Big AI for Small Devices

Yiran Chen

Duke University, USA
E-mail: yiran.chen@duke.edu

Abstract

As artificial intelligence (AI) transforms industries, state-of-the-art models have exploded in size and capability. However, deploying them on resource-constrained edge devices remains extremely challenging. Smartphones, wearables, and IoT sensors face tight limits on compute, memory, power, and communication. This gap between demanding AI models and edge hardware capabilities hinders onboard intelligence. In this talk, we will re-examine the techniques to bridge this gap and embed big AI on small devices. First, we will boost single-device efficiency via model compression. We will discuss how the properties of different hardware platforms impact the quantization and pruning strategies of deep neural network (DNN) models, benefiting actual system throughput and memory usage when considering the execution process of the models. Second, we will discuss the designs aimed at reducing the communication, computation, and storage overheads for distributed edge AI systems. We will also delve into the underlying design philosophies and their evolution toward efficient, scalable, robust, and secure edge computing systems.