CaffeOnACL

Performance Report

2017-9-22

OPEN AI LAB

Revision Record

Date	Rev	Change Description	Author
2017-9-22	0.30		

catalog

1 Purpose

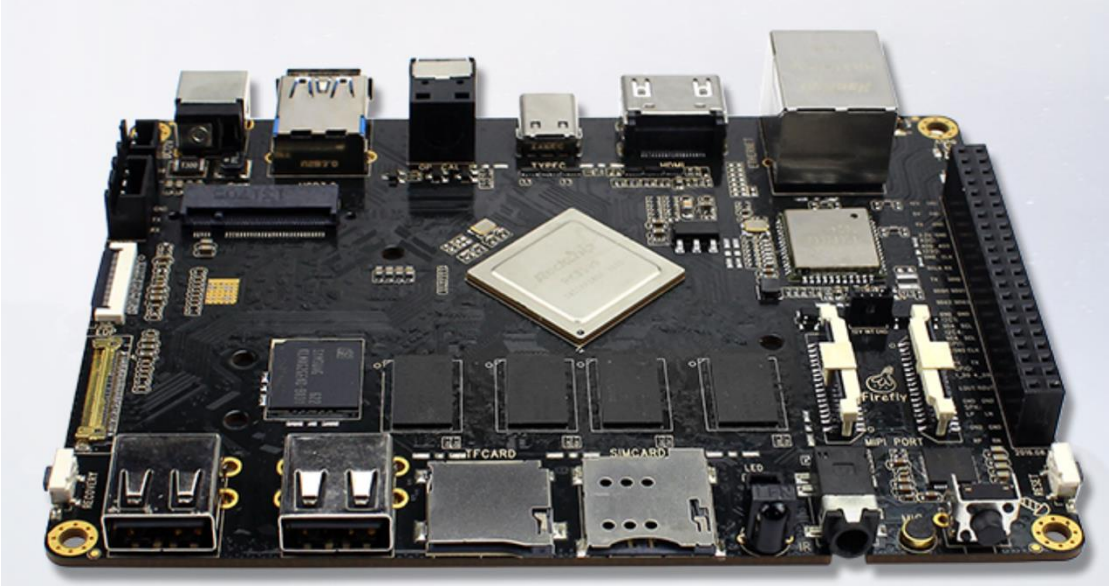
This Report is tested on RK3399 platform, including both CPU data and GPU data. We collected the data on AlexNet, GoogLeNet, SqueezeNet and MobileNet. Note that the CPU data is on a single A72 core. And we found the mixed mode can improve performance 2.78X for the best case.

2 Test Environment

Hardware SoC : Rockchip RK3399

- GPU: Mali T864 (800MHz)
- CPU: Dual-core Cortex-A72 up to 2.0GHz (real frequency is 1.8GHz); Quad-core Cortex-A53 up to 1.5GHz (real frequency is 1.4GHz)

Operating System : Ubuntu 16.04



3 Performance Improvement Achievement

The ACL_NEON's LRN and POOLING are better , and ACL_CL(GPU) has the better performances on large FC while OpenBLAS has better on CONV. It's possible to gain better performance on mixing the calculation on different comment, for example, using OpenBLAS layers (Softmax, RELU, FC, CONV) and ACL_NEON layers (LRN, Pooling) in neural network.

After we mixed the layers calculation on OpenBLAS and ACL, it's very easy to mix the layers calculation by exporting environment variable BYPASSACL, details in User Guide 5.2. We have achieved about 2.78X performance in best case.

	Original Caffe(ms)	Mixed Mode(ms)	Performance Gain
AlexNet	990	538	1.84X
GoogleNet	1388	498	2.78X
SquezzeNet	146	131	1.11X
MobileNet	231	281	0.82X

4 Performance

For GPU, the OpenCL driver need compile CL kernel for the first time running, but after 2nd time, the CL kernel may not be compiled. This will impact performance. Here we list the 1st data separately. We tested total 10 times from 2nd to 11th and calculated the average time. The data in the below tables are in the unit of second.

