

TVM optimization to speed up NN inference on ARMv7-M+ (Cortex-M4, M7, M33, M55)



INTERESTED PARTIES:

Apache TVM, Arm

Apache TVM is an open-source deep-learning compiler framework that empowers engineers to optimize and run computations efficiently on any hardware backend.

Arm Ltd. is a British semiconductor and software design company based in Cambridge, England. Its primary business is Arm CPU design, as well as other chips design.



AREA:

Machine Learning

CHALLENGE:

ARMv7-M and the following architectures (Cortex-M4, M7, M33, M55) contain SIMD extensions. At the beginning of the project, only one of the most popular NN inference operations – 'conv2d' – has already been adopted in TVM to using those extended instructions. In addition, there was a restriction set as input channel count must be a multiple of four, etc.

Thus, the power of ARMv7-M+ was not fully utilized, and the neural networks have not been executed as quickly as they could.

SOLUTION:

In the project's boundaries, we optimized more inference operations (conv1d, dense, max_pool1d, max_pool2d, avg_pool1d, avg_pool2d) and removed existing restrictions: added a 16-bit version of the convolution, expanded the solution for any number of input channels. We tested the solution on two neural networks adapted for processors with low SRAM capacity (up to 256K).

At the last step, we carried out a series of performance measurements.

RESULTS:

As a result, we improved TVM to optimize NN inference on ARMv7-M+ (Cortex-M4, M7, M33, M55) and got an increase in the speed of neural network execution of up to 300%. We added tests to be used when run on a specific ARM hardware emulator Corstone-300. We performed refactoring of an optimization support defining block for various architectures (IsaAnalyzer).

The solution is presented as 2 contributions in TVM repository: <u>PR 1 Update support for ARMv7m</u> intrinsic & PR 2 Cortex-M DSP.

TOOLS & TECHNOLOGIES

- TVM, microTVM
- ARMv7-M, ARMv8-M
- C/C++, Python, Assembler
- Zephyr
- OEMU

CONTACT US:



grovety.com



hi@grovety.com