PERFORMANCE REPORT OF MXNETONACL

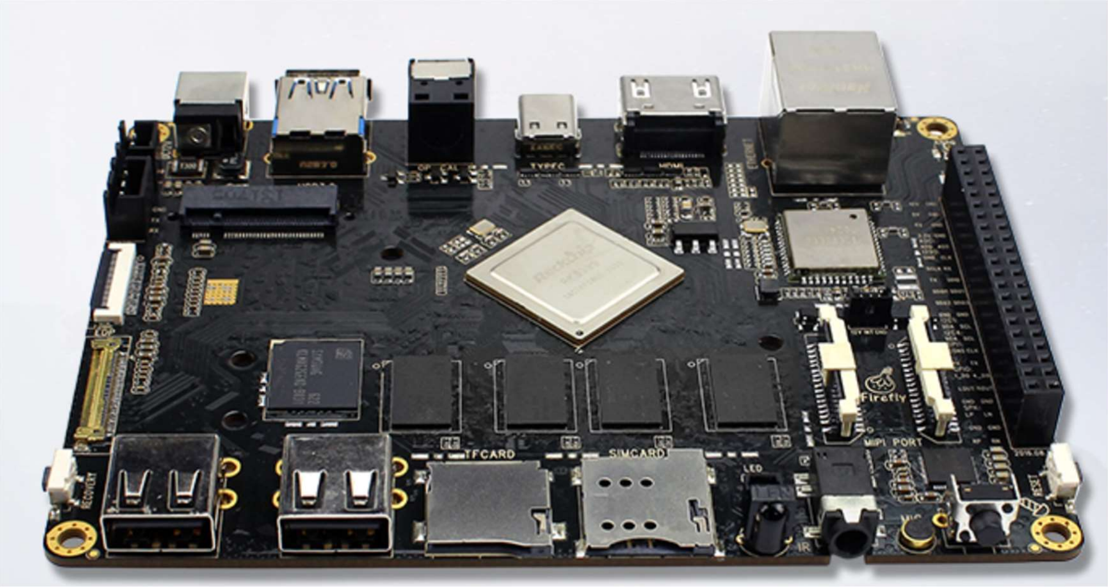
OPEN AI LAB
开放智能实验室

Index

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.There is no performance improvement for mixed mode on MXNetOnACL while on the CaffeOnACL the mixed mode can improve performance 2.78X for the best case. The reason is to be determined, but a potential reason is that Caffe matrix data is stored as row by row and MXNet's is column by column.

1. Test Environment

Board：http://wiki.t-firefly.com/index.php/Firefly-RK3399
Chip: Rockchip RK3399
System: Ubuntu 16.04
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)



2. Original MXNet has better Performance

ACL layers CONV, .CONV, FC, LR, Pooling, RELU, SOFTMAX are worse than OpenBLAS on CPU, only FC on GPU has better performance. This is different with CaffeOnACL. The reason is to be determined, but potential reason is that Caffe matrix data is stored as row by row and MXNet’s is column by column.

We almost can’t get any performance improvement by mixed mode.

	Original MXNet(ms)	Mixed Mode(ms)	Performance Gain
AlexNet	481	469	1.03X
GoogleNet	450	1251	0.36X
SquezzeNet	82	136	0.60X
MobileNet	191	296	0.64X

3. 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.

