Acceleration of Machine Learning Algorithms for Big Data Applications

Thursday, August 23, 2018, 3:00 p.m.

ENGR 3202

Katayoun Neshatpour

Advisors: Dr. Houman Homayoun, Dr. Avesta Sasan

Abstract:

Advances in various branches of technology - data sensing, data communication, data computation, and data storage have significantly changed machine learning in recent years. With new computing technologies allowing vast amount of data to be stored and creating opportunities to learn from the aggregated data, implementation of classical machine-learning algorithms including K-means, KNN, has been improved to meet the big data requirements of such applications. Moreover, the enhanced processing power of today's computing nodes enables training of more sophisticated learning models, paving the way for deep learning algorithms including convolutional neural networks (CNN).

In big data domain, various frameworks have been developed that allow the processing of large data sets with parallel and distributed algorithms, with MapReduce being an example of such frameworks. The first section of this research focuses on classical Machine-Learning applications and their hardware acceleration in MapReduce platform. Profiling of applications in this research shows that the map and/or reduce phase of these applications take up most of the execution time. Subsequently, the map functions were accelerated to evaluate how a cluster of CPUs equipped with FPGAs uses the accelerated mapper to enhance the overall performance of MapReduce. Moreover, this research studies how the type of FPGA (low-end vs. high-end), and its integration with the CPU (on-chip vs. off-chip) along with the choice of CPU (high performance big vs. low power little servers) affects the speedup yield and power reduction.

While FPGA acceleration of MapReduce offers superior energy efficiency, but MapReduce platform is not suitable for implementation of deep neural networks including CNNs. In the second part of this research an iterative approach is proposed to break down large CNNs into a sequence of smaller networks (uCNN), each processing a sub-sample of the input image, providing the ability to terminate the classification early or carry the classification to the next iteration in case of non-satisfactory confidence levels. Moreover, the contextual information resulted from early iterations of ICNN can be used to reduce the complexity of the subsequent iterations. To explore the complexity-accuracy tradeoff of ICNN, a dynamic deadline-driven exit policy for real-time applications, a confidence thresholding policy for dynamic complexity reduction, a context-aware pruning policy for parameter reduction and two hybrid pruning and thresholding policies for simultaneous parameter and complexity reduction were introduced. Simulation results on a case study with iterative AlexNet shows that with intelligent selection of the pruning and/or thresholding policies, ICNN reduces the average FLOP and parameter counts, and execution time across 50K validation images in ImageNet database by more than 25%, 80% and 38%, respectively, with negligible accuracy loss. Moreover, the real-time systems could exploit the dynamic structure of the ICNN by reducing the execution time by upto 12× by trading off accuracy with execution time.