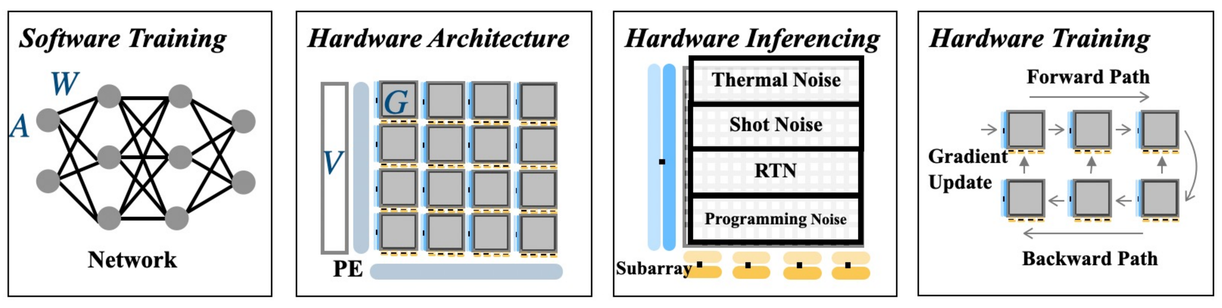
Improving the Efficiency and Robustness of In-Memory Computing in Emerging Technologies

Ph.D. Candidate: Xiaoxuan Yang
Advisors: Dr. Hai (Helen) Li, Dr. Yiran Chen
Duke University

Duke | CENTER *for* COMPUTATIONAL EVOLUTIONARY INTELLIGENCE

Background and Overview



Challenges of In-Memory Computing Systems for Neural Networks:

- Software training: existing tradeoff between **generalization and robustness**.
- Hardware architecture: some of **network functionalities** cannot be efficiently supported by existing designs.
- Hardware inferencing: device **stochastic noise** will decrease the inferencing accuracy.
- Hardware training: each training iteration will rewrite the cells on crossbar and may **wear out** the hardware.

My Ph. D. works and contributions:

- Generalized algorithm enhances robustness against weight perturbation.
- Architecture design enables efficient Transformer in PIM system.
- Systematic framework builds robust and efficient PIM System.
- Hardware-software co-design helps reliable in-memory training design.

Develop Robust Preserving Optimization [DAC' 22]

Highlights:
- Unify and improve generalization and quantization performance under bounded weight perturbation.

Methods:
- Hessian-enhanced regularization optimization (HERO)

Hessian eigenvalue regularization

$$L_r^i(W^i) = ||\nabla L(W^i + hz^i) - \nabla L(W^i)||^2,$$
$$z^i = \frac{W^{i^2}}{||W^i||_2}\frac{\nabla L(W^i)}{||\nabla L(W^i)||_2}.$$

- Finite difference approximation along high curvature direction
- Adaptive perturbation strength across different layers

Gradient of regularization

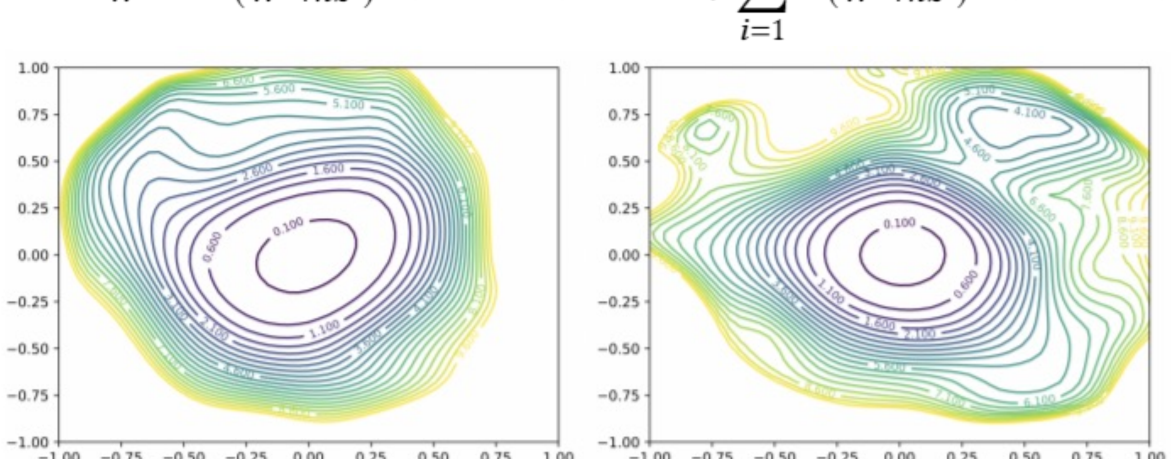
$$G(U) := ||\nabla L(U) - \nabla L(W^i)||^2$$

$$\nabla L_r^i(W^i) = \nabla_{(W^i+hz^i)} G(W^i + hz^i) \cdot \nabla_{W^i}(W^i + hz^i)$$
$$\approx \nabla_{(W^i+hz^i)} G(W^i + hz^i).$$

