Call for Papers

Cross-Layer Designs, Methodologies and Systems to Enable Micro AI for On-Device Intelligence

Guest Editors

- Tinoosh Mohsenin, University of Maryland Baltimore County, USA (Corresponding Guest Editor)
- Xin Zhang, IBM T. J. Watson Research Center, USA
- Houman Homayoun, University of California, Davis, USA
- Jae-sun Seo, Arizona State University, USA

Scope and Purpose

Artificial intelligence (AI) and robotics technology continue to have a major role in enabling future smart cities, transportation, surveillance, logistics, smart sensing, home health care and medical technologies. These fields rely on large data driven approaches and highly computation-intensive algorithms, and tend to compute in the cloud and data centers. Local processing on an embedded device is often required to provide low latency and less dependence on the communication link for bandwidth and privacy/security concerns. However, processing locally on the device is very challenging due to its limited memory storage and battery capacity. This proposal invites researchers to submit state-of-the-art computing methods through cross-layer design approaches in algorithms, architecture, hardware and system integration, which will enable micro-intelligent systems to perform on-device sensor data analytics and various AI tasks at extremely low power.
In particular, this special issue seeks contributions encompassing the following broad scopes:

**AI Model Design:** Novel reinforcement learning based neural architecture search (NAS) methods generally aim to automatically explore the least computationally complex AI models that meet a targeted accuracy. Nevertheless, these searching methods seldom take into account the challenges of deploying the computationally-reduced AI model at the edge (hardware). This topic covers NAS methods that jointly explore efficient AI models including artificial neural networks (ANNs), deep neural networks (DNNs), recurrent neural networks (RNNs), spiking neural networks (SNNs), etc., for feasible implementation on resource-bound hardware.

**Hardware Design methodology:** This topic covers novel approaches of task scheduling, tiling scheme, and data movement with various implementation styles and goals such as energy, power and timing. Implementation styles such as in-memory computing, near-memory processing, and systolic array architecture, as well as techniques considering tradeoffs among computation and communication, arrangement of processing engines, selection of fixed/ floating-point/flexible hardware datawidth, and their impact on timing, power, performance, and energy efficiency are worthy of consideration in this topic.

**Memory hierarchy and Data Movement:** In high dimensional AI models, the memory footprint dominates the silicon area and the energy and delay overheads resulting from memory communication for implementing the AI models is significantly higher as compared to computation. Methods of efficiently implementing pruned or compressed AI models as well as different methods of data movement through different levels of memory hierarchy (e.g., from DRAMs to SRAMs), precise/flexible size allocation to different memory levels are among the issues to be addressed in this topic.

**Flexibility and Reliability:** While small AI models can be more conveniently implemented at the edge, the implementation of large models is still a challenging task for the resource-bound hardware. Meanwhile, susceptibility of spiking neural networks to noise and event upsets is another challenge. This topic calls for novel methods considering flexible and reliable hardware design for versatile modern AI applications.

**Topics of Interest**

The special issue focuses on micro/tiny machine learning hardware approaches with the following topics:

- AI applications at edge (e.g., 5G, healthcare, medical applications, industry, etc.)
- Circuits and architecture design enabling micro AI
- Applications and hardware for learning on a smart micro AI device
- Ultra-low-power memory and communication design for micro AI
- Sparse learning, feature extraction and personalization
- Emerging technologies for enabling micro AI devices
• NAS methods that search optimal ANNs for feasible hardware implementations
• Cross-layer optimization for ANN architecture and application
• Energy-, bit error- and accuracy-aware pruning and quantization mechanism of ANNs
• Communication-aware AI subsystems for future on-chip adaptive learning application
• Systolic Array Architecture, Near-memory and In-memory computing techniques
• Tradeoffs among computation vs communication considering energy/power/timing
• Efficient on-chip communication/computation of multicore-based ANNs
• ANN subsystem design based on emerging non-volatile memory devices
• Data movement optimization through task scheduling of the ANN
• Novel interconnection networks for the ANNs (e.g., DNN, RNN, SNN, etc.)
• Novel topology of on-chip communication for efficient ANN processing
• NoC design for heterogeneous ANN computing
• Applications and Artificial General Intelligence (AGI)
• Flexible, reliable and robust computing methods
• Wearable AI devices for health monitoring/detection
• Algorithm and architecture co-design of energy-efficient ANNs

Submission Procedure

Prospective authors are invited to submit their papers following the instructions provided on the JETCAS website: https://mc.manuscriptcentral.com/jetcas. The submitted manuscripts should not have been previously published nor should they be currently under consideration for publication elsewhere.

Important Dates

• Manuscript submissions due: 2021-05-01
• First round of reviews completed: 2021-06-30
• Revised manuscripts due: 2021-08-15
• Second round of reviews completed: 2021-09-15
• Final manuscripts due: 2021-10-10
• Target publication date: 2021-12-01

Request for Information

Tinoosh Mohsenin (tinoosh@umbc.edu)