TinyML on STM32

Student

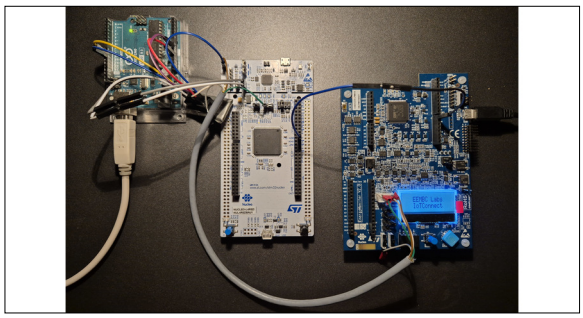
Andri Trottmann

Introduction: TinyML, a sub-genre of Edge-AI, is getting more and more attention, especially in the field of IoT and embedded real-time systems. It is the place where the world of deep learning and embedded systems meet each other. The diversity of hardware, frameworks and applications is huge and TinyML systems face different challenges due to their low power characteristic and constrained resources with just a couple of kB RAM and CPU frequency in the lower MHz area. During this research project the first step into exploring TinyML is made and the MLPerf Inference Tiny benchmark is evaluated.

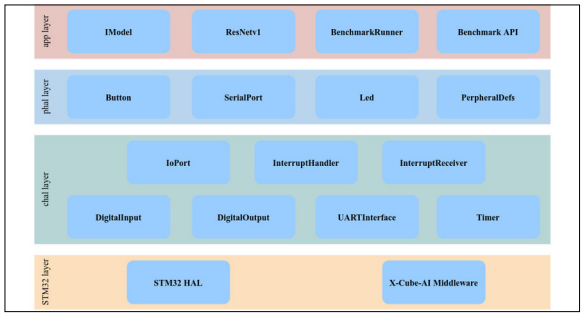
Approach: MLCommons developed a benchmark (MLPerf Inference Tiny) with the goal to provide an unbiased, flexible and reliable as well as representative benchmark for TinyML systems. The benchmark runs in two modes, the Performance and the Energy mode. As an introduction, the image classification benchmark is reproduced and evaluated. The model for this benchmark is a modified version of ResNet8 (see figure). MLCommons provides a reference implementation, which serves as a starting point for other submissions utilizing different frameworks and hardware. Secondly, a solution is implemented with ST's X-Cube-AI framework. This framework fully integrates into STM32CubeMX environment. Finally, the results of three implementations are compared: the reference implementation by MLCommons, a submitted solution by ST and the own solution, both utilizing X-Cube-AI framework.

Conclusion: The connection between IoT and TinyML is natural, because TinyML is the answer to the need for privacy, energy efficiency, responsiveness and autonomy of AI applications. With MLCommons' MLPerf Inference Tiny benchmark a tool for contextualizing future TinyML applications is found. ST's X-Cube-AI framework performs well and excels in its ease of use and integration into the STM32Cube ecosystem. Inference on ResNetv1 with X-Cube-AI took ~170ms, which out-performs the ~664ms of the reference implementation. By optimizing the MCUs peripheral component, up to 10mW in power consumption can be saved.

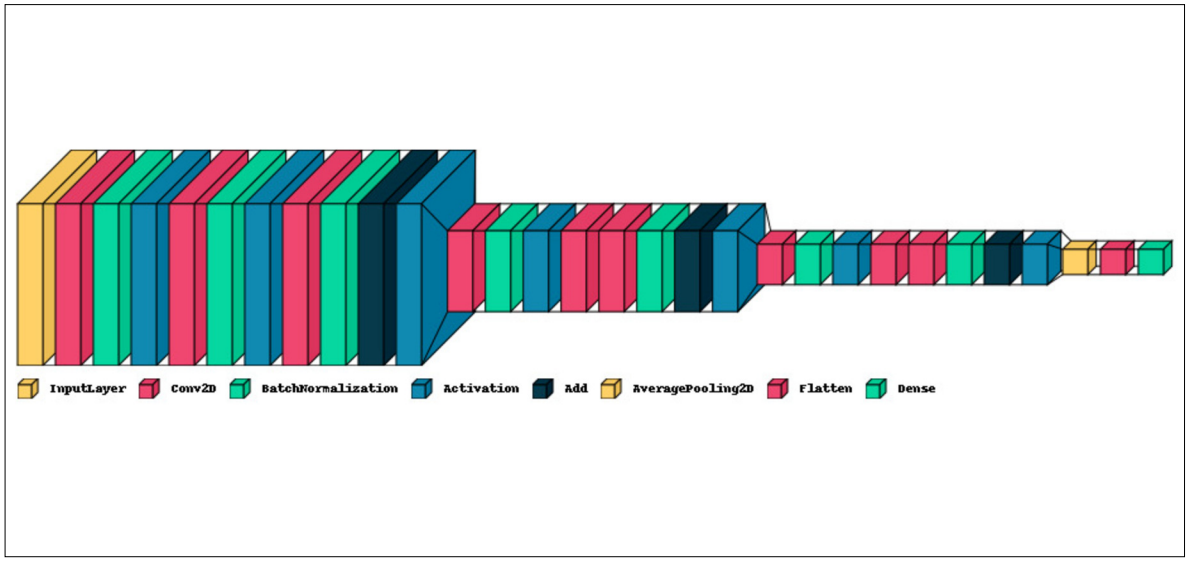
Setup for running the Energy mode benchmark.
Own presentment



Architecture of the Own-Implementation's Firmware
Own presentment



ResNetv1 model, used for image classification in the MLPerf Inference Tiny benchmark.
Own presentment



Advisor
Prof. Dr. Mitra Purandare

Subject Area
Computer Science