BATMANN: A Binarized-All-Through Memory-Augmented Neural Network for Efficient In-Memory Computing

Dr. Ngai WONG

Oct. 2021

Department of Electrical and Electronic Engineering
The University of Hong Kong

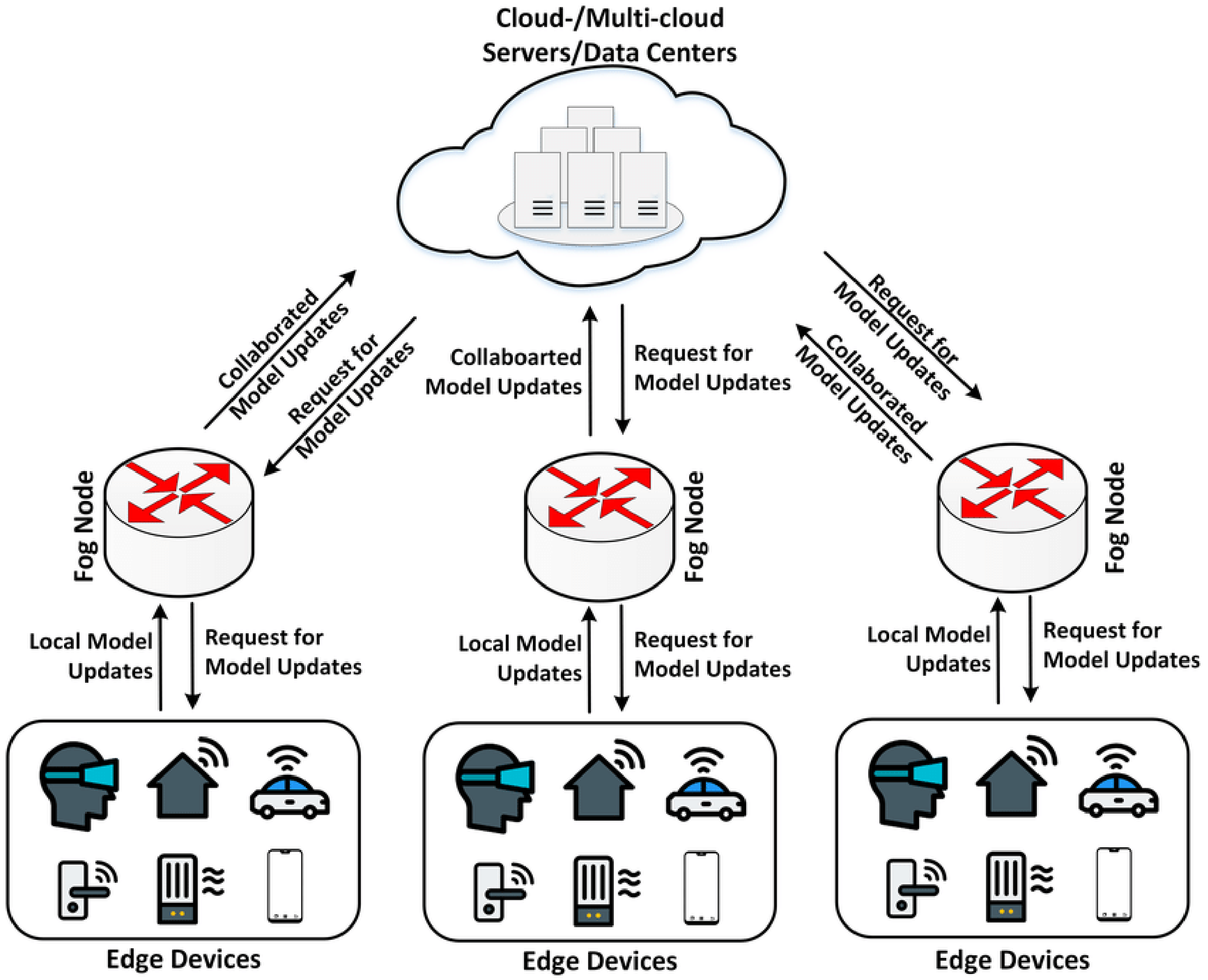
Outline

- Introduction
  - AIoT & Machine Learning
  - Requirement for Structure Evolution – In-Memory Computing
- Proposed Binarized-All-Through Memory-Augmented Neural Network (MANN)
  - Software: Design Algorithm
  - Hardware: RRAM Crossbars & Bipolar Synaptic Weights Implementation
- Experimental Results
- Conclusion

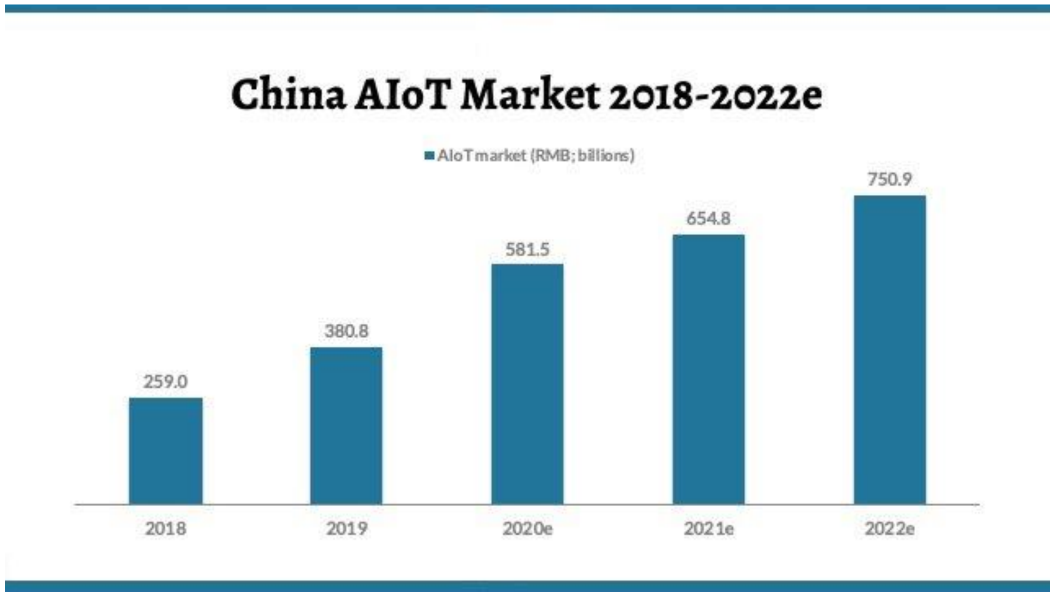
Outline

- Introduction
  - AIoT & Machine Learning
  - Requirement for Structure Evolution – In-Memory Computing
- Proposed Binarized-All-Through Memory-Augmented Neural Network (MANN)
  - Software: Design Algorithm
  - Hardware: RRAM Crossbars & Bipolar Synaptic Weights Implementation
- Experimental Results
- Conclusion

Introduction



ur Rehman M. H., et al. "Towards blockchain-based reputation-aware federated learning." IEEE INFOCOM 2020.



