Fast neural network inference on xcore.ai

Laszlo Kindrat and Andrew Cavanaugh, XMOS

Over the last few years, several hardware architectures have been developed or extended to perform neural network inference on low cost embedded devices. The approaches involving new architectures show high performance in AI tasks, but tend to lack features for general purpose computing. At the same time some of the more established general-purpose architectures have been extended with vector units or SIMD instructions that delivered only limited performance on AI tasks. Moreover, I/O performance and real time guarantees have rarely been taken into consideration during the design and evaluation of such MCU platforms. XMOS developed the xcore.ai platform which will deliver state-of-the-art performance on edge inference tasks, high speed I/O, and flexible, software-defined connectivity.

The xcore.ai platform implements XMOS' next generation architecture, featuring a novel vector unit designed for fast neural network inference on edge devices, including support for binarized networks. The platform is supported with tools that: 1) perform network optimization for the platform; 2) allow easy prototyping and deployment using TensorFlow Lite for Microcontrollers; 3) generate high performance code that can be compiled directly or modified as needed. These tools provide a seamless experience of testing and evaluating models on the device, even for users without much experience in the embedded device world.

We present benchmarks on quantized and binarized neural network architectures for image classification problems. We compare the xcore.ai platform to competitors and show that on the same benchmark it performs inference 7 to 30 times faster. We explain the model transformations done by the tools that allow optimal network performance on xcore.ai and illustrate the model deployment workflow.