MXNet-HRT

Performance Report

2018-02-09

OPEN AI LAB

Reversion Record

Date	Rev	Change Description	Author
2017-9-22	0.1.0	Initial version	Joey
2017-10-11	0.2.0	Test on ACL v17.09	Joey
2018-01-26	0.3.0	Test on ACL v17.12	Huifang
2018-02-09	0.3.1	Test on ACL v17.12	Huifang

catalog

1 Purpose

This Report is tested on RK3399 platform and the Arm Compute Library is version 17.12. The report includes both CPU data and GPU data. We collected the data on AlexNet, GoogleNet, SqueezeNet ,MobileNet,ResNet18,ResNet34,ResNet50. Note that the CPU data is on a single A72 core. There is no performance improvement for mixed mode on MXNet-HRT while on the Caffe-HRT the mixed mode can improve performance 2.8X for the best case. The reason is to be determined, but a potential reason is that Caffe's matrix data is stored as row by row and MXNet's is column by column.

2 Test Environment

Hardware SoC: firefly

http://www.t-firefly.com/product/rk3399.html

- GPU: Mali T864 (800MHz)
- RAM: 2G
- 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

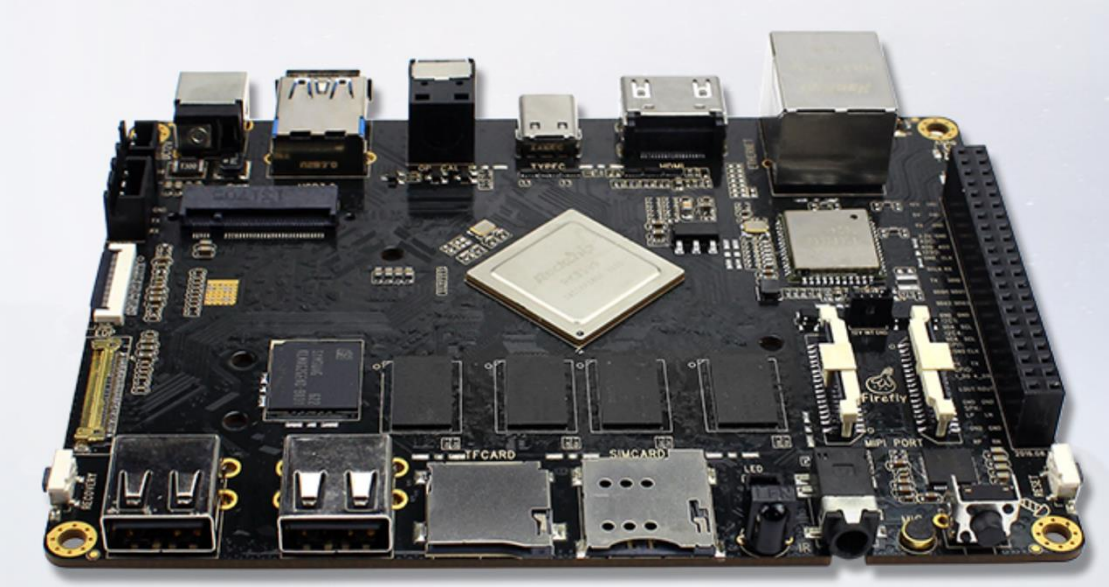

Figure 1 firefly board

3 Performance Improvement Achievement

ACL layers CONV, CONV, FC, LR, Pooling, RELU, SOFTMAX are worse than OpenBLAS on CPU, only FC on GPU has similiar performance. This is different with CaffeOnACL. The reason

is to be determined, but potential reason is that Caffe's matrix data is stored as row by row and MXNet's is column by column.

For the total time spent per inference, achieved about 1.1X performance in the best case.

Table 1 Performance comparation

	Original MXNet (ms)	Mixed Mode (ms)	Performance Gain (ms)
AlexNet	577.30	524.20	1.10
GoogleNet	566.70	508.40	1.11
SquezzeNet	116.80	116.80	1.00
MobileNet	246.20	246.20	1.00
ResNet18	470.20	470.20	1.00
ResNet34	896.50	896.50	1.00
ResNet50	977.10	977.10	1.00

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

`

Table 2 AlexNet performance for configuration

	TPI (s)	Allocate (s)	Run (s)	Config (s)	Copy (s)
1st					
ACL/NEON	1.9737	0.1743	1.4123	0.2159	0.1692
OpenBLAS	0.5883				
ACL/GPU	2.3838	0.1651	0.0638	1.3804	0.7723
MIXED	0.5413	0.0034	0.0316	0.0015	0.0050
Dynamic	2.3903	0.1684	0.0638	1.3917	0.7642
Avg. Time					
ACL/NEON	0.3294		0.3185		0.0103
OpenBLAS	0.5773				
ACL/GPU	0.2180		0.0121		0.2049
MIXED	0.5242		0.0307		0.0042
Dynamic	0.1759		0.0126		0.1622

Table 3 AlexNet performance for each layer

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	1.9737	0.2780	1.6563	0.0244	0.0071	0.0078	0.0002
OpenBLAS	0.5883	0.1579	0.3385	0.0873	0.0034	0.0011	0.0001
ACL/GPU	2.3838	1.0377	0.9754	0.0776	0.0994	0.0421	0.1515
MIXED	0.5413	0.1612	0.3368	0.0358	0.0062	0.0011	0.0001
Dynamic	2.3903	1.0400	0.9788	0.0776	0.0996	0.0425	0.1518
Avg. Time							
ACL/NEON	0.3294	0.1081	0.1940	0.0192	0.0043	0.0037	0.0001
OpenBLAS	0.5773	0.1495	0.3370	0.0862	0.0034	0.0011	0.0001
ACL/GPU	0.2180	0.1402	0.0515	0.0075	0.0082	0.0099	0.0005
MIXED	0.5242	0.1502	0.3377	0.0307	0.0045	0.0011	0.0001
Dynamic	0.1759	0.1105	0.0405	0.0072	0.0080	0.0093	0.0005

