IEEE JOURNAL ON
EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS

CALL for PAPERS

Customized sub-systems and circuits for deep learning

Guest editors

Chia-Yu Chen, IBM T.J. Watson Research Center, USA (cchen@us.ibm.com)

Boris Murmann, Stanford University, USA (murmann@stanford.edu)

Jae-sun Seo, Arizona State University, USA (jaesun.seo@asu.edu)

Hoi-Jun Yoo, KAIST, Korea (hjyoo@kaist.ac.kr)

Scope and purpose

[Rationale and Motivation]

This special issue is dedicated to recent technical advances in emerging hardware technologies that will enable deep learning across various application areas. Over the past decade, deep learning has emerged as the dominant machine learning approach, revolutionizing a wide spectrum of challenging application domains ranging from image processing, machine translation, speech recognition and many others. This rapid progress has been enabled through the availability of massive amounts of labeled data, coupled with enhanced computational capabilities of advanced hardware platforms such as Graphics Processing Units (GPUs). Despite these impressive advances, it still takes significant time and energy to train and deploy these models on leading edge hardware. Furthermore, the complexity of these models makes it challenging to perform inference using deep learning algorithms on resource-constrained IoT devices. As deep learning models become more complex, emerging hardware platforms are critical for future Artificial Intelligence (AI) breakthroughs. In this special issue, we aim to address these emerging areas and provide a comprehensive perspective on various aspects of hardware system and circuits research for future deep learning applications.

[Scope]

To cover the rapid progress of emerging areas we plan to organize papers in three topics:

1. *Digital deep learning processor.* This session aims at digital DNN processing hardware; this includes temporal parallelism architectures (such as GPU, parallel threads, SIMD), as well as partial parallelism and data-flow architectures (such as FPGA, customized SoC, and ASIC). This session will also include software platform topics such as programming models, firm-ware, accelerator evaluation tools, EDA tools for digital deep learning processors.
2. *Analog and in-memory computing approaches to deep learning.* This session highlights integration of computation into memory to save energy by reducing data movement; it also includes analog computation, ADC/DAC design, and SRAM modifications. For deep learning workloads, the communication between memory units and the location of computation can dominate the energy consumption and impact computation throughput. In-memory computing is an architecture design approach that integrates some forms of memory and compute to reduce data transfer costs and improve chip efficiency. In addition to in-memory computing, custom analog circuit design for deep-learning workloads is also included in this special issue.
3. *Algorithm-hardware interaction for deep learning.* This special issue plans to publish papers presenting novel quantization schemes, pruning, sparsity exploration, compression techniques, and distribution strategies (data- and model-parallelisms, synchronization etc.) for deep neural networks: hardware-centric deep learning algorithms. This session also intended to discuss different reinforcement learning methods amenable to hardware efficient AI models and accelerators architectures.

.

Topics of interest

- Hardware-efficient deep learning model architectures for training and inference
- Energy-efficient deep learning inference accelerators

IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS

- Quantization, pruning, and sparsification techniques for hardware-efficient deep learning algorithms
- Distributed and parallel learning algorithms, systems, and demonstrations
- Deep learning system demonstrations integrating sensors, cloud, Internet of Things, wearable devices, device-cloud interactions, and home-intelligence devices.
- Customized digital deep learning processors, FPGA, CGRA, dataflow and specific temporal architectures
- Analog and in-memory computing approaches to deep learning
- Brain-inspired non von Neumann architectures
- Accelerator evaluation tools, EDA tools for deep learning accelerator development
- Customized hardware/software co-designs for deep learning
- Machine learning system interfaces
- Deep reinforcement learning for hardware efficient AI models and hardware designs

Submission Procedure

Prospective authors are invited to submit their papers following the instructions provided on the JETCAS web-site: http://ieee-cas.org/pubs/jetcas/submit-manuscript. The submitted manuscripts should not have been previously published nor should they be currently under consideration for publication elsewhere. Note that the relationship to screen content video technologies should be explained clearly in the submission.

Important dates

• Manuscript submissions due	Nov. 19, 2018
• First round of reviews completed	Jan. 7, 2019
• Revised manuscripts due	Feb. 18, 2019
• Second round of reviews completed	March 18, 2019
• Notification of acceptance:	March 25, 2019
• Final manuscripts due	April 18, 2019

Request for information

Corresponding Guest Editor: Chia-Yu Chen, , IBM Research AI, IBM T.J. Watson Research Center, USA (cchen@us.ibm.com)