everyday genius

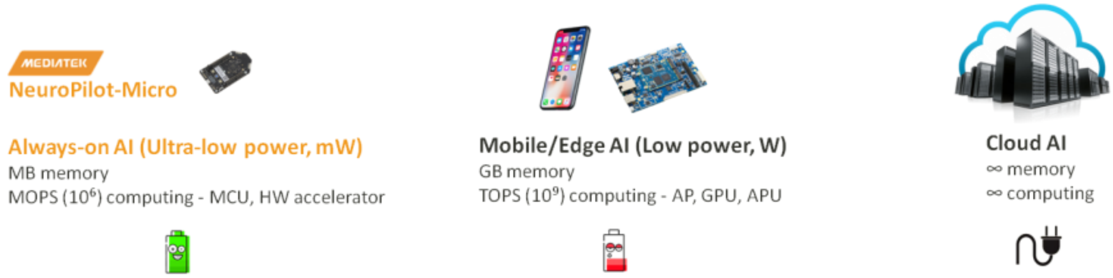
NeuroPilot-Micro Introduction

Version: 0.2 (Beta Release)

Release date: 2020-07-10

NeuroPilot-Micro

TinyML brings you a World of AI that's Ultra-Efficient and Always-On


MEDIATEK
NeuroPilot-Micro
Always-on AI (Ultra-low power, mW)
MB memory
MOPS (10^6) computing - MCU, HW accelerator
Mobile/Edge AI (Low power, W)
GB memory
TOPS (10^9) computing - AP, GPU, APU
Cloud AI
∞ memory
∞ computing

TinyML (Tiny Machine Learning) is an emerging software technology that enables machines to learn and become **smarter** through use. With the ability to deploy compressed machine learning algorithms into small embedded devices, such as a microcontrollers (MCUs), new AI applications such as vision, audio, and sensor-related applications can now operate in an **always-on** mode and with ultra-low power.

The ability to incorporate machine-learning models into microcontrollers will enable a whole new ecosystem. Challenges remain, however, and include the following:

- Applications (algorithms)
  Machine learning software algorithms must be small enough to fit into microcontrollers and the reference results must still yield good accuracy.
- The hardware/software design
  A good use of hardware/software co-operation is needed. Unlike cloud servers, microcontrollers usually offer limited resources such as smaller memory sizes, lower energy, and fewer computational capabilities.

In acknowledgement of the challenges above, as well as recognition of the benefits that can be achieved, MediaTek has created a TinyML framework offering called NeuroPilot-Micro. The main objective of NeuroPilot-Micro is to simplify and optimize the deployment of custom, specific machine learning models on microcontrollers. This, coupled with MediaTek AIoT solutions, can enable more streamlined and optimized high-performing products.

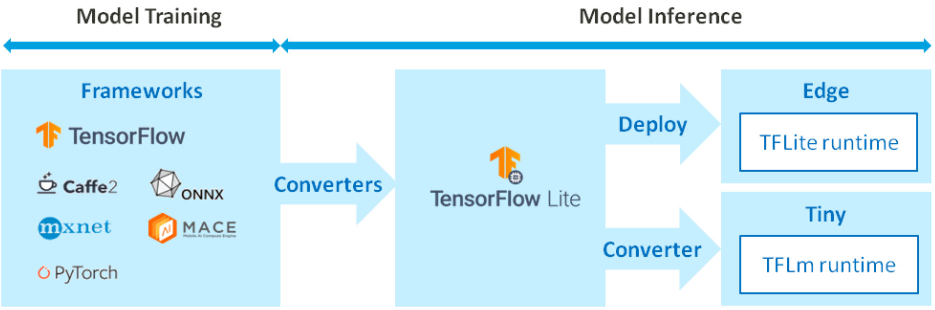
The above video shows some always-on applications including object detection; handwriting recognition; shelf detection; person detection; and dual-core processing demo. These have been implemented using MediaTek's NeuroPilot-Micro framework and deployed on the MT3620, a high performance, tri-core AIoT microcontroller.

While Flash memory is normally used in microcontrollers, one of its limitations is that the access speed is generally slow. Response times for detection can often take several seconds if the machine learning models directly run using the Flash memory. With NeuroPilot-Micro optimizations deployed on the MT3620, though, this is not an issue. As shown in the video above, response times show great improvement, decreasing the response time from several seconds to approximately 1 second.