Source: iResearch Jan 2020

- Integration of AI and IoT in practical applications
- Implementation of AIoT ↑ => intelligent terminal devices ↑ => edge computing ↑
- The ability of efficient computing power, local autonomous decision-making and response
- This ML calculation must occur on the device side rather than the cloud

Introduction

Machine learning at the edge device

creates benefits and challenges in the meantime

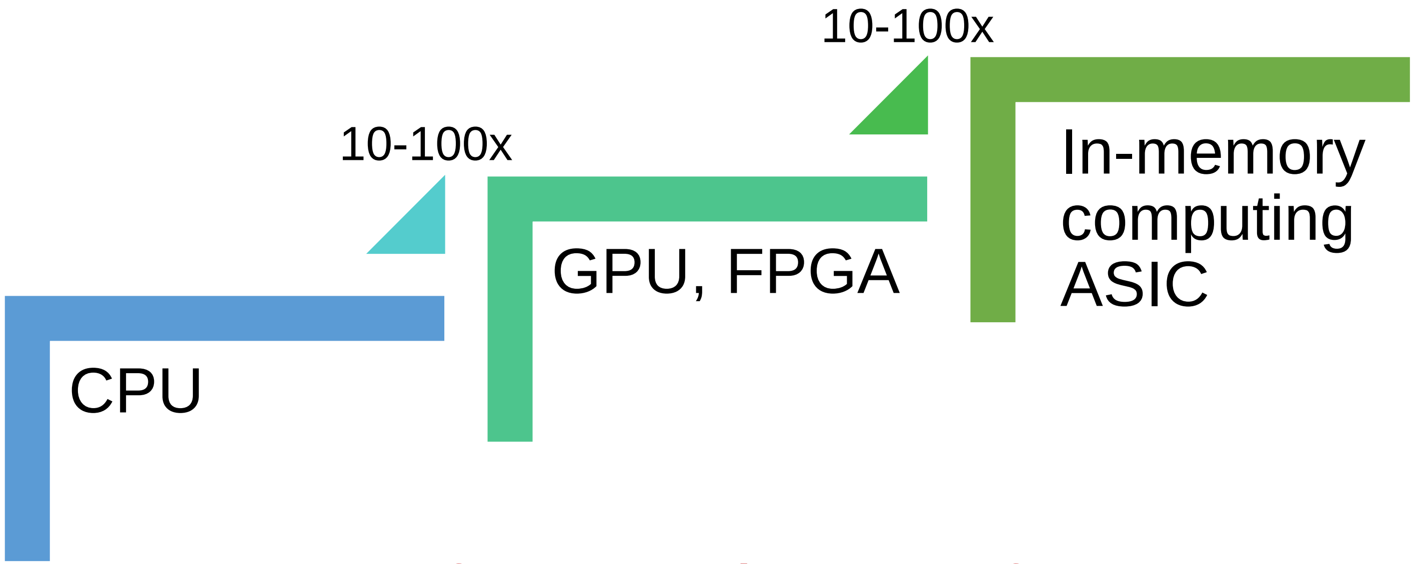
Benefits

- ✓ User-specific data
- ✓ Quick responsiveness
- ✓ Low power
- ✓ Good privacy

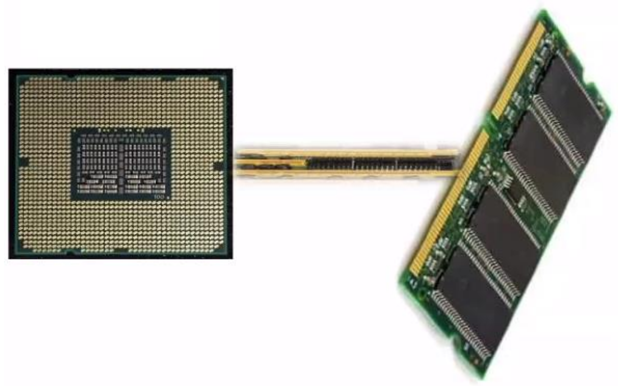
Challenges

- ➢ High energy-efficient
- ➢ High throughput
- ➢ On-device training
- ➢ Large neural network model

