Oral S04

Artificial Intelligence Circuits and Systems (I)

Date/Time	8/3(三)14:30-15:30
Chair(s)	吳崇賓教授 / 國立中興大學電機工程學系 鄭湘筠教授 / 中央研究院資訊科技創新研究中心

S04.1 🕓 14:30 — 14:42

Hardware-Friendly Weight Pruning Algorithm for CNN Model Size Reduction using Universal Pattern Sets

Wei-Cheng Chou, Cheng-Wei Huang, Yung-Han Chen, and Juinn-Dar Huang
Institute of Electronics, National Yang Ming Chiao Tung University

Pattern-based weight pruning on CNNs has been proven an effective model reduction technique. In this paper, we first present how to select hardware friendly pruning pattern sets that are universal to various models. We then propose a progressive pruning framework, which produces more globally optimized outcomes. Moreover, to the best of our knowledge, this is the first paper dealing with the pruning issue of the first and also the most sensitive layer of a CNN model through a twostaged pruning strategy. Experiment results show that the proposed framework achieves 2.25x/2x computation/model reduction while minimizing the accuracy loss.

S04.2 🕓 14:42 — 14:54

A VLSI Implementation of a Local Binary Convolutional Neural Network

Yu-Tong Shen[1], Shan-Chi Yu[2], and Ing-Chao Lin[1]
[1]Department of Computer Science and Information Engineering, National Cheng Kung University
[2]Department of Engineering Science, National Cheng Kung University

In order to reduce the computational complexity of convolutional neural networks (CNNs), the local binary convolutional neural network (LBCNN) is proposed. In this work, we propose a platform that includes a weight preprocessor and layer accelerator for the LBCNN. Our weight preprocessor takes advantages of sparsity in the LBCNN and encodes the weight offline. The layer accelerator effectively uses the encoded data to reduce computational complexity and memory accesses for an inference. When compared to the state-of-the-art design, the experimental results show that the number of clock cycles is reduced 7.9 times, and memory usage is reduced 1.7 times. The synthesized results show that the clock period is reduced 5.7%; the cell area is reduced 19.7%, and the power consumption is reduced 15.2%. The total execution time is 8.38 times better, and the inference accuracy is not affected.

S04.3 🕓 **14:54 — 15:06**

A 62.45 TOPS/W Spike-Based Convolution Neural Network Accelerator with Spatiotemporal Parallel Data Flow and Sparsity Mechanism

Chen-Han Hsu, Yu-Hsiang Cheng, Zhao-fang Li, Ping-Li Huang, and Kea-Tiong Tang
National Tsing Hua University

Convolutional neural networks (CNNs) have been widely used for image recognition and classification in recent years. Low energy consumption is crucial in the circuit design of edge devices, and data reuse is one method of reducing energy consumption. In addition, spiking neural networks (SNNs) are receiving increasing attention due to their low power use. However, the temporal characteristic of SNNs causes repeated data access at different time steps, leading to high energy consumption. In this paper, a spiked-based CNN accelerator that can support various inference time steps and models is proposed. Spatiotemporal parallel data flows are employed to reuse data from different time steps and, convolution operations are used to reduce energy consumption. Furthermore, the accelerator is designed for high sparsity and event driven SNNs. The synthesis achieves power efficiency of 62.45 TOPS/W and area efficiency of 7.58 TOPS/kmm2

S04.4 🕓 **15:06 — 15:18**

Automated Quantization Range Mapping for DAC/ADC Non-linearity in Computing-In-Memory

Chi-Tse Huang, Yu-Chuan Chuang, Ming-Guang Lin, and An-Yeu (Andy) Wu
Graduate Institute of Electronics Engineering, National Taiwan University

Computing-in-memory (CIM) has demonstrated the great potential of analog computing in improving the energy efficiency of matrix-vector multiplications for deep learning applications. Albeit low-power feature of CIM, the non-linearity of digital-to-analog converters (DACs)/analog-to-digital converters (ADCs) causes deviation between the computed outputs and desired values, thus degrading classification accuracy. This paper proposes Automated Quantization Range Mapping (A-QRM) mechanism to mitigate the negative effect of non-linearity on model accuracy. Instead of fixing the quantization range for quantized deep learning models, the proposed A-QRM automatically finds a better quantization range that balances the model capability and quantization errors caused by the non-linearity. Experimental results show that our proposed A-QRM achieves 89.02% and 86.93% of top-1 accuracy in ResNet20 and VGG8 on Cifar-10, respectively, under the non-linearity of DACs/ADCs.

S04.5 🕓 15:18 — 15:30

Efficient Dilated Convolution Engine with Group-reordering and Normalization

Tong Wu, Kai-Ping Lin, Shih-Yi Sun, and Chao-Tsung Huang
Department of Electrical Engineering, National Tsing Hua University

Dilated convolution neural network models have demonstrated the ability to increase the receptive field without increasing the model size and complexity. However, direct implementation of dilated convolution is inefficient in terms of the input buffer size and hardware utilization with inserting zeros in the weight kernels. In this paper, we propose a power/area-efficient dilated convolution engine, which supports arbitrary dilation rates ranging from 1 to 16 between different layers. Compared to the baseline architecture, this design has 53.9% and 69.9% of area and power reduction, respectively.