Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M22-262P
Published: 5/2/2023

Inventors

Jae-Sun Seo
Jian Meng
Li Yang
Deliang Fan

Contact

Physical Sciences Team

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

Deep neural networks (DNNs) have shown extraordinary performance in recent years for various applications, including image classification, object detection, speech recognition, etc. Accuracy-driven DNN architectures tend to increase model sizes and computations in a very fast pace, demanding a massive amount of hardware resources. Frequent communication between the processing engine and the on-/off-chip memory leads to high energy consumption, which becomes a bottleneck for the conventional DNN accelerator design. To overcome such challenges, in-memory computing (IMC) has been proposed as a promising solution for energy-efficient DNN acceleration.

Regarding memory technologies for IMC, SRAM and DRAM are both volatile and suffer from leakage power in CMOS devices. These disadvantages promote nonvolatile memory as an attractive solution for IMC-based DNN acceleration. Resistive random-access memory (RRAM), a nonvolatile memory, can store multiple levels in one cell, resulting in dense storage as well as high multiply-and-accumulate (MAC) throughput. On the other hand, having a high amount of computation in a small area can increase the power density, which can, in turn, elevate the temperature. When the temperature increases, the ability to hold the programmed values becomes weaker and the RRAM conductance starts to drift away. Such variation will affect the macro-level IMC results, layer-by-layer computations, and eventually the final output of the DNN, leading to incorrect inference predictions. There is a need for a temperature-resilient solution for RRAM-based IMC that considers critical retention failure issues.

Researchers at Arizona State University have developed a temperature-resilient scheme for resistive random-access memory (RRAM) in-memory computing (IMC) that considers critical retention failure issues against temperature variations. This scheme includes a deep neural network (DNN) training algorithm that can efficiently recover the hardware inference accuracy against temperature variations. The scheme also includes a system-level hardware design to resolve temperature-dependent retention issues with one-time DNN deployment. Without retraining or updating any DNN weights after the initial RRAM programming, this system largely improves the inference accuracy across all experiments and enhances the robustness of RRAM hardware against a wide range of temperature differences.

Related publication: Temperature-Resilient RRAM-Based In-Memory Computing for DNN Inference

Potential Applications:

- Deep neural networks (DNNs)

- Resistive random-access memory (RRAM) devices

Benefits and Advantages:

- New DNN training algorithm and system-level hardware design for RRAM-based IMC
  - Training algorithm considers thermal-changes and time-variations of conductance drifting, leading to highly robust DNN models
  - Thermal-aware RRAM-based inference engine design
- Experimentation on a 256 x 256 RRAM array with the circuit-level benchmark simulator NeuroSim produced robust RRAM IMC-based DNN inference with >30% CIFAR-10 accuracy and >60% TinyImageNet accuracy with temperature variations