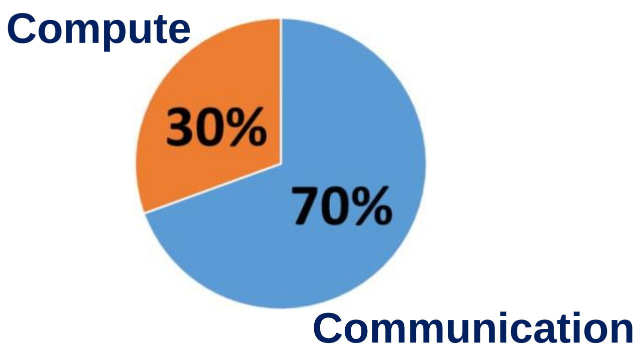
Requirement for Structure Evolution



Performance Enhancement for AI



Von-Neumann memory bottleneck

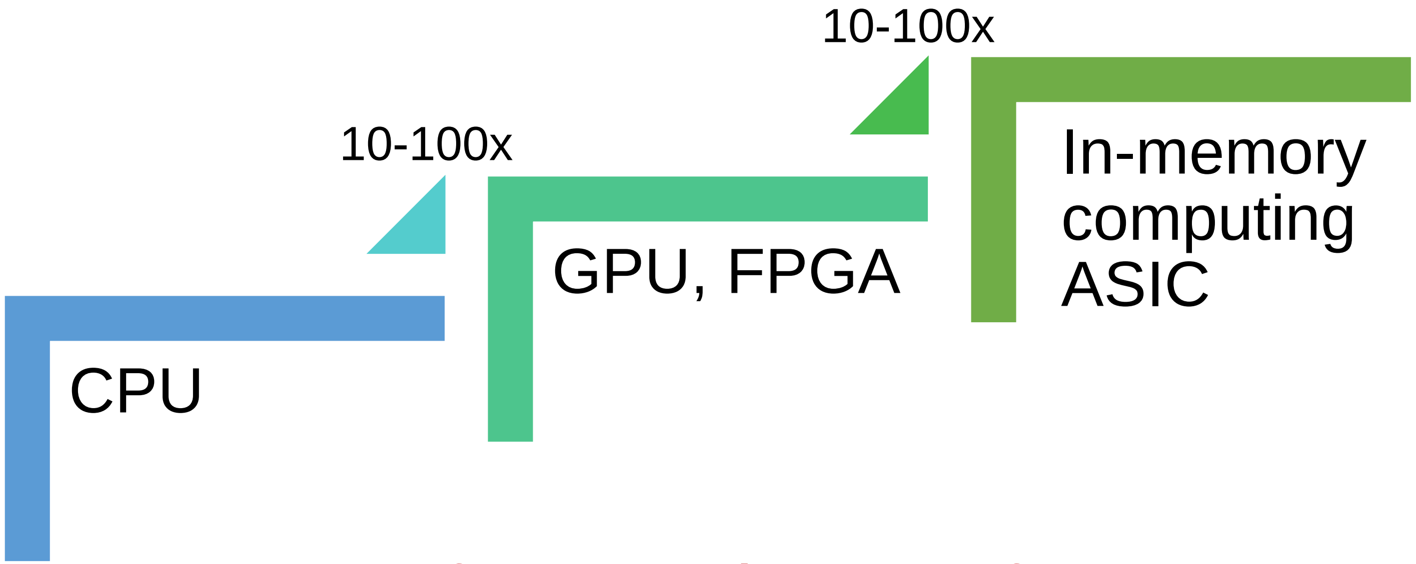


Moving data costs around 70% energy

- Traditional Von-neumann computing system consumes extensive energy and latency in moving the data rather than computation itself.

Limitation !

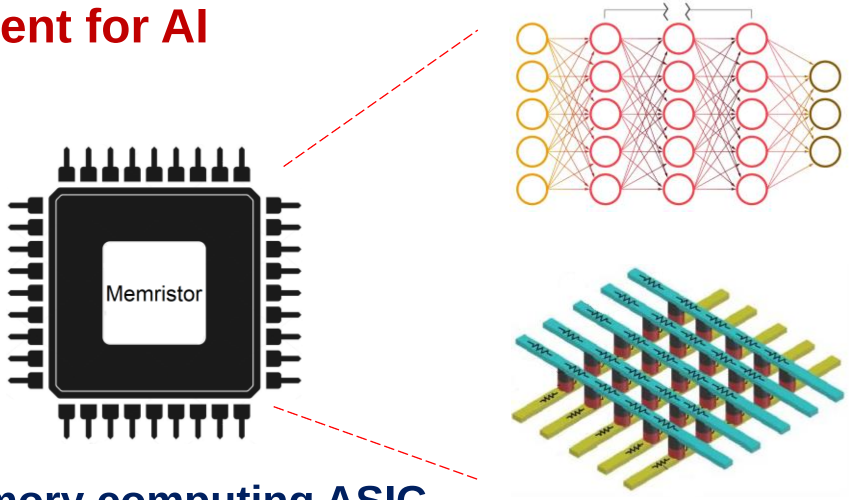
Requirement for Structure Evolution



Performance Enhancement for AI

Can we get inspiration from our brain?

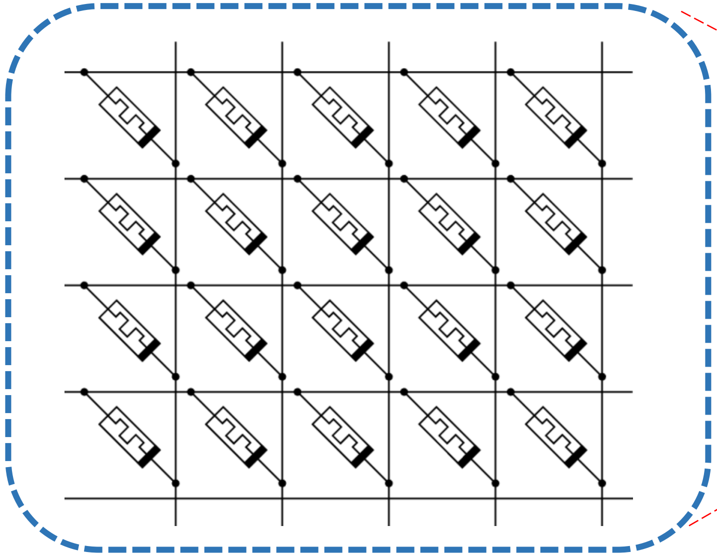
- Hugh number of neurons and synapses
- Highly parallel
- Analog domain
- Noisy environment
- Low power (20 W)

