Oral S13

Emerging Memory and Computing Technologies for AI Accelerators

Date/Time	8/3(四)　11:30-12:30
Chair(s)	李宇軒／元智大學電機工程學系 林書彥／元智大學電機工程學系

S13.1　│ 11:30—11:45

Reliable and High Accurate Stochastic Computing Architecture for Artificial Neural Networks

Kun-Chih (Jimmy) Chen[1] and Wei-Ren (Tony) Syu[2]

[1] *Institute of Electronics, National Yang Ming Chiao Tung University*

[2] *Department of Computer Science and Engineering, National Sun Yat-Sen University*

In recent years, security issue has become one of the major consideration when designing hardware implementation. However, hardware security often comes with increment in hardware cost, so it is necessary to trade-off between security considerations and hardware cost. Therefore, the fault tolerant computing approach with low hardware overhead is now receiving increasing attention. The Stochastic Computing (SC) has been proven as a low hardware overhead and fault tolerant computing approach. Thus, many SC-based Artificial Neural Network (ANN) designs were proposed in recent years. However, due to the stochastic bit stream computing, the conventional SC-based architectures suffer from low computing accuracy. On the other hand, the most common security issue on the hardware security is tempering attack. As a result, it is suitable to apply stochastic computing approach to implement the high reliable hardware architecture. In this work, we propose a novel adder and input data pre-processing method to improve the accuracy of conventional SC-based ANN design. Compared with the conventional binary architecture, the proposed SC-based ANN architecture can maintain the accuracy of 94% even when the rate of tampering attack on neurons in hidden layer is 30%, while the accuracy of the conventional binary architecture dropped to 51%. Moreover, in terms of power consumption and area, compared to conventional binary architecture, the proposed SC-based ANN architecture can reduce power consumption by 76% and area cost by 92%.

S13.2 | 11:45—12:00

CIMR-V: An End-to-End SRAM-based CIM Accelerator with RISC-V for AI Edge Device

Yan-Cheng Guo[1], Tian-Sheuan Chang[1], Chih-Sheng Lin[2], Bo-Cheng Chiou[2], Chih-Ming Lai[2], Shyh-Shyuan Sheu[2], Wei-Chung Lo[2] and Shih-Chieh Chang[2]
[1] Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2] Industrial Technology Research Institute

Computing-in-memory (CIM) is renowned in deep learning due to its high energy efficiency resulting from highly parallel computing with minimal data movement. However, current SRAM-based CIM designs suffer from long latency for loading weight or feature maps from DRAM for large AI models. Moreover, previous SRAM-based CIM architectures lack end-to-end model inference. To address these issues, this paper proposes CIMR-V, an end-to-end CIM accelerator with RISC-V that incorporates CIM layer fusion, convolution/max pooling pipeline, and weight fusion, resulting in an 87.96% reduction in latency for the keyword spotting model. Furthermore, the proposed CIM-type instructions facilitate end-to-end AI model inference and full stack flow, effectively synergizing the high energy efficiency of CIM and the high programmability of RISC-V. Implemented using TSMC 28nm technology, the proposed design achieves an energy efficiency of 2546.67 TOPS/W and 26.21 TOPS at 50 MHz.

S13.3 | 12:00—12:15

Optimizing RRAM-Based Neural Network Accelerators with A Variation-Aware Framework Considering Practical Designs

Fang-Yi Gu, Cheng-Han Yang, Ing-Chao Lin
Department of Computer Science and Information Engineering, National Cheng Kung University

Emerging resistive random access memory (RRAM) has garnered significant interest in computing-in-memory (CIM) due to its high efficiency in multiply-accumulate operation (MAC), which is a key computation in Neural Networks (NN). However, RRAM cells suffer from variations caused by imperfect fabrication, resulting in deviations from target values. This significantly degrades the accuracy of RRAM-based NN accelerators. Additionally, practical hardware designs of RRAM-based NN accelerators require high-resolution analog-to-digital converters (ADCs) when activating a large number of wordlines and bitlines in a crossbar array, leading to increased power consumption. In this paper, we propose a unary-based non-uniform quantization method to mitigate variations and consider practical hardware design aspects. Our method quantizes the NN model, ensuring that each RRAM cell has equivalent significance to mitigate the impact of variations. Subsequently, we introduce a variation-aware operation unit (OU) based framework. In this framework, only the RRAM cells in the same OU are activated

simultaneously, eliminating the need for a high-resolution ADC. Furthermore, we present three methods: OU skipping, OU recombination, and OU compensation, to further mitigate the impact of variations. Experimental results demonstrate that our methods exhibit lower accuracy loss under variations compared to the state-of-the-art methods across four NN models (AlexNet, VGG16, ResNet34, and ResNet50) on two datasets (CIFAR-10 and CIFAR-100) with cell resolutions ranging from 2 bits to 4 bits.

S13.4 | 12:15—12:30

Defect-Aware Weight Mapping Framework for Reliable and Efficient ReRAM-Based Neuromorphic Computing on Edge Devices

Fang-Yi Gu, Yi-Hong Hsieh, Ing-Chao Lin
Department of Computer Science and Information Engineering. National Cheng Kung University

ReRAM (Resistive Random-Access Memory) is a non-volatile resistive memory with the potential to perform Vector Matrix Multiplication (VMM), a time-consuming operation in neuromorphic computing, using ReRAM crossbars. The demand for neuromorphic computing on edge devices has led to the widespread adoption of ReRAM-based compute-in-memory systems (RCS) due to their ability to significantly reduce computing time and energy consumption. However, reliability issues, such as stuck-at faults (SAFs) and resistance variations, can impact the performance of ReRAM in neural network computing. To address these challenges, we propose a defect-aware weight mapping framework comprising two methods: SAF-aware value mapping and readjustment, and variation-aware value decomposition and finetuning. These methods leverage defective ReRAM cells, eliminate the need for redundant hardware, and achieve fault tolerance without requiring model retraining. Experimental results demonstrate that our proposed framework effectively mitigates the inference accuracy loss, reducing it from 76% to just 16% for ResNet50 on the Cifar-100 benchmark. These results are obtained considering a variation σ of 1.0 and approximately 10% of SAF ReRAM cells.