RRAM Based Convolutional Neural Networks for High Accuracy Pattern Recognition and Online Learning Tasks

Z. Dong, Z. Zhou, Z.F. Li, C. Liu, Y.N. Jiang, P. Huang, L.F. Liu, X.Y. Liu, J.F. Kang*

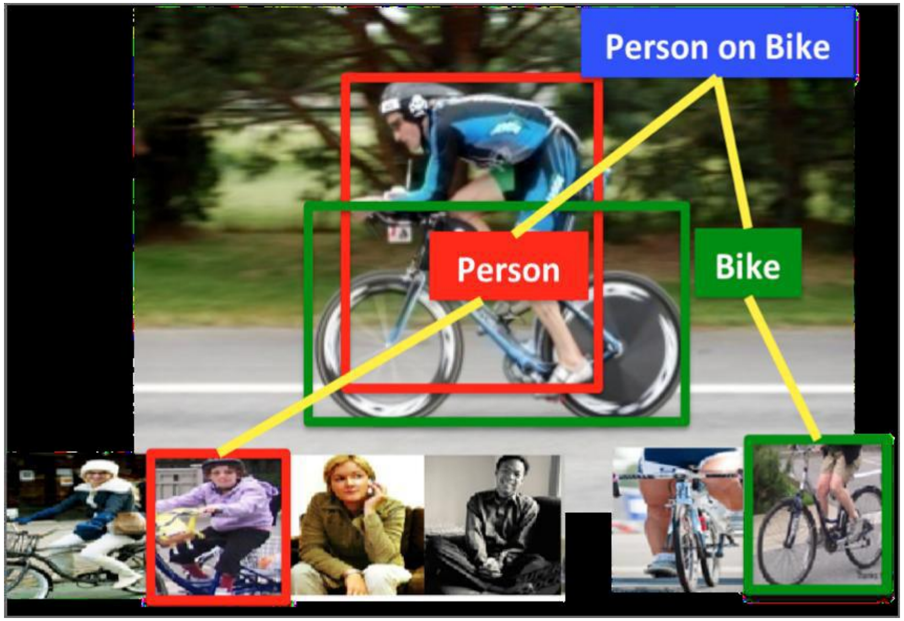
Institute of Microelectronics, Peking University, Beijing 100871, China

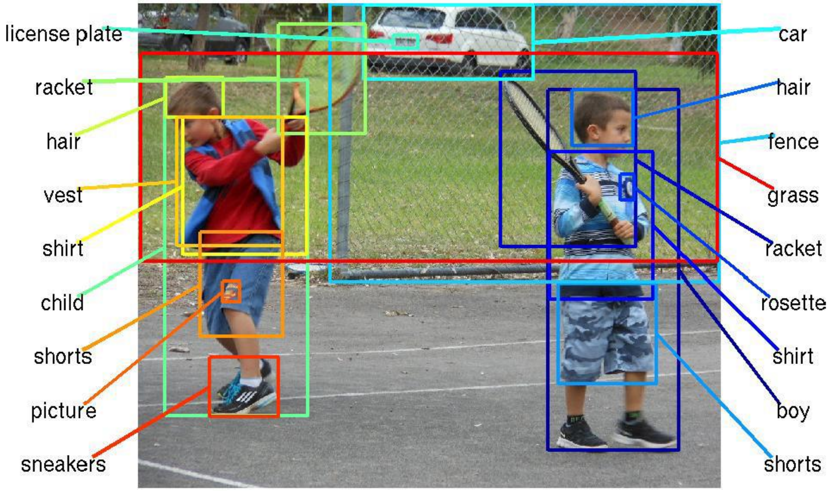
Silicon Nanoelectronics Workshop, June 4-5, 2017

Outline

- Introduction
- CNN Architecture and Circuits
- Device Fabrication and Characteristics
- Results and Discussion
- Online Learning
- Conclusion

The Success of Neural Network Algorithms:





Pictures are from Stanford CS231n class and from the Internet.

The Success of Neural Network Algorithms:

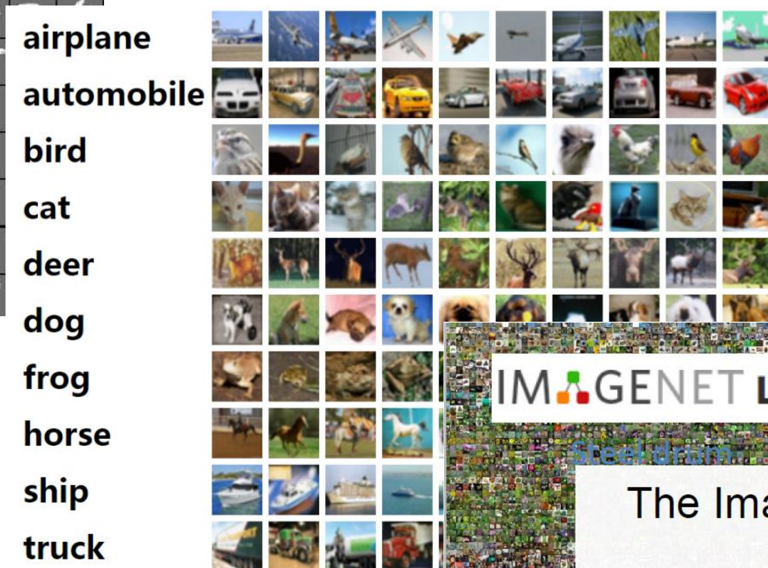
Pictures are from the Internet.

The Success of Neural Network Algorithms:

Pictures are from the Internet.

Deep Learning is Driven by Data:

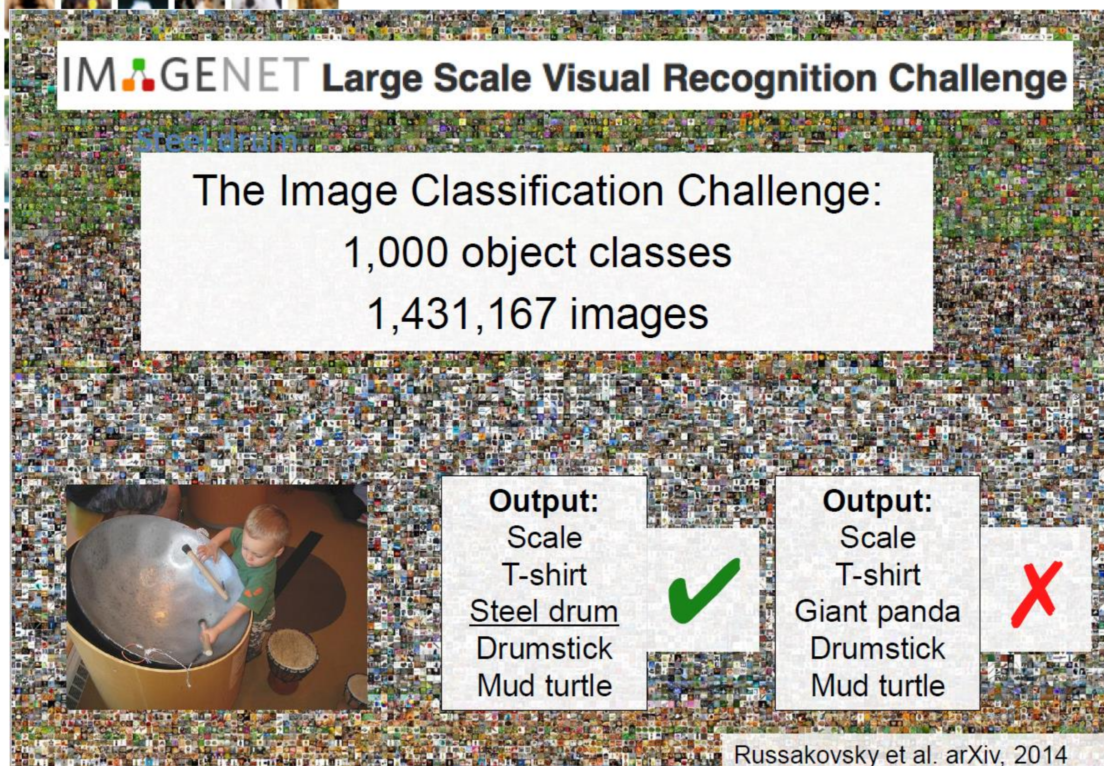
SNW2017



LSVRC (a part of ImageNet)
Training set:
About 1200000 256*256
Validation set:
50000 256*256
Test set:
100000 256*256
RGB