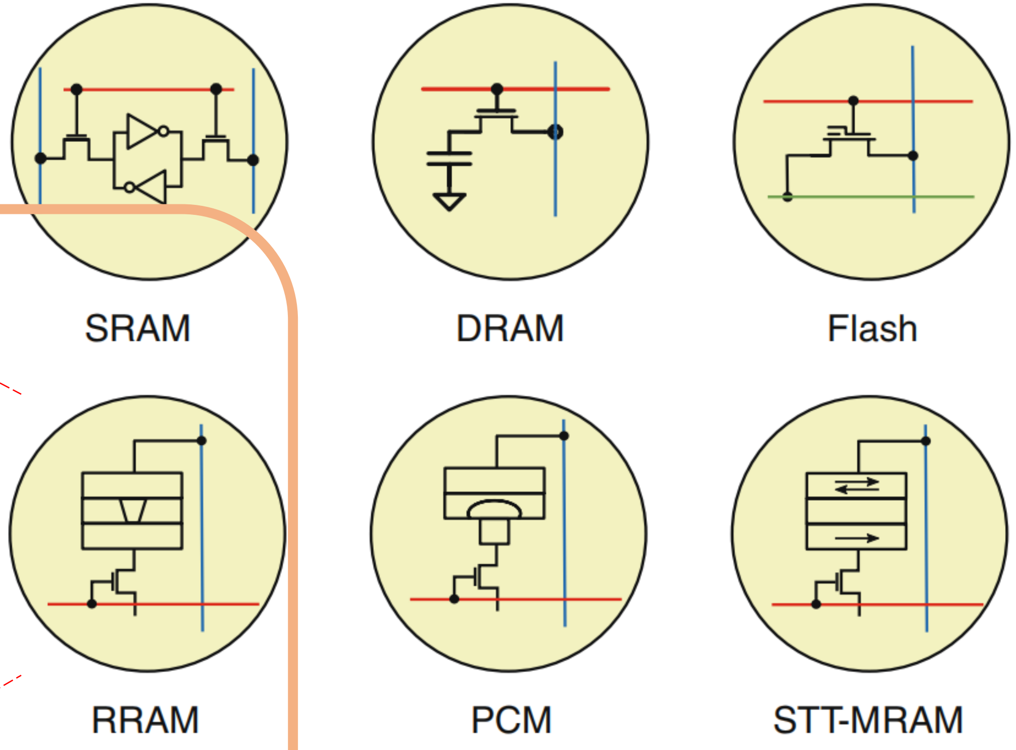


In-memory computing ASIC (Memory + Data processing)

Good solution !

Devices for In-Memory Computing





Charge-based Memory Cell

Resistance-based Memory Cell

Sebastian, Abu, et al. "Memory devices and applications for in-memory computing." Nature nanotechnology (2020)

Today's topic

A Proposed BATMANN based on binary neural network (BNN) design techniques

- Non-volatile memory characteristic
- Reliability and compatibility with CMOS fabrication process
- RRAM-based crossbars for performing MVM
- Hardware-friendly neural network - BNN

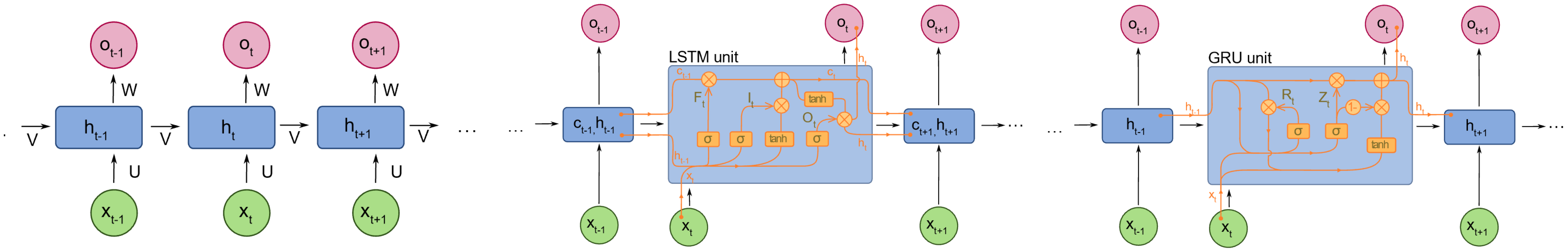
Outline

- Introduction
  - AIoT & Machine Learning
  - Requirement for Structure Evolution – In-Memory Computing
- Proposed Binarized-All-Through Memory-Augmented Neural Network (MANN)
  - Software: Design Algorithm
  - Hardware: RRAM Crossbars & Bipolar Synaptic Weights Implementation
- Experimental Results
- Conclusion

Memory-Augmented Neural Network (MANN)

Limitations of existing networks that aim to handle temporal dependencies in sequential prediction problems.

1. **Recurrent Neural Networks (RNNs)**
   - ➢ vanishing gradients
   - ➢ exponential growth in the number of parameters
   - ➢ costly computation due to increased memory size
2. **Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)**
   - ➢ have difficulties when searching through past memories



(a) RBNN (b) LSTM (c) GRU

The figures are from https://en.wikipedia.org/wiki/Recurrent_neural_network.

Memory-Augmented Neural Network (MANN)

MANN is proposed to alleviate the gradient vanishing problem as it satisfies two criteria:

1. the information stored in the memory is stable and element-wise addressable
2. the number of learnable network parameters are not tied with the size of the memory

There are two main components in a MANN:

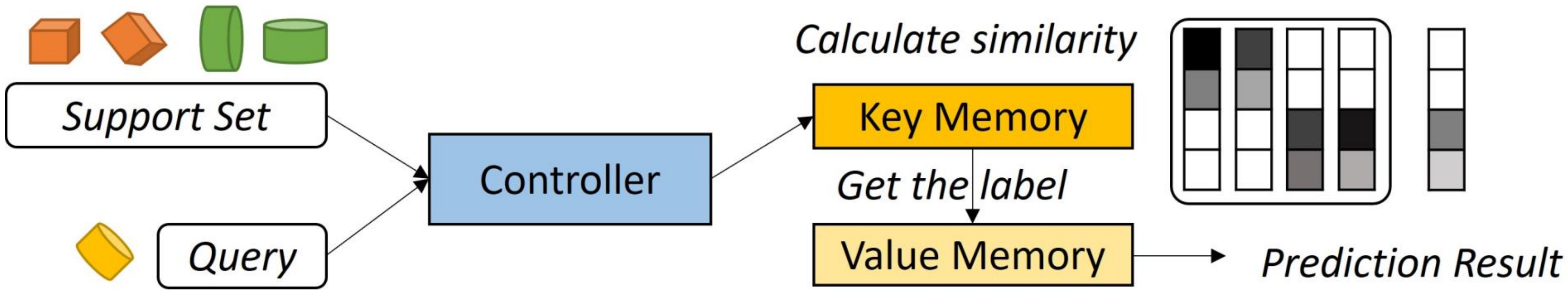
1. a controller

   The controller can learn how to read from and write to the memory.