The items(TPI, Allocate, Run, Config, Copy, FC, CONV, LRN, Pooling, RELU, SOFTMAX) in the below tables:

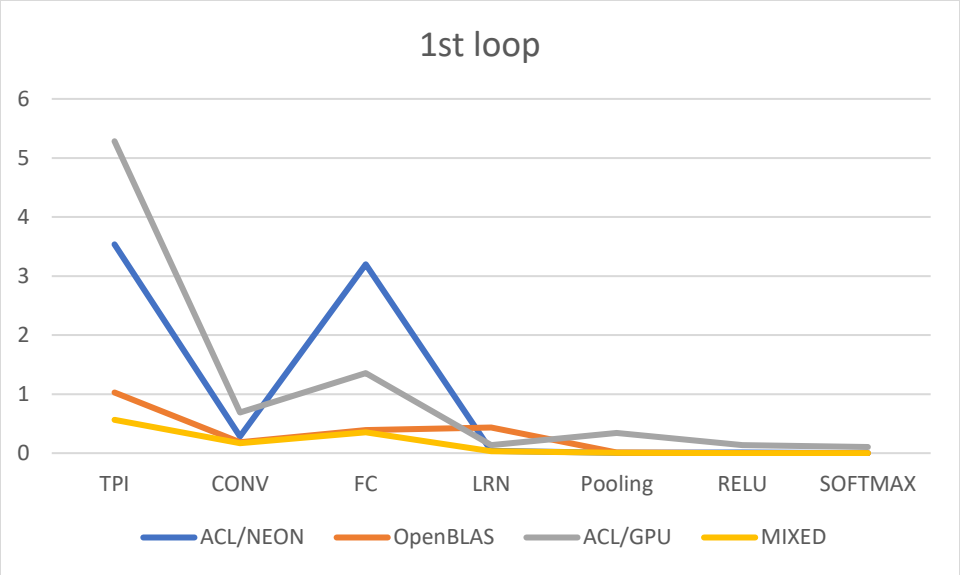
TPI : The total time for per inference
Avg. Time : tested total 10 times from 2nd to 11th and calculated the average time. The unit of all the data columns in tests below is second.

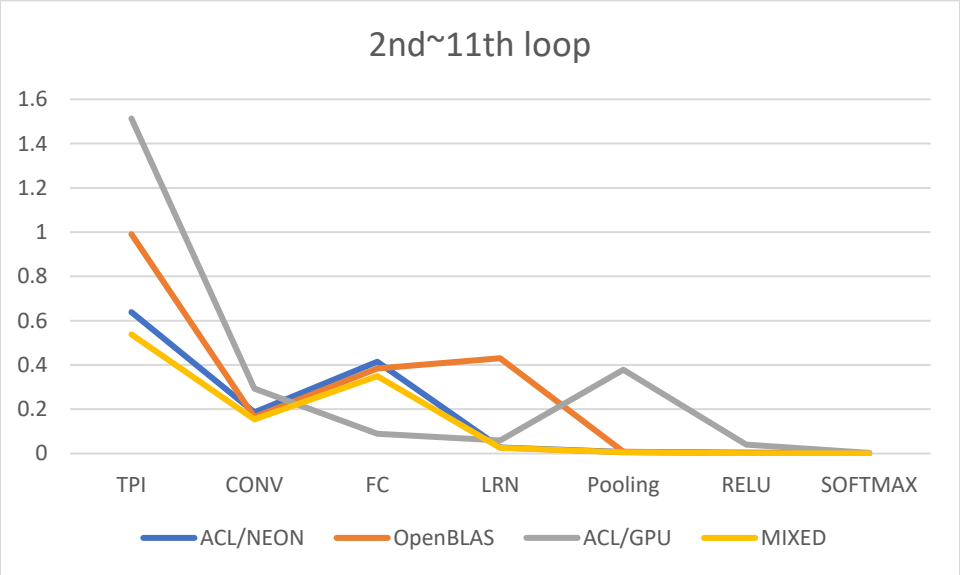
The details see user manual section “Use Cases”.