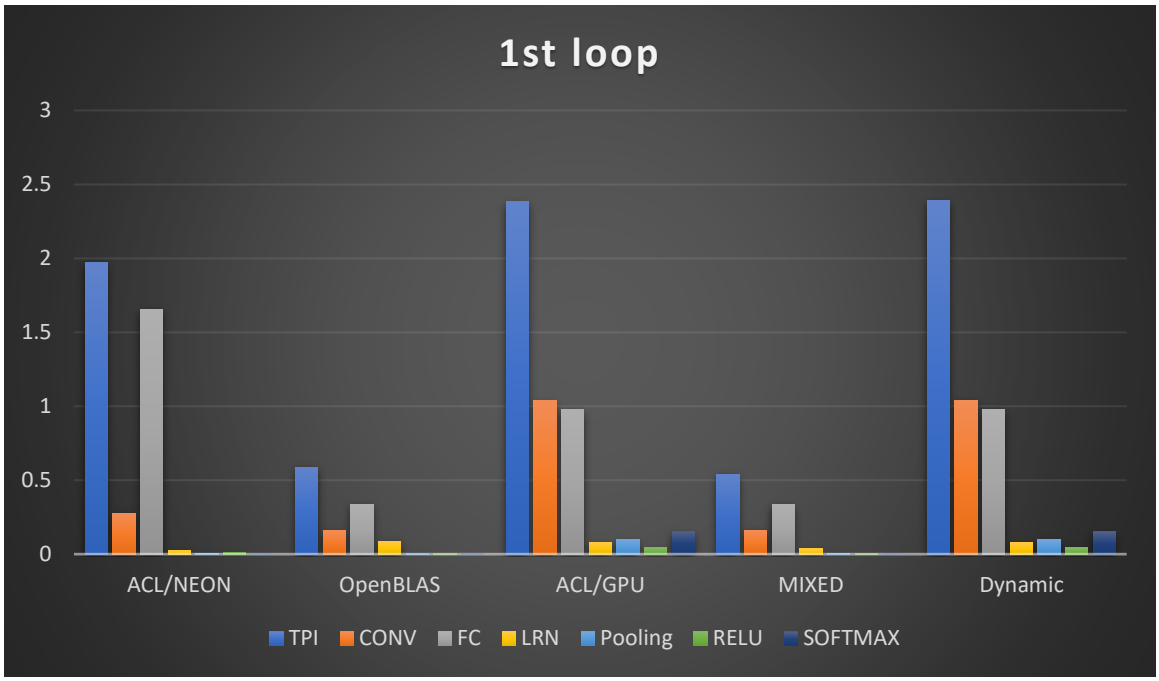

Figure 2 AlexNet 1st loop

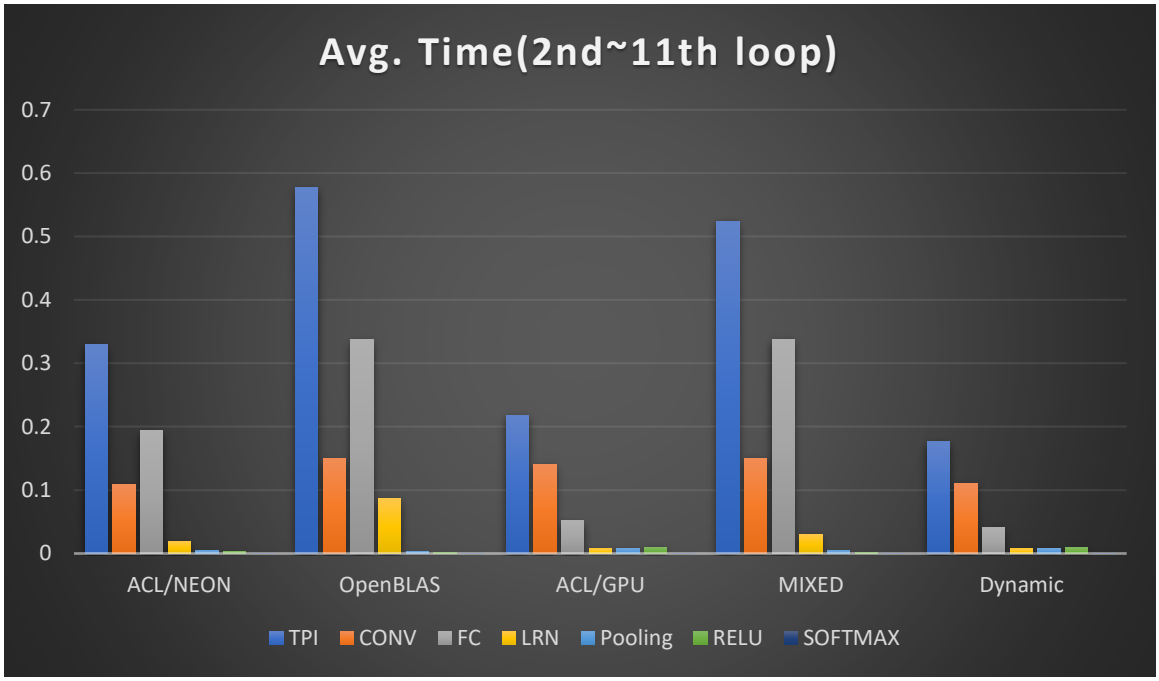

Figure 3 AlexNet Avg. Time(2nd~11th loop)

4.2 GoogleNet

Table 4 GoogleNet performance for configuration

	TPI (s)	Allocate (s)	Run (s)	Config (s)	Copy (s)
1st					
ACL/NEON	1.3799	0.0861	0.6392	0.2256	0.4201
OpenBLAS	0.5829				
ACL/GPU	1.3655	0.0860	0.6373	0.2257	0.4071
MIXED	0.5515	0.0252	0.0818	0.0038	0.0354
Dynamic	5.2026	0.1137	0.1198	3.5770	1.3799
Avg. Time					
ACL/NEON	0.6521		0.5727		0.0757
OpenBLAS	0.5667				
ACL/GPU	0.6500		0.5709		0.0755
MIXED	0.5084		0.0820		0.0316
Dynamic	1.0024		0.0633		0.9339

Table 5 GoogleNet performance for each layer

	TPI (s)	CONV (s)	FC (s)	LRN (s)	Pooling (s)	RELU (s)	SOFTMAX (s)
1st							
ACL/NEON	1.3799	1.1734	0.0193	0.0629	0.0553	0.0393	0.0002
OpenBLAS	0.5829	0.3870	0.0047	0.1422	0.0394	0.0065	0.0001
ACL/GPU	1.3655	1.1595	0.0193	0.0626	0.0551	0.0391	0.0003
MIXED	0.5515	0.3921	0.0050	0.0628	0.0564	0.0059	0.0001
Dynamic	5.2026	4.3651	0.1494	0.0860	0.2355	0.1415	0.1539
Avg. Time							
ACL/NEON	0.6521	0.5146	0.0062	0.0536	0.0396	0.0176	0.0001
OpenBLAS	0.5667	0.3768	0.0045	0.1390	0.0381	0.0055	0.0001
ACL/GPU	0.6500	0.5131	0.0062	0.0535	0.0394	0.0174	0.0001
MIXED	0.5084	0.3832	0.0049	0.0537	0.0402	0.0056	0.0001
Dynamic	1.0024	0.8309	0.0022	0.0143	0.0648	0.0600	0.0005