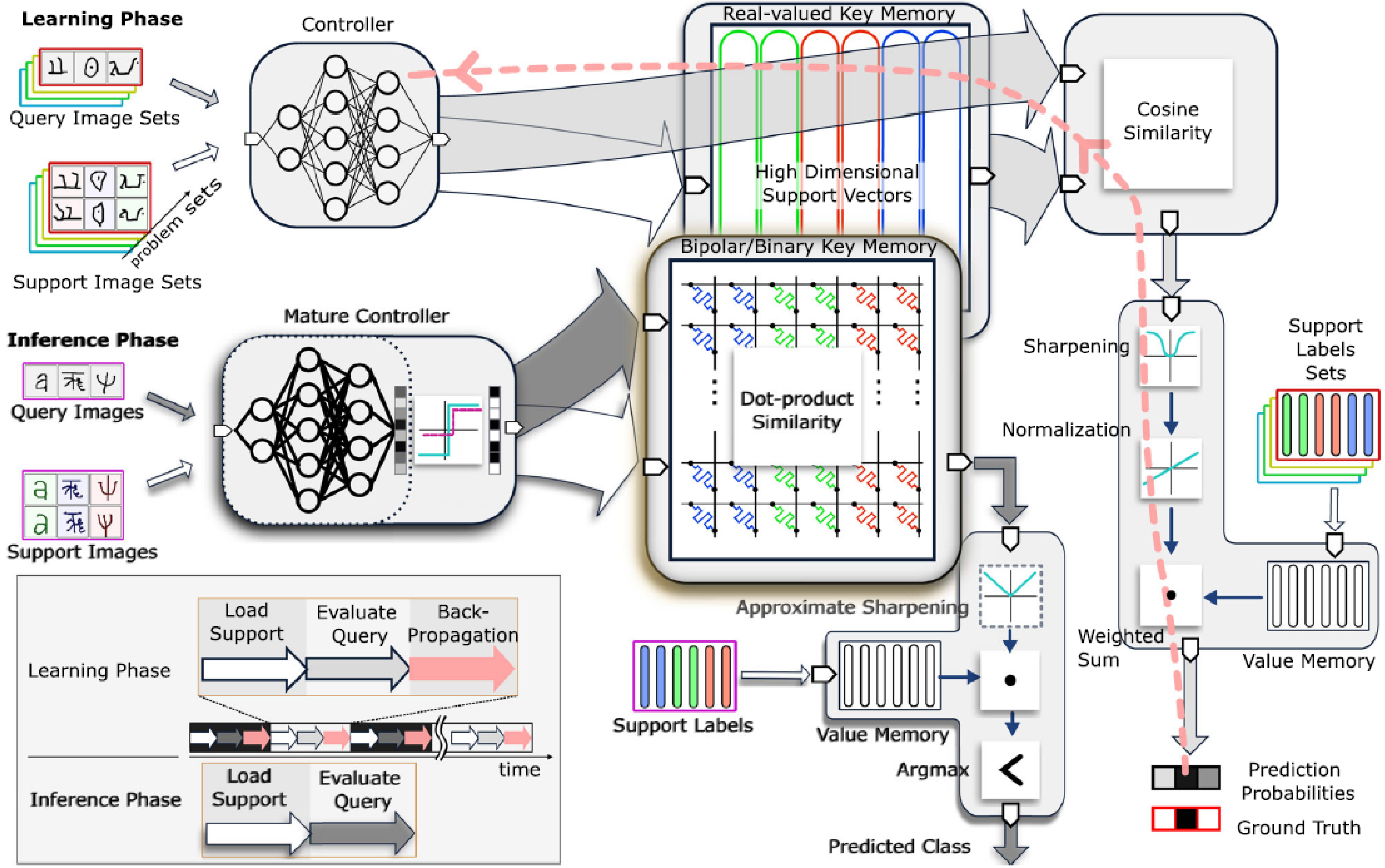
2. external memory

   Key memory: it stores and compares the learned patterns.

   Value memory: it holds the labels.



Memory-Augmented Neural Network (MANN)



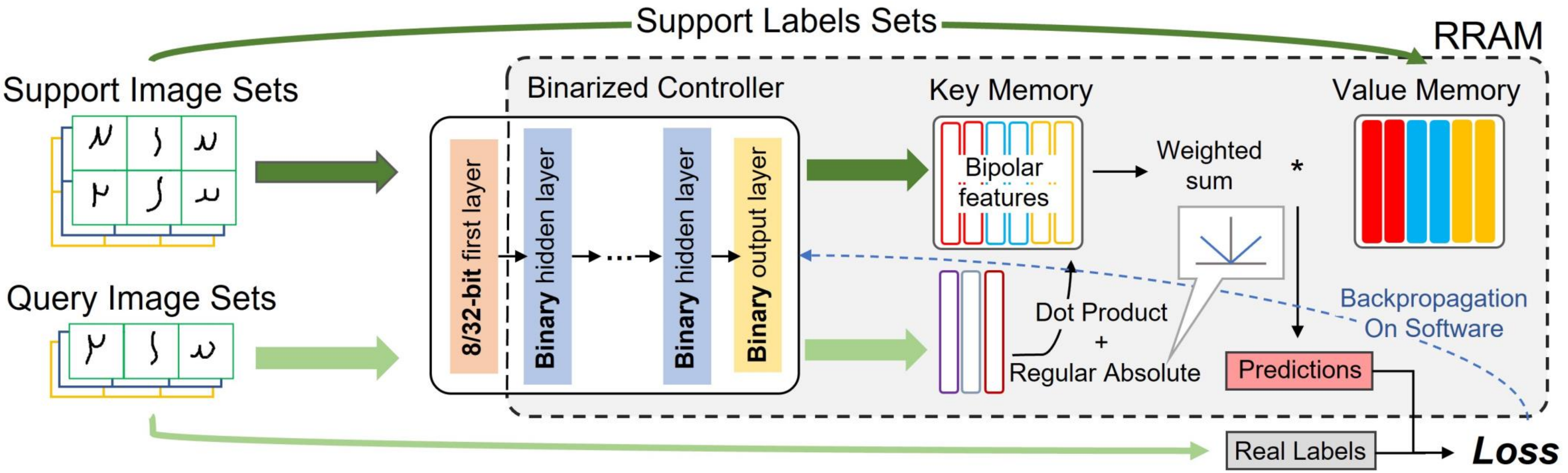
Karunaratne, Geethan, et al. "Robust high-dimensional memory-augmented neural networks." Nature communications (2021)

Limitations:

1. The controller is full-precision.
2. The similarity measure and sharpening function are inconsistent in the training and inference stages.
   (Training: cosine similarity + softabs. Inference: dot product + abs.)