- Approximate L_r gradient
- Apply SAM gradient as first-order regularization

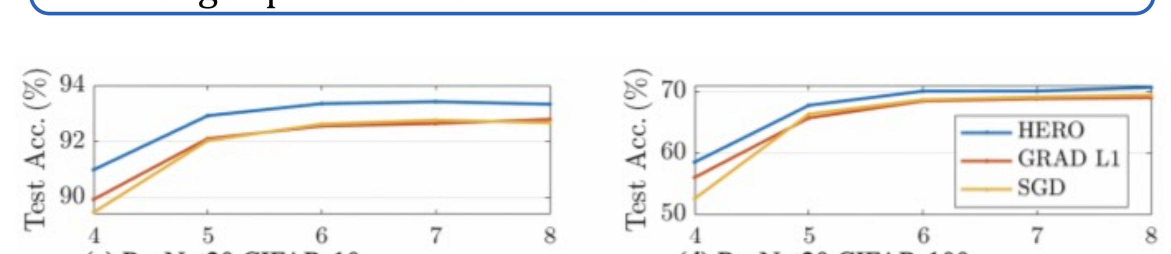
Overall optimization step

$$\nabla_{W^i} = \nabla_{(W^i+hz^i)} L(W^i + hz^i) + \alpha W + \gamma \sum_{i=1}^{N} \nabla_{(W^i+hz^i)} G(W^i + hz^i),$$



(a) HERO, test accuracy: 93.44% (b) SGD, test accuracy: 92.82%

- HERO generates smoother loss surface and provides stronger perturbation tolerance.

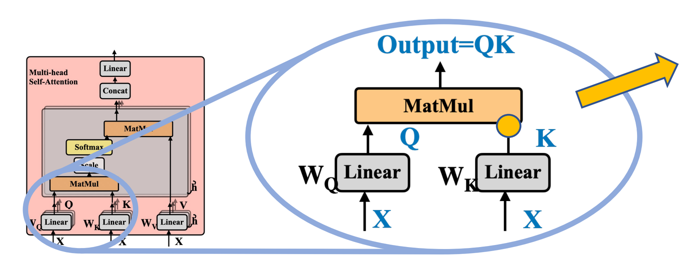


- HERO has better testing accuracy.
- HERO exhibits higher robustness against post-training quantization.

Enable Efficient Transformer with PIM [ICCAD' 20]

Highlights:
- Accelerate the scaled dot-product attention of Transformer using ReRAM-based PIM with ReTransformer design.
- Eliminate some data dependency by avoiding writing intermediate results using the proposed matrix decomposition technique.

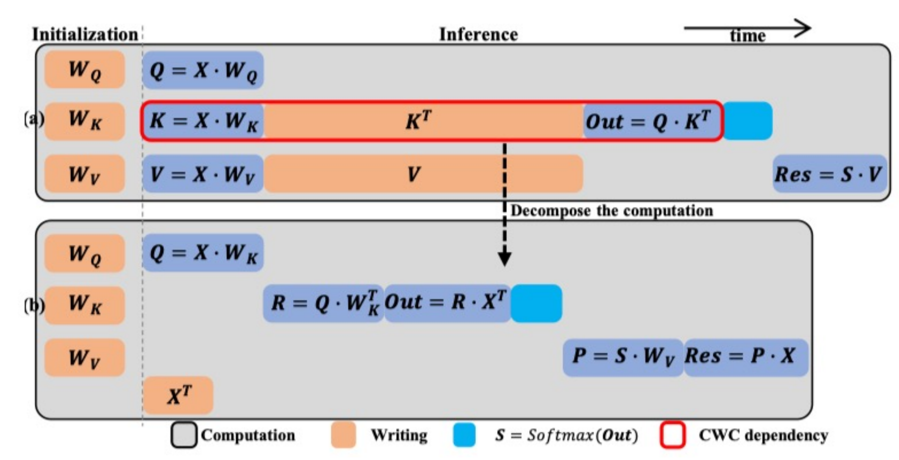
Challenge: MatMul layer deals with results from the previous step.