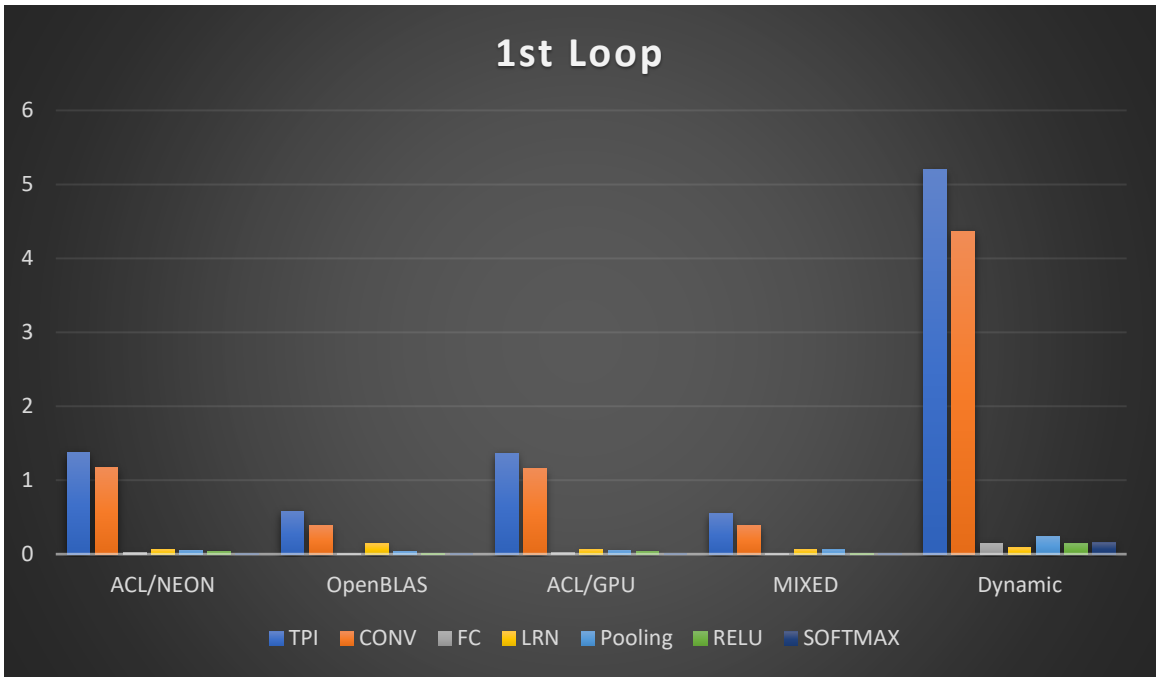

Figure 4 GoogleNet 1st Loop

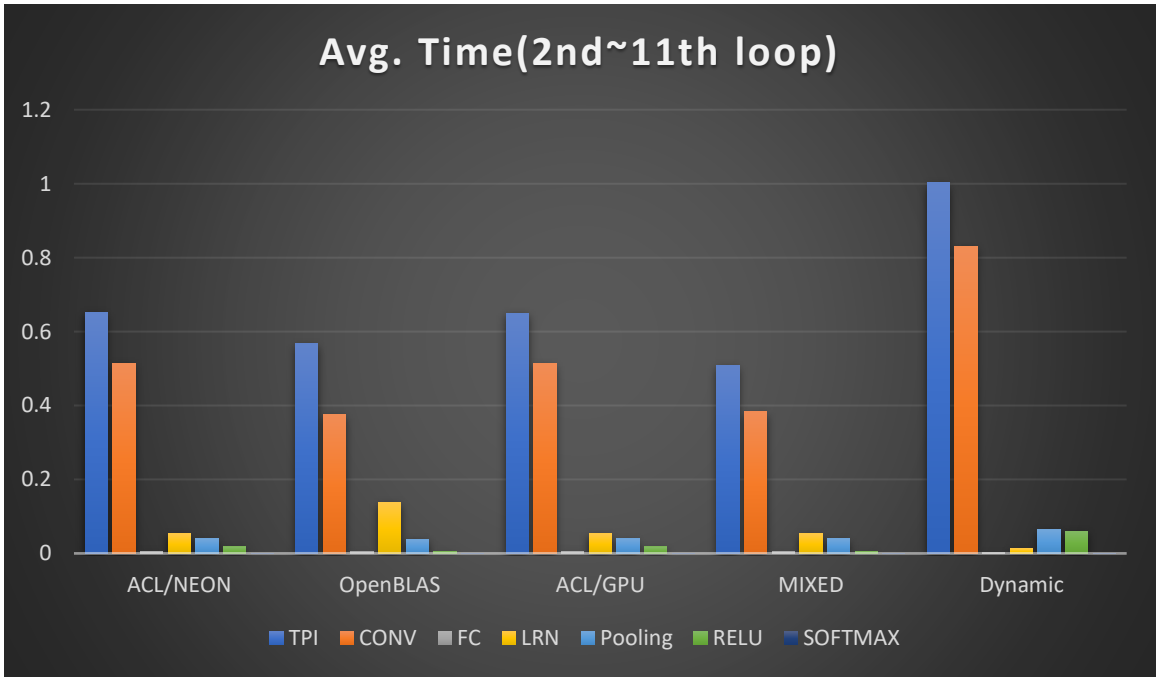

Figure 5 GoogleNet Avg. Time(2nd~11th loop)

4.3 SqueezeNet

Table 6 SqueezeNet performance for configuration

	TPI (s)	Allocate (s)	Run (s)	Config (s)	Copy (s)
1st					
ACL/NEON	0.4393	0.0444	0.1724	0.0890	0.1288
OpenBLAS	0.1286				
ACL/GPU	2.7444	0.0357	0.0470	2.2435	0.4124
MIXED	0.1648	0.0136	0.0168	0.0005	0.0164
Dynamic	2.7082	0.0353	0.0469	2.2175	0.4029
Avg. Time					
ACL/NEON	0.1962		0.1588		0.0353
OpenBLAS	0.1168				
ACL/GPU	0.3238		0.0286		0.2924
MIXED	0.1358		0.0165		0.0145
Dynamic	0.3260		0.0289		0.2944

Table 7 SqueezeNet performance for each layer

	TPI (s)	CONV (s)	FC (s)	LRN (s)	Pooling (s)	RELU (s)	SOFTMAX (s)
1st							
ACL/NEON	0.4393	0.3590			0.0168	0.0309	0.0003
OpenBLAS	0.1286	0.1092			0.0104	0.0048	0.0001
ACL/GPU	2.7444	2.3495			0.0862	0.0932	0.1536
MIXED	0.1648	0.1112			0.0168	0.0046	0.0001
Dynamic	2.7082	2.3185			0.0858	0.0900	0.1529
Avg. Time							
ACL/NEON	0.1962	0.1502			0.0102	0.0141	0.0001
OpenBLAS	0.1168	0.0997			0.0096	0.0046	0.0001
ACL/GPU	0.3238	0.2347			0.0206	0.0390	0.0009
MIXED	0.1358	0.0996			0.0105	0.0044	0.0001
Dynamic	0.3260	0.2355			0.0209	0.0397	0.0009

