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:M23-094P^

Published: 3/18/2024

Inventors

Ayushi Dube
Ankit Wagle
Gian Singh
Sarma Vrudhula

Contact

Physical Sciences Team

In-Memory Hardware-Software Co-design for Image Processing

A substantial part of high energy consumption (> 60%) and large latency (> 90%) of conventional von-Neumann architectures can be attributed to the unavoidable data movement between the processor and main memory (DRAM). This is perhaps the major limiting factor for big data and machine learning applications whose usage is permeating into practically all domains of computing. This has reinvigorated and accelerated the development of processing in-memory (PiM) architectures, which involve integrating processing elements inside the memory architecture so that data can be processed in-place, avoiding data transfers between processor and memory.

Image filtering tasks belong to a category of error tolerant applications because the precision of the computation can be reduced without substantially degrading that perceived quality of the image or even affecting any subsequent decision that is made based on the image. For instance, cameras used for traffic surveillance could operate in a low precision mode to detect objects like cars, and subsequent processing could operate in high precision mode to capture details such as a license plate number. The ability to dynamically switch between different levels of precision is extremely valuable in edge computing with smart cameras. Current technologies describe computing architectures that include hardware for both exact and approximate computations and control logic (software) that allows for switching between the two at runtime. Although current technologies suffer from substantial penalties incurred in area, latency, and energy for switching.

Researchers at Arizona State University have developed a hardware-software co-design consisting of a processing in-memory (PiM) architecture with embedded neural processing elements that are highly reconfigurable. This PiM architecture allows for precision tuning without incurring area, latency, or energy overhead.

Related publication: Tunable Precision Control for Approximate Image Filtering in an In-Memory Architecture with Embedded Neurons

Potential Applications:

- Image processing operations (e.g., image filtering, digital image filtering)

Benefits and Advantages:

- User provided fine-tuned dynamic control
- Peak Signal to Noise Ratio (PSNR, output quality metric for image filtering applications) increased from 25 dB to 50 dB without incurring any extra cost in terms of energy or latency
- Maximum improvement in energy efficiency and throughput is 2X, gains in energy efficiency against a MAC-based Processing Element array is 3X-6X and

corresponding throughput is 2.26X-4.52X