IEEE/ACM Workshop on Hardware and Algorithms for Learning On-a-chip (HALO)
Summary Report

1. Background

Machine learning algorithms are being developed and will fundamentally alter the way individuals and organizations live, work, and interact with each other. However their computational complexity still challenges the state-of-the-art computing platforms, especially when the application of interest is tightly constrained by the requirements of low power, high throughput, small latency, etc. In recent years, there have been enormous advances in implementing machine learning algorithms with application-specific hardware (e.g., FPGA, ASIC, etc.). There is a timely need to map the latest learning algorithms to physical hardware, in order to achieve orders of magnitude improvement in performance, energy efficiency and compactness. Recent progress in computational neurosciences and nanoelectronic technology, such as resistive memory devices, will further help shed light on future hardware-software platforms for learning on-a-chip.

The workshop on Hardware and Algorithms for Learning On-a-chip (HALO) is organized to explore the potential of on-chip machine learning, to reveal emerging algorithms and design needs, and to promote novel applications for learning. It aims to establish a forum to discuss the current practices, as well as future research needs in the fields. HALO is collocated with International Conference on Computer-Aided Design (ICCAD) 2016. As the premium conference on IC design automation, ICCAD attracts many experts from industry and academia.

2. Workshop Statistics

The HALO Workshop was successfully held at Doubletree Hotel in Austin on November 5th, 2015. Because this is the very first time, the workshop organizers decided to invite eleven speakers from universities and leading companies. In addition, a poster session was organized with 31 posters by university students and researchers. There were 73 attendees participated the workshop. Four sessions were presented on the hardware acceleration, learning algorithms and neuromorphic hardware design. The detailed agenda was posted at http://nimo.asu.edu/halo.

3. Technical Summary

The speakers presented a comprehensive view on hardware acceleration, learning algorithms, and neuromorphic computing. The development of deep learning algorithms, such as convolutional neural network (CNN), have led to the breakthroughs in the accuracy of many computer vision and information processing tasks. However, their multi-layer structure is enormous and complex, requiring substantial computing resources to train and evaluate. Indeed, machine learning with big data has become one of the most important workloads in both data centers and mobile platforms. Even with state-of-the-art CPUs/GPUs, an acceleration factor of 10^2-10^4 is critically needed to perform real-time video analytics. While advanced hardware solutions, such as IBM TrueNorth, help bring expensive learning and classification to a low-power processor, they are still much less efficient compared to the human brain. In this context, we need to fundamentally re-think the learning algorithm and hardware architecture, inspired by mathematical and neuromorphic principles at multiple levels.

During the workshop, several presenters discussed acceleration of large-scale learning algorithms (e.g., deep CNNs, logistic regression, and specialized image recognition) on FPGAs and ASICs. Heterogeneous hardware offers a promising path towards major improvement in processing capability while achieving high energy efficiency. It was demonstrated that by co-optimizing the algorithm and the architecture, these implementations effectively reduce the latency in model training and evaluation. The

results enable a broad range of applications, including human-like visual capability, driving assistance, and human-machine symbiotic data management.

In addition to hardware architecture, the development of next-generation learning on-a-chip requires novel algorithms that are able to organize the objects in a hierarchy, transfer previous knowledge, and train the model dynamically and on line. Traditional deep learning algorithms, with a flattened model and off-line training, fail in many mobile applications, such as automatic target acquisition and tracking, robotics, and autonomous vehicle. The focus of the discussion was how to evolve CNN, recurrent neural network, Bayesian optimization, and other algorithms towards these capabilities, while still satisfying the constraints of latency and power in hardware implementation.

The research on on-chip learning further outreaches to computation models and hardware that are beyond conventional digital design. Examples include recent advances of the spiking TrueNorth system, as well as various neuromorphic machines that are inspired by the brain. These approaches mimic the principles of the cortical and sensory systems at the device, circuit, and network levels. Initial results illustrate significant opportunities to perform learning, inference, memory and other important tasks.

4. Looking Forward

With ever-increasing complexity of machine learning algorithms and the need of information analytics on a mobile platform, the speed and power consumption of today's computing hardware are greatly challenged. Recent advances in sensing technology further exacerbate the situation with a larger volume of data. Therefore, a holistic approach of concurrent innovations in algorithms and hardware is essential to accelerate learning and classification on a chip under stringent power constraints.

Following this conclusion, we plan to continue our workshop next year, with possible joint presentation from learning algorithms and hardware design for mobile applications.

5. Acknowledgement

We would like to thank National Science Foundation (NSF) and ACM Special Interest Group on Design Automation (SIGDA) for their support to our workshop, as well as the technical support by IEEE Council on Electronic Design Automation (CEDA). We also would like to acknowledge the support by ICCAD organization committee to help manage the workshop.