MNIST:
Handwritten
digits dataset
Training set:
60000 28*28
Test set:
10000 28*28
Grayscale

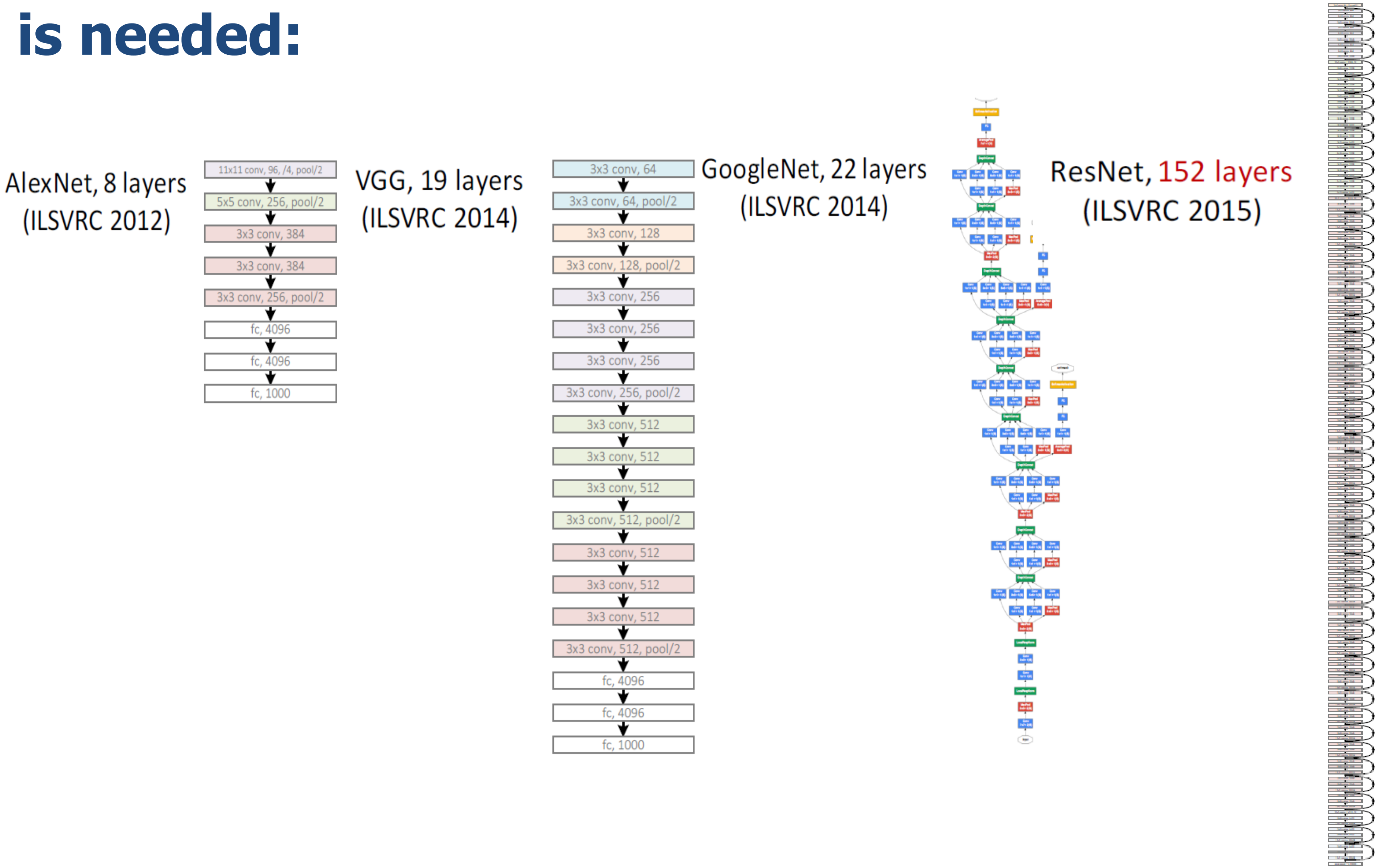
CIFAR-10:
Training set:
50000 32*32
Test set:
10000 32*32
RGB



Pictures are from the Internet.

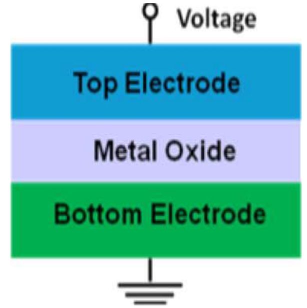
More complicated network structure is needed:

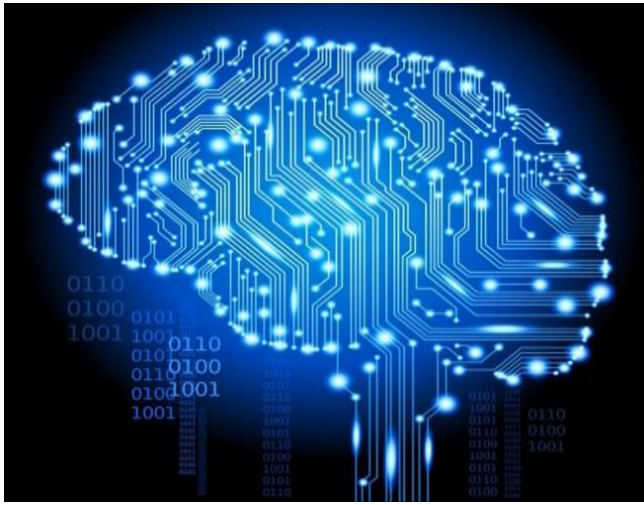
SNW2017



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. CVPR 2016.

RRAM can be used in brain-inspired computing architecture:





Advantages:

1. Non-volatile memory
2. High read/write speed
3. Low area cost
4. MIM structure
5. CMOS Compatibility
6. Potential for 3D integration

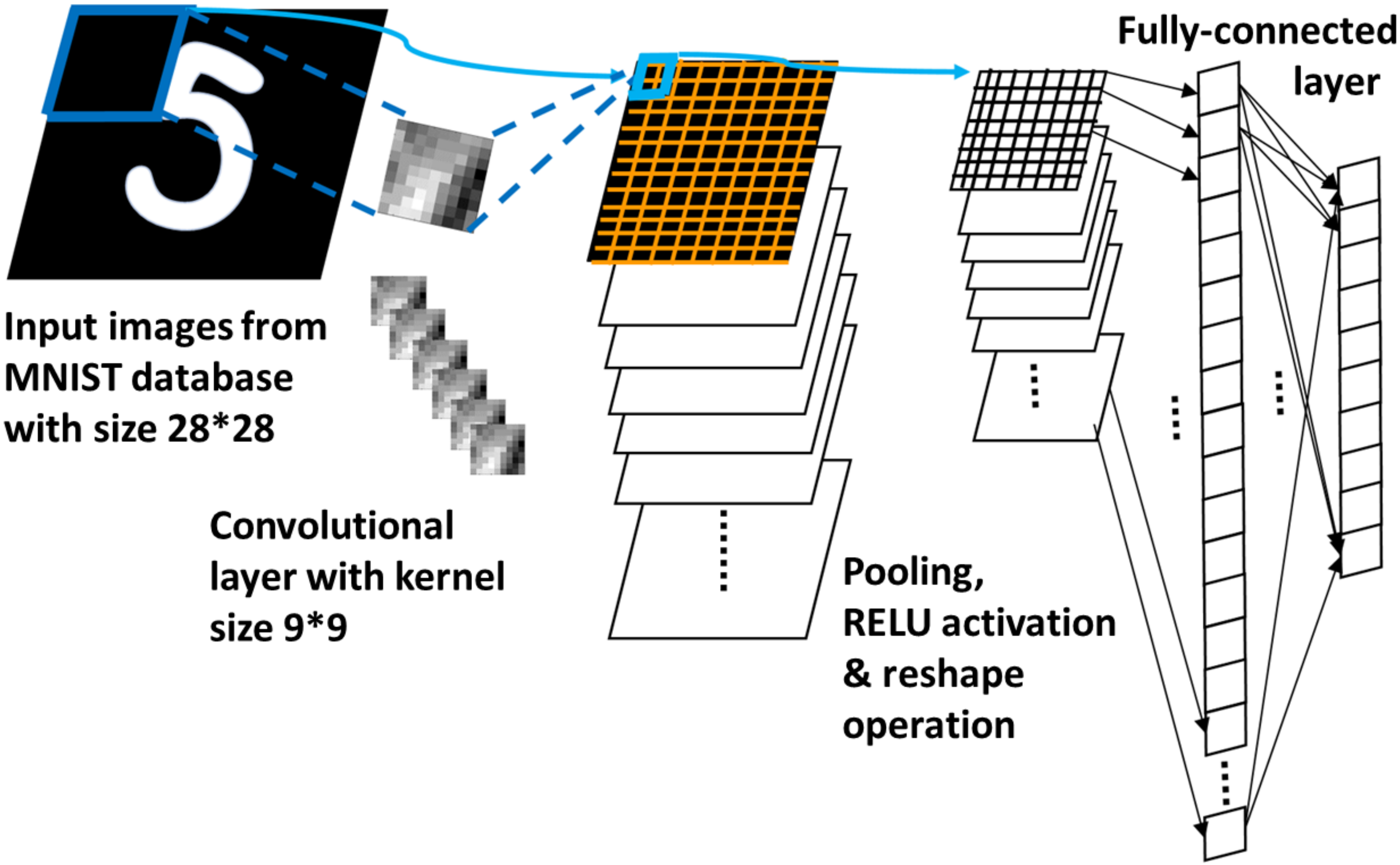
Two pictures are from the Internet.