Software: Design Algorithm

The architecture of the proposed BATMANN implemented on RRAM, which contains not only the key-value memory but also a binarized controller, all realizable with 2-level RRAM cells.

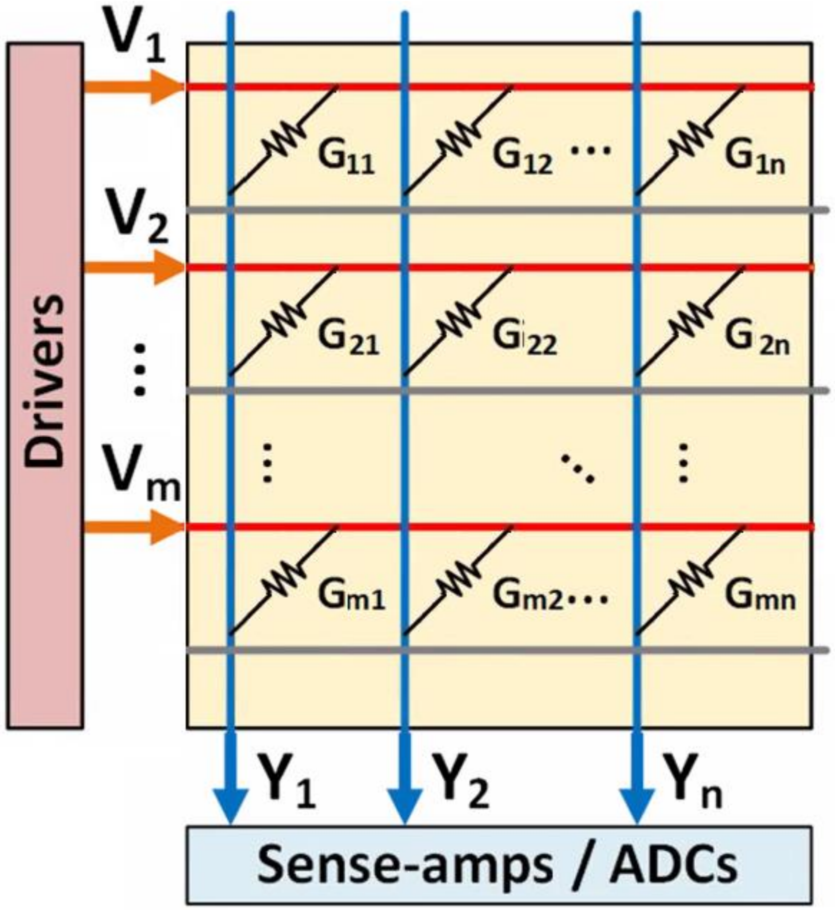


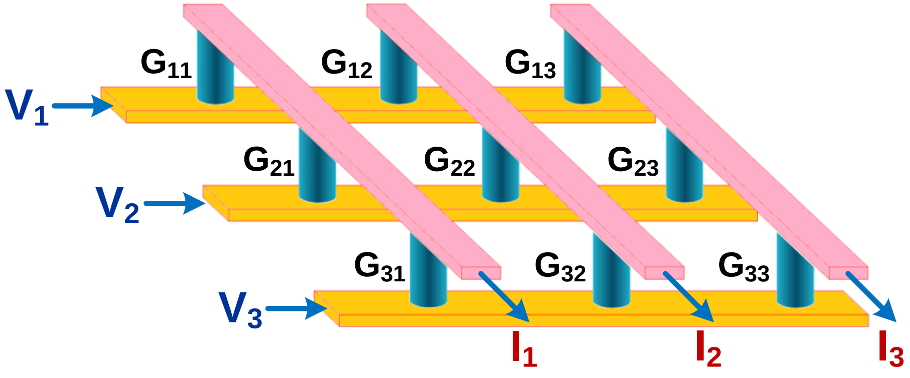
Remarks:

1. Only the first layer of the controller is 8/32-bit, whereas the remaining layers (including the last FC layer) are all binarized.
2. The similarity measure and the sharpening function are consistent during learning and inference phases, without gradient approximation during backpropagation.
3. BNN training scheme:
   1) XNOR-Net: it attempts to minimize the quantization error arising from mapping the FP weights to their quantized levels with a learnable per-channel scaling factor.
   2) RBNN: it further accounts for the angular bias between the FP and bipolar weightsand tries to minimize it during training.

Hardware: RRAM Crossbars

Analog Vector-Matrix Multiplication (VMM)





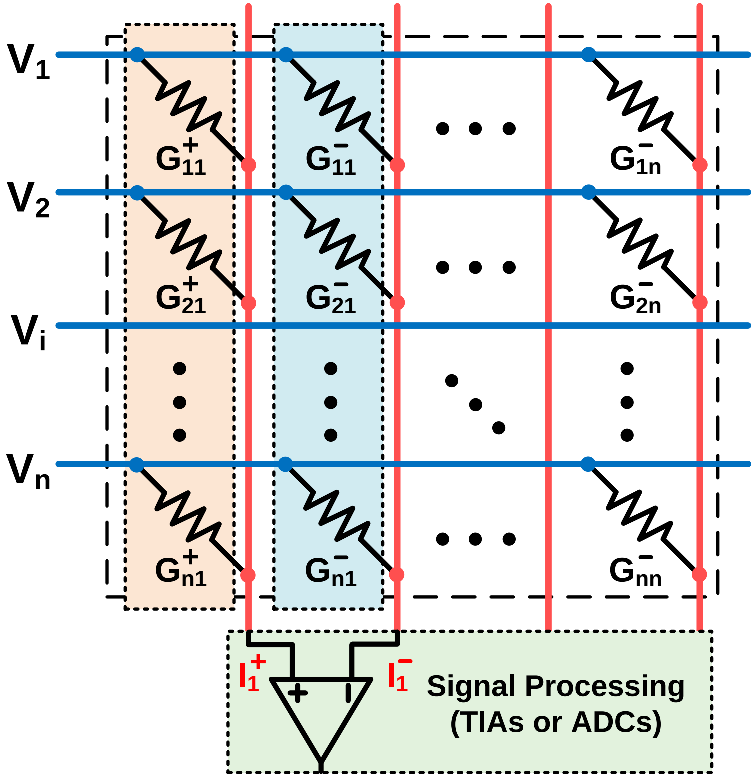
$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}^T \begin{bmatrix} G_{11} & G_{12} & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & G_{32} & G_{33} \end{bmatrix} = \begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix}^T$$

