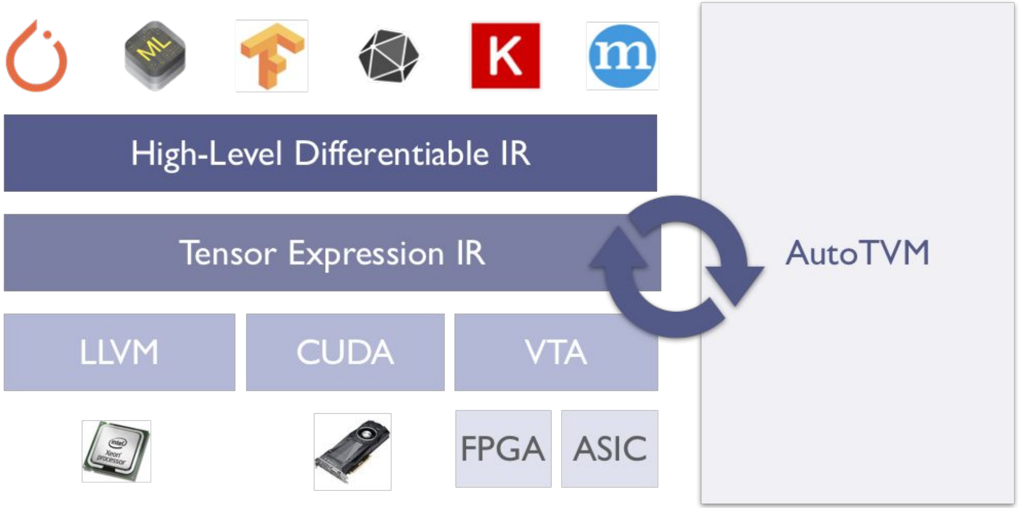
OctoML

µTVM: TVM on Bare-Metal Devices

TVM Conference 12/5/2019
Logan Weber

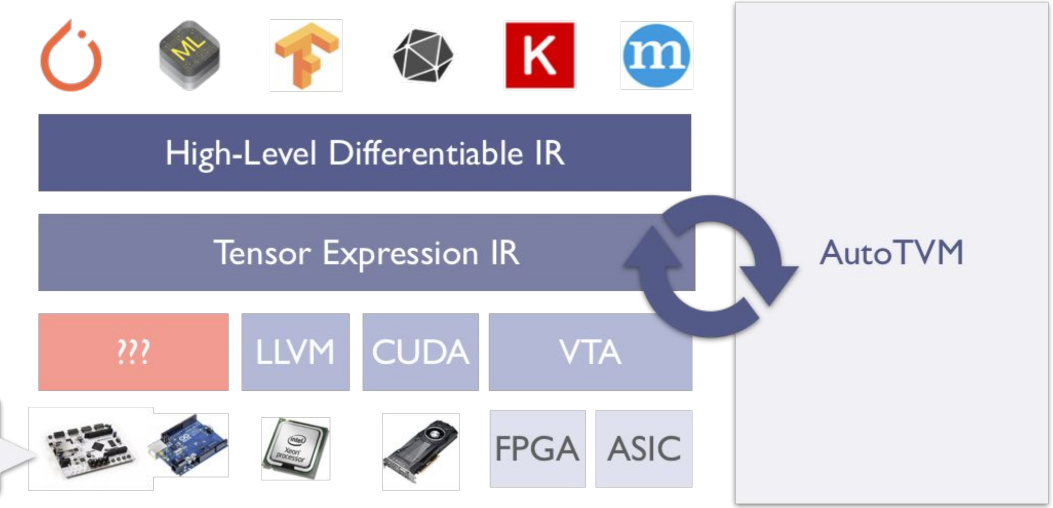
Motivation

Many hardware targets already enjoy speedups from TVM



Motivation

Except for microcontrollers...



High-Level Differentiable IR

Tensor Expression IR

AutoTVM

??? LLVM CUDA VTA

FPGA ASIC

What about us? ☹

Enter µTVM

Device Checklist:

❏ GCC Cross-Compiler

❏ JTAG Support

Enter µTVM

Device Checklist:

