Optimizing inference efficiency for tiny DNNs

Harris Teague, Principal Engineer, Qualcomm AI Research

In this talk, I will explore some of the ways that we are working on improving model inference efficiency for tiny devices – where power, area, memory, compute resources are limited. I will present results for a few of these: compute scheduling optimization, model compression, quantized inference, and in-memory computing. Finally, I will discuss our plans for next research steps to further understand and develop the technology.