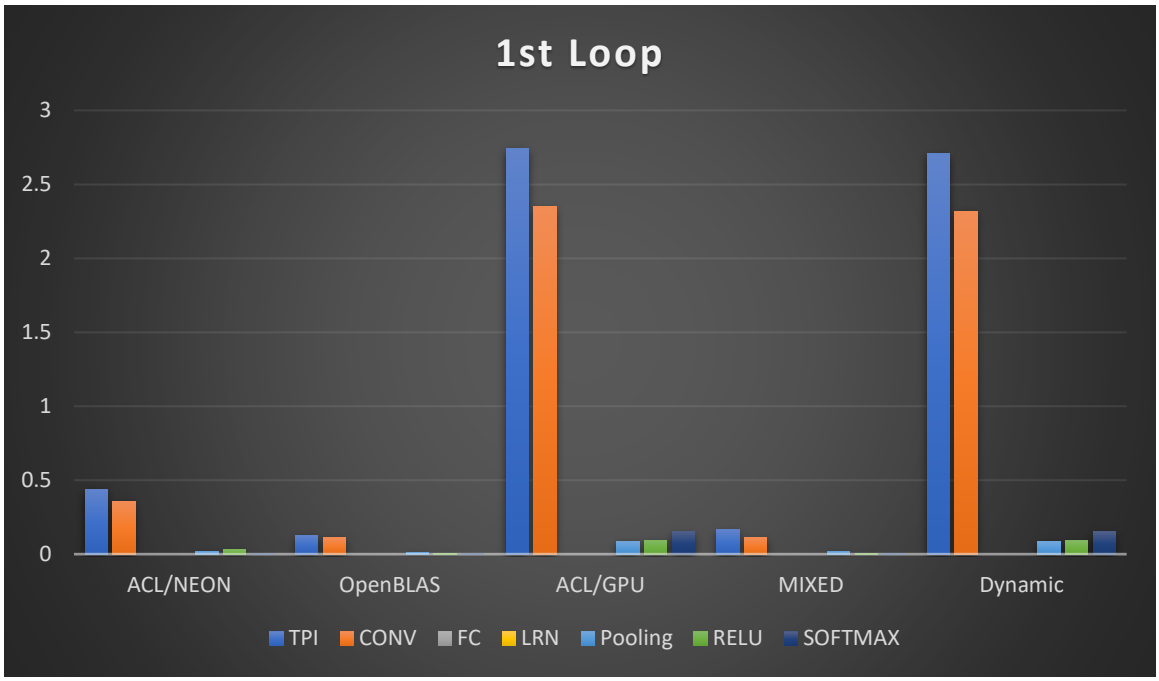

Figure 6 SqueezeNet 1st Loop

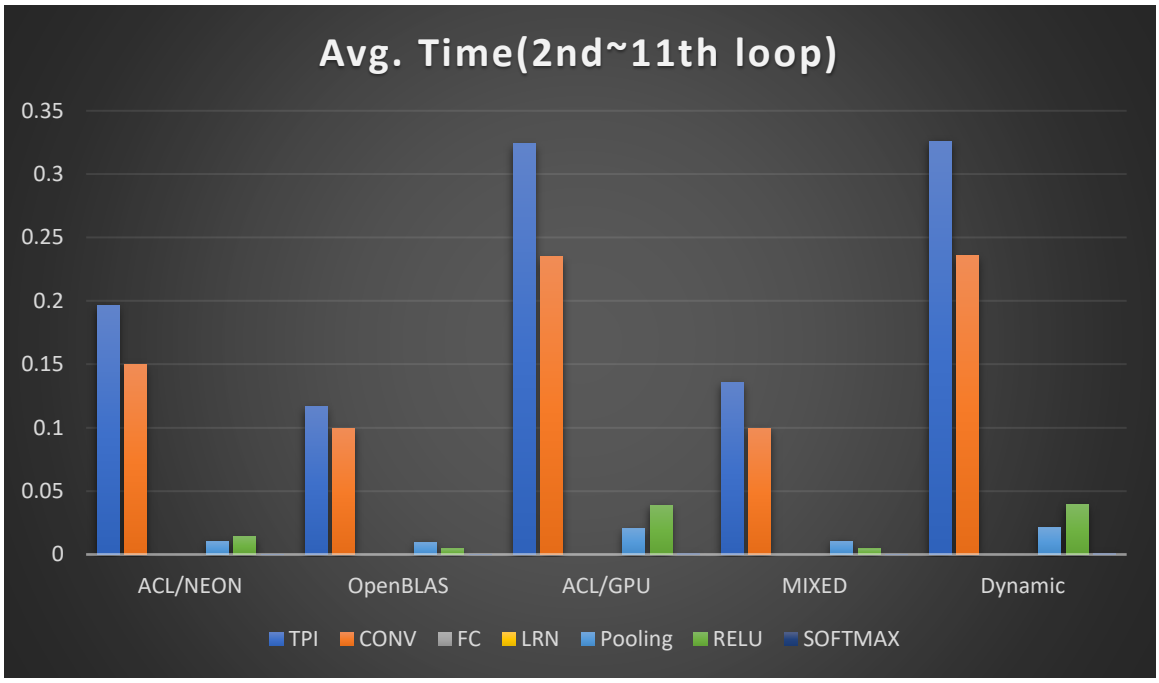

Figure 7 SqueezeNet Avg. Time(2nd~11th loop)

4.4 MobileNet

Table 8 MobileNet performance for configuration

	TPI (s)	Allocate (s)	Run (s)	Config (s)	Copy (s)
1st					
ACL/NEON	0.6838	0.0804	0.2483	0.0731	0.1853
OpenBLAS	0.2654				
ACL/GPU	0.4340	0.0099	0.0119	0.0674	0.0899
MIXED	0.3430	0.0289	0.0246	0.0006	0.0372
Dynamic	2.2918	0.0708	0.0377	1.2743	0.8092
Avg. Time					
ACL/NEON	0.3657	0.0003	0.2224	0.0005	0.0589
OpenBLAS	0.2462				
ACL/GPU	0.3416		0.0177		0.0875
MIXED	0.2933		0.0234		0.0351
Dynamic	0.5892	0.0003	0.0393	0.0161	0.4479

Table 9 MobileNet performance for each layer

	TPI (s)	CONV (s)	FC (s)	LRN (s)	Pooling (s)	RELU (s)	SOFTMAX (s)
1st							
ACL/NEON	0.6838	0.5372			0.0001	0.0599	
OpenBLAS	0.2654	0.2439			0.0001	0.0089	
ACL/GPU	0.4340	0.2422			0.0001	0.0090	
MIXED	0.3430	0.2400			0.0001	0.0088	
Dynamic	2.2918	1.9959			0.0001	0.1225	
Avg. Time							
ACL/NEON	0.3657	0.2867			0.0001	0.0285	
OpenBLAS	0.2462	0.2282			0.0001	0.0080	
ACL/GPU	0.3416	0.2265			0.0001	0.0085	
MIXED	0.2933	0.2249			0.0001	0.0087	
Dynamic	0.5892	0.4235			0.0001	0.0641	

