Getting started with Arm NN on Android, in just 5 minutes
Welcome!

Tweet us: @ArmSoftwareDev -> #AIVTT

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks
## Our upcoming Arm AI Tech Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Title</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>November 30th</td>
<td>Getting started with Arm NN on Android, in just 5 minutes</td>
<td>Arm</td>
</tr>
<tr>
<td>December 14th</td>
<td>Improve PyTorch App Performance with Android NNAPI Support</td>
<td>Arm</td>
</tr>
<tr>
<td>January 25, 2022</td>
<td>Peaks, Valleys and Thresholds: The art of segmenting real-time sensor data for tinyML classification, regression and anomaly detection</td>
<td>Reality AI</td>
</tr>
<tr>
<td>February 8th, 2022</td>
<td>Faster time-to-production for computer vision AI on Arm powered edge devices, but just how fast is fast?</td>
<td>Deeplite</td>
</tr>
<tr>
<td>February 22, 2022</td>
<td>* New Arm ML Quarterly Research Special * Federated Learning Based on Dynamic Regularization to Debias Model Updates</td>
<td>Arm ML Research</td>
</tr>
</tbody>
</table>

Visit: developer.arm.com/techtalks
Presenters

Ronan Naughton
Senior Product Manager – ML, Arm

Jan Eilers
Arm NN Software Engineer, Arm

Gian Marco Iodice
ACL Staff Software Engineer, Arm
Introduction - Arm NN and Arm Compute Library

Large Install Base:

- Powering ML in billions of devices:
  - Mobile
  - DTV
  - Automotive
  - Laptop

Superior Performance:

- Full Heterogenous execution (CPU, GPU) using a single interface
- Uses advanced network optimization techniques, workload tuning and GEMM heuristics

Arm Specific Optimizations:

- Outperforms generic ML libraries due to Arm specific optimization (e.g. dot product for Armv8.2A)
- Quick adoption of new Arm technologies e.g. Armv9-A features
Heterogeneous ML Software and Tools Platform

Applications

High Level Frameworks
(e.g. TensorFlow Lite, PyTorch, Android NNAPI)

Tools

Dynamic Workload Compilers & Drivers

CPU Cortex-A Neoverse

GPU Mali

NPU Ethos-N

© 2021 Arm
Arm NN on Android – using Android NNAPI

- Pre-installed by vendor on many Android devices
- Support for TensorFlow Lite and PyTorch networks
- Full CTS / VTS compliance
- 72 Operators accelerated with Arm NN backends
- NNAPI model caching supported
- Arm NN HAL driver:
  - android.hardware.neuralnetworks HAL interfaces (1.0 – 1.3)
Arm NN on Android - using TensorFlow Lite Delegate

- All TensorFlow Lite Networks can be supported:
  - Key ops (70) – accelerated
  - Unsupported ops – fall back to TensorFlow Lite implementation
- Quick setup time with Android Library (AAR)
- Configurable Optimizations
  - Backend selection
  - fast math
  - fp32->fp16
  - num-of-threads
  - memory-import
- Flexible build options (Cmake, Bazel)
Android Sample applications

https://www.tensorflow.org/lite/examples

Image Segmentation

Pose Estimation
Android Demo – Image Segmentation

Using TensorFlow Lite GPU Delegate
Android Demo – Image Segmentation
Using Arm NN and Arm Compute Library

Android Application

Arm NN TFLite Delegate

Arm NN

ACL

TFLite Interpreter

Arm AAR file

TFLite GPU Delegate

Mali GPU

TensorFlowLite

Input Image Size: 257 x 257
Arm NN enabled: true
Number of threads: 4
Pre-process execution time: 6 ms
Model execution time: ?? ms
Mask flatten time: 119 ms
Full execution time: 136 ms

Labels Found:
potted plant  background  person
chair
Android Demo – Pose Estimation
Using Arm NN and Arm Compute Library

Android Application

Arm NN TFLite Delegate

Arm NN

ACL

TFLite Interpreter

Arm AAR file

TFLite GPU Delegate

Mali GPU

TensorFlowLite

Fps: 22
Score: 0.44
Device: ArmNN-GpuAce
Model: PoseNet
Pose Classification
Performance Summary

**Image Segmentation**

- **Input Image Size:** 257 x 257
- **GPU enabled:** true
- **Number of threads:** 4
- **Pre-process execution time:** 9 ms
- **Model execution time:** 37 ms
- **Mask flatten time:** 137 ms
- **Full execution time:** 163 ms

**Labels Found:**
- potted plant
- background
- person
- chair

**Pose Estimation**

- **Input Image Size:** 257 x 257
- **GPU enabled:** true
- **Number of threads:** 4
- **Pre-process execution time:** 6 ms
- **Model execution time:** 11 ms
- **Mask flatten time:** 119 ms
- **Full execution time:** 136 ms

**Labels Found:**
- potted plant
- background
- person
- chair

---

<table>
<thead>
<tr>
<th>Accelerator</th>
<th>Approximate Execution time</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFLite GPU Delegate</td>
<td>37ms</td>
</tr>
<tr>
<td>Arm NN Delegate</td>
<td>11ms</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Accelerator</th>
<th>Average FPS achieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFLite GPU Delegate</td>
<td>17</td>
</tr>
<tr>
<td>Arm NN Delegate</td>
<td>24</td>
</tr>
</tbody>
</table>
What’s next??

• Continuous performance improvement on CPU & GPU
• Support for new Arm technologies
• Addition to Maven Central
• Sync Fence support
• NNAPI Support Library interface
• AHardwareBuf support
• NNAPI FL6 features
Thank You
Danke
Merci
谢谢
ありがとうございます
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
tודה
Thank you!

Tweet us: @ArmSoftwareDev -> #AIVTT

Check out our Arm Software Developers YouTube channel

Signup now for our next AI Virtual Tech Talk: developer.arm.com/techtalks