Open AID

By **OPEN** AI LAB

Feb 10 2018

Abstract

Open AID simplifies the application development of Vision and Speech on the edge. It brings Domain Libraries for Vision and Speech inferences along with a unified API for developers, Tengine with improved DL frameworks on Caffe, MXNet and TensorFlow for inference, and Heterogeneous Computing Library (**HCL**) for optimized Arm CPU and Mali GPU utilization. Developers can use it for fast implementation of the Vision and/or Speech applications with AI capabilities.

Applications provided within Open AID include face detection/recognition, gesture recognition, tracking and speech recognition. More applications will be supported in the near future. Applications developed on Open AID can be deployed to Arm-based edge heterogeneous computing platform.

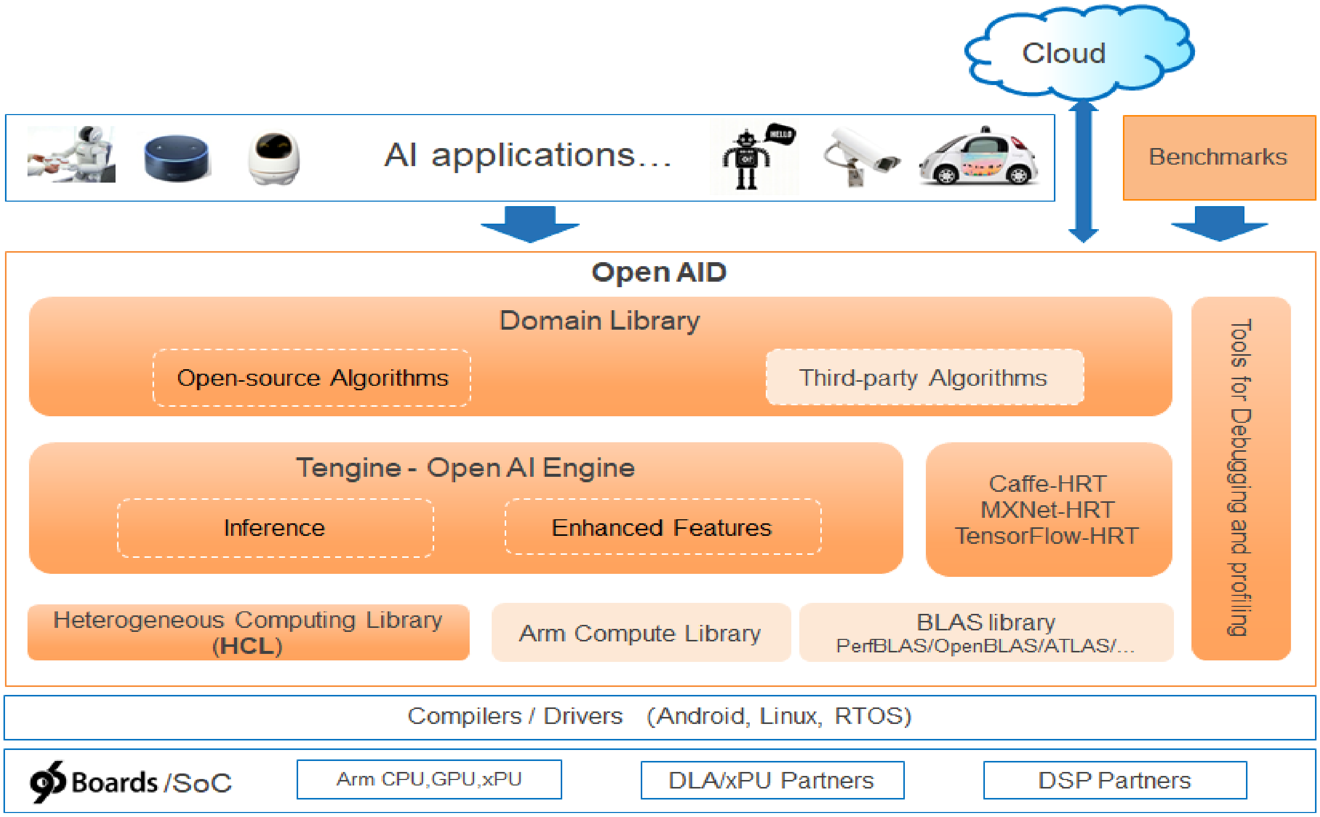
OPEN AI LAB

Contents

1 Introduction

AID (AI Distro) is an AI core software development kit developed by **OPEN** AI LAB. Combined with optimized or hardware accelerated AI algorithms, it provides a unified, open and friendly interface to utilize the best available resources on the platform for AI capable application developers.