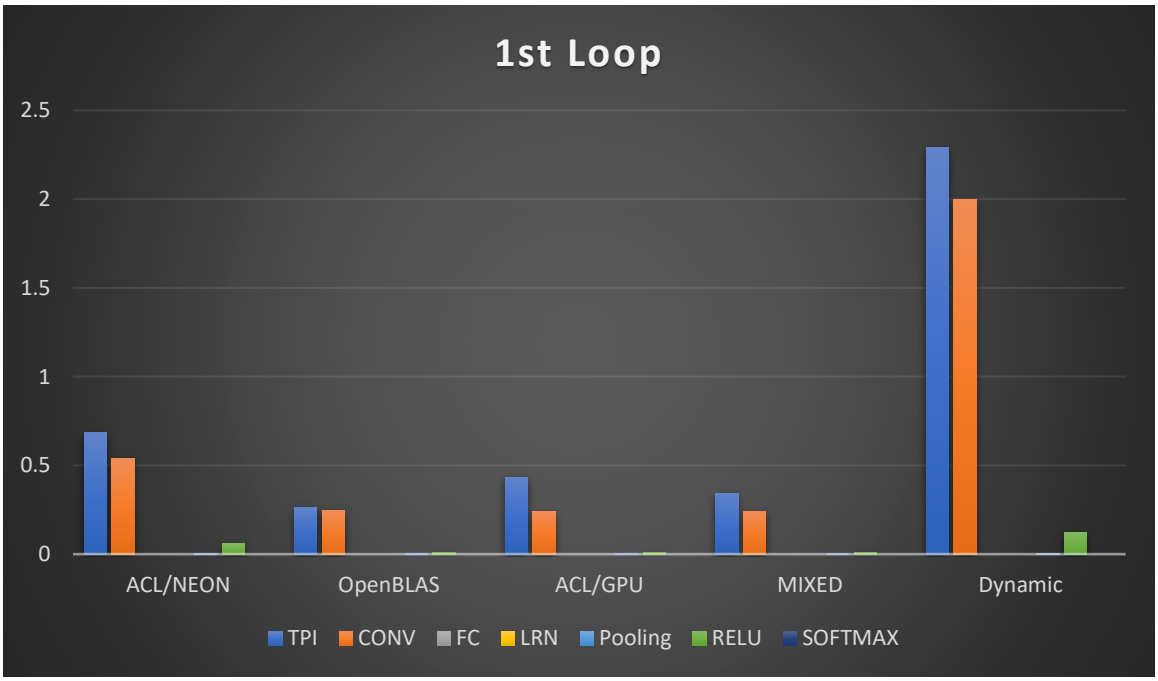

Figure 8 MobileNet 1st Loop

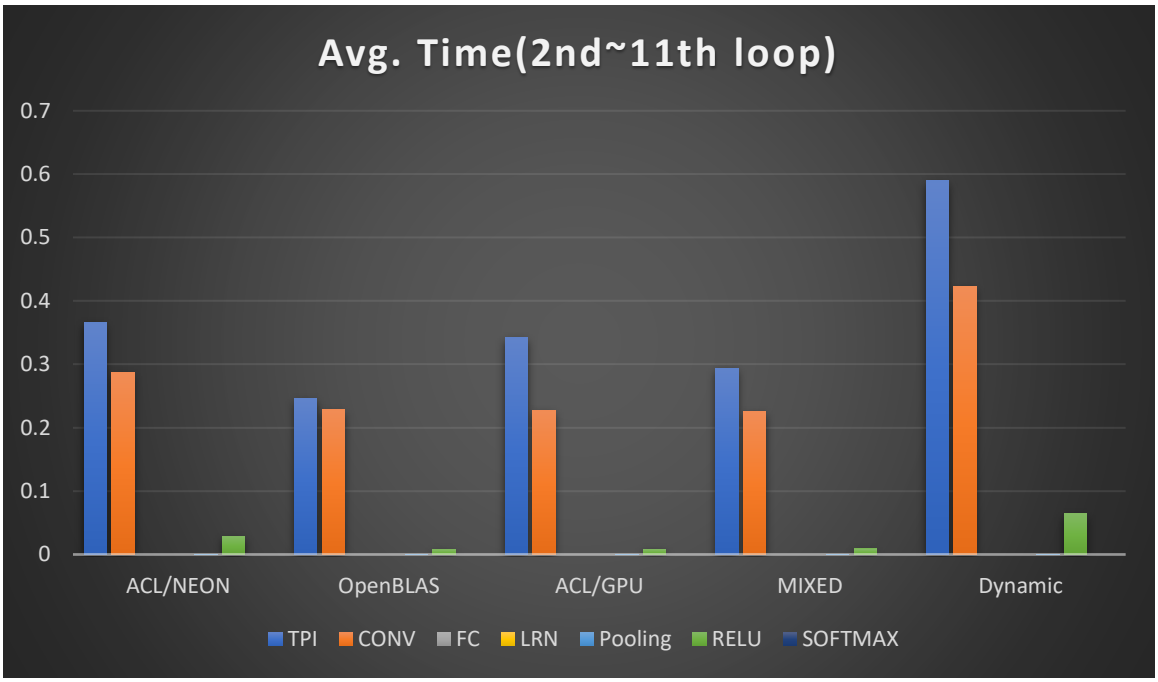

Figure 9 MobileNet Avg. Time(2nd~11th loop)

4.5 ResNet18

Table 10 ResNet18 performance for configuration

	TPI (s)	Allocate (s)	Run (s)	Config (s)	Copy (s)
1st					
ACL/NEON	1.1953	0.0765	0.7013	0.2226	0.1898
OpenBLAS	0.4812				
ACL/GPU	3.5631	0.0812	0.0559	1.9086	1.5105
MIXED	0.5254	0.0194	0.0203	0.0024	0.0239
Dynamic	3.3236	0.0809	0.0570	1.8953	1.2837
Avg. Time					
ACL/NEON	0.6252		0.5854		0.0380
OpenBLAS	0.4702				
ACL/GPU	0.8823		0.0327		0.8471
MIXED	0.4858		0.0174		0.0195
Dynamic	0.4525		0.0408		0.4083

Table 11 ResNet18 performance for each layer

	TPI (s)	CONV (s)	FC (s)	LRN (s)	Pooling (s)	RELU (s)	SOFTMAX (s)
1st							
ACL/NEON	1.1953	1.1038	0.0099		0.0086	0.0268	0.0003
OpenBLAS	0.4812	0.4618	0.0021		0.0056	0.0039	0.0001
ACL/GPU	3.5631	2.9931	0.1419		0.0745	0.0791	0.1522
MIXED	0.5254	0.4533	0.0098		0.0090	0.0040	0.0002
Dynamic	3.3236	2.7641	0.1397		0.0739	0.0778	0.1511
Avg. Time							
ACL/NEON	0.6252	0.5759	0.0029		0.0057	0.0136	0.0001
OpenBLAS	0.4702	0.4531	0.0022		0.0056	0.0039	0.0001
ACL/GPU	0.8823	0.7787	0.0013		0.0099	0.0329	0.0005
MIXED	0.4858	0.4443	0.0027		0.0057	0.0038	0.0001
Dynamic	0.4525	0.3698	0.0008		0.0083	0.0281	0.0004

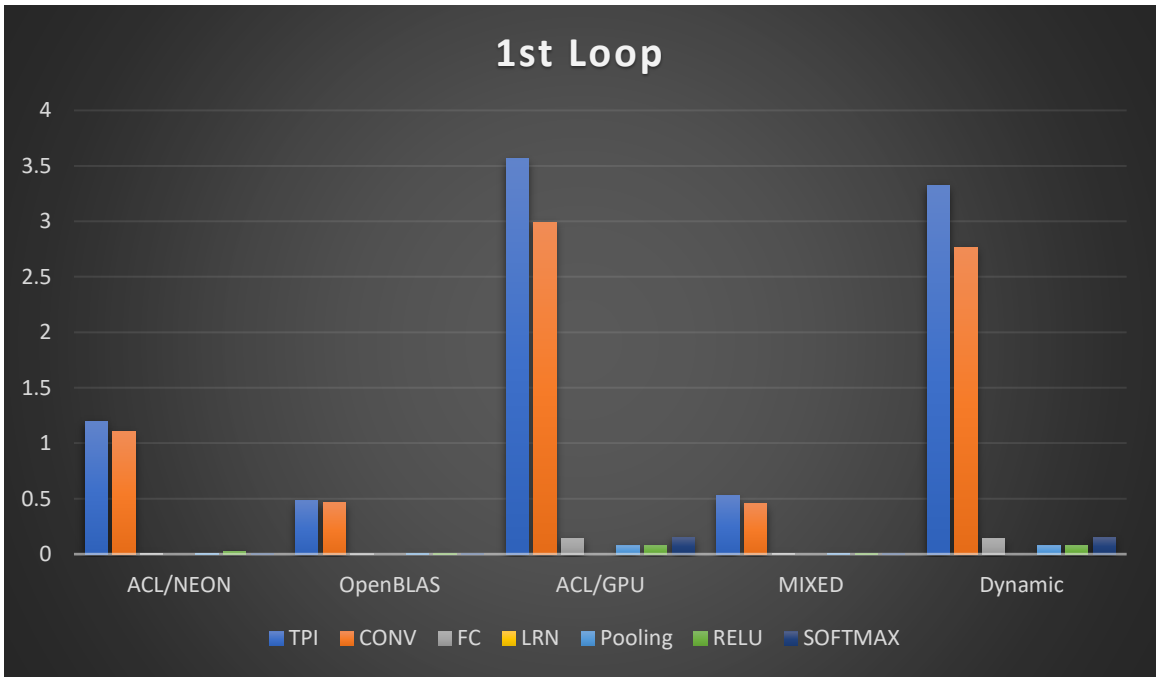

Figure 10 ResNet18 1st Loop

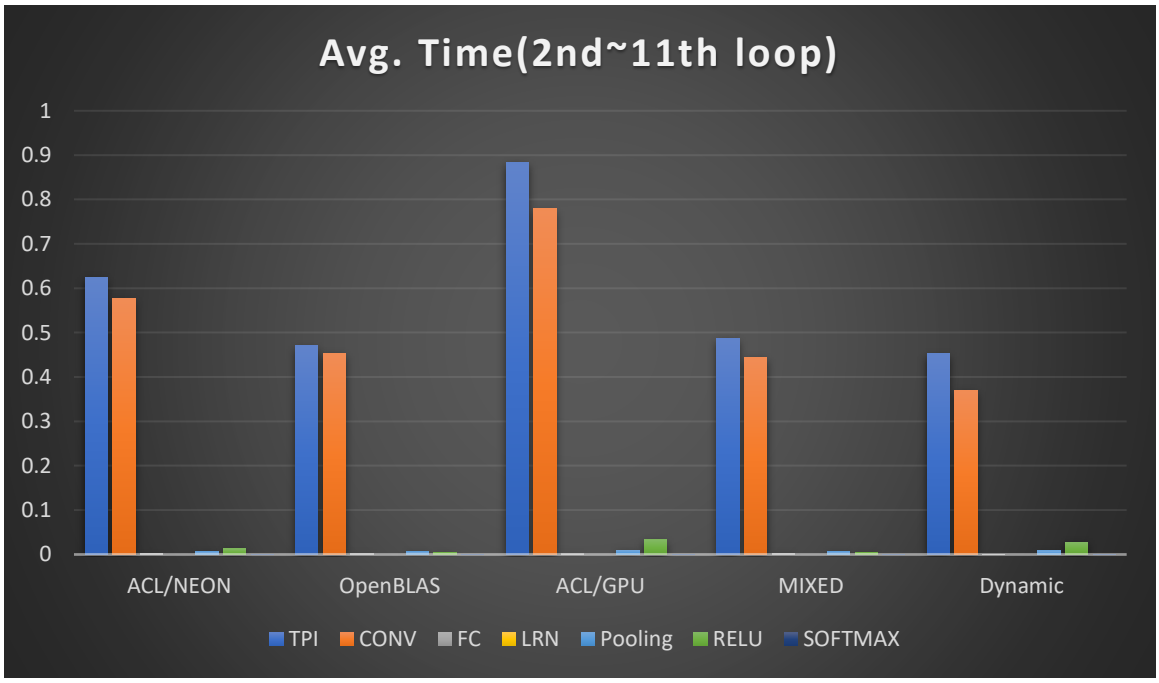

Figure 11 ResNet18 Avg. Time(2nd~11th loop)

