Tutorial 5: Synapses, Circuits, and Architectures for Analog In-Memory Computing-Based Deep Neural Network Inference Hardware Acceleration

Irem Boybat, IBM Research Europe

Abstract:

The advent of deep neural networks (DNNs) has revolutionized numerous fields, including computer vision and natural language processing. These powerful models have showcased remarkable capabilities in solving complex problems, but their training and inference procedures often require significant computational resources. To address this challenge, notable activity was centered around specializing or developing digital hardware accelerators for DNNs, such as graphics processing units (GPUs) and tensor processing units (TPUs). However, this talk will go beyond digital acceleration and instead focus on the emerging field of analog in-memory computing (AIMC). More specifically, it will delve into devices, synapses, circuits, and architectures for building energy efficient yet accurate AIMC-based inference accelerators for DNNs. We will explore how the AIMC paradigm can overcome the limitations of traditional digital computing approaches and offer better energy efficiency by blurring the distinction between memory and computing. We will investigate the impact of device characteristics and their organization into synapses on DNN inference accuracies. Furthermore, we will discuss the role of peripheral circuits and accelerator architectures on energy efficiency and performance.

Bio:

Dr. Irem Boybat is a Research Scientist at IBM Research Europe, Zurich, Switzerland. She received her Ph.D. degree in Electrical Engineering from Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland, in 2020. Previously, she had obtained an M.Sc. degree in Electrical Engineering from EPFL, Switzerland, in 2015, and a B.Sc. degree in Electronics Engineering from Sabanci University, Turkey, in 2013. Her research is primarily centered around analog in-memory computing for accelerating deep neural networks using phase-change memory devices. She has co-authored over 50 scientific papers in journals and conferences, received four best conference presentation/paper/poster awards and holds 8 granted patents. She was a co-recipient of the 2018 IBM Pat Goldberg Memorial Best Paper Award and 2020 EPFL PhD Thesis Distinction in Electrical Engineering.