Towards Efficient Architectural Support for AI-based IoT Applications

Prof. Tao Li
Professor
Department of Electrical and Computer Engineering
University of Florida
USA

Date: 2 May 2018 (Wednesday)
Time: 11:00 a.m. – 12:00 noon
Venue: Room PQ703, 7/Floor, PQ Core, Mong Man Wai Building,
The Hong Kong Polytechnic University

Abstract
In recent years, the artificial intelligence (AI) techniques, represented by deep neural networks (DNN), have demonstrated transformative impacts to modern Internet-of-Things (IoT) applications such as smart cities and smart transportation. With the increasing computing power and energy efficiency of mobile devices, there is a growing interest in performing AI-based IoT applications on mobile platforms. As a result, we believe the next-generation AI-based applications are pervasive across all platforms, ranging from central cloud data center to edge-side wearable and mobile devices.

However, we observe several architectural gaps that challenge the pervasive AI. First, the diversity of computing hardware resources and different end-user requirements present challenges to AI-based applications deployment on various IoT platforms, which results in inferior user satisfaction. Second, the traditional statically trained DNN model could not efficiently handle the dynamic data in the real IoT environments, which leads to low inference accuracy. Lastly, the training of DNN models still involves extensive human efforts to collect and label the large-scale dataset, which becomes impractical in IoT big data era where raw IoT data is largely un-labeled and un-categorized.

In this talk, I will introduce our recent research which enables pervasive AI-based IoT applications to become high-efficient, user-satisfactory, and intelligent. I will first introduce Pervasive AI, a user satisfaction-aware deep learning inference framework, to provide the best user satisfaction when migrating AI-based applications from Cloud to all kinds of platforms. Next, I will describe In-situ AI, a novel computing paradigm tailored to AI-based IoT applications. Finally, to achieve real intelligent (support autonomous learning) in IoT nodes, I will introduce an unsupervised GAN-based deep learning accelerator.

About the Speaker
Prof. Tao Li is a Full Professor in the Department of Electrical and Computer Engineering at the University of Florida. He received a Ph.D. in Computer Engineering from the University of Texas at Austin. His research interests include computer architecture, microprocessor/memory/storage system design, virtualization technologies, energy-efficient/sustainable/dependable data center, cloud/big data computing platforms, the impacts of emerging technologies/applications on computing, and evaluation of computer systems. Prof. Tao Li received 2009 National Science Foundation Faculty Early CAREER Award, 2008, 2007, 2006 IBM Faculty Awards, 2008 Microsoft Research Safe and Scalable Multi-core Computing Award and 2006 Microsoft Research Trustworthy Computing Curriculum Award. Prof. Tao Li co-authored two papers that won the Best Paper Awards in ICCD 2016, HPCA 2011 and seven papers that were nominated for the Best Paper Awards in HPCA 2018, HPCA 2017, ICPP 2015, CGO 2014, DSN 2011, MICRO 2008 and MASCOTS 2006. Prof. Tao Li is one of the College of Engineering winners, University of Florida Doctor Dissertation Advisor/Mentoring Award for 2013-2014 and 2011-2012. During 2015-2017, Prof. Tao Li served at the National Science Foundation as a Program Director within the CISE directorate. He is a Fellow of IEEE.

All are welcome!
Enquiries:
Professor George Baciu
Email: csgeorge@comp.polyu.edu.hk
Tel: 2766 7295 / 2766 7272

We drive innovation through SMART COMPUTING