Session Oral 3 (8/7 Wed. 13:30 – 15:00)

Session Topic: AI Computing and Acceleration

Session Chair: Shih-Hsu Huang (Chung Yuan Christian University) and Wei-Kai Cheng (Chung Yuan Christian University)

Room: 6F 茗廳

1. 13:30 – 13:43 (SB11) Low Accuracy Loss and Hardware-Friendly Model Compression Technique for DNNs

   Ya-Chu Chang and Juinn-Dar Huang

   National Chiao Tung University

   Deep neural networks (DNNs) are broadly utilized in numerous machine learning applications nowadays. In a large DNN with several hidden layers, the number of weights (or coefficients) required to complete an entire perceptron are indeed huge. However, these excessive number of weights not only require a big

chunk of memory but also create a significant memory access traffic, which incurs a heavy burden especially for small-scale embedded systems and edge devices. Therefore, several techniques have been proposed during the past few years. We present a new hardware-friendly model compression technique in this paper. It can achieve a compression rate of 20X~30X while keeping the accuracy loss below 1%.

2. 13:43 – 13:56 (SB12) A Simulator for Evaluating the Fault-Tolerance Capability of DNNs

Yung-Yu Tsai and Jin-Fu Li

National Central University

Deep neural network (DNN) is considered as one effective technique for the artificial intelligence applications. A DNN is constituted by a large amount of neurons arranged in a form of multilayers. Typically, a DNN has overprovisioning neurons such that it has the property of fault tolerance [1][2]. However, how to evaluate the fault tolerance capability of DNNs is an important issue. In this paper, we propose a

simulator to estimate the loss of inference accuracy due to the faults in a DNN model or hardware accelerator. The simulator is implemented based on the platforms of Keras and Tensorflow. It can evaluate the fault-tolerance capability of a DNN at model and hardware levels. The proposed simulator can estimate the accuracy loss of a DNN model caused by a faulty neuron, a faulty link, or faulty input. The fault injection mechanism is done through the bitwise operation at the parameter of Tensorflow. The simulator integrates the Tensorflow and Keras platform to evaluate the accuracy of a DNN model with faulty elements. Also, the simulator can estimate the inference accuracy loss of a DNN accelerator caused by the faulty buffers. Simulation results of accuracy with respect to different fault rates for the LeNet and 4C2F models are conducted.

3. 13:56 – 14:09 (SB13) Approximate Systolic Array-based Processor for AI Computation

   Wei-Kai Tseng, Huan-Jan Chou, Ning-Chi Huang, and Kai-Chiang Wu

   National Chiao Tung University

Approximate computing is an emerging strategy which trades computational accuracy for computational cost in terms of performance, energy, and/or area. We propose a novel sensor-based approximate adder, Carry Truncate Adder (CTA), for high-performance energy-efficient arithmetic computation, while considering the accuracy requirement of error-tolerant applications. On top of a fully-optimized ripple carry adder, the performance of our adder is enhanced by 2.17X. When applied in error-tolerant applications such as image processing and handwritten digit recognition, our approximate adder leads to very promising quality of results compared to the case when an accurate adder is used. Systolic arrays are widely used as matrix multiplication accelerators for DNNs. To improve the performance and energy efficiency of a systolic array, we apply the idea of timing speculation based on our proposed CTA into a systolic array. By using in-situ sensors for approximate multiplier-accumulator (MAC) computation in a systolic array, the computation which needs longer propagation time will drop the next result of multiplication and occupy two (adjacent) MACs to complete the current multiplication and accumulation.

In the experiments, compared to the original systolic array (without any approximation), our proposed approximate systolic array can reduce the clock period from 8.57 to 5.39 (ns) with only 1% accuracy loss on MNIST dataset.

4. 14:09 – 14:22 (SB14) Approximate Logic Circuit Design for AI Applications

Wei-Hung Lin, Hsu-Yu Kao, and Shih-Hsu Huang

Chung-Yuan Christian University

To reduce the power consumption of an embedded system, the design of approximate logic circuits appears as a promising solution for many error-resilient applications. In this paper, we will introduce the approximate logic circuit design for AI applications. We have developed a circuit library of approximate logic circuits for AI applications. Moreover, we also have constructed a neural network (NN) design framework for the users to utilize the circuit library to develop their AI applications. So far, we have

implemented ICNet, which is a famous semantic segmentation NN, by using the proposed NN design framework. Experimental results show that, compared with the original ICNet model, even if all the multiplications and activation functions are replaced by our approximate logic circuits, the accuracy loss is only 0.1%. Our future works is to provide more approximate logic circuits in the NN design framework for the trade-off. We will also try to develop more AI applications based on the NN design framework.

5. 14:22 – 14:35 (SB15) Dataflow Exploration Framework for Data Reuse of CNN Computation

Xiang-Yi Liu, Yuan-Chih Lo, Tsai-Yu Tsai, and Wei-Kai Cheng

Chung-Yuan Christian University

In the edge intelligence system, due to the limited hardware resources, memory accesses become the bottleneck of DNN hardware accelerator. Different from CPU or GPU architecture, dataflow processing is an effective method to speed-up the efficiency of data migration in the DNN hardware accelerator.

However, memory accesses consume a high percentage of energy in this type of DNN architecture. In this paper, we propose a modified dataflow approach based on Eyeriss to reduce data volume of external memory access. Experimental results show that our dataflow approach can reduce data migration of kernel and input feature map between external DRAM and internal buffer.

6. 14:35 – 14:48 (S0168) AIP: Saving the DRAM Access Energy of CNNs Using Approximate Inner Products

Cheng-Hsuan Cheng and Ren-Shuo Liu

National Tsing Hua University

In this work, we propose AIP (Approximate Inner Product), which approximates the inner products of CNNs' fullyconnected (FC) layers by using only a small fraction (e.g., onesixteenth) of parameters. We observe that FC layers possess several characteristics that naturally fit AIP: the dropout training strategy, rectified linear units (ReLUs), and top-n operator. Experimental results show that 48% of DRAM access energy can be reduced at the cost of only 2% of top-5 accuracy loss (for VGG-f).

7. 14:48 – 15:01 (S0030) Filter Pruning based on Dynamic Convolutional Neural Network for Surveillance Video

Chun-Ya Tsai, De-Qin Gao, and Shanq-Jang Ruan

National Taiwan University of Science and Technology

The large-scale surveillance videos analysis becomes important as the development of the intelligent city; however, the heavy computational resources necessary for the state-of-the-art deep learning model makes real-time processing hard to be implemented. As the characteristic of high scene similarity generally existing in surveillance videos, we propose an effective compression architecture called dynamic convolution, which can reuse the previous feature maps to reduce the calculation amount; and combine with filter pruning to further speed up the performance.