4.6 ResNet34

Table 12 ResNet34 performance for configuration

	TPI (s)	Allocate (s)	Run (s)	Config (s)	Copy (s)
1st					
ACL/NEON	2.5576	0.1261	1.3160	0.3663	0.7409
OpenBLAS	0.9097				
ACL/GPU	5.0484	0.1456	0.1008	2.0855	2.7041
MIXED	0.9806	0.0273	0.0269	0.0027	0.0339
Dynamic	4.6647	0.1455	0.1050	2.0826	2.3200
Avg. Time					
ACL/NEON	1.1666		1.0981		0.0654
OpenBLAS	0.8965				
ACL/GPU	0.9678		0.0729		0.8890
MIXED	0.9305		0.0240		0.0297
Dynamic	0.8492		0.0731		0.7703

Table 13 ResNet34 performance for each layer

	TPI (s)	CONV (s)	FC (s)	LRN (s)	Pooling (s)	RELU (s)	SOFTMAX (s)
1st							
ACL/NEON	2.5576	2.4228	0.0096		0.0086	0.0412	0.0002
OpenBLAS	0.9097	0.8853	0.0021		0.0056	0.0062	0.0001
ACL/GPU	5.0484	4.3755	0.1418		0.0740	0.1221	0.1515
MIXED	0.9806	0.8802	0.0100		0.0090	0.0063	0.0002
Dynamic	4.6647	3.9981	0.1402		0.0733	0.1215	0.1515
Avg. Time							
ACL/NEON	1.1666	1.0904	0.0028		0.0056	0.0208	0.0001
OpenBLAS	0.8965	0.8744	0.0023		0.0056	0.0062	0.0001
ACL/GPU	0.9678	0.8309	0.0008		0.0081	0.0475	0.0004
MIXED	0.9305	0.8696	0.0028		0.0059	0.0060	0.0001
Dynamic	0.8492	0.7201	0.0007		0.0081	0.0477	0.0004

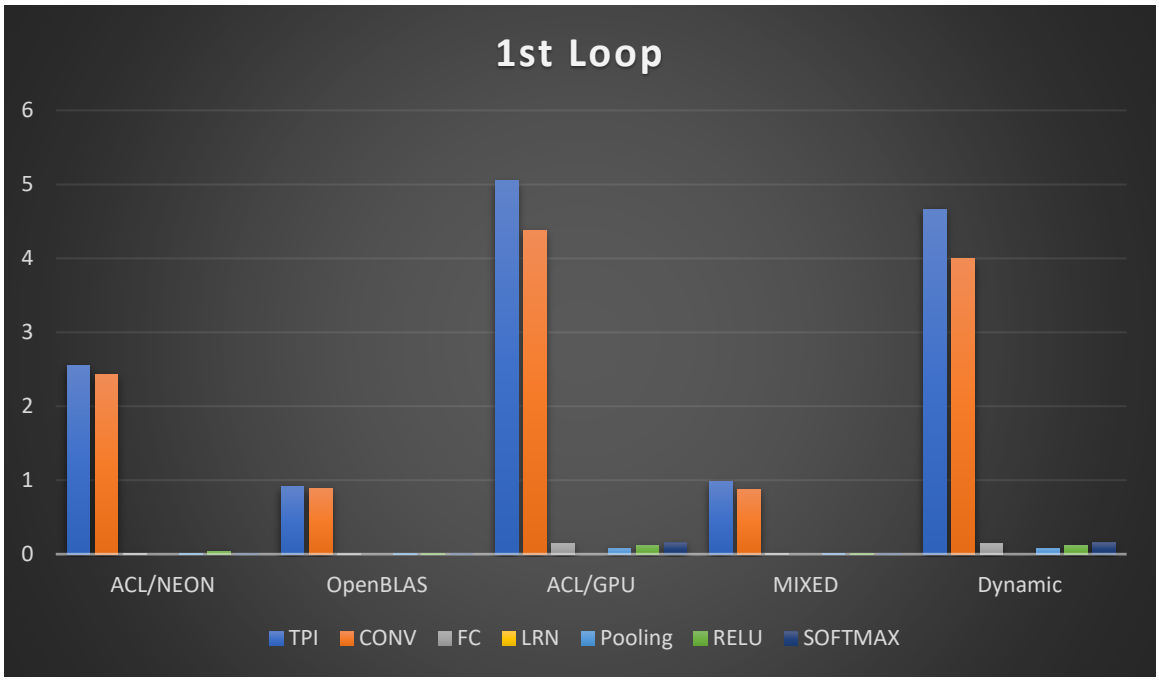

Figure 12 ResNet34 1st Loop

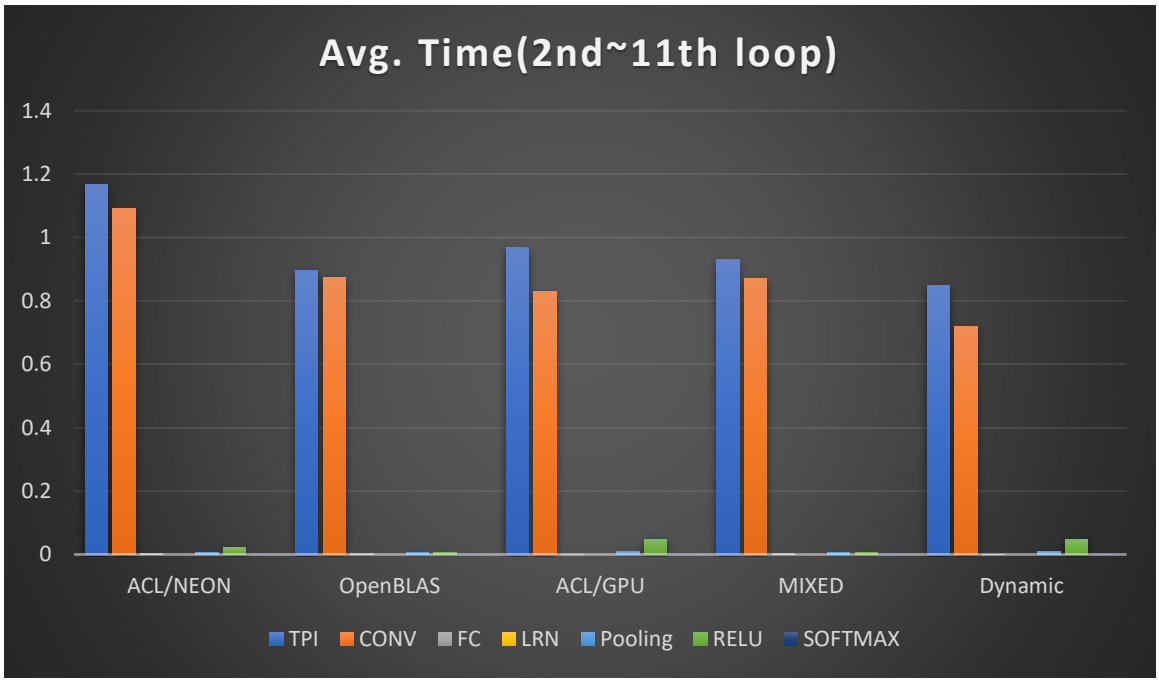

Figure 13 ResNet34 Avg. Time(2nd~11th loop)

