OPEN AI LAB

- Initiated by Arm, Allwinner, Horizon Robotics and Arm Accelerator on Dec 1\textsuperscript{st} 2016.
- Focus on open AI software at edge computing
- Based at Shanghai, Beijing and Shenzhen, China.
Focus on Edge Computing

Cloud Training

- ARM+GPU
- TPU2
- X86+Acc

Cloud Inference

- X86+FPGA
- TPU
- ARM+GPU+Acc

Edge Training

Long-term requirement?

Edge Inference

Our Focus

- NVDLA+GPU
- NVTensorRT
- ARM+GPU

Confidential
OAID (Open AID)

AI applications...

Open AID

Domain Library

Open-source Algorithms

Tengine - Open AI Engine

Inference

Enhanced features

Heterogeneous Computing Library (HCL)

Arm Compute Library

Caffe-HRT, MXNet-HRT, TensorFlow-HRT

PerfBLAS/OpenBLAS/ATLAS/...

Third-party Algorithms

Tools for Debugging and profiling

Compilers / Drivers (Android, Linux, RTOS)

Arm CPU, GPU

DLA/xPU

DSP
Optimized Caffe, TensorFlow, MXNet

- Caffe-HRT
- MXNet-HRT
- TensorFlow-HRT
- Highlights
  - Better performance than original Caffe or MXNet or TensorFlow
  - Dynamic scheduler embedded
  - Optimization on DL Operators
  - Supporting embedded debugging and profiling tools
Optimized Caffe performance on A53

- Optimized Caffe (Caffe-HRT) is up to 9.5x the performance of Caffe.

Running Time (ms) on A53@1.6Ghz

- **High Performance** and Low mem footprint
- Easy deployment for AI applications
- Flexible to Heterogeneous Computing Platform

**Tengine**
- Graph Inference C API
- Graph Representation
- Graph Executor
- Scheduler
- Device Executor
- HAL

**Operators**
- Plugin API

**Device Driver**
- CPU
- GPU
- xPU/DLA
- DSP

**AI Applications**
- Caffe/TensorFlow/MXNet/… API
- Python/JS/JNI API

**Caffe/MXNet/TensorFlow/ONNX/Tengine Models**
Tengine performance

- Tengine is up to 2.46x the performance of Optimized Caffe (Caffe-HRT).

<table>
<thead>
<tr>
<th>Model</th>
<th>Origin Caffe (fp32)</th>
<th>Caffe-HRT with HCL (fp32)</th>
<th>Caffe-HRT with HCL (int8)</th>
<th>Tengine (fp32)</th>
<th>Tengine (int8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SqueezeNet</td>
<td>147</td>
<td>104</td>
<td>91</td>
<td>83</td>
<td>66</td>
</tr>
<tr>
<td>MobileNet</td>
<td>306</td>
<td>251</td>
<td>234</td>
<td>116</td>
<td>95</td>
</tr>
</tbody>
</table>
Application/Algorithm

- Domain Library includes the below algorithms.
  - Open-source algorithms.
  - 3rd-party commercial Algorithms.
- The below data is based on a single A72@1.8Ghz core.

<table>
<thead>
<tr>
<th>Application</th>
<th>Algorithm</th>
<th>Platform</th>
<th>Test Configuration</th>
<th>Performance (ms/frame)</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face Detection</td>
<td>MTCNN</td>
<td>A72, Optimized Caffe</td>
<td>480P frame, 64*64 minimum face size, single face</td>
<td>43</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>A company proprietary</td>
<td>A72</td>
<td>480P frame, 64*64 minimum face size, single face</td>
<td>32</td>
<td>31</td>
</tr>
<tr>
<td>Face Recognition (including Detection)</td>
<td>MTCNN+Lightened CNN</td>
<td>A72, Optimized Caffe</td>
<td>480P frame, 64*64 minimum face size, single face</td>
<td>225</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>B company proprietary</td>
<td>A72</td>
<td>480P frame, 64*64 minimum face size, single face</td>
<td>308</td>
<td>3</td>
</tr>
<tr>
<td>Object Tracking</td>
<td>TLD-DSST</td>
<td>A72</td>
<td>480P frame</td>
<td>71</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>GOTURN+SqueezeNet</td>
<td>A72</td>
<td>480P frame, 2FC, 1x1 SqueezeNet</td>
<td>247</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>C company proprietary</td>
<td>A72</td>
<td>480P frame</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>Gesture Recognition</td>
<td>CVGesture</td>
<td>A72</td>
<td>2 static gestures</td>
<td>69</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>D company proprietary</td>
<td>A72</td>
<td>1 static gesture and 1 dynamic gesture</td>
<td>24</td>
<td>42</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>GMM-HMM (Kaldi-based)</td>
<td>A72</td>
<td>100 voice commands recognition</td>
<td>RTF(Real Time Rate)=1</td>
<td>/</td>
</tr>
</tbody>
</table>
Open AI Platform

- 96Boards + Open AID = An open AI platform
  - So far Open AID has run on HiKey 960, Dragonboard 820c and Rock960.
Thanks!

Contact us: support@openailab.com