Thinking Big with Tiny ML: Low Power High Performance DNN Accelerators for Mobile and IoT Applications

KAIST ICT Endowed Chair Professor, School of Electrical Engineering, KAIST

The artificial intelligence (AI) revolution is being widely spread even to the IoT with the help of 5G wireless communication. Compared to the Cloud-based or Edge-based AI applications, Internet-of-Things (IoT) applications require more autonomous, adaptive, and cooperative operations with extremely limited power, computing and memory resources without stable communication channels. AI, especially deep neural network (DNN), is the key technology to support such autonomy and adaptivity of the IoT machines in an unpredictable environment with limited available information. The IoT machines should contain not only inference but also training capabilities to adapt to environmental changes based on their experiences. Therefore, software and hardware co-optimization for DNN training is necessary for low-power and high-speed accelerators, in the same way it brought a dramatic increase in the performance of DNN inference accelerators. In addition, deep reinforcement learning (DRL) accelerators will be an essential part of the tide, showing a lot of benefits at making continuous decisions in an unknown environment, where labeled data is difficult to acquire.