4.7 ResNet50

Table 14 ResNet50 performance for configuration

	TPI (s)	Allocate (s)	Run (s)	Config (s)	Copy (s)
1st					
ACL/NEON	3.1125	0.2413	1.5877	0.4051	0.8642
OpenBLAS	0.9914				
ACL/GPU	7.2338	0.2845	0.1360	2.5758	4.2194
MIXED	1.1884	0.0697	0.0865	0.0072	0.0865
Dynamic	4.6647	0.1455	0.1050	2.0826	2.3200
Avg. Time					
ACL/NEON	1.4489		1.2991		0.1445
OpenBLAS	0.9771				
ACL/GPU	2.1249		0.0955		2.0224
MIXED	1.0776		0.0719		0.0806
Dynamic	0.8492		0.0731		0.7703

Table 15 ResNet50 performance for each layer

	TPI (s)	CONV (s)	FC (s)	LRN (s)	Pooling (s)	RELU (s)	SOFTMAX (s)
1st							
ACL/NEON	3.1125	2.7627	0.0437		0.0087	0.1063	0.0002
OpenBLAS	0.9914	0.9339	0.0099		0.0057	0.0157	0.0001
ACL/GPU	7.2338	6.2501	0.1608		0.0746	0.2342	0.1524
MIXED	1.1884	0.9172	0.0410		0.0094	0.0162	0.0002
Dynamic	4.6647	3.9981	0.1402		0.0733	0.1215	0.1515
Avg. Time							
ACL/NEON	1.4489	1.2614	0.0119		0.0057	0.0528	0.0001
OpenBLAS	0.9771	0.9229	0.0099		0.0057	0.0165	0.0001
ACL/GPU	2.1249	1.7829	0.0038		0.0095	0.1160	0.0005
MIXED	1.0776	0.9071	0.0121		0.0061	0.0159	0.0001
Dynamic	0.8492	0.7201	0.0007		0.0081	0.0477	0.0004

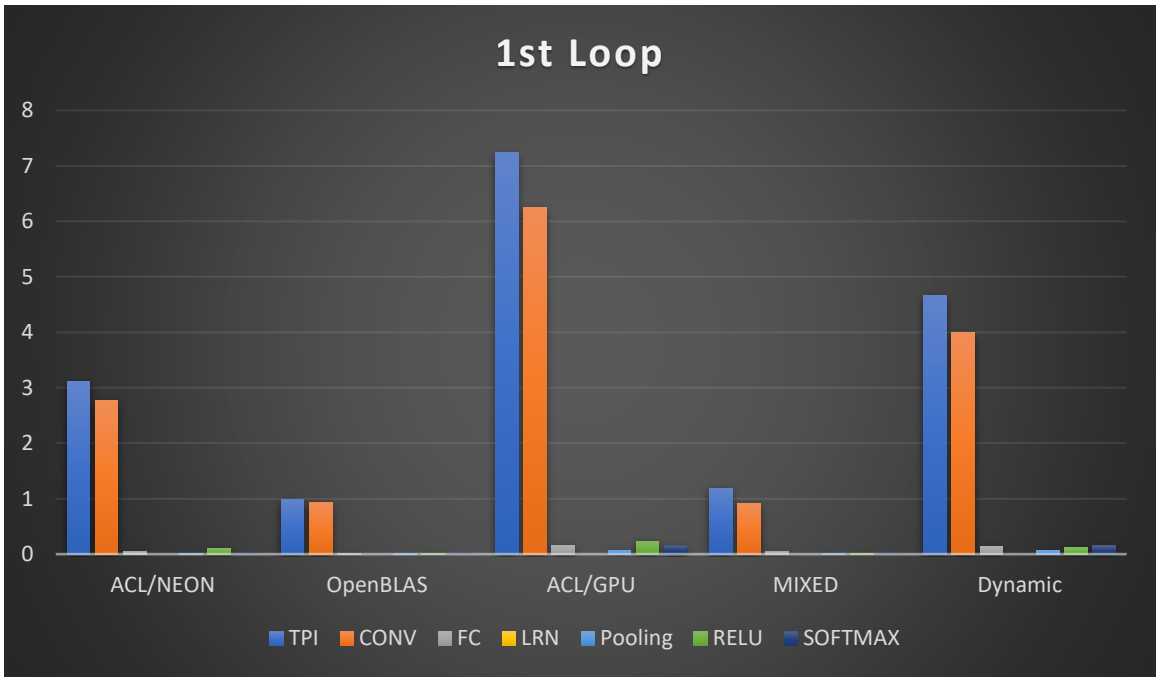

Figure 14 ResNet50 1st Loop

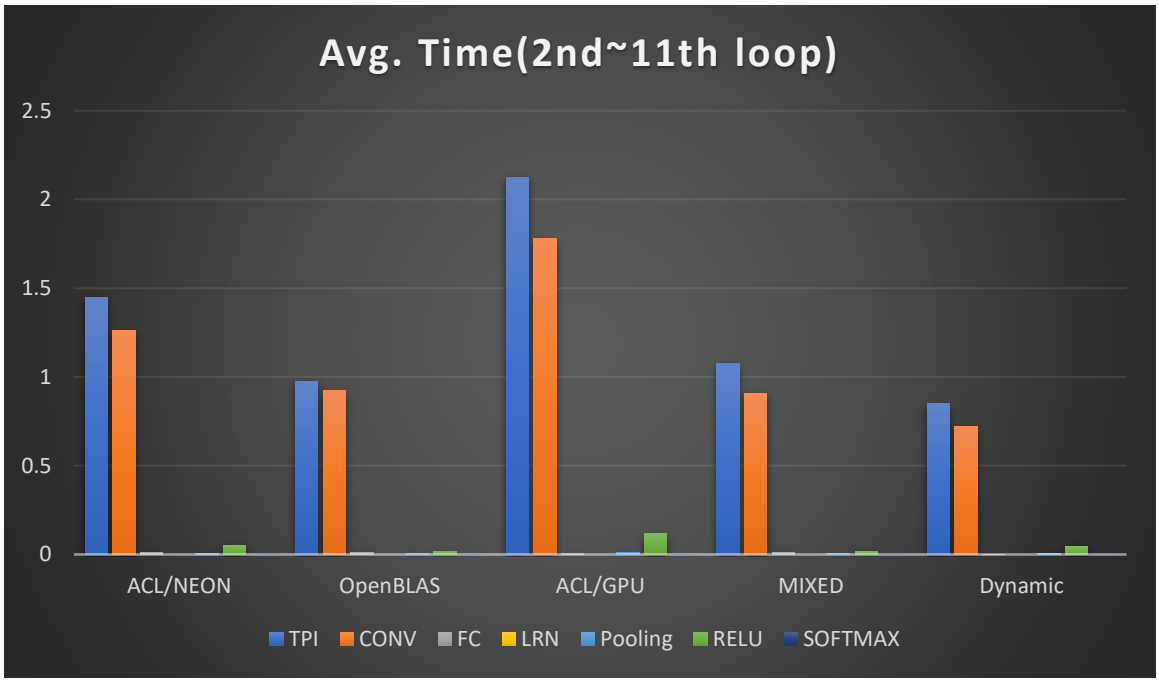

Figure 15 ResNet50 Avg. Time(2nd~11th loop)

5 Performance On Different Cores

The TPI is not very stable, it's in wide fluctuation. The data in the tables is lower limit of the range.

