Abstract

High-Efficiency Neural Network Inference using DesignWare ARC EMxD Processors & TensorFlow Lite for Microcontrollers
Jamie Campbell
Software Engineering Manager
Synopsys, Inc.

Deeply-embedded AIoT applications doing Neural Network (NN) inference need to achieve the stated real-time performance requirements on systems with limited memory and power budget. Meanwhile, developers want a convenient way of migrating their NN graphs to a deeply-embedded environment. This poster describes how we created a highly-efficient AIoT inference solution based on an industry-standard front end combined with a highly optimized hardware + software backend.

Synopsys DesignWare ARC EMxD Processors are well-suited for AIoT NN inference thanks to their small size, unrivaled performance efficiency and low power consumption. The combination of vector multiply-accumulate (MAC) DSP instructions with features like XY memory and Address Generation Units (AGUs), enables the ARC EM9D processor to achieve an effective throughput of 2 MACs every cycle, which allows it to perform NN operations, like convolutions, very efficiently.

The embARC Machine Learning Inference (MLI) Library is an open-source, C software library optimized for ARC processors. It includes a comprehensive set of kernels targeted at deeply embedded ML applications.

TensorFlow Lite for Microcontrollers (TFLM) is a port of TensorFlow Lite designed to run ML on resource-constrained microcontrollers. Developers can leverage existing TensorFlow environments for graph development, training and testing, and then use established flows to convert their graphs to TensorFlow Lite and finally to TFLM data structures. Synopsys has developed an optimized implementation for TFLM targeting EM processors, using highly-optimized kernels from the MLI library as the backend.

Himax Technologies has created the WE-I Plus AI accelerator ASIC targeting low-power AIoT applications.   In order to provide optimal performance of tinyML applications on their accelerator, Himax selected the ARC EM9D processor for its high MAC utilization and low power consumption, and will provide a software solution which leverages the optimized MLI library working as a backend to the industry-standard TensorFlow Lite for Microcontrollers framework. This integrated hardware + software solution provides developers with a high-performance, low-power and easy-to-use solution for their ML/AIoT development needs. Target AIoT applications for the WE-I Plus include smart home applications, battery-powered surveillance applications, and tiny AI value-added applications like AI TVs and AI air conditioners.