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...

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
Preliminary Int-8 Conv2D Results

**Fast Int-8 Conv2D**

<table>
<thead>
<tr>
<th></th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MicroTVM Untuned</td>
<td>321</td>
</tr>
<tr>
<td>MicroTVM Tuned</td>
<td>122</td>
</tr>
<tr>
<td>CMSIS-NN</td>
<td>50</td>
</tr>
</tbody>
</table>

**RGB Int-8 Conv2D**

<table>
<thead>
<tr>
<th></th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MicroTVM Untuned</td>
<td>125</td>
</tr>
<tr>
<td>MicroTVM Tuned</td>
<td>55</td>
</tr>
<tr>
<td>CMSIS-NN</td>
<td>35</td>
</tr>
</tbody>
</table>

arm_convolve_HWC_q7_fast in CMSIS-NN  

arm_convolve_HWC_q7_RGB in CMSIS-NN
Coming Soon to μTVM (Self-Hosted Models)
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
Questions?