Oral S01

Advanced Neural Network Processors and Systems

Date/Time 8/3 (三) 13:30-14:30

Chair(s) 李國君教授 / 成功大學電機工程學系

S01.1 13:30 — 13:45

An Ultra-Low-Power Neural Signal Processor for Seizure Prediction

Yi-Yen Hsieh, Yu-Cheng Lin, and Chia-Hsiang Yang
Graduate Institute of Electronics Engineering, National Taiwan University

This work proposes the world's first integrated neural signal processor for closed-loop neuromodulation. The area cost of the energy operator is reduced by 28% with an approximated energy operator (AEO). The proposed scaling-based Newton-Raphson divider achieves a 2.7x higher convergence speed. For the alternating direction method of multipliers (ADMM)-based SVM training, the proposed pointer-based matrix multiplication (PBMM) reduces 99.9% of operations. With LDL decomposition, the required number of multiplications is reduced by up to 82%. For seizure prediction, the chip achieves a sensitivity of 92.0% and a false alarm rate (FAR) of 0.57/h with a training latency of 8.44ms and a power dissipation of 2.31mW at 6.05MHz. The performance of seizure detection also surpasses the existing literature with the dedicated hardware implementation. Compared with a high-end CPU, this work achieves a 2.45x10^4x higher area efficiency and a 1.32x10^6x higher energy efficiency.

S01.2 13:45 — 14:00

A Lego-based Neural Network Design Methodology by using Flexible NoC

Kun-Chih Chen, Yi-Sheng Liao, and Cheng-Kang Tsai
Department of Computer Science and Engineering, National Sun Yat-sen University

Deep Neural Networks (DNNs) have shown superiority in solving the problems of classification and recognition in recent years. However, DNN hardware implementation is challenging due to the high computational complexity and diverse dataflow in different DNN models. A large body of research has focused on accelerating specific DNN models or layers and proposed dedicated designs to mitigate this design challenge. However, dedicated designs for specific DNN models or layers limit the design flexibility. This work takes advantage of the similarity among different DNN models and proposes a novel Lego-based Deep Neural Network on a Chip (DNNoC) design methodology. We work on common neural computing units (e.g., multiply-accumulation and pooling) and create some neuron computing units called NeuLego processing elements (NeuLegoPEs). These NeuLegoPEs are then interconnected using a flexible Network-on-Chip (NoC) to construct different DNN models. To support large-scale DNN models, we enhance the reusability of each NeuLegoPE by proposing a Lego placement method. The proposed design methodology allows leveraging different DNN model implementations, helping to reduce implementation cost and time-to-market. Compared with the conventional approaches, the

proposed approach can improve the average throughput by 2,802% for target DNN models. Besides, the corresponding hardware is implemented to validate the proposed design methodology, showing on average 12,523% hardware efficiency improvement by considering the throughput and area overhead simultaneously.

S01.3 14:00 — 14:15

Sparse Compressed Spiking Neural Network Accelerator for Object Detection

Hong-Han Lien[1] and Tian Sheuan Chang[2]
[1]Program of Artificial Intelligence, National Yang Ming Chiao Tung University
[2]Institute of Electronics, National Yang Ming Chiao Tung University

Spiking neural networks (SNNs), which are inspired by the human brain, have recently gained popularity due to their relatively simple and low-power hardware for transmitting binary spikes and highly sparse activation maps. However, because SNNs contain extra time dimension information, the SNN accelerator will require more buffers and take longer to infer, especially for the more difficult high-resolution object detection task. As a result, this paper proposes a sparse compressed spiking neural network accelerator that takes advantage of the high sparsity of activation maps and weights by utilizing the proposed gated one-to-all product for low power and highly parallel model execution. The experimental result of the neural network shows 71.5$\%$ mAP with mixed (1,3) time steps on the IVS 3cls dataset. The accelerator with the TSMC 28nm CMOS process can achieve 1024$\times$576@29 frames per second processing when running at 500MHz with 35.88TOPS/W energy efficiency and 1.05mJ energy consumption per frame.

S01.4 14:15 — 14:30

Real-Time Block-Based Embedded CNN for Gesture Classification on an FPGA

Ching-Chen Wang[1], Yu-Chun Ding[2], Ching-Te Chiu[3], Chao-Tsung Huang[2], Yen-Yu Cheng[3], Shih-Yi Sun[2], Chih-Han Cheng[3], and Hsueh-Kai Kuo[3]
[1]Ambarella
[2]Department of Electrical Engineering, National Tsing Hua University
[3]Department of Computer Science, National Tsing Hua University

This paper presents a block-based embedded convolutional neural network (CNN) for gesture classification on field-programmable gate array (FPGA) in real time. Gesture recognition is an important tool to spontaneous interact with human machine interface. Many CNN architectures using RGB images have been proposed for gesture classification. RGB based gesture classification may cause incorrect results under insufficient light or similar gestures. In addition, most of the CNN architectures cannot run in real time on edge devices due to their large number of parameters and DRAM data access. In this paper, a block-based CNN using RGB-D data is proposed for gesture classification. Adding depth images to RGB images boots the classification accuracy. A CNN architecture with block-based feature maps is built for embedded FPGA implementations. The total number of parameters of the proposed RGB-D embedded CNN (eCNN) model is only 0.17M and it achieves 99.96% and 99.88% accuracy with 32-bit floating point and 8-bit fixed point implementation for America Sign Language (ASL) data set. The RTL simulation of the proposed eCNN model has the average inference speed of 0.171 milliseconds at frequency of 250MHz for a single pair RGB-D image. Implemented on a FPGA integrated with Microsoft Kinect v2 achieve an inference time in 19.42 ms which achieves high accuracy and real-time performance.