Open AID is an open source project hosted by https://github.com/OAID



2 Deep Learning Framework support

Open AID includes the high-performance *Heterogeneous Run Time*(**HRT**) version of the popular Deep Learning Frameworks, such as Caffe, MXNet, in which inference operators are accelerated by Heterogeneous Computing Library (**HCL**). These Deep Learning Frameworks also support acceleration via BLAS library and Arm Compute Library[*REF*].

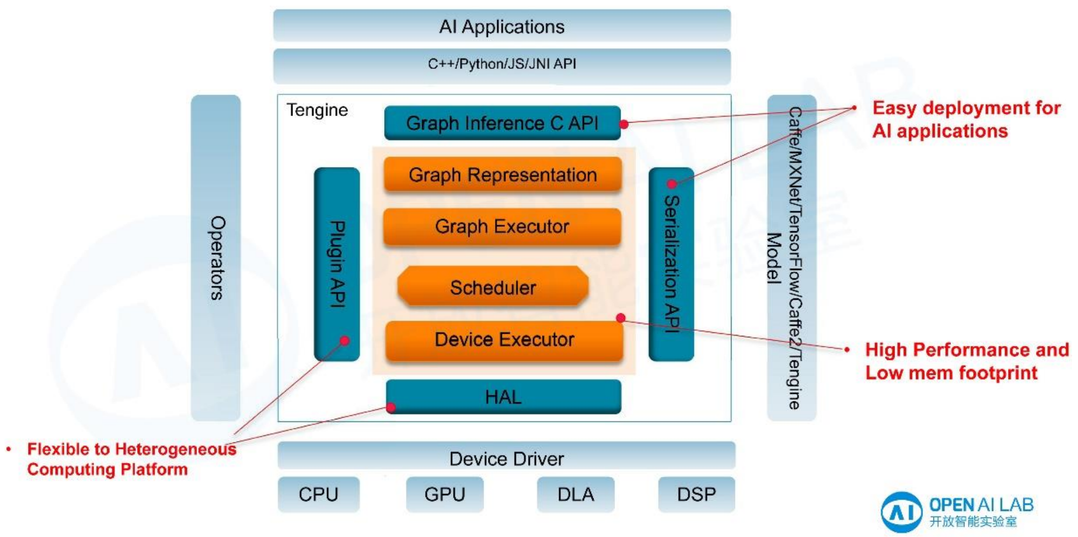
Open AID also integrates Tengine designed by **OPEN** AI LAB to accelerate diverse inferences for higher performance.

HCL(*Heterogeneous Computing Library*) provides the unified interface to use heterogeneous hardware computing resources, such as CPU, GPU, DSP, DLA, etc. The library provides high performance implementation of neural network operators, as well as implementation of frequently used data pre-process/post-process operators. Each supported H/W has its specific optimization in library to achieve maximum performance of the available computing resource.

2.1 Tengine

It provides best-in-class optimized Arm-based heterogeneous platform implementation for neural computing operators, such as convolution, pooling, etc. In the meantime Tengine's framework is designed to support various Deep Learning Accelerators (DLAs) as well.