- In practice, network model pretrained on software.
- Each layer can be mapped onto RRAM crossbar.
- Device variation on the network lead to classification accuracy loss

Hardware: Bipolar Synaptic Weights Implementation (1)

Challenges:

The synaptic weights in each layer of DNN can be either positive or negative, but the conductance of an RRAM cell is always positive and cannot be programmed to be a real negative value.



Double-Column (DC) Approach

$$I_1^+ = \sum_{i=1}^{n} V_i G_{i1}^+ \quad (1a)$$

$$I_1^- = \sum_{i=1}^{n} V_i G_{i1}^- \quad (1b)$$

$$I_1^+ - I_1^- = \sum_{i=1}^{n} V_i (G_{i1}^+ - G_{i1}^-) \quad (2)$$

$$G_{ij}^+, G_{ij}^- \in [G_{\min}, G_{\max}]$$

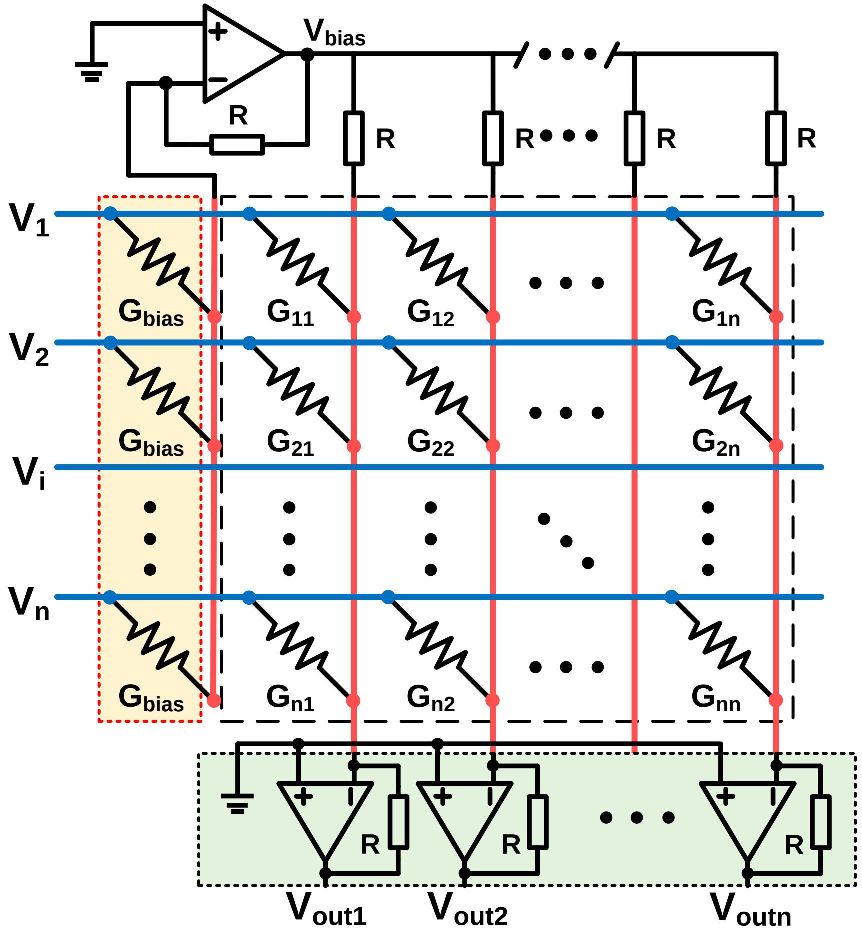
- Each synaptic weight is mapped to an RRAM-cell pair
- The subtraction of the differential pair column *(I^+-I^-)* can be obtained from *Eq.(2)* that includes *(G_{i1}^+ - G_{i1}^-)* factor
- Each conductance value will belong to a set of *[Gmin, Gmax]*

➡ bipolar synaptic weight

Hardware: Bipolar Synaptic Weights Implementation (2)

Challenges:

The synaptic weights in each layer of DNN can be either positive or negative, but the conductance of an RRAM cell is always positive and cannot be programmed to be a real negative value.



Single-Column (SC) Approach

$$V_{outj} = -(\frac{V_{bias}}{R} + \sum_{i=1}^{n} V_i G_{ij}) R \quad (3)$$

$$\text{where } V_{bias} = -\sum_{i=1}^{n} V_i G_{bias} R$$

$$V_{outj} = \sum_{i=1}^{n} V_i W_{ij} \quad (4)$$

$$\text{where } W_{ij} = (G_{bias} - G_{ij}) R$$

- Half of RRAM devices is saved
- The extra peripheral circuits and one additional bias column

1. The inputs are delivered into bias RRAM cells.
2. The opposite terminals of bias RRAMs are collected together and connected to the negative terminal of the amplifier.
3. On the back-end, the summation of the currents is delivered into another amplifier for I-to-V conversion.
4. The output of each column contains a defined W_{ij} factor

bipolar synaptic weight

Outline

- Introduction
  - AIoT & Machine Learning
  - Requirement for Structure Evolution – In-Memory Computing
- Proposed Binarized-All-Through Memory-Augmented Neural Network (MANN)
  - Software: Design Algorithm
  - Hardware: RRAM Crossbars & Bipolar Synaptic Weights Implementation
- Experimental Results
- Conclusion

Experimental Results

Controller evaluation results under different schemes:

