Integrate TVM With MediaTek Neuropilot for Mobile Devices

Robert Lai\(^1\), Chun-Ping Chung\(^{(1,2,*)}\), Sheng-Yuan Cheng\(^2\), Jenq-Kuen Lee\(^2\)

1 – MediaTek Inc.
2 – Dept. CS, National Tsing Hua University, Taiwan
* - Speaker
Outline

• Previous presentation: Applying TVM Bring-Your-Own-Codegen to Android NNAPI
• Introducing MediaTek Neuron
• Utilizing MediaTek Neuron with TVM Bring-Your-Own-Codegen
• Experiment
• Summary
Previous Presentation: Applying TVM BYOC to Android NNAPI

TVM RFC - Relay to NNAPI
https://discuss.tvm.apache.org/t/rfc-byoc-android-nnapi-integration/9072
Introducing MediaTek Neuron

- MediaTek Neuron is an AI Solution for MediaTek platform
  - **Neuron Compiler**
    - Support multiple frontend
    - Provide high level graph IR
    - Common graph optimization
    - Partition graph to heterogeneous target like cpu, gpu, apu
  - **Neuron Runtime**
    - Provides APIs to load low-level binary generated by Neuron Compiler
    - Handle device switch, control flow, memory allocate, etc
Utilizing MediaTek Neuron with TVM BYOC

• BYOC allow developers to define their own partition rules and offload subgraph to the external compiler.

• Subgraphs left in the TVM compilation flow can still benefit from AutoTVM / AutoScheduler.
Utilizing MediaTek Neuron with TVM BYOC – Detail

**External Compiler**

- Partitioned Relay Function
- Handle different Type of Relay node
- Map Relay OP to Neuron OP
- Construct Neuron Graph
- Compile Neuron Graph

**ExprVisitor**

**low-level binary of APUs**

**External Runtime**

- TVM Neuron Module
- Construct Neuron Runtime
- Prepare Input/Output
- Inference

**low-level binary of APUs**
Experiment

Environment: MTK Dimensity 800 5G chip

• Up to 6X speedup comparing to native TVM and NNAPI Flow

![Speedup of BYOC on Neuron (Models are of FP32)](chart)

![TBS Pattern in MobileBert](chart)

**Neuron Advantage**
- can use APU to accelerate inference on android device

**TVM Advantage**
- faster support to new OP and model
- more frontend supporting
Summary

• We enabled TVM BYOC flow to MediaTek Neuron for mobile devices
  – With this flow, developer can use MediaTek APU to accelerate inference.
  – Experiment shows that we have 6X speedup compared to native TVM and NNAPI Flow.
  – How to effectively find the best partition method is a topic that’s worth to research in the future.
Thank you

Questions and Discussions