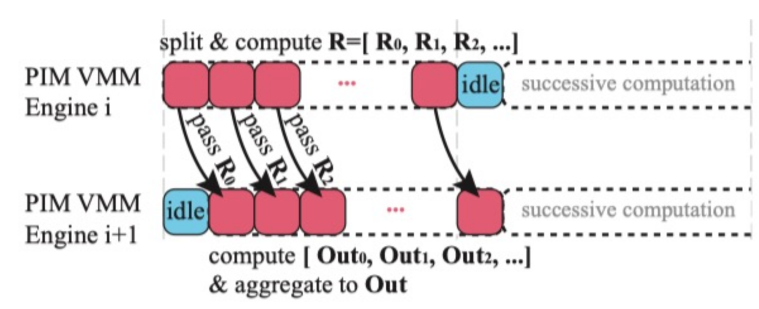
Write intermediate results onto ReRAM.

Negative factor to latency, energy, and endurance.

Optimized MatMul: use matrix decomposition in scaled dot-product attention to eliminate the data dependency and reduce the computation latency.



Sub-matrix pipeline: design a fine granularity for Transformer inference.



✓Compared to GPU and Pipelayer, ReTransformer improves computing efficiency by 23.21× and 3.25×, respectively.

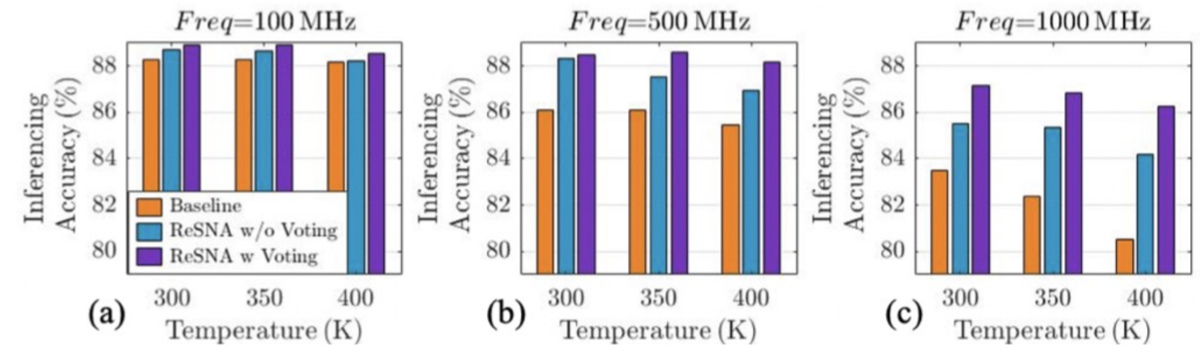
Explore Robust and Efficient PIM System [ICCAD' 21]

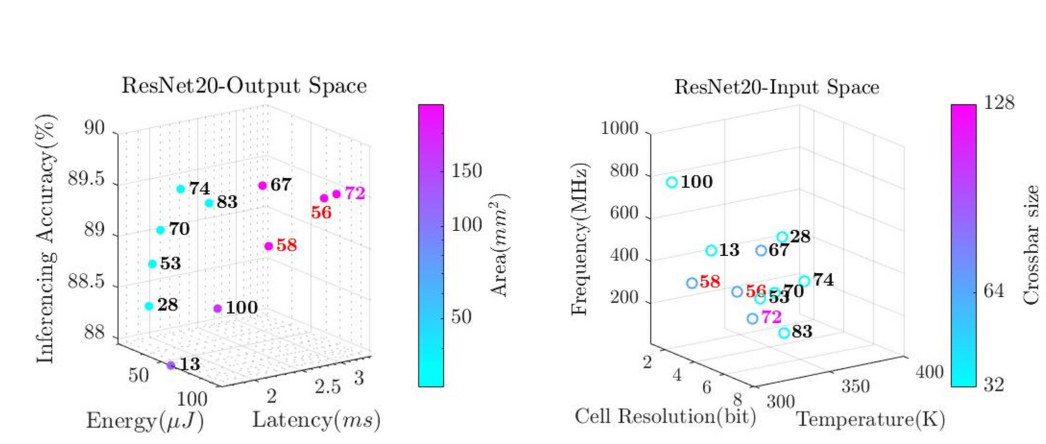
Highlights:
- Achieve high inferencing accuracy under stochastic noise.
- Effectively explore area-, energy-, latency-efficient designs.

Methods:
- ReSNA: ReRAM-based **Stochastic-Noise-Aware** Training
- CF-MESMO: **Continuous Fidelity** Max-value Entropy Search Multi-objective Optimization

Cell Resolution	*Programming Noise*	*Inference Accuracy*
Crossbar Size	*Thermal Noise*	*Area*
Frequency	*Shot Noise*	*Energy*
Temperature	*RTN*	*Latency*

MO Input, **Fidelity Selection** → Evaluation Process → MO Objectives → Next Candidate MO Input, **Fidelity Selection**





✓Our CF-MESMO is capable of finding the Pareto optimal results.
✓We can avoid high-latency or high-energy design based on our criteria and budget.
✓From the Pareto set, we can see that high-cell resolution setting or high-frequency setting appears in the Pareto front due to the ReSNA method.