Wong, H. S. Philip, et al. "Metal–Oxide RRAM." Proceedings of the IEEE 100.6(2012):1951-1970.

Outline

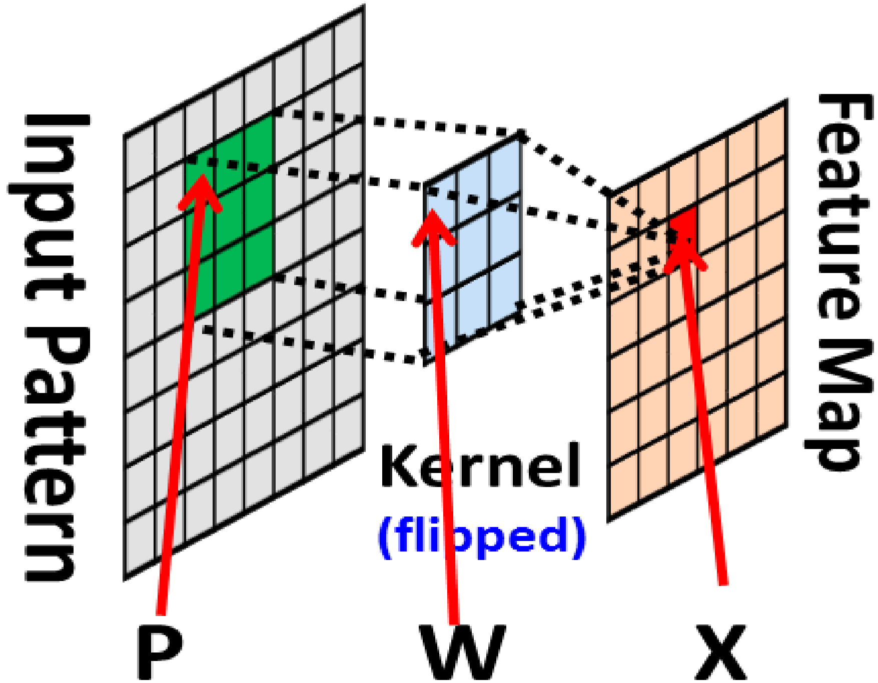
- Introduction
- CNN Architecture and Circuits
- Device Fabrication and Characteristics
- Results and Discussion
- Online Learning
- Conclusion

The architecture of RRAM-based convolutional neural network:



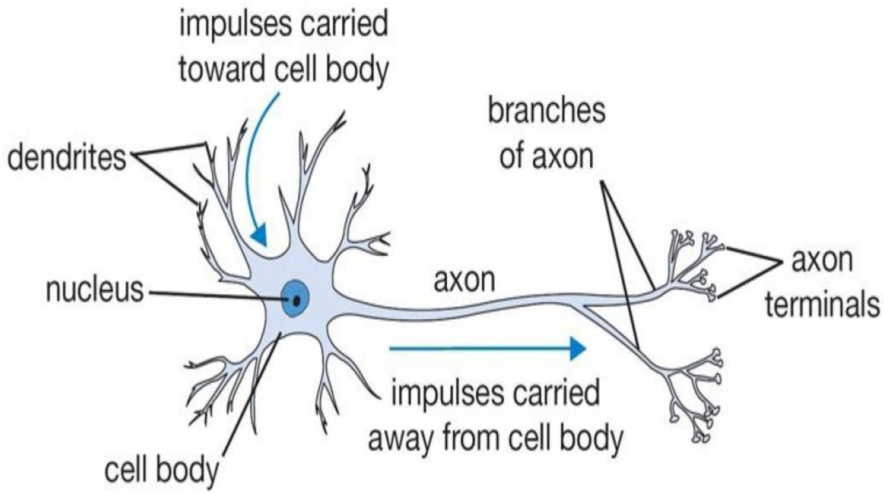
Schematic of Convolutional Operation:

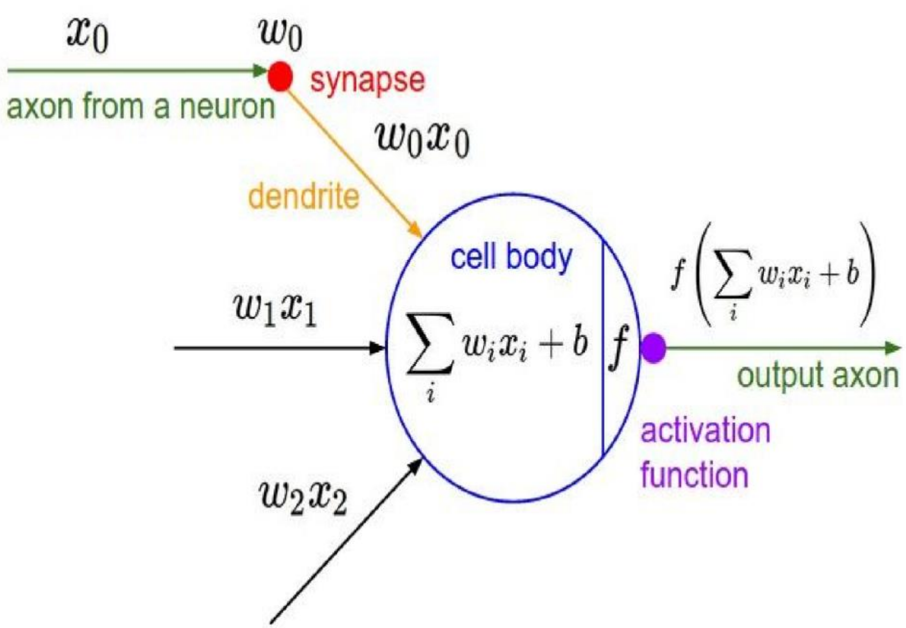
$$X = \sum_{j=1}^{n} \sum_{i=1}^{m} P_{i,j} W_{i,j}$$

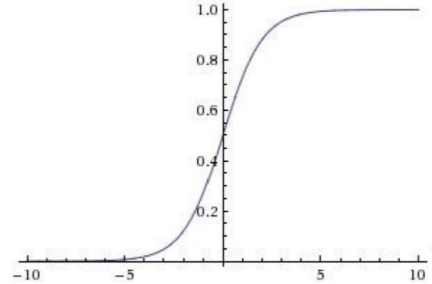


- **P** is one area which has the same size as W.
- **W** is the convolutional kernel.
- **X** is one pixel of the output feature map which is equal to the sum of every product of corresponding pixels in P and W.
- Move the area P then X becomes the whole feature map.