4.1 AlexNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	3.5360	0.1931	2.8732	0.2061	0.1383
OpenBLAS	1.0270	0	0	0	0
ACL/GPU	5.2829	0.1672	0.6906	1.3657	0.2982
MIXED	0.5640	0.0030	0.0251	0.0007	0.0060
Avg. Time					
ACL/NEON	0.6386	0	0.5236	0	0.0085
OpenBLAS	0.9907	0	0	0	0
ACL/GPU	1.5132	0	0.4754	0	0.1722
MIXED	0.5381	0	0.0249	0	0.0046

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	3.5360	0.2846	3.1980	0.0365	0.0069	0.0086	0.0004
OpenBLAS	1.0270	0.1856	0.3922	0.4349	0.0101	0.0029	0.0002
ACL/GPU	5.2829	0.6887	1.3543	0.1377	0.3414	0.1355	0.1037
MIXED	0.5640	0.1707	0.3516	0.0321	0.0067	0.0016	0.0002
Avg. Time							
ACL/NEON	0.6386	0.1862	0.4147	0.0269	0.0063	0.0040	0.0001
OpenBLAS	0.9907	0.1644	0.3840	0.4311	0.0089	0.0018	0.0001
ACL/GPU	1.5132	0.2936	0.0898	0.0595	0.3788	0.0408	0.0030
MIXED	0.5381	0.1544	0.3496	0.02616	0.0062	0.0016	0.0001

