Huawei HiAI Unlock the Future

Eric Zhou
Kirin Chipset Solution Planning Director
zhouchen@huawei.com
AI is coming round

- Get AI for your apps
- 25 Apps, 9 with AI

TIME The 25 Best Apps of the Year So Far (2017.6)
Why On-device AI?

Real-time

Privacy
AI Knowledge in an instant burst
AI Knowledge processing in real time is key point
A real APP needs More Resource

- **CPU**
  - Task scheduling
  - Load balance
  - Memory allocation

- **GPU**
  - UI rendering
  - Graphics processing

- **NPU**
  - AI Computing

- **ISP/DSP**
  - Camera 3A
  - Image processing
Work load for a AI APP

- CPU little core
- CPU big core
- GPU
- NPU
Two Key Questions for AI APPs

● What Performance we need?
● Who will make the APPs?
Real-time Processing is the Requirement

5fps  
15fps  
>25fps
To Match Real-time AI Processing Requirement

- The NPU supports a dedicated set of AI instructions for neural network model operations that allow more efficient parallel execution of more neural network operators within minimal clock cycles.

- **CPU = Scalar**
  - General-purpose computing
  - Logic control

- **GPU = Vector**
  - Image processing
  - Large-scale parallel computing

- **NPU = Tensor**
  - Knowledge model processing
  - Dedicated AI Instruction-Set
  - Large-scale parallel computing
High Performance Density

5.5Billion Transistors

<table>
<thead>
<tr>
<th></th>
<th>Die Size</th>
<th>NN Perf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPU</td>
<td>3%</td>
<td>100%</td>
</tr>
<tr>
<td>CPU</td>
<td>6%</td>
<td>5%</td>
</tr>
<tr>
<td>GPU</td>
<td>18%</td>
<td>60%</td>
</tr>
</tbody>
</table>
What developers care about?

- Development Cost
  - The porting cost: operators, tools, documents and FAE

- Benefits
  - Users, downloads, marketing

- Intellectual property protection
  - Model (algorithm) encryption
HiAI Mobile Computing Platform

- Open AI ecosystem
- 3rd parties APPS empowered by HiAI mobile computing platform

<table>
<thead>
<tr>
<th>APP</th>
<th>APP</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>TensorFlow/TensorFlow Lite</td>
<td>Caffe/Caffe2</td>
<td></td>
</tr>
</tbody>
</table>

HiAI APIs

HiAI Model and Resouse Management/Runtime

<table>
<thead>
<tr>
<th>CPU</th>
<th>GPU</th>
<th>DSP</th>
<th>NPU</th>
</tr>
</thead>
</table>
What HiAI provides?

- Provides frequently used artificial intelligence function APIs which can efficiently run on mobile devices.
- Supports popular frameworks.
- Provides processor-independent acceleration APIs to allow the Kirin hardware to accelerate model and operator computing.
- Schedules heterogeneous resources flexibly to enable developers to accelerate neural network model and operator computing.

Tool chain
Abundant APIs
Complete documents
Easy-to-use source code
<table>
<thead>
<tr>
<th>HiAI SDK V100</th>
<th>HiAI SDK V150</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Framework</strong></td>
<td>Caffe, TensorFlow</td>
</tr>
<tr>
<td><strong>Operators</strong></td>
<td>42</td>
</tr>
<tr>
<td><strong>Tools</strong></td>
<td>Command line</td>
</tr>
<tr>
<td><strong>Documents and Support</strong></td>
<td>User manual Source code FAE</td>
</tr>
<tr>
<td>What kinds of APP that HiAI can unlock</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>Short videos and live streaming</strong></td>
<td>Facial recognition, gesture recognition, portrait segmentation, posture recognition, and video stylization</td>
</tr>
<tr>
<td><strong>Social platform</strong></td>
<td>Photo classification, image recognition, and super-resolution imaging</td>
</tr>
<tr>
<td><strong>AR</strong></td>
<td>Depth estimation, light estimation, environmental understanding, and simultaneous localization and mapping (SLAM)</td>
</tr>
<tr>
<td><strong>Photo taking and retouching</strong></td>
<td>Facial beautification and image enhancement</td>
</tr>
<tr>
<td><strong>Shopping</strong></td>
<td>Shopping through image recognition</td>
</tr>
<tr>
<td><strong>Translation and words processing</strong></td>
<td>Instant camera translation, optical character recognition (OCR), word splitting, named entity recognition (NER), textual emotion recognition, and intelligent textual response</td>
</tr>
</tbody>
</table>
Portrait Segmentation
Shopping through image recognition on device
The HiAI mobile computing platform supports sparse model acceleration. The NPU can skip the multiply-add algorithms with a coefficient of zero, which can greatly improve the computing efficiency and reduce the bandwidth while maintaining computing precision.

