 mJ	150 mJ	88 mJ
Area	100 μm²	1190 μm²	3657 μm²	1072 μm²
Latency	2×10^4 s	3.36×10^4 s	5.6×10^7 s	4.2×10^8 s

Table 2. The summary of system-level performance on 1M images of MNIST.