Workshop on Hardware and Algorithms for Learning On-a-chip (HALO) 2019

The Westin Westminster, Westminster, CO, Thursday, November 7, 2019

Time	Schedule
8:15am – 8:30am	Introduction and opening remarks
8:30am – 9:20am	**Keynote talk 1** Allen Rush (AMD)
9:20am – 10:10am	**Keynote talk 2**: Machine Learning on Social Network Platforms Hsien-Hsin Lee (Facebook)
10:10am – 10:30am	Coffee Break
Session 1: Hardware Acceleration of Machine Learning Session Chair:	
10:30am – 10:50am	Neural Networks Accelerator Design from the User Perspective Yu Wang (Tsinghua University)
10:50am – 11:10am	Algorithm/Hardware Co-design for Energy/Area efficient In-Memory Neural Network Computing Jae-Joon Kim, POSTECH
11:10am – 11:30am	Bringing Powerful Machine-learning Systems to Daily-life Devices via Algorithm-hardware Co-design Yingyan Lin (Rice University)
11:30am – 11:50am	Toward Next-Generation Acceleration for AI: A Heterogeneous Computing Approach Jaewoong Sim (Intel)
11:50am – 12:10pm	Neural Networks Accelerator Design from the User Perspective Zheng Zhang (UCSB)
12:10pm – 1:10pm	Lunch
1:10pm – 2:00pm	**Keynote talk 3**: On-Device AI for Augmented Reality (AR) Systems Vikas Chandra (Facebook)
Session 2: Intelligent Mobile Applications Session Chair:	
2:00pm – 2:20pm	A Product Engine for Energy-Efficient Execution of Binary Neural Networks Using Resistive Memories Pierre-Emmanuel GAILLARDON (University of Utah)
2:20pm – 2:40pm	Evolutionary Optimization for Neuromorphic Systems Catherine Schuman (Oak Ridge National Lab)
2:40pm – 3:00pm	Micro AI: When Intelligence Moves to the Low Power Sensors Tinoosh Mohsenin (UMBC)
3:00pm –3:20pm	Beyond Energy-Efficiency: Enabling fault-aware Learning On-a-chip Siddharth Garg (New York University)
3:20pm – 3:40pm	Coffee Break

Poster Session	
3:40pm – 4:10pm	Poster presentations (2mins each) P1. How to Obtain and Run Light and Efficient Deep Learning Networks Fan Chen (Duke University) P2. Achieving Super-Linear Speedup across Multi-FPGA for Real-Time DNN Inference Weiwen Jiang (University of Notre Dame) P3. On Neural Architecture Search for Resource-Constrained Hardware Platforms Qing Lu (University of Notre Dame) P4. CIMAT: A Transpose SRAM-based Compute-In-Memory Architecture for Deep Neural Network On-Chip Training Hongwu Jiang (Georgia Institute of Technology) P5. HR3AM: a Heat Resilient design for RRAM based neuromorphic computing Xiao Liu (University of California San Diego) P6. ACG-Engine: An Inference Accelerator for Content Generative Neural Networks Haobo Xu (Institute of Computing Technology, Chinese Academy of Sciences) P7. Mixed Precision Neural Architecture Search for Energy Efficient Deep Learning Zhixuan Jiang (The University of Texas at Austin) P8. Enhanced Error-Correcting DNN Classifier Towards Robust Machine Learning On-a-chip Tao Liu (Lehigh University) P9. PCONV: A Desirable Sparsity Dimension for Real-time Execution -- From Algorithm to Framework Xiaolong Ma (Northeastern University) P10. 2.5ms MobileNet-V2 Execution on Mobile Phone -- A Compiler-Assisted Block Pruning Framework Zhengang Li (Northeastern University) P11. INA: Incremental Network Approximation Algorithm for Limited Precision Deep Neural Networks Zheyu Liu (Tsinghua University) P12. Approximating Backpropagation for a Biologically Plausible Local Learning Rule in Spiking Neural Networks Haowen Fang (Syracuse University) P13. Leveraging Model Diversity for High QoS Deep Learning Inference in the Clouds Jeff (Jun) Zhang (New York University)
4:10pm – 5:00pm	Poster discussions