NeuroPilot-Micro

NeuroPilot-Micro is a framework for executing ultra-low power, machine learning applications on microcontrollers. It includes a micro-runtime as well as op libraries that can optimize machine learning inference to be executed efficiently with low-power consumption (mW); limited memory capacity (MB); and low computing power (in MHz).

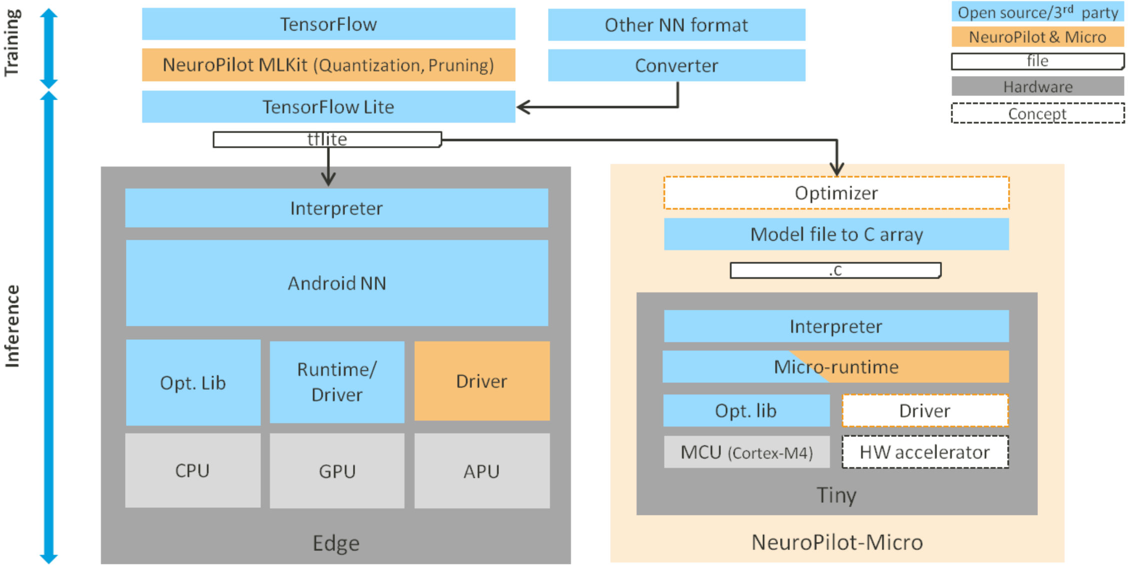
TFLite NN Model Training, Inference



[TensorFlow Lite for Microcontrollers (TFLm)](#) is the input model for NeuroPilot-Micro. If you use other AI frameworks, like TensorFlow, Caffe, ONNX, etc., they will need to be converted into the TensorFlow Lite (TFLite) model. Some useful converter tools can be obtained from open source or third-party software. When considering the development activities on microcontrollers, though, TFLm model supports fewer operations than TensorFlow Lite.

The TFLm NN model training and inference flow is depicted in the above Figure.

NeuroPilot-Micro Flow



The figure above illustrates the working flow of NeuroPilot-Micro. The box labeled Edge is the [NeuroPilot](#) inference framework and the box labeled Tiny is the NeuroPilot-Micro inference framework, where NeuroPilot-Micro framework leverages some functionalities from the NeuroPilot SDK.

The components in the orange colored boxes are the major functionalities provided by **NeuroPilot-Micro framework**.

- **NeuroPilot ML Kit** (Quantization, Pruning)
  For model optimization, the TensorFlow model can be quantized into fixed point format or can reduce model size by pruning tools first. This part can use *Neuro Pilot ML Kit* tools, Model Converter and Quantization Tool and Network Reduction Tool.
- **Micro-runtime**
  The micro-runtime offers memory and low power optimization mechanisms to enable the trained inference to run quickly and efficiently on microcontrollers.
- **Optimized library** (Optimized lib)
  The NeuroPilot-Micro SDK supports optimized operations by different microcontrollers. The MT3620 platform includes an ARM CMSIS NN Software Library that delivers optimized performance.

Next Steps

The Getting Started and User Guide are also useful when learning how to use the SDK.

Some interested showcases can be found at Demo.