Towards Optimal Deployments of Machine Learning Models on Selectable Target Hardwares

MSE Project 2

Student

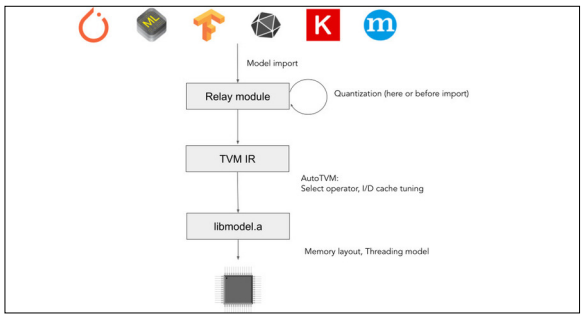
Andri Trottmann

Introduction: Tiny machine learning (TinyML) systems are highly diverse, requiring in-depth knowledge of both embedded systems and machine learning – fields where gaining expertise is a long process. Because of its diversity, it is particularly challenging to find optimal model-to-hardware mappings. To assist companies and people with no or little TinyML experience, a design space exploration software, which identifies the Pareto-optimal hardware-model pairs, is proposed. This project marks the first step towards that goal.

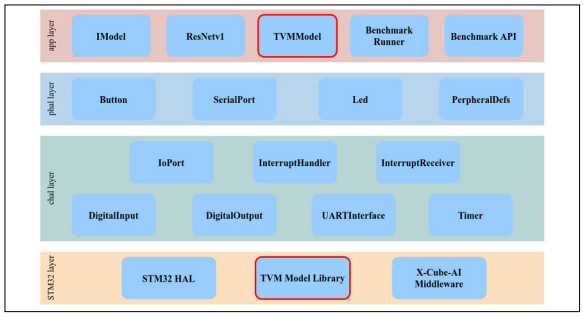
Approach: Performing design space exploration on different hardware targets requires a flexible machine learning framework. Tensor Virtual Machine (TVM) is such a framework. To gain experience with TVM on microcontrollers, the image classification benchmark with Resnetv1 from MLPerf Tiny is taken as target project. The generated TVM model is compiled utilizing Zephyr's HAL and deployed on a Nucleo-L4R5ZI (Arm Cortex-M4). For testing TVM's flexibility when it comes to target hardwares, the model is deployed on two other boards, the Nucleo-F746ZG and NXP's MIMXRT1170-EVK (both Arm Cortex-M7). Additionally, preliminary work includes the development of a basic software structure and the completion of a conceptual software design for a TinyML design space exploration tool.

Conclusion: TVM's strength is its flexibility, but this also makes TVM complex. The achieved runtime performance (measured as inferences per second or inf./s) in a system setup with Zephyr is 1.079 inf./s on Nucleo-L4R5ZI, 2.758 inf./s on Nucleo-F746ZG and 5.412 inf./s on MIMXRT1170-EVK (at four times higher CPU frequency). Despite the lower runtime performance compared to STM32Cube.AI, TVM is well-suited for integration into a design space exploration tool due to its flexibility not only in supporting diverse hardware targets but also in optimizing the model's computational graph. The work conducted in this project provides a foundation for future developments regarding design space exploration.

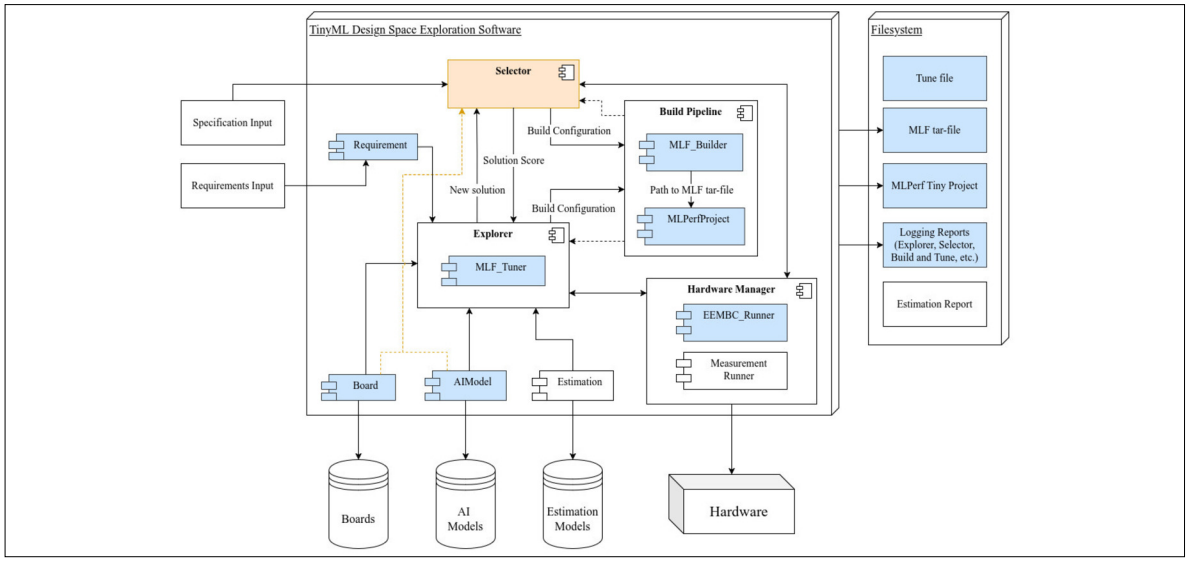
Workflow for creating C code from a TVM model.
https://tvm.apache.org/docs/topic/microtvm/index.html



Firmware architecture newly based on the TVM model.
Own presentment



Proposed software structure of the TinyML design space exploration tool.
Own presentment



Advisor
Prof. Dr. Andreas Breitenmoser

Subject Area
Computer Science