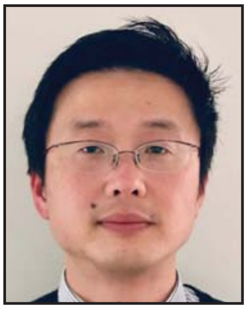
Session 9 Overview: *ML Processors From Cloud to Edge*

MACHINE LEARNING SUBCOMMITTEE



Session Chair:
SukHwan Lim
Samsung, Suwon, Korea

Session Co-Chair:
Luca Benini
ETH Zurich, Zurich, Switzerland

Session Moderator:
Vivienne Sze
MIT, Cambridge, MA

2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9365814

Significant progress has been made in machine learning processor design in two different but important topic areas. The first addresses flexible accelerators for inference and training in the most advanced CMOS technology nodes (e.g. 5nm and 7nm) for mobile and the cloud. The second topic area covers application-specific acceleration engines for ultra-low-power applications, including wearable devices. This session comprises nine papers, covering a diverse set of neural networks targeted at a wide range of applications, including gesture recognition, smart cameras, speech-to-text and keyword spotting.

7:00 AM

9.1 **A 7nm 4-Core AI Chip with 25.6TFLOPS Hybrid FP8 Training, 102.4TOPS INT4 Inference and Workload-Aware Throttling**

Ankur Agrawal, IBM Research, Yorktown Heights, NY

In Paper 9.1, IBM Research presents a 7nm 4-core AI chip that offers separate floating-point and fixed-point pipelines to enable performance improvements without model accuracy degradation, while a high-bandwidth on-chip ring maintains high compute utilization. Workload-aware throttling is used to maximize performance within a specified power envelope. Their 19.6mm^2 chip demonstrates up to 3.5TFLOPS/W and up to 25.6TFLOPS hybrid fp8 iso-accuracy training, up to 16TOPS/W and 102TOPS int4 inference, as well as support for fp16, fp32 and int2 computation.

7:08 AM

9.2 **A 28nm 12.1TOPS/W Dual-Mode CNN Processor Using Effective-Weight-Based Convolution and Error-Compensation-Based Prediction**

Huiyu Mo, Institute of Microelectronics of Tsinghua University, Beijing, China

In Paper 9.2, Tsinghua University describes a 28nm 12.1TOPS/W CNN processor employing effective-weight convolution and error-compensation prediction, eliminating >90% multiplications compared to prior CNN implementations with <1% additional overhead. A residual pipeline mode is used to avoid the need for off-chip memory accesses within residual blocks, allowing hardware utilization to be maintained near 100%. The processor consumes 1.9mm^2 and 131.6mW at 470MHz, achieving improvements in TOPS/W energy-efficiency, GOPS/mm^2 area-efficiency, and energy-per-frame with respect to state-of-the-art 8b CNN processors.

7:16 AM

9.3 **A 40nm 4.81TFLOPS/W 8b Floating-Point Training Processor for Non-Sparse Neural Networks Using Shared Exponent Bias and 24-Way Fused Multiply-Add Tree**

Jeongwoo Park, Seoul National University, Seoul, Korea

In Paper 9.3, Seoul National University presents an 8b floating point training processor in 40nm technology for state-of-the-art non-sparse neural networks, achieving 69.0% ResNet-18 Top-1 accuracy on ImageNet and requiring 43% less memory access by use of 8b tensors. By combining a 24-way fused multiply-add tree with a flexible 2D routing scheme, their 7.29mm^2 chip achieves 4.81TFLOPS/W energy efficiency and 2.48× higher training efficiency than prior work.

978-1-7281-9549-0/21/$31.00 ©2021 IEEE

7:24 AM

9.4 PIU: A 248GOPS/W Stream-Based Processor for Irregular Probabilistic Inference Networks Using Precision-Scalable Posit Arithmetic in 28nm

Nimish Shah, KU Leuven - MICAS, Leuven, Belgium

In Paper 9.4, KU Leuven introduces a 3.8mm^2 probabilistic inference unit (PIU) in 28nm technology, targeted at exact inference of irregular probabilistic sum-product networks (SPNs). To accelerate these workloads, PIU decouples compute/load/store streams with a co-optimized memory hierarchy, employs precision-scalable posit arithmetic for accurate manipulation of low probability values, and aligns the parallel 8,16, and 32b floating-point computations with compiler optimizations. The resulting energy efficiency is 1173× and 271× higher than RTX2080 GPU and CPU, respectively, reaching a peak of 248GOPS/W, 33.7GOPS and 8.9GOPS/mm^2.