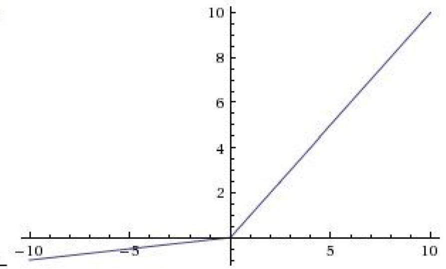
Neural Model and Activation Functions:



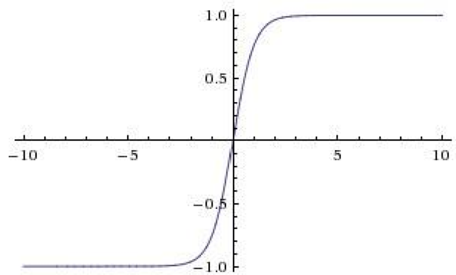




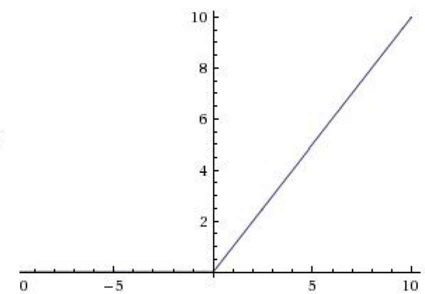
Sigmoid



Leaky ReLU



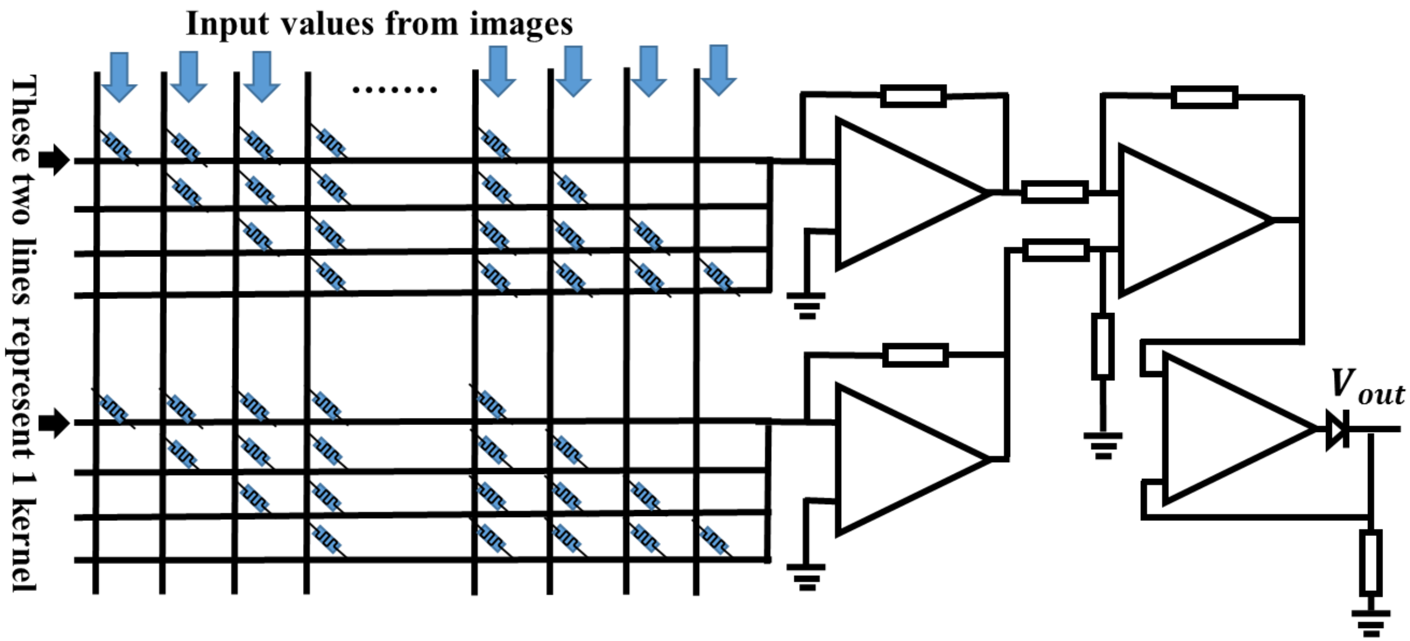
tanh(x)



ReLU

Pictures are from the Internet and Stanford CS231n class.

The Circuit of Convolutional Operation, Pooling and Activation:

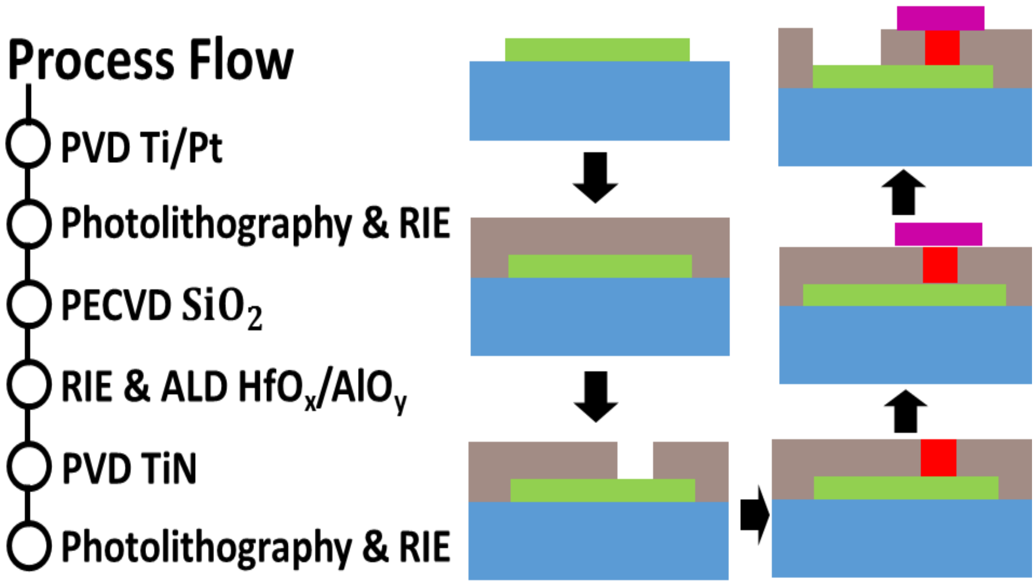


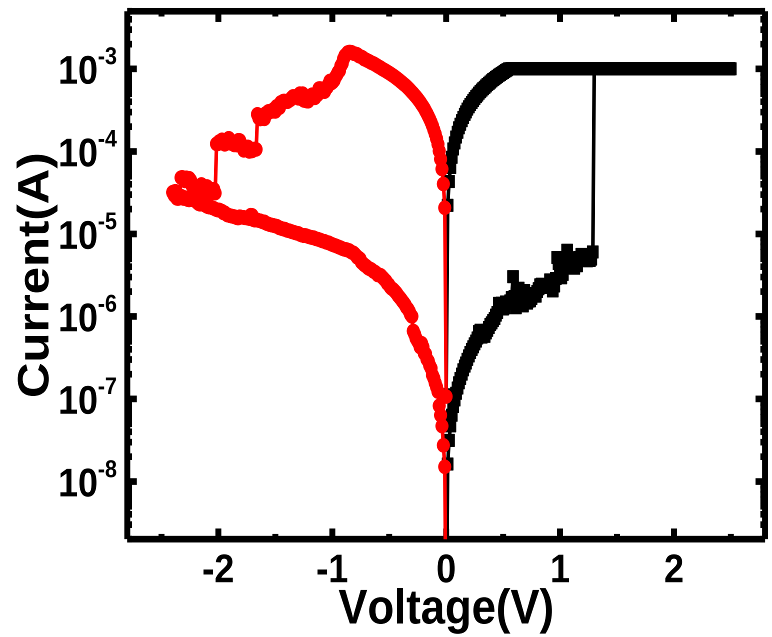
- Two rows of RRAM represent one kernel.
- The blue RRAMs in the first row correspond to one selected area, and the blue RRAMs in the second row correspond to an adjacent area.
- The blank points in the RRAM crossbar array represent RRAMs without forming process, which have extremely high resistance and thus invalid.
- The integrated currents are then be activated and converted to output voltage.