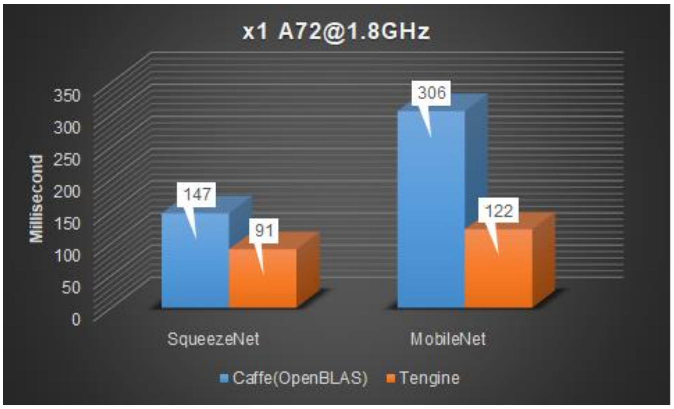
Block Diagram

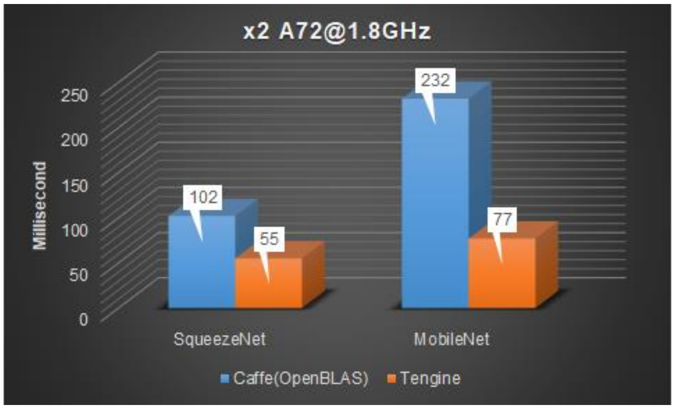


HIGHLIGHTS:

High-Performance	Manually optimized neon code for NN operators
Heterogeneous-Computing	Support different silicon fabrics, such as CPU, GPU, DLA
Lite-Weight	No dependency on third-part S/W packages except for C/C++ run-time library
Caffe-Wrapper	Application based on Caffe just needs to recompile the source to use Tengine
Direct-Load	Caffe model and MXNet model can be loaded directly by Tengine
Extendability	The model loading, operator definition and execution, and device driver are all built as plug-ins. It is easy for user to develop his own operator or to add driver to support his own DLA

For example, Tengine includes an optimized A72 CPU device implementation. The performance data as follows (on Rockchip rk3399 SoC based platform):



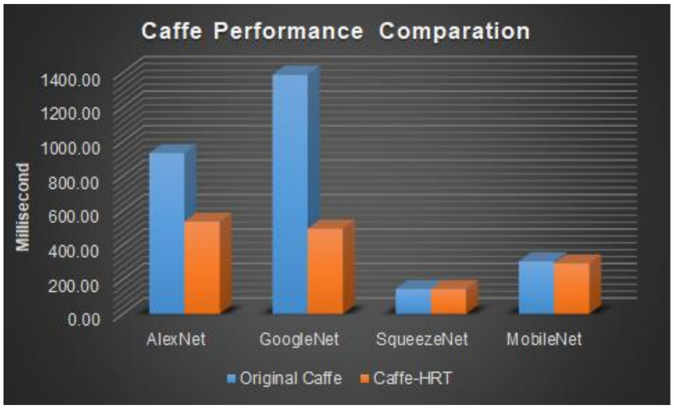


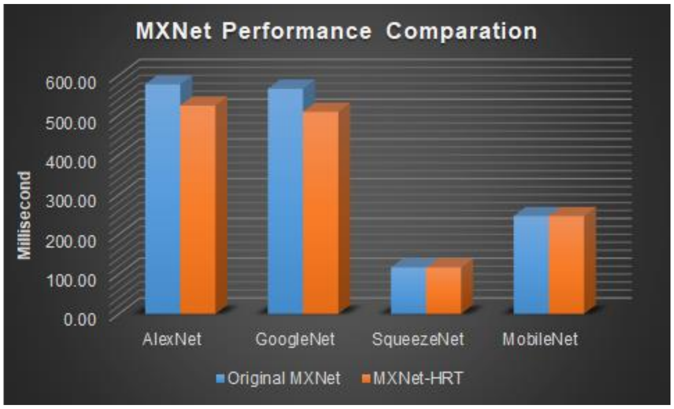
2.2 Caffe-HRT/MXNet-HRT

OPEN AI LAB further developed the Caffe/MXNet and added heterogeneous capabilities to the frameworks and created Caffe-HRT[*REF*]/MXNet-HRT[*REF*], heterogeneous computing infrastructure framework to speed up Deep Learning on heterogeneous embedded platform.