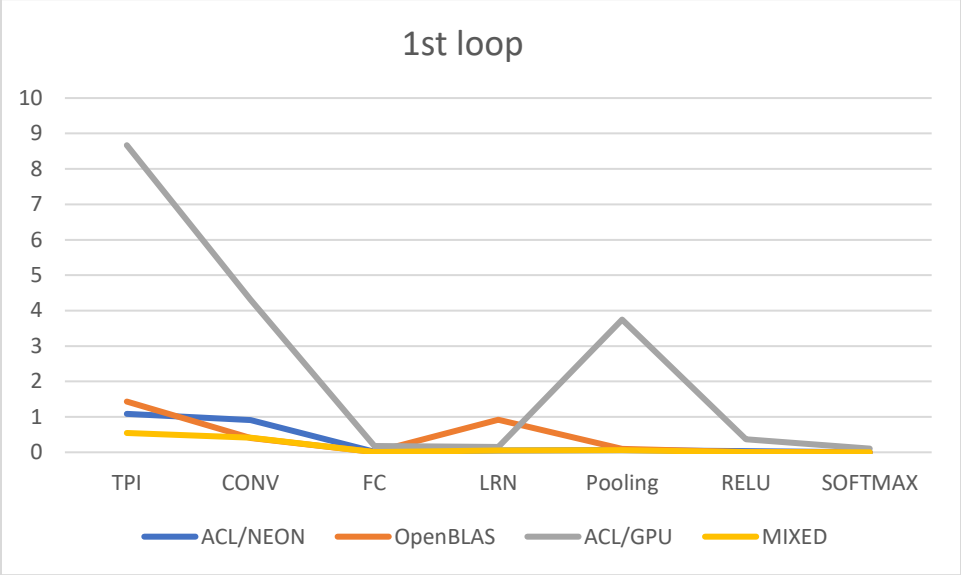


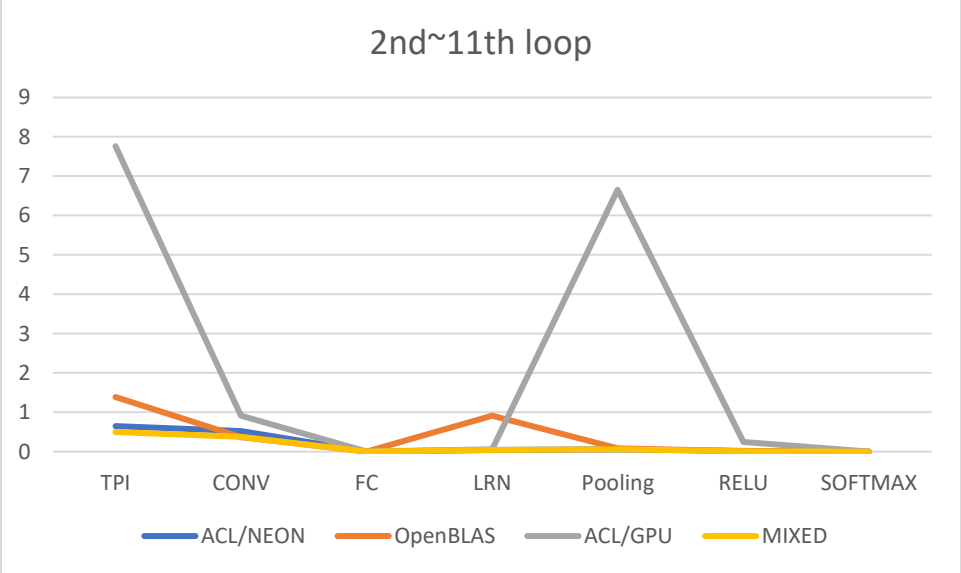


4.2 GoogleNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	1.0850	0.0662	0.6817	0.1173	0.1718
OpenBLAS	1.4321	0	0	0	0
ACL/GPU	8.6689	0.0956	4.1021	2.7105	1.6307
MIXED	0.5500	0.0045	0.0609	0.0025	0.0231
Avg. Time					
ACL/NEON	0.6506	0	0.5713	0	0.0475
OpenBLAS	1.3888	0	0	0	0
ACL/GPU	7.7620	0	4.9121	0	2.6528
MIXED	0.4985	0	0.0602	0	0.0185

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	1.085	0.9078	0.0172	0.0545	0.0613	0.0366	0.0002
OpenBLAS	1.4321	0.4023	0.0045	0.923	0.0943	0.0077	0.0002
ACL/GPU	8.6689	4.3098	0.1743	0.1481	3.7466	0.3631	0.1043
MIXED	0.545	0.4104	0.004131	0.0553	0.0642	0.0078	0.0002
Avg. Time							
ACL/NEON	0.6506	0.525	0.0043	0.0455	0.0564	0.0165	0.0001
OpenBLAS	1.3888	0.3728	0.0045	0.917	0.0866	0.0077	0.0001
ACL/GPU	7.762	0.9142	0.0032	0.0601	6.6507	0.2501	0.0025
MIXED	0.4985	0.3783	0.0041	0.0462	0.0587	0.0078	0.0001