- ☑ GCC Cross-Compiler
- ❑ JTAG Support

Enter µTVM

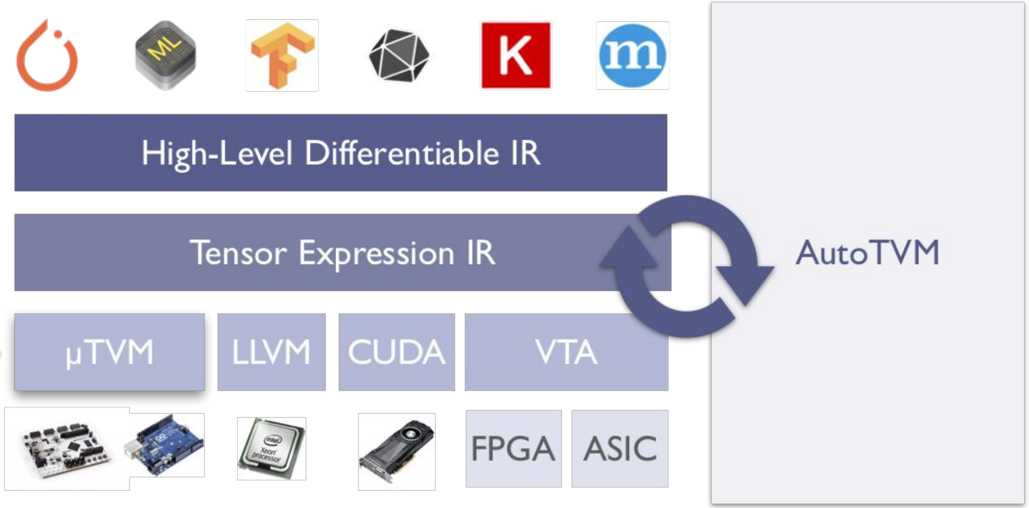
Device Checklist:

☑ GCC Cross-Compiler

☑ JTAG Support

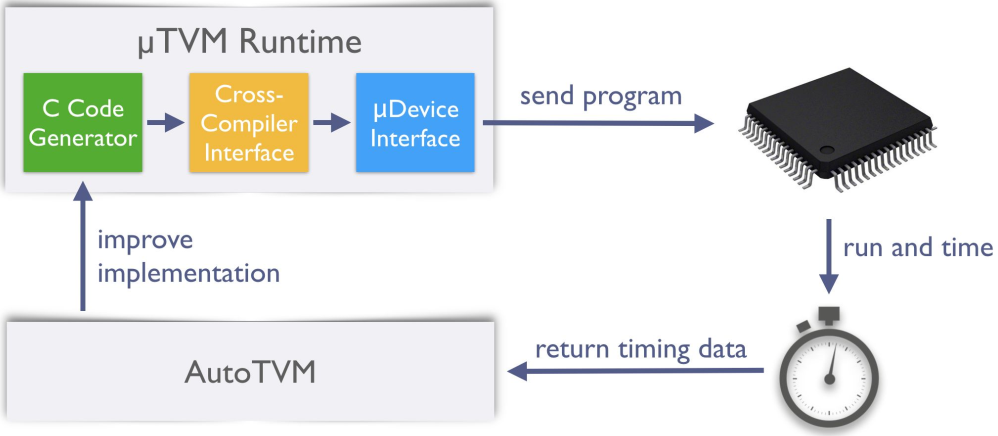
Enter µTVM

- Generate C for operators and feed into cross-compiler
- Use JTAG to read/write memory and execute

