Temperature-Resilient RRAM-based In-Memory Computing for DNN Inference

[1]Jian Meng, [2]Wonbo Shim, [1]Li Yang, [1]Injune Yeo, [1]Deliang Fan, [2]Shimeng Yu, [1]Jae-sun Seo

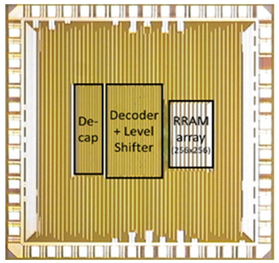
[1]Arizona State University, [2]Georgia Institute of Technology

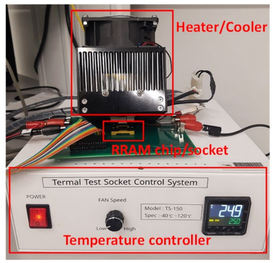
Introduction

Resistive RAM (RRAM) based in-memory computing (IMC) has emerged as a promising paradigm for efficient deep neural network (DNN) acceleration. However, the multi-bit RRAMs often suffer from non-ideal characteristics such as drift and retention failure against temperature changes, leading to significant inference accuracy degradation. In this work, we present a new temperature-resilient RRAM-based IMC scheme for reliable DNN inference hardware.

On-chip RRAM Characteristics Measurement

RRAM prototype chip and Temperature control equipment





On-chip RRAM Conductance Measurement

- **90nm RRAM chip fabricated by Winbond©**
- **Array size:** 256×256
- **Device type:** 1T1R HfO_2 based 2-bit-per-cell RRAM.

Measurement setup:

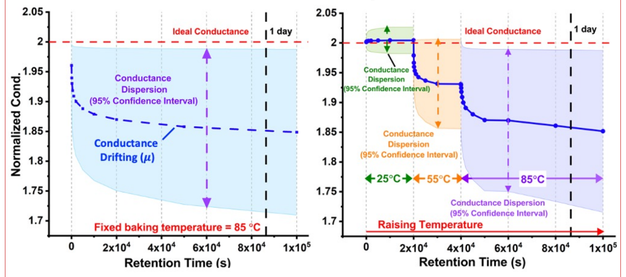
- **Temperature measurement:** National Instrument© PXIe system.
- **Temperature Control:** TS-150 equipment from Semicon Advance Technology (SAT©)
- **Measurement temperature:** 25°C to 120°C
- **Measurement time:** 20 to 80,000 seconds (1 day).

Measurement outcomes:

- Raw RRAM conductance values.
- **Thermal characteristics + other non-ideal effects** (e.g., random telegraph noise).

Temperature-Dependent RRAM Conductance Modeling

Modeled Static retention variation and Dynamic retention variation



Static Retention Variations

Based on the measurement results, the retention variation can be characterized as the average **conductance drifting μ** and **standard deviation σ.** For the given temperature K and retention time t, we have:

$$\Delta\mu^K = \mu^K(t) - \mu^K_{init} = A^K_\mu \times \log t$$
$$\Delta\sigma^K = \sigma^K(t) - \sigma^K_{init} = B^K_\sigma \times \log t$$

Drifting rates A^K_μ and B^K_σ are modeled based on the linear regression from the chip measurements.

Dynamic Retention Variations

If the temperature is increased from K_1 to K_2, it means that the initial condition at K_2 is the variation at the transition point from K_1:

$$\Delta\mu^{K_2} = \mu^{K_2}(t) - \mu^{K_2}_{init} = A^{K_2}_\mu \times \log(t + T') - A^{K_2}_\mu \times \log T'$$
$$T' = 10^{\Delta\mu^{K_1}_T / A^{K_2}_\mu}$$

Challenges of noise injection training

1. Efficiently injecting the cell-level noises to DNN

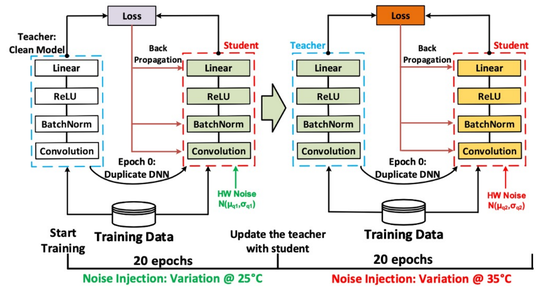
Decompose the low precision weights then inject the noise will largely slow down the training process.