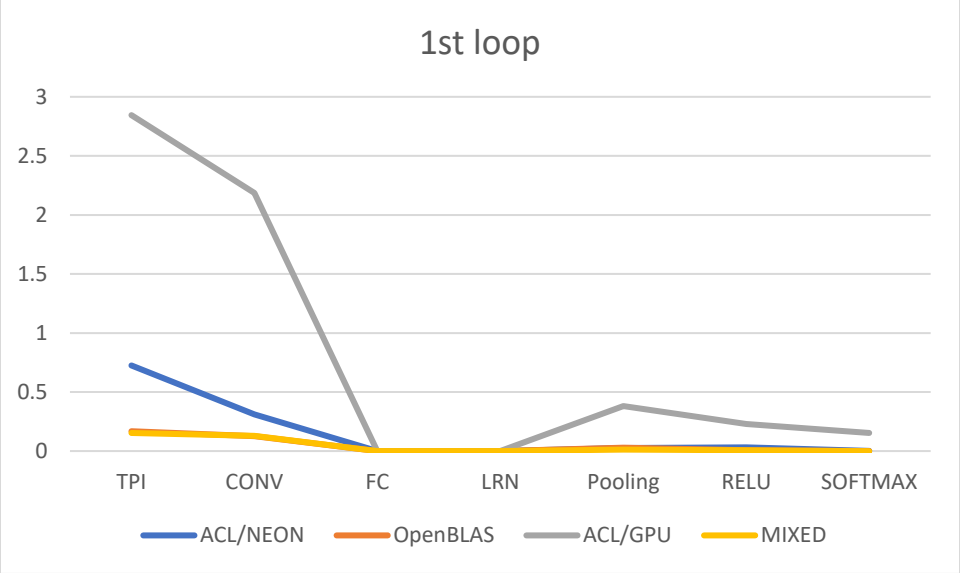


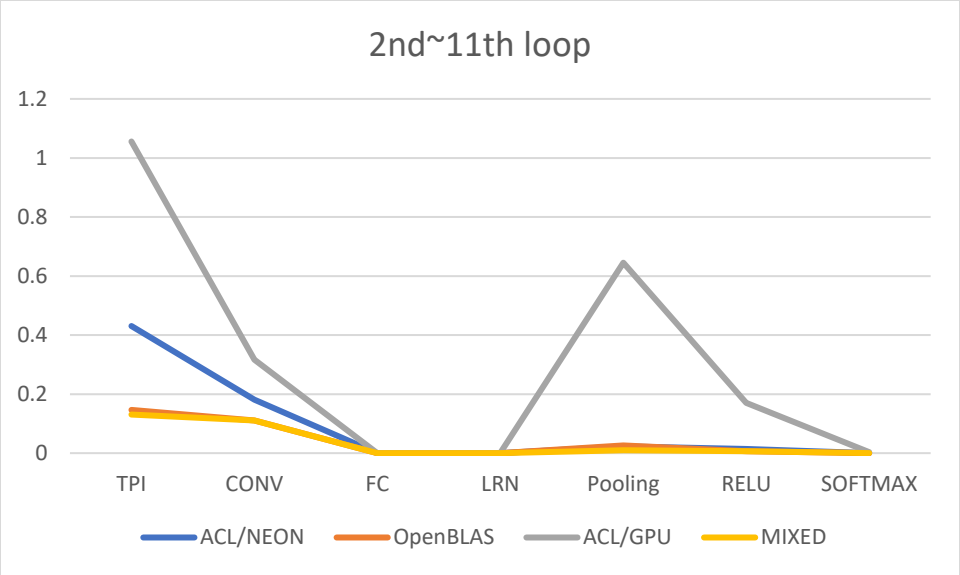


4.3 SqueezeNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.7249	0.0329	0.2233	0.037	0.0643
OpenBLAS	0.168	0	0	0	0
ACL/GPU	2.63	0.1766	0.7343	1.2149	0.3204
MIXED	0.153	0.0001	0.0034	0.0001	0.0041
Avg. Time					
ACL/NEON	0.4306	0	0.1899	0	0.0229
OpenBLAS	0.1464	0	0	0	0
ACL/GPU	1.0557	0	0.6845	0	0.3405
MIXED	0.1307	0	0.0033	0	0.0033

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	0.7249	0.311	0	0	0.0249	0.0313	0.0003
OpenBLAS	0.168	0.1273	0	0	0.028	0.0067	0
ACL/GPU	2.845	2.1851	0	0	0.3798	0.2297	0.1543
MIXED	0.153	0.1282	0	0	0.0111	0.0067	0.0002
Avg. Time							
ACL/NEON	0.4306	0.1813	0	0	0.0225	0.014	0.0001
OpenBLAS	0.1464	0.1103	0	0	0.0259	0.0066	0
ACL/GPU	1.0557	0.3169	0	0	0.645	0.1699	0.0037
MIXED	0.1307	0.1107	0	0	0.0097	0.0067	0.0001