Outline

- **Introduction**
- **CNN Architecture and Circuits**
- **Device Fabrication and Characteristics**
- **Results and Discussion**
- **Online Learning**
- **Conclusion**

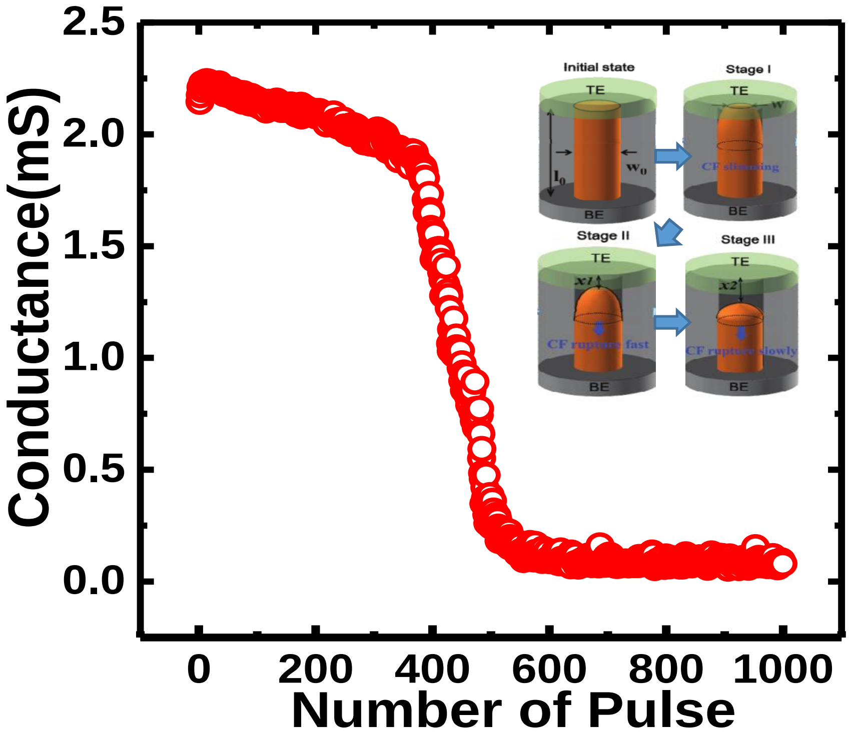
The Process Flow and Measured Typical I-V curve:

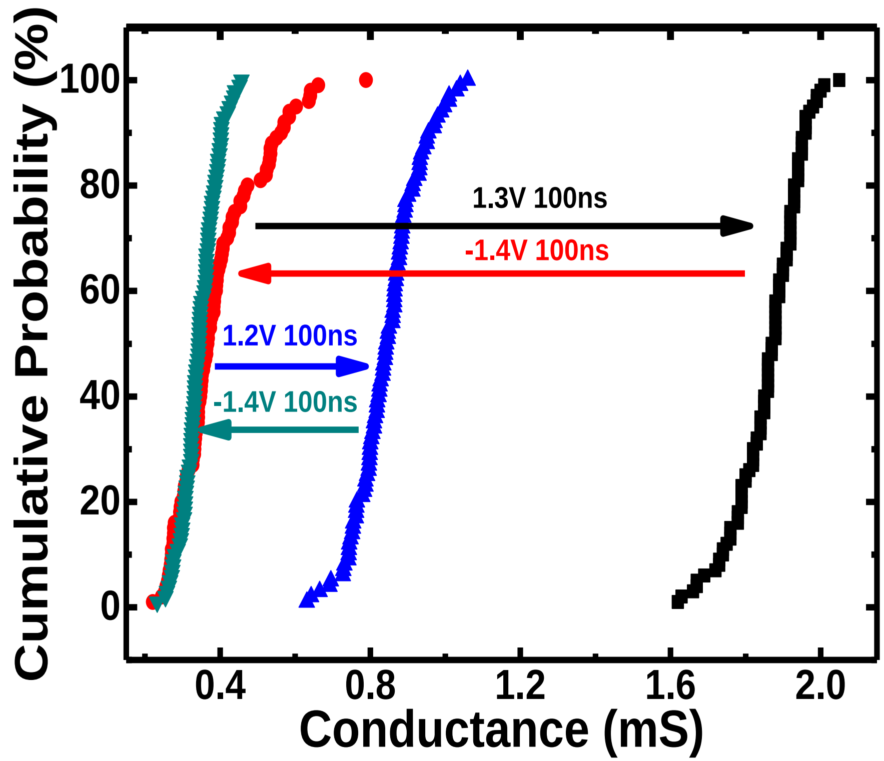




- ◆ The black curve in the right figure is corresponding to Set process.
- ◆ The red curve in the right figure is corresponding to Reset process.

Measured Multilevel Resistance States of RRAM devices:





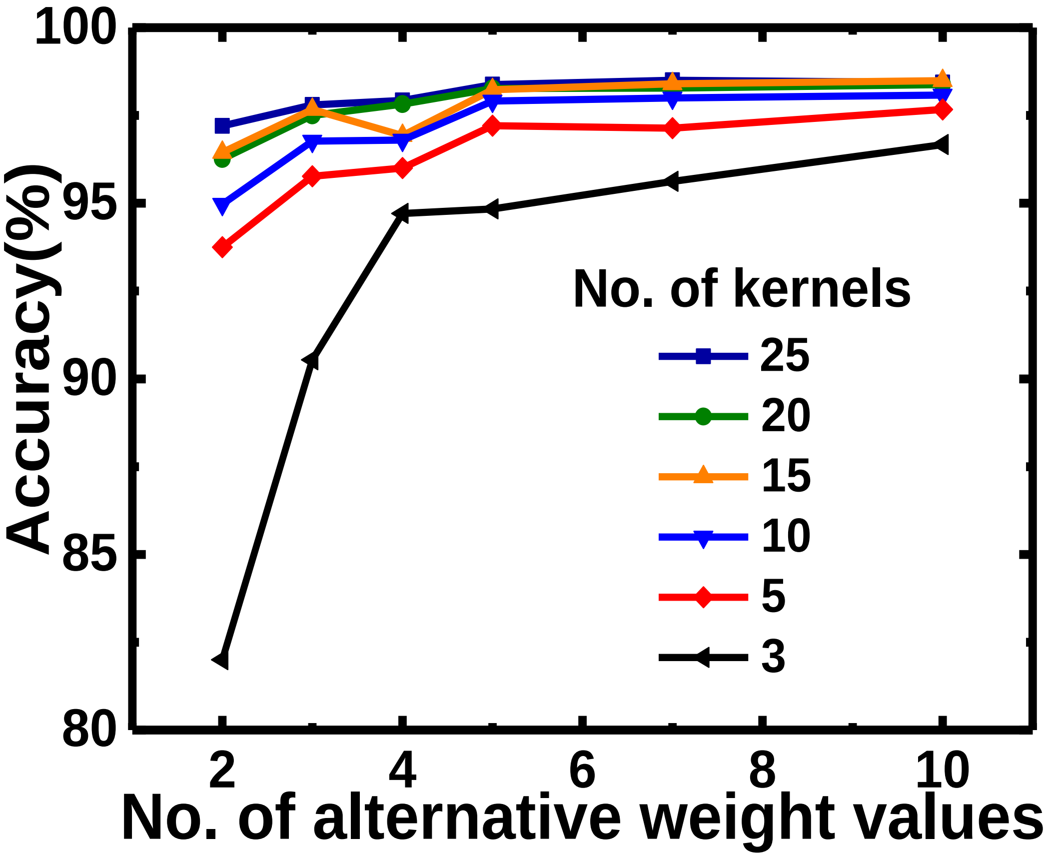
- By applying continuous small reset pulses, up to hundreds of resistance states could be reached.
- The proportion between two conductance states is around 3, since the difference of two RRAMs represent one value, 0, ±1, ±2, ±3 can be obtained using only two conductance states.