2. The "amnesia" of DNN model

Training the model while injecting the selected noise can only recover the inference performance at the corresponding temperature and time.

Proposed Temperature-Resilient IMC Framework

Phase 1: Progressive Knowledge Distillation



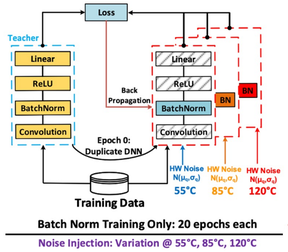
Phase 1: Progressive Knowledge Distillation (PKD)

Initial step: Injecting the low-temperature noises to the student model while the clean model is employed as a teacher

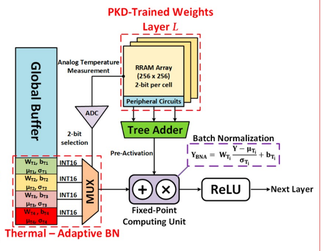
Noise injection: Convert the cell-level noises to the weight levels noise → Noise injection during quantization

Progressive distillation: Injecting the high temperature noises to the student while using the previous student model (low-temperature noise injected) as the new teacher

Phase 2: BatchNorm Adaptation



Phase 3: Hardware Implementation



Phase 2 & 3: BatchNorm Adaptation & HW Implementation

Initial step: Inherit the PKD-trained model then freeze the all the parameters updating process **except** the BatchNorm.

Noise injection: Injecting the high temperature noises to the student model.

Hardware Implementation:

Individually trained BN for different high temperatures.

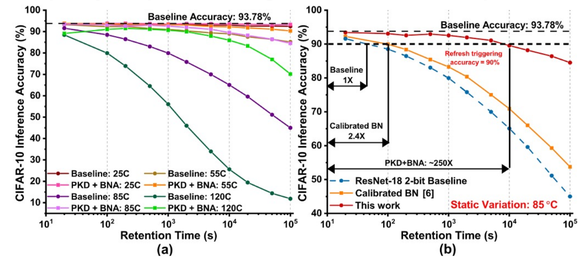
Switch to the corresponding BN parameters when the threshold temperature reached.

Experimental Results

Experimental Setup

- **Operating temperature:** 25°C to 120°C
- **Operating time:** 20 sec to 1×10^5 sec
- **Model & Dataset:** 2-bit ResNet-18 on CIFAR-10 dataset
- **HW Inference:** 256 ×256 1T1R HfO_2 based 2-bit per cell RRAM implemented by NeuroSim (Peng, IEDM, 2019)

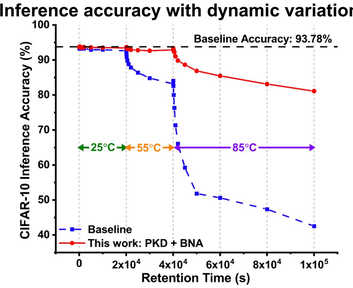
(a) Inference accuracy with static variations (b) Accuracy and refresh frequency comparison among the baseline model and prior work



Experimental Results: Static Thermal Variation

- **Over 40% accuracy improvements**, compared with 2-bit ResNet-18 baseline model.
- **Around 250X and 100X refreshing frequency reduction** compared to the baseline and prior work (Tsai, arXiv, 2020).

Inference accuracy with dynamic variations



Experimental Results: Dynamic Thermal Variation

- **Over 45% accuracy improvements**, compared with 2-bit ResNet-18 baseline model.

Acknowledgement

We thank Winbond Electronics for RRAM chip fabrication support. This work is partially supported by JUMP CBRIC, JUMP ASCENT, SRC AIHW program, and NSF grants 1652866/1715443/1740225.