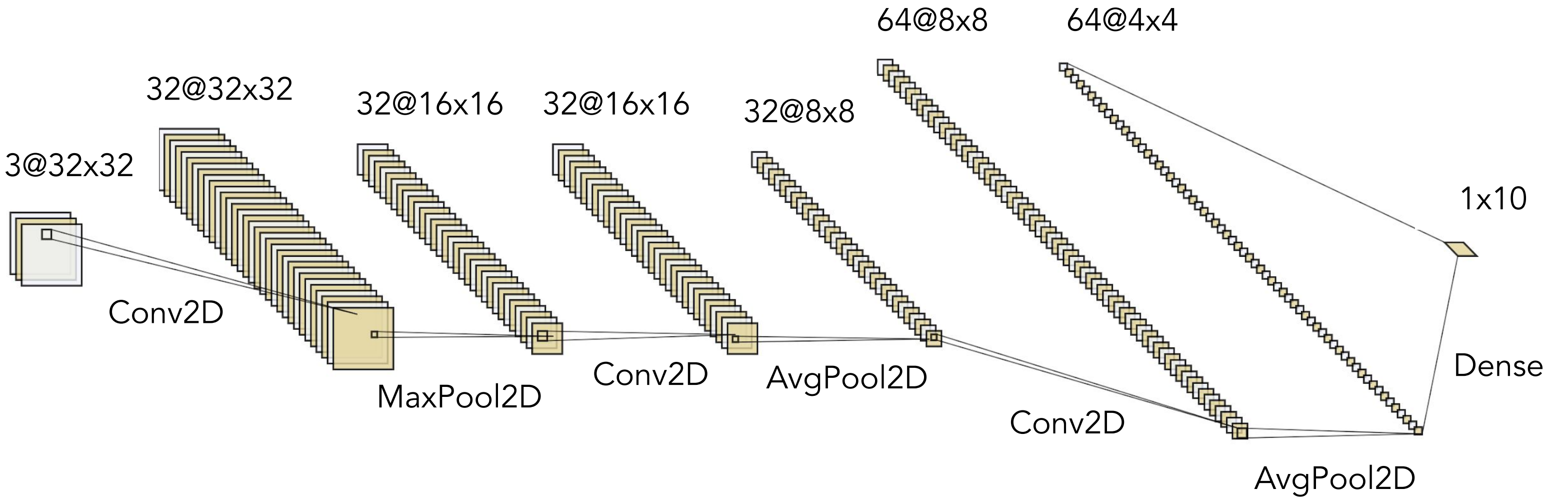


AutoTVM on µTVM

- Same pipeline as usual
- Load kernels into RAM instead of flash



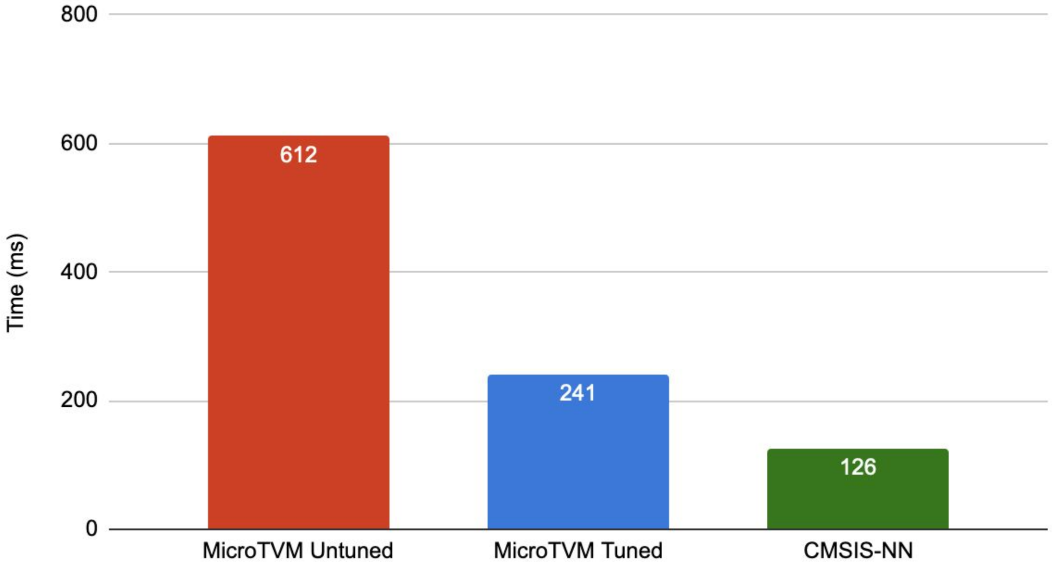
End-to-End CIFAR-10 Evaluation



Replicated an int8-quantized CNN from an ARM Mbed tutorial

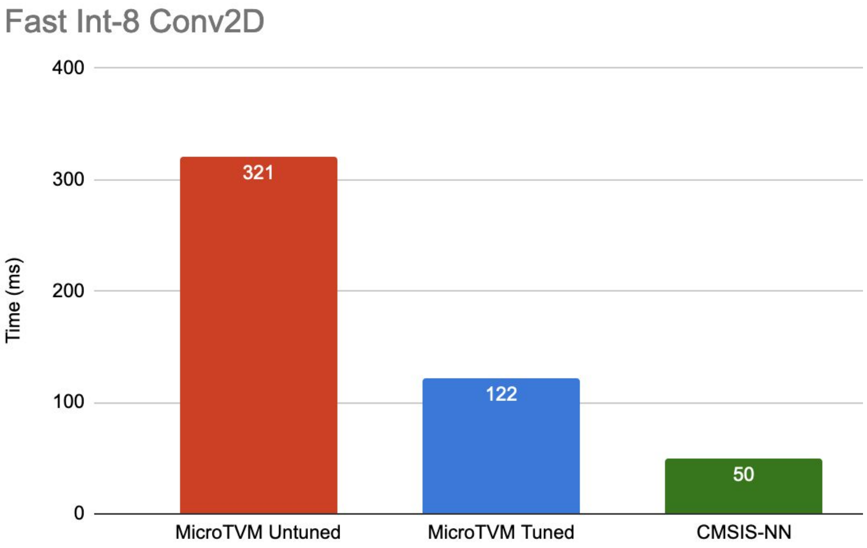
Preliminary CIFAR-10 CNN Results

- Ran on ARM Cortex-M7
- Compared against CMSIS-NN
- Vanilla template
- ~5 hours of tuning
- No vectorization