Outline

- **Introduction**
- **CNN Architecture and Circuits**
- **Device Fabrication and Characteristics**
- **Results and Discussion**
- **Online Learning**
- **Conclusion**

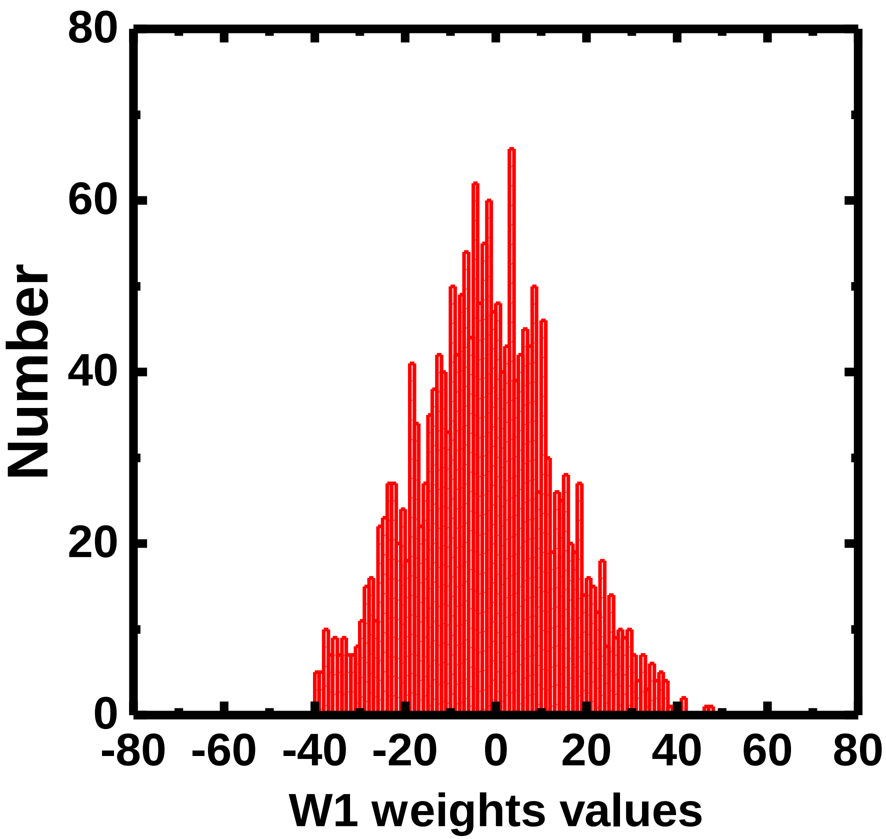
The influence of alternative weight values and the number of kernels:

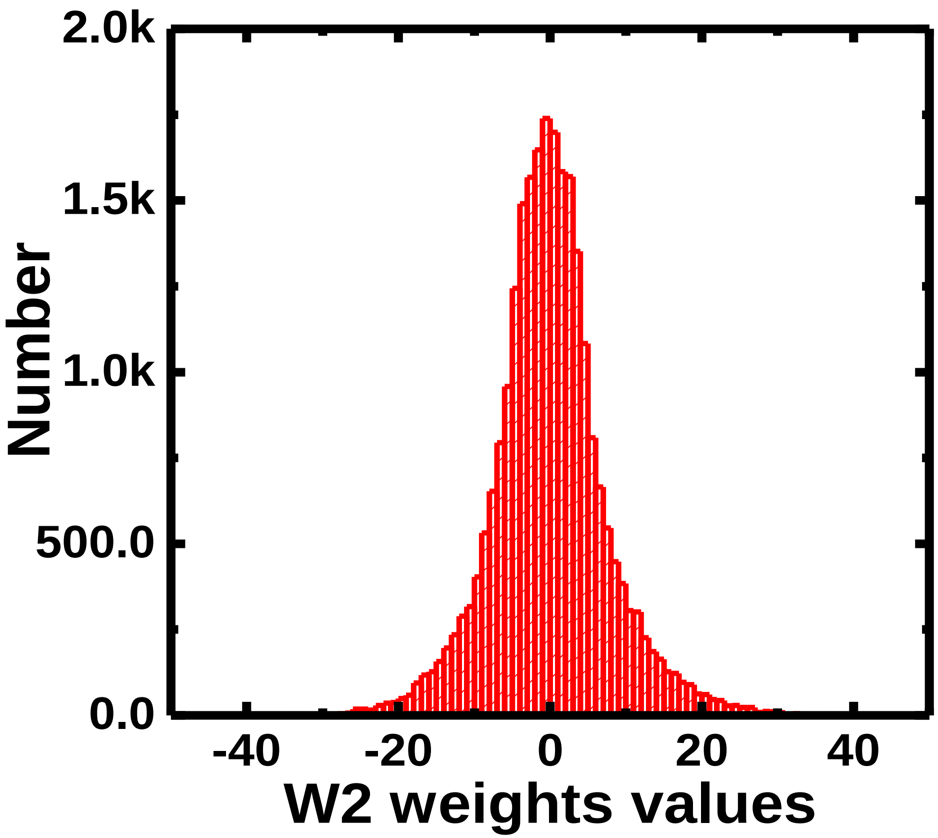
SNW2017



- If the alternative weight number is larger than 4, there won't be considerable reduction of recognition accuracy.
- When the number of kernels is larger than 10, increasing kernel number could not make recognition accuracy higher.
- The maximum recognition accuracy on MNIST is 98%.

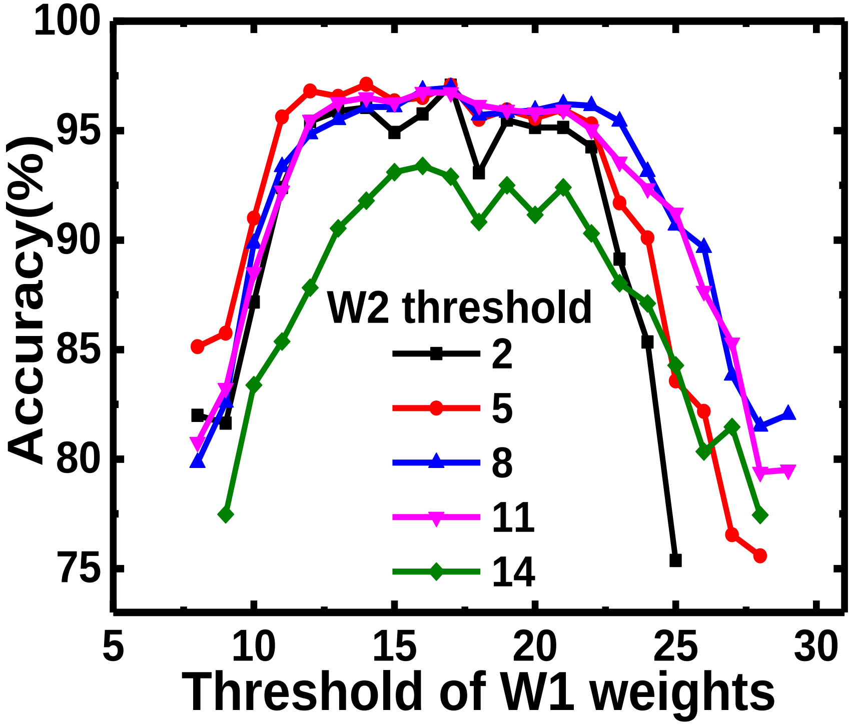
The frequency histogram of full-precision weights:

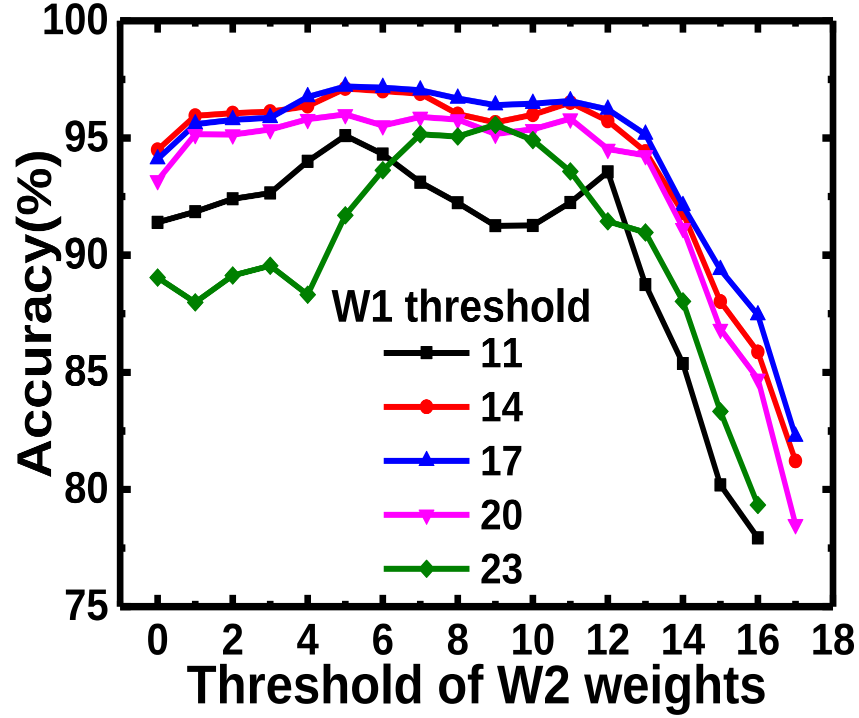




- The total number of weights in W2 is much larger than that of W1, thus the distribution curve of W2 is more smooth.
- A good quantification scheme is supposed to maintain the standard deviation close to that of the original weights.