The HiAI mobile computing platform supports 8-bit and 1-bit quantization, effectively reducing the computing bandwidth and storage consumption and improving energy efficiency.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Energy [pJ]</th>
<th>Relative Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 bit int ADD</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>32 bit float ADD</td>
<td>0.9</td>
<td>9</td>
</tr>
<tr>
<td>32 bit Register File</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>32 bit int MULT</td>
<td>3.1</td>
<td>31</td>
</tr>
<tr>
<td>32 bit float MULT</td>
<td>3.7</td>
<td>37</td>
</tr>
<tr>
<td>32 bit SRAM Cache</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>32 bit DRAM Memory</td>
<td>640</td>
<td>6400</td>
</tr>
</tbody>
</table>
HiAI Mobile Computing Platform

Execution

As shown in the figure below, by using conversion tools, a trained neural network model is converted into an offline model that can be efficiently executed on the HiAI mobile computing platform, and output as the binary file, the offline model. The main purpose of converting the standard neural network model (such as Caffe) into an offline model is to optimize the network configuration. After conversion, an optimized target file, namely the offline model, is generated. The offline model is serialized and stored on the disk. As a result, when the inference is performed, the optimized target file is used, which is

Offline Mode Generation
As shown in the figure below, during the offline model computing, the offline model is loaded from the file, and data such as images input by users is copied to HiAI's NPU for computing. User data needs to be transferred from the DDR SDRAM to the NPU and then returned to the DDR SDRAM for each inference during computing.
How to integrate your knowledge model

1. Operator compatibility assessment
2. Model conversion
3. Model loading
4. Model execution
5. Model unloading
HiAI Mobile Computing Platform

Advantages

- HiAI: enables offline models to be compiled without occupying the runtime, which improves runtime efficiency.
- HiAI: supports multiple frameworks such as Caffe and TensorFlow/Lite, and will support Caffe2, ONNX in the future.
- HiAI: supports multiple operators and rapid iteration, and supports operator customization in later versions.
- HiAI: supports Kirin 970 and later chip platforms and rapidly increasing devices, and is widely used.
- HiAI is also integrated into HiKey970 which is the third generation of Hikey series in 96boards.
More and More
New HiKey970 AI Development Board is coming

- Huawei HiAI SDK
- Up to 25X Performance 50X Power Efficiency
- Dedicated Neural-network Processing Unit (NPU)
- Heterogeneous Resource Management

Popular AI stacks

Mainly OS

More Hardware interfaces

More stronger

Compute abilities
HiKey970 Spec

Software

- UEFI + ARM Trusted Firmware
- Kernel 4.9
- CAN driver
- CSI driver
- WiFi enable
- Video Codec enable

Hardware

- CPU: 4 x A73 2.36GHz + 4 x A53 1.8GHz
- GPU: Mali G72 MP12
- NPU: 256MAC/cycle
- 6GB 1866MHz, 4 Channel LPDDR4x
- 4 lanes CSI + 2 lanes CSI
- CAN V2.0B up to 1Mbps
- Video Dec up to H.265 3840x2160@60fps

OS Choices

- Ubuntu
- Debian
- Android Master

Stacks

- Huawei HiAI
- Android NN
- OpenGL ES
- OpenCL
Shining Star Plan

**Talent cultivation**
- Campus project
- Training
- Certification

**Development support**
- Innovation lab
- Development environment
- Testing tools

**Innovation support**
- Funded
- Cloud-Client resource
- Contest awards

**Marketing**
- User growth
- Co-Marketing
- Promotion
How to Get HiAI SDK and HiKey970?

- [developer.huawei.com](http://developer.huawei.com/consumer/en/home)
- [www.96boards.org](https://www.96boards.org)
- [www.hihope.org](http://www.hihope.org)
- Amazon channel for worldwide market, two stores in Amazon: “Hoperun” and “HiHope”, Secure96 is on sell now
- Taobao channel for China market: [shop596522926.taobao.com/index.htm?spm=2013.1.w5002-17633002258.2.32b43ca7TxdF52](https://shop596522926.taobao.com/index.htm?spm=2013.1.w5002-17633002258.2.32b43ca7TxdF52)
- [www.lenovator.com](http://www.lenovator.com)

$299
Available in APR. 2018
Thank You

#HKG18
HKG18 keynotes and videos on: connect.linaro.org
For further information: www.linaro.org