7:32 AM

9.5 A 6K-MAC Feature-Map-Sparsity-Aware Neural Processing Unit in 5nm Flagship Mobile SoC

Jun-Seok Park, Samsung Electronics, Hwaseong, Korea

In Paper 9.5, Samsung Electronics presents a 5.46mm^2 reconfigurable neural processing unit (NPU) in 5nm technology for mobile SoCs. Their NPU has 3 cores with 6k MACs in total, boosting performance and energy-efficiency by feature-map zero skipping, a reconfigurable adder-tree-based datapath, feature map compression and fast resource scheduling. With all three NPU cores deployed, their NPU achieves high performance (623 inferences/s) and high energy efficiency (13.6TOPS/W) on an 8b Inception-V3 model.

7:36 AM

9.6 A 1/2.3inch 12.3Mpixel with On-Chip 4.97TOPS/W CNN Processor Back-Illuminated Stacked CMOS Image Sensor

Ryoji Eki, Sony Semiconductor Solutions, Tokyo, Japan

In Paper 9.6, Sony Semiconductor Solutions describes a 62mm^2 stacked-chip solution, combining a back-illuminated (BI) pixel CMOS image sensor at 4056×3040 resolution and 1.55μm pixel-pitch, together with 4.97TOPS/W of 22nm digital signal processing of 8b, 16b, or 32b integers in two different DSP cores, multiple tensor direct memory access (TDMA) engines and a scheduler optimized for convolutional neural network (CNN) processing. Capable of 120fps with parallel image readout and CNN inference, their system can offer AI processing capability at reduced size, power and cost, while addressing privacy concerns.

7:40 AM

9.7 A 184μW Real-Time Hand-Gesture Recognition System with Hybrid Tiny Classifiers for Smart Wearable Devices

Yuncheng Lu, Nanyang Technological University, Singapore, Singapore

In Paper 9.7, Nanyang Technological University presents a real-time ultra-low-power hand gesture recognition system for wearable and IoT devices, combining a recognition core with hybrid compact classifiers for static gesture recognition and an error-tolerant sequence analyzer for dynamic gesture recognition. Their 1.5mm^2 chip in 65nm technology achieves 184µW at 0.6V and can recognize 24 dynamic gestures with an average accuracy of 92.6% when the hand moving speed is within 30-40cm/s

7:48 AM

9.8 A 25mm^2 SoC for IoT Devices with 18ms Noise-Robust Speech-to-Text Latency via Bayesian Speech Denoising and Attention-Based Sequence-to-Sequence DNN Speech Recognition in 16nm FinFET

Thierry Tambe, Harvard University, Cambridge, MA

In Paper 9.8, Harvard University describes a 21.8mm^2 SoC in 16nm technology for noise-robust speech recognition that includes a Bayesian source separation engine optimized for unsupervised speech-denoising via Gibbs sampling using 32b fixed-point, as well as a reconfigurable processor optimized for whole-model acceleration of large vocabulary bidirectional attention-based DNNs using 8b floating-point. The proposed speech-enhancing automated speech recognition (ASR) pipeline denoises noise-corrupted speech with 7.3dB signal-to-distortion-ratio (SDR), while achieving 18ms end-to-end latency and consuming 2.24mJ of energy per frame.

7:56 AM

9.9 A Background-Noise and Process-Variation-Tolerant 109nW Acoustic Feature Extractor Based on Spike-Domain Divisive-Energy Normalization for an Always-On Keyword Spotting Device

Dewei Wang, Columbia University, New York, NY

In Paper 9.9, Columbia University presents an always-on keyword spotting system in 65nm which pairs a normalized acoustic feature extractor (NAFE) chip featuring divisive energy normalization (DN) together with a spiking neural network classifier chip. Their KWS system exhibits 570nW power dissipation and robustness to process variation while maintaining high accuracy (96.5% for 1 keyword, 90.2% for 4 keywords) across a variety of strong background-noise environments.