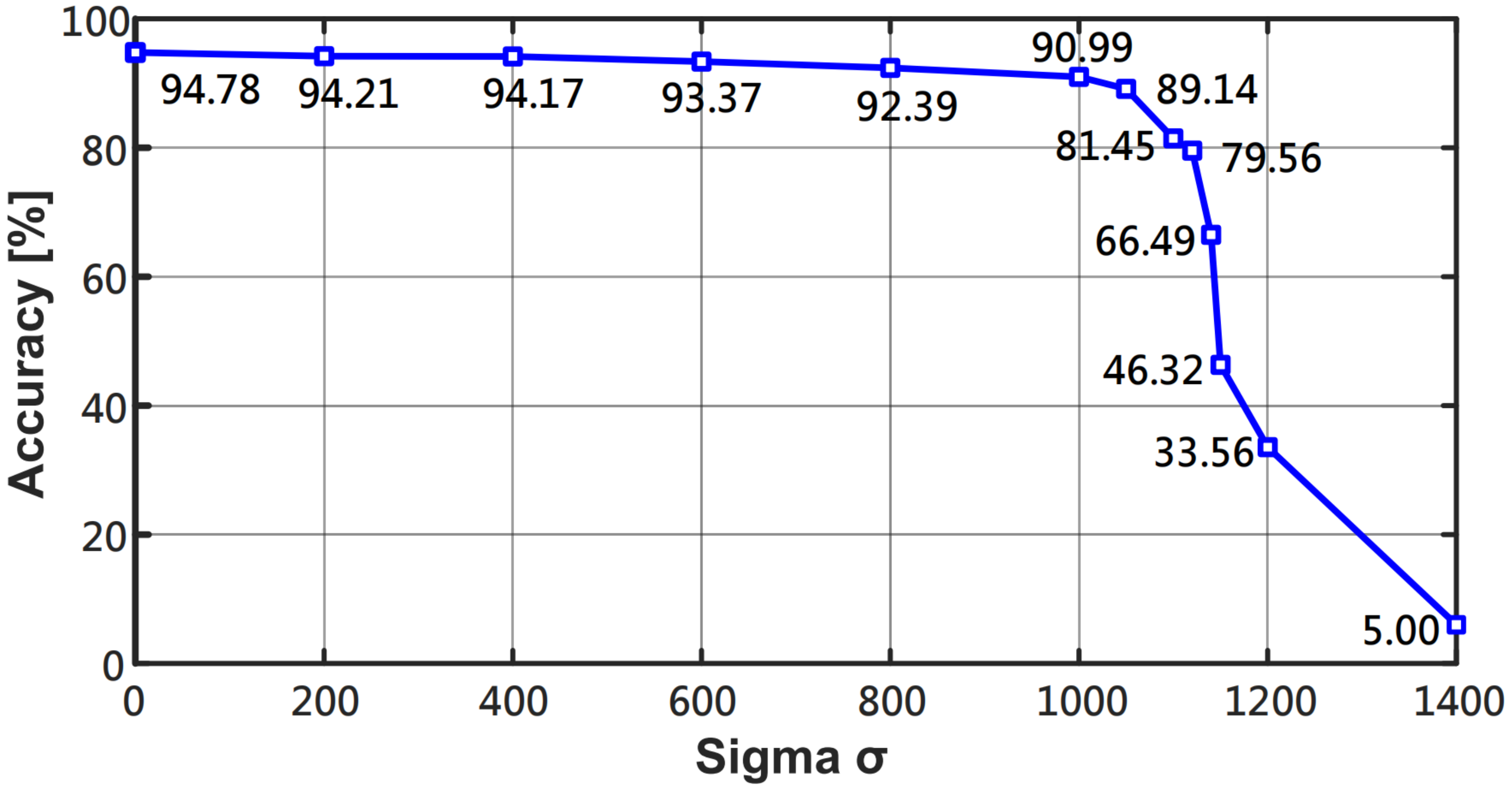
- **FP32:** full-precision controller.
- **XNOR:** binarized controller with 8-bit first CONV layer and 8-bit last FC layer training in the XNOR scheme.
- **RBNN:** binarized controller with FP32 first CONV layer and 8-bit last FC layer training in the RBNN scheme.
- **BATMANN$_X$:** binarized controller with 8-bit first CONV layer and binarized last FC layer training in the XNOR scheme.
- **BATMANN$_R$:** binarized controller with FP32 first CONV layer and binarized last FC layer training in the RBNN scheme.

No.	Learning				Inference				Acc. (%)
	Controller	Key	Similarity	Func	Controller	Key	Similarity	Func	
1	FP32	FP32	Cosine	Softabs	FP32	Bipolar	Dot	Abs	95.56
2	XNOR [9]	Bipolar	Cosine	Softabs	Bipolar	Bipolar	Dot	Abs	95.49
3	BATMANN$_X$	Bipolar	Dot	Abs	Bipolar	Bipolar	Dot	Abs	**96.53**
4	RBNN [10]	Bipolar	Cosine	Softabs	Bipolar	Bipolar	Dot	Abs	96.30
5	BATMANN$_R$	Bipolar	Dot	Abs	Bipolar	Bipolar	Dot	Abs	5.00

Best Performance!

Experimental Results

Deploy on RRAM crossbars with considering device variations.



Main parameter setting	
R_{on}	1 $k\Omega$
R_{off}	10 $k\Omega$
Dataset	Omniglot
Mapping Approach	DC
Sigma	[0, 1400]

- It shows the trend of accuracy degradation with an increased standard deviation σ from 0 to 1400 with respect to the memristance of each RRAM cell.
  (remains almost unchanged for σ up to 600, and drops sharply when σ goes beyond 1000)

Outline

- Introduction
  - AIoT & Machine Learning
  - Requirement for Structure Evolution – In-Memory Computing
- Proposed Binarized-All-Through Memory-Augmented Neural Network (MANN)
  - Software: Design Algorithm
  - Hardware: RRAM Crossbars & Bipolar Synaptic Weights Implementation
- Experimental Results
- Conclusion

Conclusion

- ✓ A binarized-all-through memory augmented neural network (BATMANN) is proposed.
- ✓ Both the encoder and memory units are end-to-end trained and realized with RRAM crossbars using simple 2-level cells.
- ✓ BATMANN provides a promising solution for in-memory AI computing on the edge application.

Acknowledgement

- This work is support in part by the General Research Fund (GRF) project 17206020, and in part by ACCESS – AI Chip Center for Emerging Smart Systems, Hong Kong SAR.

Thank you for your attention
Q&A