HRT frameworks take full advantage of the Arm hardware calculation capabilities in two levels. It supports heterogeneous computing with GPU and CPU, and hybrid computing for OpenBLAS and Arm Compute Library. It retains all the features of the original Caffe/MXNet architecture which users can use to deploy their applications seamlessly.

HRT frameworks are user-friendly, fast, modularized and open. As well as the performance improvement, it also offers users the slick application development experience on Arm-based SoC embedded system.





The performance speedup ratio :

Caffe-HRT	: 1.0~2.8
MXNet-HRT	: 1.0~1.1

3 Domain Library

Domain Library is a key component of Open AID. As an AI algorithm library, Domain Library includes advanced support for specific domain applications, such as vision and speech. It runs on popular Deep Learning Framework, like Caffe, MXNet, and Tengine, to provide on-device deep learning algorithm services to applications. With Tengine together, it is capable to support and services to other traditional machine learning algorithm services as well.

3.1 FaceRecognition

FaceRecognition [*REF*] implements face detection and recognition, using MTCNN to detect, and Lightened CNN to recognize.

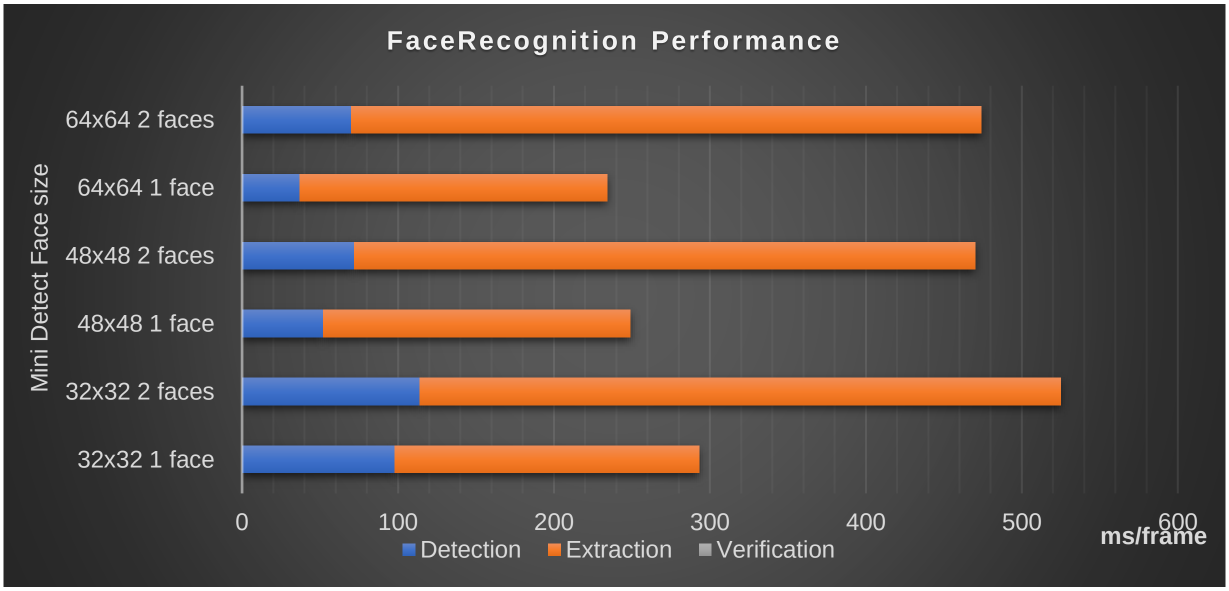
Features

Face Detection	Predicting face and landmark location.
Features Extraction	Extracting 256-dimension face features.
Features Matching	Computing the similarity distance for a pair of 256-dimension features.