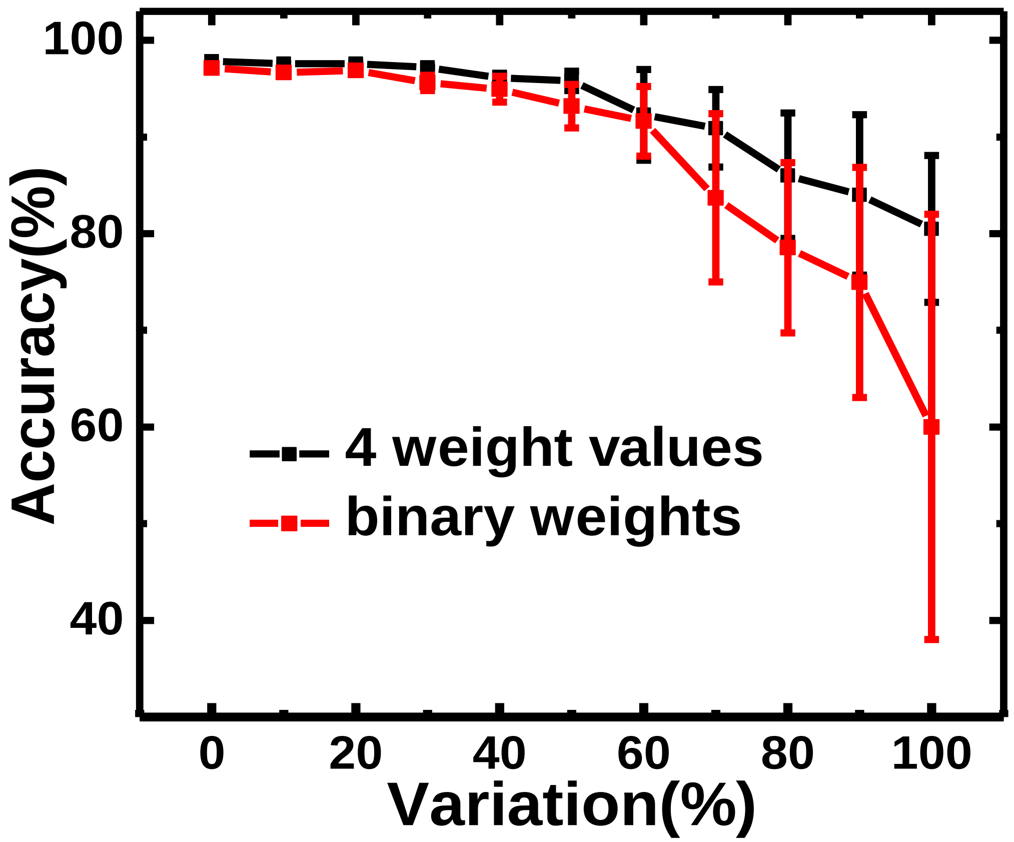
The influence of W1 and W2 threshold on recognition accuracy:





- The left figure shows the relationship between W1 threshold and recognition accuracy with W2 threshold fixed. (12 – 22 are acceptable)
- The right figure shows the relationship between W2 threshold and recognition accuracy with W1 threshold fixed. (1 – 11 are acceptable)

The influence of RRAM conductance variation on recognition accuracy:

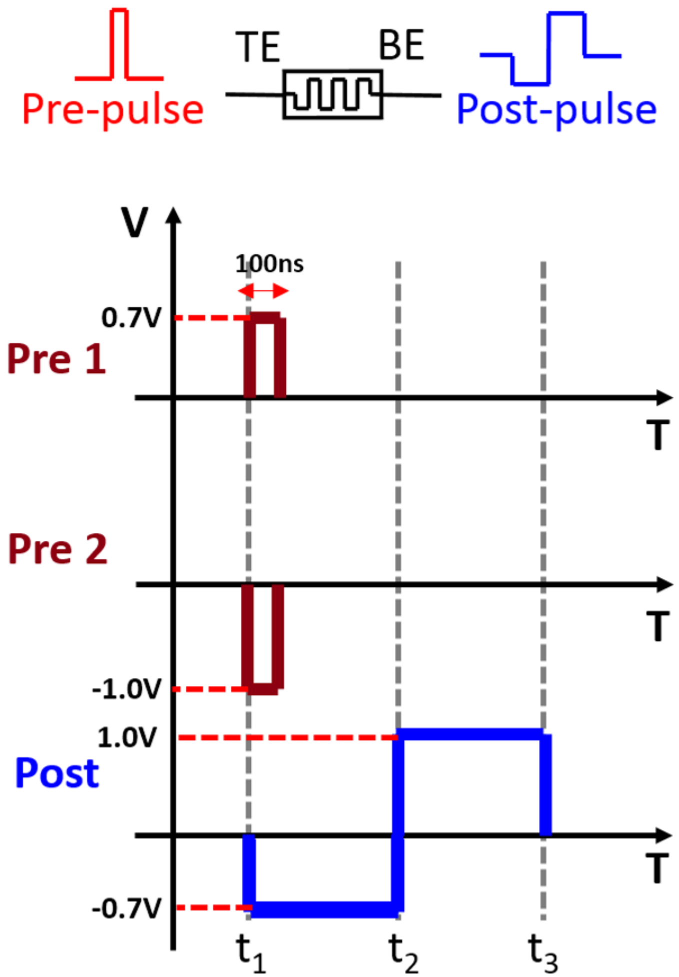


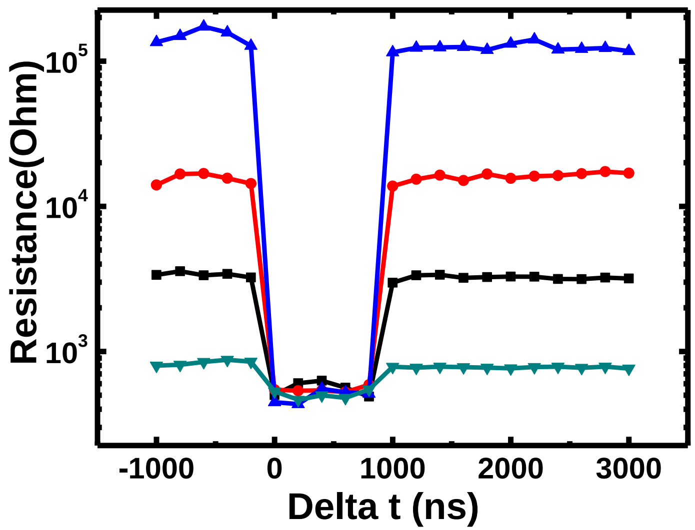
- As the variation increases, the average recognition accuracy begins to fall and the fluctuation gets larger.
- The 4-weight-value scheme is more stable than the binary scheme when variation is large.
- Keeping the variation lower than 50% is necessary for recognition system.