5.1 The TPI Data For ACL/NEON, OpenBLAS And Mixed Mode

AlexNet TPI data for ACL/NEON, OpenBLAS and mixed mode

Table 16 AlexNet TPI data for mixed mode

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)	Dynamic(s)
1xA53	0.3315	0.5785	0.9277	0.3804
1xA72	0.3294	0.5773	0.5242	0.1759
2xA72	0.3279	0.5101	0.4480	0.1616
4xA53	0.3294	1.8793	0.6366	0.3832
2xA72+4xA53	0.3263	2.3872	0.5176	0.2389

GoogleNet TPI data for ACL/NEON, OpenBLAS and mixed mode

Table 17 GoogleNet TPI data for mixed mode

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)	Dynamic(s)
1xA53	1.2108	1.3855	1.2986	1.0651
1xA72	0.6521	0.5667	0.5084	1.0024
2xA72	0.4320	0.4265	0.3513	0.8030
4xA53	0.6029	0.7352	0.6481	1.0775
2xA72+4xA53	0.3838	0.4359	0.3836	1.0856

SqueezeNet TPI data for ACL/NEON, OpenBLAS and mixed mode

Table 18 SqueezeNet TPI data for mixed mode

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)	Dynamic(s)
1xA53	0.7995	0.6376	0.7330	0.8184
1xA72	0.3657	0.2462	0.2933	0.5892
2xA72	0.3169	0.1914	0.2298	0.5649
4xA53	0.5719	0.4099	0.5053	0.8312
2xA72+4xA53	0.2868	0.1944	0.2520	0.7610

MobileNet TPI data for ACL/NEON, OpenBLAS and mixed mode.

Table 19 MobileNet TPI data for mixed mode

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)	Dynamic(s)
1xA53	0.3829	0.3104	0.3673	0.3795
1xA72	0.1962	0.1168	0.1358	0.3260
2xA72	0.1383	0.0795	0.1008	0.3080
4xA53	0.2495	0.1549	0.2106	0.3872
2xA72+4xA53	0.1267	0.0879	0.1989	0.3816

ResNet18 TPI data for ACL/NEON, OpenBLAS and mixed mode.

Table 20 ResNet18 TPI data for mixed mode

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)	Dynamic(s)
1xA53	1.1542	1.2120	1.2602	0.9158
1xA72	0.6252	0.4702	0.4858	0.4525
2xA72	0.4436	0.3051	0.3300	0.5235
4xA53	0.5535	0.5033	0.5478	0.8611
2xA72+4xA53	0.4262	0.3371	0.3620	0.7003

ResNet34 TPI data for ACL/NEON, OpenBLAS and mixed mode.

Table 21 ResNet34 TPI data for mixed mode

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)	Dynamic(s)
1xA53	2.1409	2.3703	2.4444	1.2478
1xA72	1.1666	0.8965	0.9305	0.8492
2xA72	0.8056	0.5722	0.6174	1.0873
4xA53	1.0123	0.9070	0.9845	1.3392
2xA72+4xA53	0.7341	0.6278	0.6589	1.1232

ResNet50 TPI data for ACL/NEON, OpenBLAS and mixed mode.

Table 22 ResNet50 TPI data for mixed mode

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)	Dynamic(s)
1xA53	2.8004	2.5506	2.7960	2.1781
1xA72	1.4489	0.9771	1.0776	1.4304
2xA72	1.0232	0.6487	0.7409	1.6328
4xA53	1.4085	1.0185	1.2712	2.1556
2xA72+4xA53	0.9560	0.6758	0.8133	2.0884

5.2 The TPI In Mixed mode

The TPI data for different CPU cores in mixed mode:

Table 23 1.1 The TPI In Mixed mode

	AlexNet (s)	GoogleNet (s)	MobileNet (s)	SqueezeNet (s)	ResNet18 (s)	ResNet34 (s)	ResNet50 (s)
1xA53	0.9277	1.2986	0.7330	0.3673	1.2602	2.4444	2.7960
1xA72	0.5242	0.5084	0.2933	0.1358	0.4858	0.9305	1.0776
2xA72	0.4480	0.3513	0.2298	0.1008	0.3300	0.6174	0.7409
4xA53	0.6366	0.6481	0.5053	0.2106	0.5478	0.9845	1.2712
2xA72+ 4xA53	0.5176	0.3836	0.2520	0.1989	0.3620	0.6589	0.8133

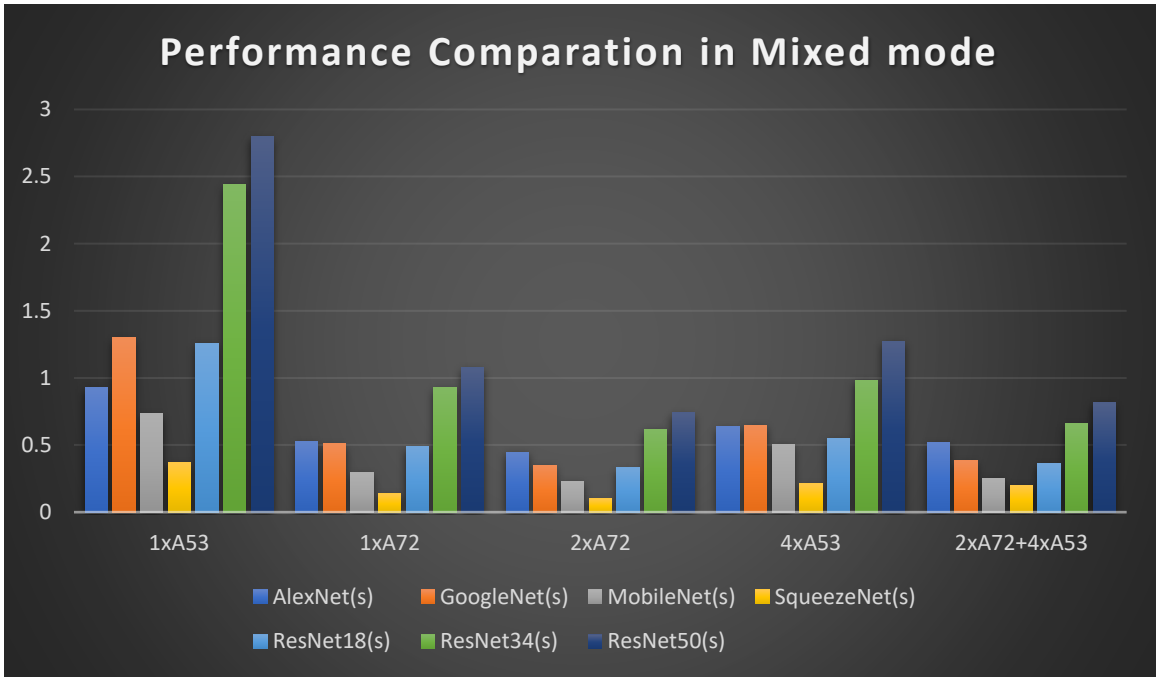

Figure 16 Performance Comparation in mixed mode

6 Conclusion

From the above test cases, we can deduce that: the performances of large FC are better under ACL_CL(GPU) than under NEON and OpenBLAS.

Table 24 Performance of FC layer for different models

	AlexNet (s)	GoogleNet (s)	SquezzeNet (s)	MobileNet (s)	ResNet18 (s)	ResNet34 (s)	ResNet50 (s)
FC/ACL/NEON	0.1942	0.0061	0	0	0.0029	0.0028	0.0119
FC/OpenBLAS	0.3356	0.0045	0	0	0.0022	0.0023	0.0099
FC/ACL/GPU	0.0530	0.0061	0	0	0.0013	0.0008	0.0038