Guest Editors' Introduction: Frontiers of Hardware and Algorithms for On-chip Learning

YU CAO, Arizona State University

XIN LI, Duke University

JAE-SUN SEO, Arizona State University

GANESH DASIKA, Arm Research

Machine learning is an important technique that has been continuously growing during the past several decades. It has now been applied to numerous emerging domains, including healthcare, smart manufacturing, business intelligence, autonomous driving, etc. Today, machine-learning algorithms are often implemented with custom integrated circuits, devices, and systems to achieve high throughput, low power, and a small form factor. These requirements, however, are mutually exclusive and very difficult to simultaneously satisfy by state-of-the-art design methodologies.

This motivates investigations from two directions. First, efficiently implementing leading-edge machine learning algorithms on hardware requires us to propose novel circuits, devices, and architectures that may substantially deviate from the conventional CPU/GPU systems. Second, the resource constraints posed by hardware platforms motivate us to re-think the fundamental theories and formulations of machine learning algorithms.

With the support of the ACM Special Interest Group in Design Automation and the IEEE Council of Electronic Design Automation, we organized the second workshop on Hardware and Algorithms for Learning On-a-chip (HALO) in 2016. Our workshop aimed to facilitate a close collaboration between the communities of hardware design and machine learning algorithms, thereby establishing and growing the interdisciplinary area of on-chip machine learning.

This special issue solicits high-quality presentations from HALO 2016. In addition, a general call-for-papers was distributed to the research community to solicit a broader range of participation. After the peer-review process, the following six articles were accepted for publication:

1. Deep Neural Network Optimized to Resistive Memory with Nonlinear Current-Voltage Characteristics
2. Energy-Efficient Neural Computing with Approximate Multipliers
3. Real-Time and Low-Power Streaming Source Separation Using Markov Random Field
4. A GPU-Outperforming FPGA Accelerator Architecture for Binary Convolutional Neural Networks

Authors' addresses: Y. Cao, 781 E Terrace Road, ISTB4 563, Arizona State University, Tempe, AZ 85287; email: Yu.Cao@asu.edu; X. Li, 130 Hudson Hall, Duke University, Box 90291, Durham, NC 27708; email: xinli.ece@duke.edu; J. Seo, 781 E Terrace Road, ISTB4 551B, Arizona State University, Tempe, AZ 85287; email: jaesun.seo@asu.edu; G. Dasika, Arm, 5707 Southwest Parkway, Bldg. 1, Suite 100, Austin, TX 78735; email: Ganesh.Dasika@arm.com.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

2018 Copyright is held by the owner/author(s).

1550-4832/2018/07-ART14

<https://doi.org/10.1145/3205944>

5. A Study of Complex Deep Learning Networks on High Performance, Neuromorphic, and Quantum Computers
6. Online Adaptation and Energy Minimization for Hardware Recurrent Spiking Neural Networks

We thank all authors who submitted their papers to this special issue. We also express our sincere gratitude to the reviewers for providing high-quality reviews in a timely fashion to enhance the quality of the accepted papers. We are indebted to Dr. Yuan Xie, the Editor-in-Chief of ACM Journal of Emerging Technologies in Computing Systems, and his editorial team for their consistent support on this special issue. We hope that the readers will enjoy the articles selected for this special issue and that it will facilitate further research in this exciting area.