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:M21-070P
Published: 2/7/2022

Inventors

Jae-Sun Seo

Shihui Yin

Mingoo Seok

Bo Zhang

Contact

Shen Yan
shen.yan@skysonginnovations.com

Programmable In-Memory Computing Accelerator for Low-Precision Deep Neural Network Inference

-Background In the era of artificial intelligence, various deep neural networks (DNNs), such as multi-layer perceptron, convolutional neural networks, and recurrent neural networks, have emerged and achieved human-level performance in many recognition tasks. These DNNs usually require billions of multiply-and-accumulate (MAC) operations, soliciting energy-efficient and high-throughput architecture innovation for on-device DNN workloads. Among a variety of solutions, in-memory computing (IMC) has widely attracted research interests, owing to high computation parallelism, reduced data communication, and energy-efficient analog accumulation for low-precision quantized DNNs. Single-macro-level or layer-level IMC designs have been recently demonstrated with high energy efficiency. However, due to the limited number of IMC macros integrated on-chip, it is difficult to evaluate system-level throughput and energy efficiency. Recent works hard-wired the data flow of both IMC and non-IMC operation, exhibiting limited flexibility to support layer types other than batch normalization and activation layers. To complicate matters further, hardware loop support is often omitted, incurring large overhead in latency and instruction counts. Invention Description Researchers at Arizona State University and Columbia University have developed a programmable in-memory computing (IMC) accelerator for low-precision deep neural network inference. A key feature of the IMC accelerator is its ability to integrate over 100 capacitive-coupling-based IMC static random-access memory (SRAM) macros and demonstrate large-scale integration of IMC SRAM macros. In addition, a flexible single-instruction-multiple-data (SIMD) processor with a custom instruction set architecture is designed/integrated to support a range of vector operations beyond multiply-and-accumulate, enabling the programmable IMC accelerator to execute various layers such as pooling and residual layers. The instruction set architecture also features hardware loop support, largely reducing instruction count and latency. Potential Applications • Deep neural networks • In-memory computing accelerators Benefits and Advantages • Programmable DNN inference accelerator integrates a very large number of in-memory computing (IMC) SRAM macros, which eliminates the need for reloading weights for small networks or allows hiding of the weight reloading latency • Instruction set architecture (ISA) is designed for pipelined IMC and SIMD processors, supporting more layer structures such as max/average pooling layers, addition with short-cut layers, 5x5 convolution layers, and stride other than 1 • ISA features hardware loop support, reducing instruction count and latency Related Publication: PIMCA: A 3.4-Mb Programmable In-Memory Computing Accelerator in 28nm for On-Chip DNN InferenceResearch Homepage of Professor Jae-sun Seo