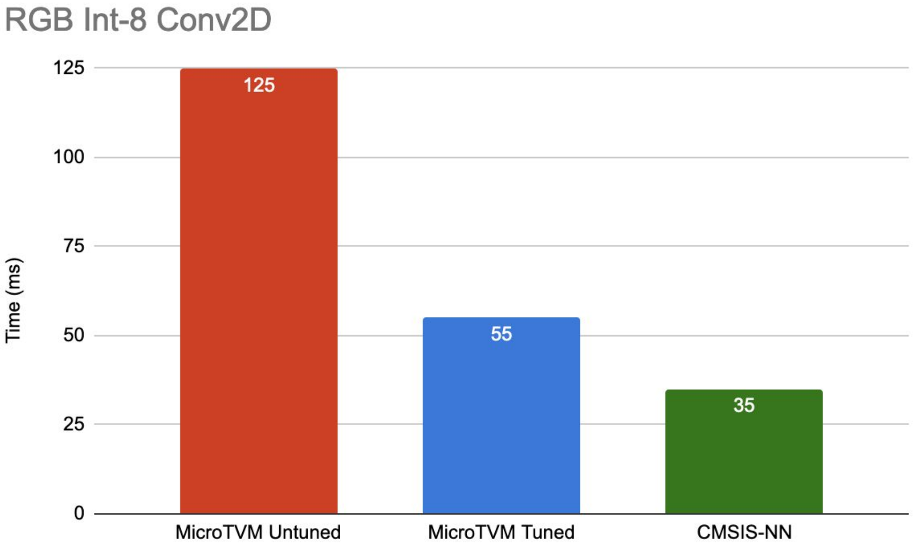


OctoML

Preliminary Int-8 Conv2D Results

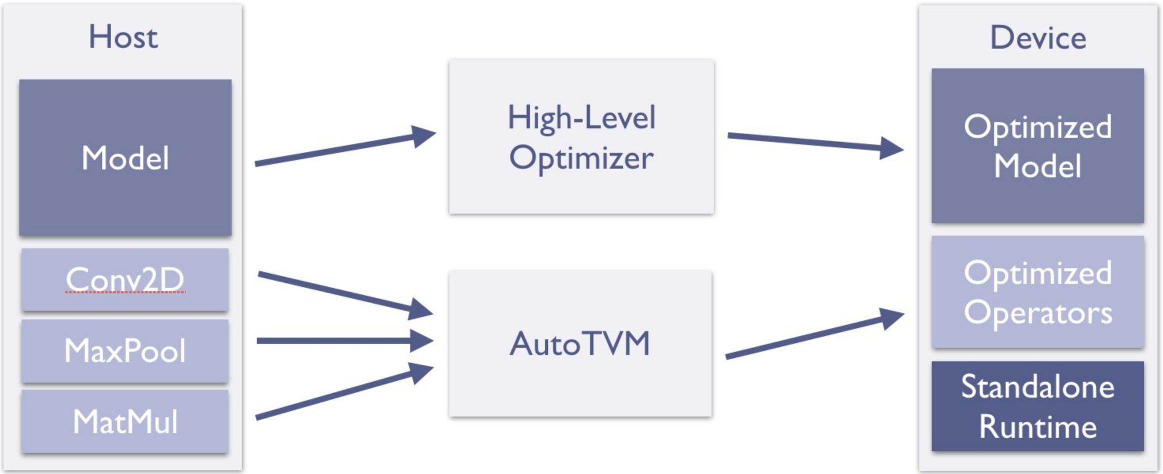


`arm_convolve_HWC_q7_fast` in CMSIS-NN



`arm_convolve_HWC_q7_RGB` in CMSIS-NN

Coming Soon to µTVM (Self-Hosted Models)



Host

Model

Conv2D

MaxPool

MatMul

High-Level Optimizer

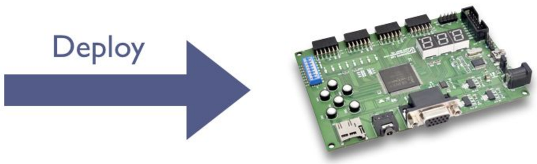
AutoTVM

Device

Optimized Model

Optimized Operators

Standalone Runtime



Stay Tuned!

- An in-depth writeup will be coming soon to the TVM blog

Acknowledgments

- Tianqi Chen, who has provided invaluable mentorship on this project
- OctoML, for allowing me to continue my work on MicroTVM under an internship
- Pratyush Patel, for collaborating on early prototypes

OctoML

Questions?