Oral S08

Advanced AI Technologies for Specific Applications

Date/Venue	8/1(四) 11:10-12:10 [海棠廳]
Chairs(s)	呂仁碩 /清華大學電機系

S08.1 | 11:10-11:21

ESSR: An 8K@30FPS Super-Resolution Accelerator With Edge Selective Network

Chih-Chia Hsu, and Tian-Sheuan Chang

Institute of Electronics, National Chiao Tung University Hsinchu, Taiwan

Deep learning-based super-resolution (SR) is challenging to implement in resource-constrained edge devices for resolutions beyond full HD due to its high computational complexity and memory bandwidth requirements. This paper introduces an 8K@30FPS SR accelerator with edge-selective dynamic input processing. Dynamic processing chooses the appropriate subnets for different patches based on simple input edge criteria, achieving a 50% MAC reduction with only a 0.1dB PSNR decrease. In conjunction with hardware-specific refinements, the model size is reduced by 84% to 51K, but with a decrease of less than 0.6dB PSNR. Additionally, to support dynamic processing with high utilization, this design incorporates a configurable group of layer mapping that synergizes with the structure-friendly fusion block, resulting in 77% hardware utilization and up to 79% reduction in feature SRAM access. The implementation, using the TSMC 28nm process, can achieve 8K@30FPS throughput at 800MHz with a gate count of 2749K, 0.2075W power consumption, and 4797Mpixels/J energy efficiency, exceeding previous work.

S08.2 | 11:22-11:33

An Efficient Co-Processor Design for Kalman Filter Targeting at Multiple Object Tracking

Kuang-Hung Chen, and Jun-Hao Wang

Department of Electronic Engineering, Feng -Chia University, Taichung, Taiwan

Kalman filter consists of several formulas involving a substantial amount of matrix multiplications and matrix inverses. Implement these formulas in a straightforward way is not a good strategy because these matrices exhibit lots of sparsity and regularity resulting in many zeros and repetitive calculations. Therefore, a VLSI architecture that reduces the redundant matrix multiplications in Kalman filter is crucial. In this paper, we propose a low-power, co-processor design for realizing Kalman filter needed in multi-target tracking systems. We optimize the efficient co-processor design by leveraging the expansion of Kalman filter formulas, matrix sparsity and special characteristics, matrix data

preprocessing, add-subtract based multipliers and dividers, and adaptive instruction set design. These optimizations enable the design to operate with low power consumption and succinct area cost. Unlike the state-of-the-art (SOTA) ASIC designs customizing the hardware design for each formula one by one, we propose a programmable co-processor to satisfy the computation needs of all formulas in a flexible and efficient manner. Implementation results show that the proposed design costs 2991 gates, reducing the area by 96% compared to SOTA ASIC designs, and consumes only 14mW of power.

S08.3 | 11:34-11:45

SegTransformer: Revolutionizing Softmax Efficiency via Segmentation with ReRAM-based PIM Acceleration

Yu-Cheng Wang, Ing-Chao Lin, Yuan-Hao Chang

Currently, many well-known neural networks primarily adopt the Transformer architecture for natural language processing (NLP) applications. However, the Transformer focuses on handling long-distance relationships between words, leading to relatively long execution times. Several ReRAM-based Processor-In-Memory (PIM) architectures have been proposed to take advantage of in-memory computing capabilities of ReRAM-based crossbar and accelerate the attention mechanism in Transformers. They, however, often shift the performance bottlenecks from the attention mechanism to Softmax computations. Another problem is the inefficient Softmax computations when computing extremely small values. Therefore, we propose SegTransformer, a ReRAM-based PIM accelerator that leverages segmentation techniques to address the Softmax bottleneck and propose SparseAware Softmax Acceleration (SASA) to reduce inefficiencies in Softmax computation. Our experimental results demonstrate that SegTransformer outperforms the state-of-the-art Transformer accelerators.

S08.4 | 11:46-11:57

A New Number Representation Format and Its Hardware Support for Accurate Low-Bit Quantization in Recommendation Systems

Yu-Da Chu, Pei-Hsuan Kuo, Lyu-Ming Ho, and Juinn-Dar Huang
Institute of Electronics , National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Deep-learning-based recommender systems with huge embedding tables have become crucial for web content recommendation. However, the growing size of those tables, reaching hundreds of gigabytes or even terabytes, presents a tough challenge for inferencing on resource-constrained hardware. In this paper, we present an innovative 6-bit fixed-point number representation format for more precise quantization on recommender models. The proposed format is designed to accommodate the nonuniform weight distribution inside those huge embedding tables. To further minify the model, K-means quantization technique is used for 4-bit quantization and beyond. Besides, we also propose dedicated hardware decoder architectures for both 6-bit and 4-bit quantization to ensure efficient

runtime inference. Experimental results show that the proposed low-bit (8~3-bit) quantization techniques on embedding tables yield 4~10.7x model size reduction with little accuracy loss with comparison to original FP32 model. Hence, the proposed number representation format and low-bit quantization techniques can effectively and drastically reduce the size of large recommender systems models with low area cost while still keeping the accuracy loss minimized.

S08.5 | 11:58-12:09

An Efficient Deformable Convolutional Network Co-Processor Design for Image Semantic Segmentation

Kuang-Hung Chen, Reng-Jie Chen, and Tai-En Chung
Department of Electronic Engineering , Feng-Chia University , Taichung, Taiwan

Deformable Convolutional Network (DCN) improves recognition accuracy under object deformations but introduces much higher computational complexity. Based on the deformation principle, the DCN co-processor demands more flexible and efficient hardware architecture. Accordingly, we reduce hardware cost by elaborating word-length of the DCN co-processor and replacing multipliers in bilinear interpolation with shift-add operating units. The implementation platform is ZCU-104 FPGA. This design surpasses state-of-the-art DCN accelerators in both hardware efficiencies, i.e., LUT efficiency and BRAM efficiency. Additionally, this design achieves 73.7% mIoU on the Cityscapes dataset which outperforms well-known AI models.