Oral S07

Artificial Intelligence Circuits and Systems (II)

Date/Time 8/4 (四) 11:00-12:00

Chair(s) 黃柏蒼教授 / 國立陽明交通大學國際半導體產業學院
湯松年教授 / 中原大學資訊工程學系

S07.1 11:00 — 11:12

A Multi-Precision Neural Network Inference Acceleration System Based on FPGA

Yu-Tung Liu, Guo-Yang Zeng, and Tzi-Dar Chiueh
Graduate Institute of Electronics Engineering, National Taiwan University

Neural networks have achieved excellent results in many fields, such as image recognition, natural language processing, etc. With the increasing penetration of intelligent edge devices, fast and efficient neural network inference has become the focus of many research groups. This paper proposes a mixed-precision neural network architecture and low-power circuit implementation. The accelerator circuit supports weights in three precisions and achieves hardware usage flexibility through the im2col algorithm. Finally, we integrated the algorithm and hardware control flow with PyTorch to create a complete solution for users to perform quantized NN training and FPGA acceleration deployment under one framework. At a clock speed of 150MHz, our U250 FPGA accelerated system is 10.57 times more energy-efficient than the CPU. Another DLA solution using the SoC-based KV260 FPGA module is portable and consumes only a few Watts and achieves similar higher energy efficiency.

S07.2 11:12 — 11:24

An Integration-Friendly CNN Accelerator Architecture with High Utilization and Scalability

Chia-Heng Hu, I-Hao Tseng, Pei-Hsuan Kuo, Yu-Hsiang Huang, and Juinn-Dar Huang
Institute of Electronics, National Yang Ming Chiao Tung University

In this paper a highly scalable VLIW-driven CNN accelerator architecture is proposed. A new VLIW instruction, which specifies all settings of an entire convolution layer and natively supports layer concatenation, is defined. A multi-mode input aligner (MMIA) is developed to efficiently organize input data for various convolution modes. A zero-initial-latency (ZIL) buffer is created to further boost the performance. A strip-based dataflow is adopted to drastically minimize external DRAM accesses. The accelerator is also equipped with an AXI4 on-chip bus interface, an instruction queue, ping-pong DRAM I/O buffers, and is thus ready for efficient and easy SoC integration. An accelerator instance with 576 MACs has been implemented using TSMC 40nm process. The core logic only requires 490K gates and the total internal memory size is merely 286KB. The peak performance is 1440 GOPS @1.25GHz and the core power efficiency is 8.71 TOPS/W. Moreover, the proposed accelerator has also enabled a real-time image semantic segmentation system for autonomous driving on an FPGA system.

S07.3 11:24 — 11:36

A Linear Quantization Training Method for Hardware Constraints of In-Memory Computing Architecture

Hao-Wen Kuo, Zhaofang Li, Yu-Hsiang Cheng, Shih-Ting Lin, Meng-Fan Chang, and Kea-Tiong Tang
Department of Electrical Engineering, National Tsing Hua University

Analog computing is used in in-memory computing (IMC) to improve the energy efficiency of accelerators. However, because of hardware constraints, such as the interleaved structure of the components, and nonideal effects caused by the peripheral circuits, the performance of neural networks is hindered. Therefore, a five-dimensional searching direction method for backpropagation is proposed that can increase top-1 accuracy by 1.62% after quantization in ResNet-50 on ImageNet; in addition, a matrix-vector multiplication quantizer and deviation estimator are proposed that reduce the noise caused by nonideal effects and further improve the top-1 accuracy of ResNet-50 by 1.72% and of Mobile-Net v2 by 21.08%. The proposed methods improve the performance and show state-of-the-art results of neural networks in IMC environments.

S07.4 11:36 — 11:48

Tiling Strategy for CNN Zero-Skipping Accelerator

Yen-Xun Chen, Chih-Hung Kuo, Shi-An Zhan, and Kuan-Hung Chen
Department of Electrical Engineering National Cheng Kung University

For AI edge devices, input data should be partitioned into smaller tiles due to limited on-chip memory. Different tile sizes significantly affect data access times and on-chip memory size. In this paper, we explore different tile strategies to find the appropriate one with the best energy efficiency. We design a Zero-Skipping Accelerator (ZSA) that adopts row stationary data flow to reduce the data movement and achieve highly parallel computing. The Pre-processing module handles sparse features to save energy. Experiments show that a suitable tile size can achieve better performance and energy efficiency.

S07.5 11:48 — 12:00

Static Effective Weight Convolution for Deep learning Accelerators

Tz-Yuang Su, Chun-Yuan Chen, and Tian-Sheuan Chang
Dept. of Electric Engineering, National Yang Ming Chiao Tung University

Eliminating unnecessary operations with effective weight-based convolution (EWC) is a promising approach to accelerate convolutional neural network (CNN) processing. However, a previous approach with a dynamic search for effective weights results in complex control and a high area overhead. This paper presents a static effective weight-based convolution (SEWC) that uses four static effective weights instead of six dynamic effective weights to enable simple control and offline decomposition. Furthermore, the control overhead is reduced with simple weight information encoding. The multiplication after decomposition is transformed to simple addition and shift with a modified Booth's algorithm to completely remove multiplications. The experimental results show that this approach can reduce 32.3% of the area and 36.4% of the power consumption for the PE implementation compared to the previous design.