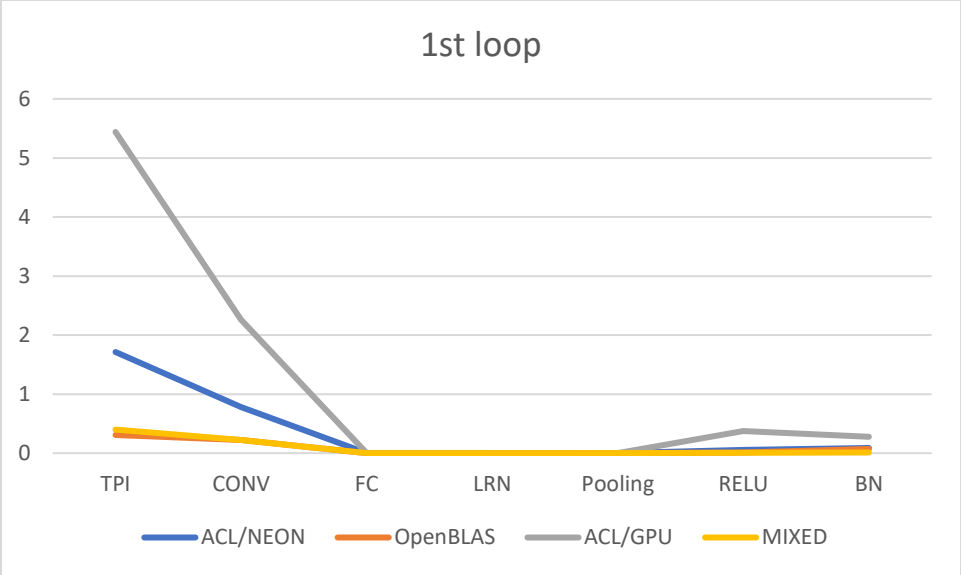


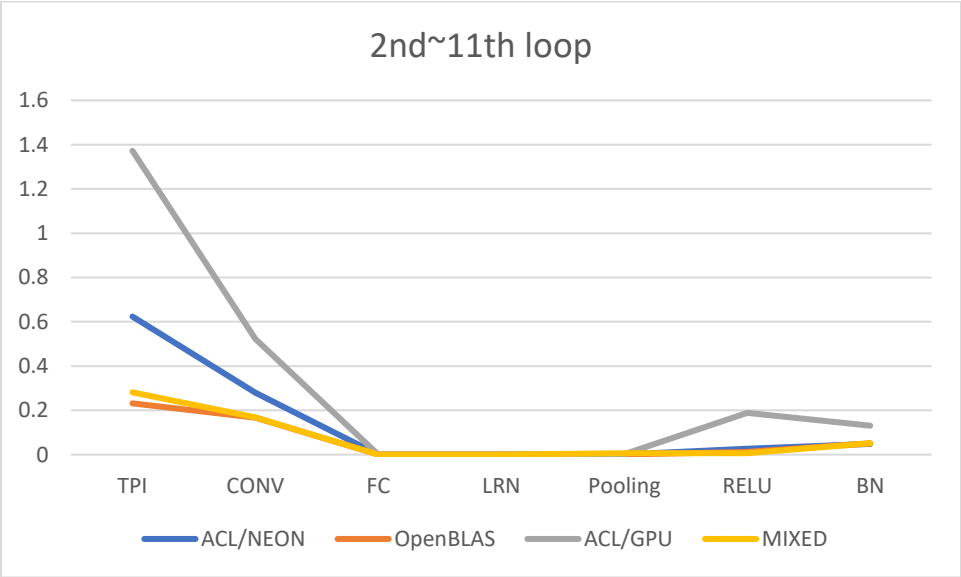


4.4 MobileNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	1.7126	0.0834	0.3556	0.0458	0.2993
OpenBLAS	0.3074	0	0	0	0
ACL/GPU	5.44	0.125	0.456	1.5329	0.4132
MIXED	0.399	0.0001	0.0034	0.0001	0.0041
Avg. Time					
ACL/NEON	0.6236	0	0.2136	0	0.0553
OpenBLAS	0.2314	0	0	0	0
ACL/GPU	1.3722	0	0.3639	0	0.1656
MIXED	0.2812	0	0.0232	0	0.0265

	TPI	CONV	FC	LRN	Pooling	RELU	BN
1st							
ACL/NEON	1.7126	0.7834	0	0	0.0009	0.0568	0.0875
OpenBLAS	0.3074	0.2239	0	0	0.0007	0.0125	0.0702
ACL/GPU	5.44	2.2583	0	0	0.0002	0.3752	0.2769
MIXED	0.399	0.224	0	0	0.0007	0.0067	0.0126
Avg. Time							
ACL/NEON	0.6236	0.2789	0	0	0.0007	0.0263	0.0488
OpenBLAS	0.2314	0.1673	0	0	0.0007	0.0125	0.0509
ACL/GPU	1.3722	0.5219	0	0	0.0018	0.1882	0.1309
MIXED	0.2812	0.1675	0	0	0.007	0.0067	0.0506





5 Conclusion

From the above test cases, we can deduce that :

- the performances of LRN and POOLING are better under ACL_NEON than under OpenBLAS
- the performances of large FC are better under ACL_CL(GPU) than under NEON and OpenBLAS

	AlexNet(s)	GoogleNet(s)	SquezzeNet(s)	MobileNet(s)
LRN/ACL	0.0269	0.0455	0	0
LRN/OpenBLAS	0.4311	0.917	0	0
POOLING/ACL	0.0063	0.0564	0.0225	0.0007
POOLING/OpenBLAS	0.0089	0.0866	0.0259	0.0007
FC/ACL/GPU	0.0898	0.0032	0	0
FC/ACL/NEON	0.4147	0.0043	0	0
FC/OpenBLAS	0.3840	0.0045	0	0

However, for different cases, you may see different result for different layers by using ACL or OpenBLAS. Therefore, for applications, you can select best solution by combining ACL and OpenBLAS together.