Limitations

- Distance less than 6m for detection;
- Angle less than 20 degrees for recognition;

FaceRecognition performance (Input image Size 640x480)



3.2 CVGesture

CVGesture [*REF*] implements detection and recognition to different hand gestures, based on OpenCV 3.3.0 (Open Source Computer Vision Library) .

Features

Palm Recognition Recognizing hand facing camera with five fingers open

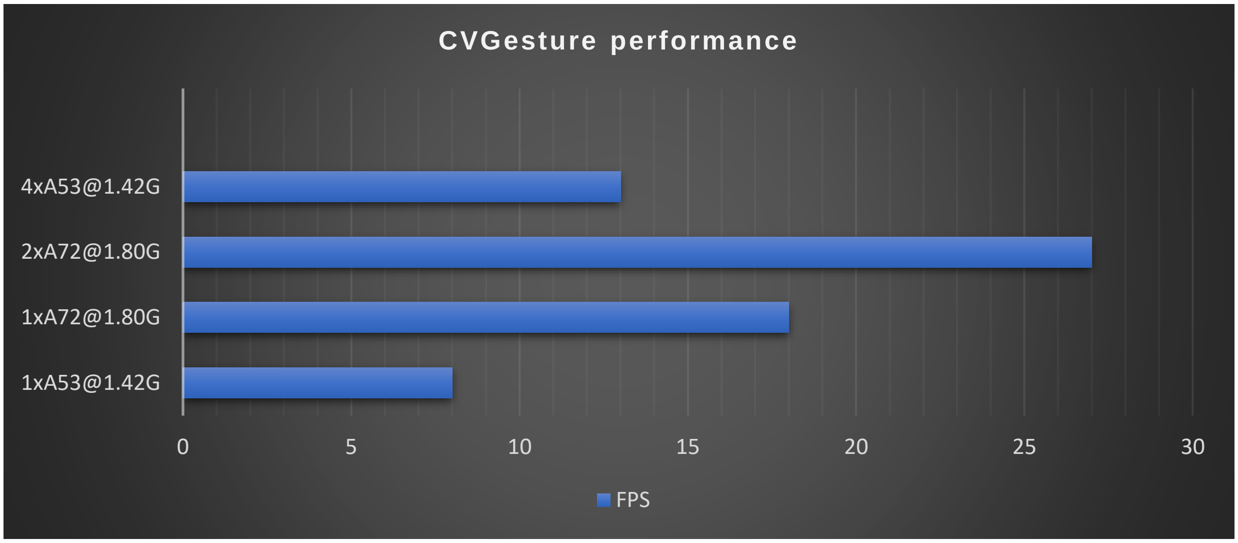
Fist Recognition Recognizing hand facing camera with five fingers clenched

Multiple Gestures Recognition Recognizing several hands appearing in the camera at the same time

Limitations

- Angle less than 30 degrees for the direction of front and back;
- Angle less than 45 degrees for the direction of left and right;

CVGesture performance



4 Conclusion

Open AID maximizes the computation capability from existing and upcoming SoCs to the full extent. It provides an optimized unified APIs for the application, and integrates continuously improved application algorithm library to achieve high performance application support.

Open AID focuses on the application of algorithm on SoC (algorithm validation and Optimization). It provides the professional technical reference and guidance which can help users launch their AI products easily.

Open AID quickly builds platform for users. The unity and integrity of Open AID also support the rapid application for developers. Integration our continuous optimization work to meet users' all kinds of requirement e.g. completion of chip, platform and system verification, performance testing and evaluation.

References

Arm Computer Library	https://github.com/ARM-software/ComputeLibrary
Tengine	https://github.com/OAID/Tengine
Caffe-HRT	https://github.com/OAID/Caffe-HRT
MXNet-HRT	https://github.com/OAID/MXNet-HRT
FaceRecognition	https://github.com/OAID/FaceRecognition
CVGesture	https://github.com/OAID/CVGesture