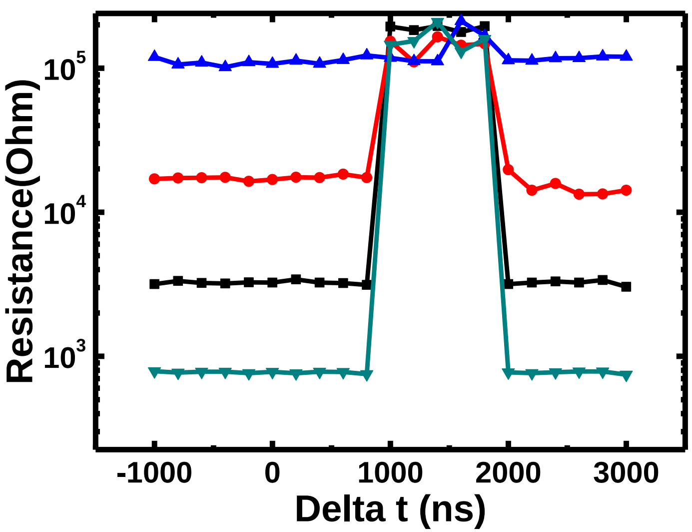
Outline

- Introduction
- CNN Architecture and Circuits
- Device Fabrication and Characteristics
- Results and Discussion
- Online Learning
- Conclusion

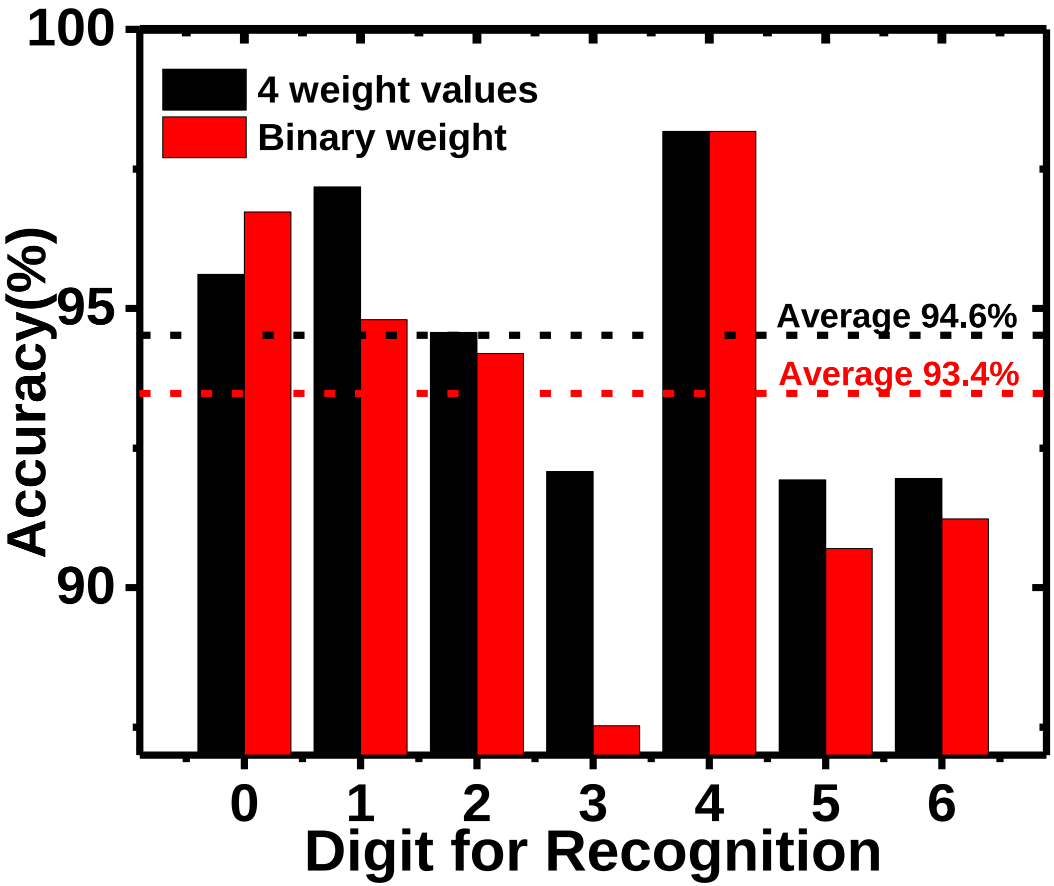
The binary protocol for online operation:





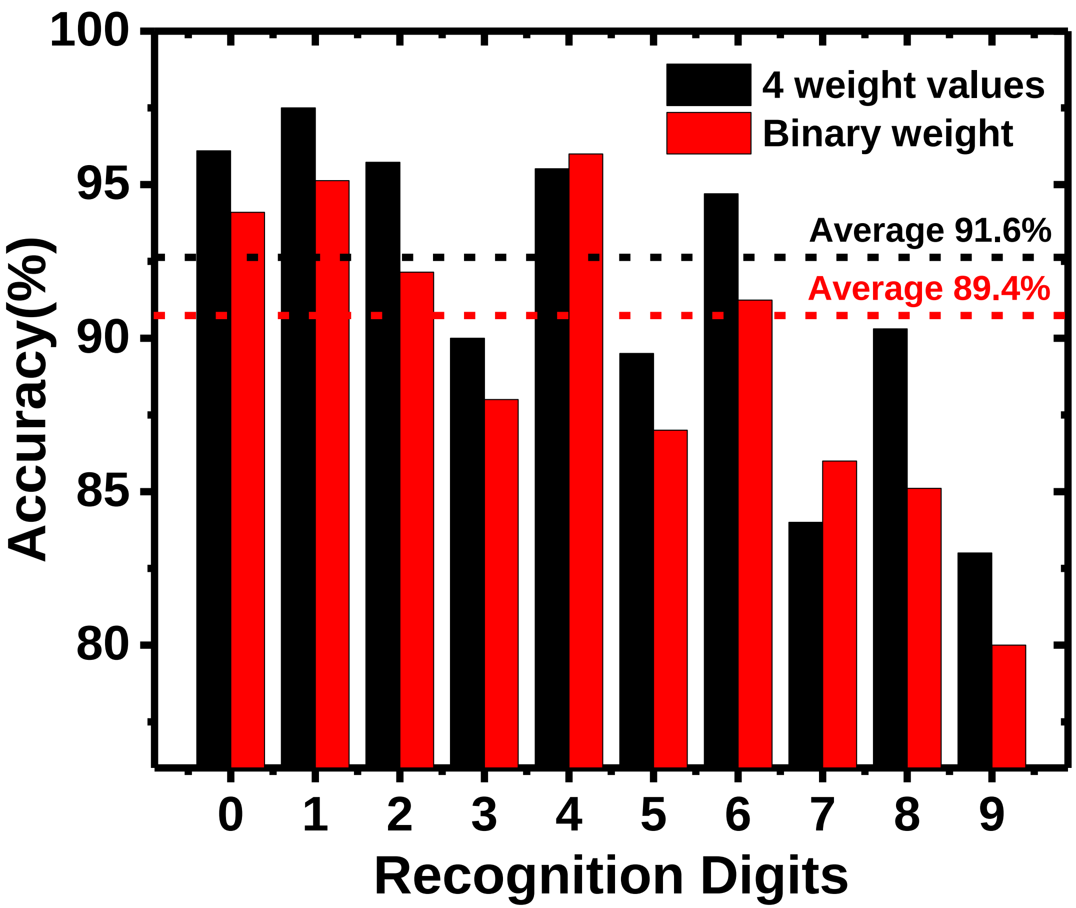


Online Learning Results of both schemes:



- The average recognition accuracy of 4-weight-value scheme is 94.6%.
- The average recognition accuracy of binary weight scheme is 93.4%.
- 4-weight-value scheme has a little better performance.

Online Learning Accuracy on the whole MNIST dataset:



- The average recognition accuracy of 4-weight-value scheme is a bit higher than the binary scheme.
- The average recognition accuracy on 10-digit task is always lower than that on 7-digit task.

Outline

- Introduction
- CNN Architecture and Circuits
- Device Fabrication and Characteristics
- Results and Discussion
- Online Learning
- Conclusion

Conclusion:

- CNN architecture and circuits are designed and optimized.

- Quantification methods for utilizing binary or multilevel RRAM as synapses are proposed, and our CNN performs with 98% accuracy using multilevel RRAM and 97% accuracy using binary RRAM.

- Factors that affect recognition accuracy are studied in detail.

- Online learning is performed based on the developed CNN system.

Thanks for your attention!

Question & Answer