Build Endurance-aware Training [TCAD In Revision]

Highlights:
- Propose ESSENCE framework with an endurance-aware structured stochastic gradient pruning method.
- Dynamically adjust the probability of gradient update based on the current update counts to reduce the number of rewrite accesses.

- Row-wise or column-wise update
- Reduce the update probability if the gradient amplitude is low
- Reduce the update probability if the existing number of update in that row/column is large.

The expectation of the endurance-aware structured stochastic pruning gradient remains an **unbiased** gradient towards the minimization target

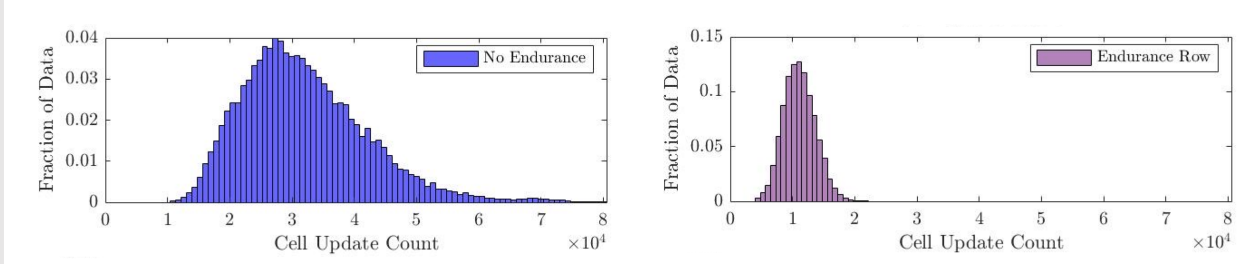
EXPERIMENT RESULTS ON THE CIFAR-10 DATASET.

(a) ResNet20

Methods	Mean Update Counts (Savings)	Accuracy
SGD	117.30×10^3 (1×)	90.67%
No Endurance	33.26×10^3 (3.52×)	90.49%
Essence Row	11.45×10^3 (10.24×)	90.51%
Essence Column	11.40×10^3 (10.29×)	90.93%

(b) VGG19-BN

Methods	Mean Update Counts (Savings)	Accuracy
SGD	117.30×10^3 (1×)	92.66%
No Endurance	25.64×10^3 (4.57×)	92.43%
Essence Row	11.45×10^3 (10.24×)	92.61%
Essence Column	11.58×10^3 (10.13×)	92.41%



Update count distribution in the last convolution layer in the ResNet20 network: Left: No Endurance, Right: Essence Row.

Summary

My works covers the following topics:
- design efficient PIM-based architecture for state-of-the-art models.
- guarantee the performance under the noise of the real hardware.
- enable the reliable and durable PIM-based hardware training.

My works contribute to the goal of achieving efficient and robust PIM designs and implementations with algorithmic/architectural/systematic innovations.

Reference

Publications presented in this poster:

[1] **X. Yang***, H. Yang*, N. Gong, and Y. Chen, "HERO: hessian-enhanced robust optimization for unifying and improving generalization and quantization performance," in DAC 2022.
[2] **X. Yang**, B. Yan, H. H. Li, and Y. Chen, "ReTransformer: ReRAM-based processing-in-memory architecture for transformer acceleration," in ICCAD 2020.
[3] **X. Yang**, S. Belakaria, B. K. Joardar, H. Yang , J. R. Doppa, P. P. Pande, K. Chakrabarty, and H. H. Li, "Multi-objective optimization of ReRAM crossbars for robust DNN inferencing under stochastic noise," in ICCAD 2021.
[4] **X. Yang**, H. Yang, J. R. Doppa, P. P. Pande, K. Chakrabarty, and H. H. Li, "ESSENCE: Exploiting structured stochastic gradient pruning for endurance-aware ReRAM-based in-memory training systems," TCAD in Revision.

Other publications on this topic:

[5] **X. Yang**, C. Wu, M. Li, and Y. Chen, "Tolerating noise effects in processing-in-memory systems for neural networks: A hardware–software codesign perspective," Adv. Intell. Sys. 2022.
[6] C. Wu, **X. Yang**, H. Yu, R. Peng, I. Takeuchi, Y. Chen, and M. Li, "Harnessing optoelectronic noises in a photonic generative network," Sci. Adv. 2022.
[7] C. Wu, **X. Yang**, Y. Chen, and M. Li. "Photonic Bayesian Neural Network using Programmed Optical Noises." JSTQE Under review.

I am seeking a job as a tenure track faculty member in 2022-2023.