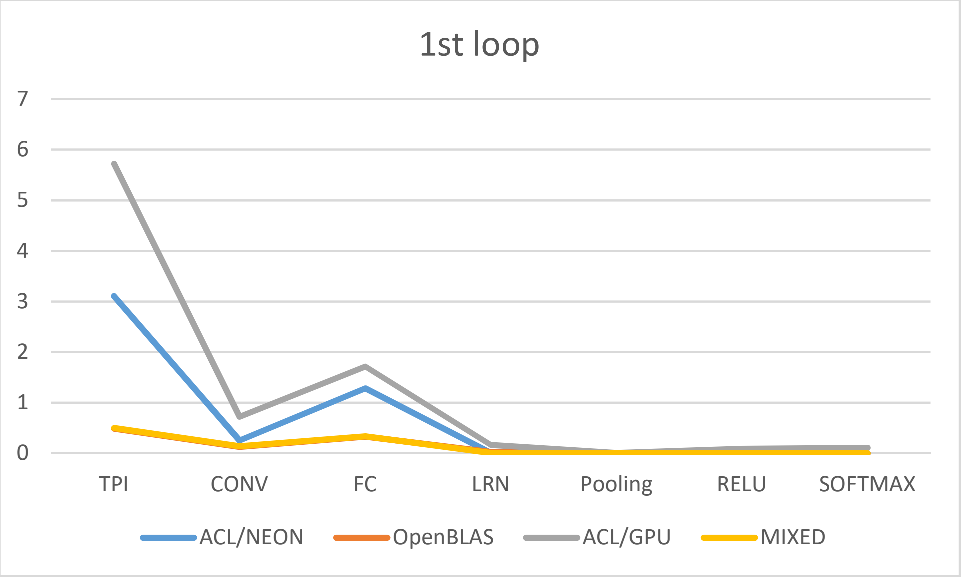
The details see user manual section “Use Cases”.

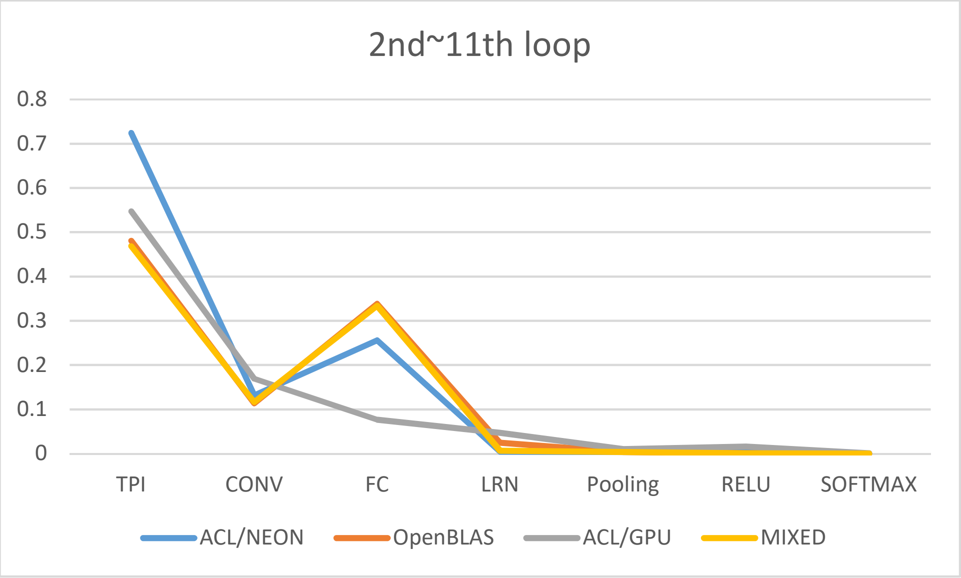
3.1 AlexNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	3.1067	0.3502	0.9049	0.1799	0.112
OpenBLAS	0.487	0	0	0	0
ACL/GPU	5.719	0.4451	0.5052	1.7378	0.2215
MIXED	0.5001	0.0009	0.0064	0.0004	0.0009
Avg. Time					
ACL/NEON	0.7239	0.0017	0.3187	0	0.0038
OpenBLAS	0.4806	0	0	0	0
ACL/GPU	0.5473	0.0089	0.2007	0	0.0171
MIXED	0.4685	0.0003	0.0059	0	0.0006

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	3.1067	0.2552	1.2849	0.0072	0.0037	0.0073	0.0015
OpenBLAS	0.4897	0.1255	0.3332	0.0259	0.0037	0.0013	0.0001
ACL/GPU	5.719	0.7207	1.7126	0.1665	0.0103	0.0891	0.1102

MIXED	0.5001	0.1401	0.3375	0.0085	0.0038	0.0016	0.0001
Avg. Time							
ACL/NEON	0.7239	0.1317	0.2559	0.0047	0.0037	0.0034	0.0002
OpenBLAS	0.4806	0.1133	0.3381	0.0247	0.0034	0.0011	0.0001
ACL/GPU	0.5473	0.1692	0.077	0.0471	0.0103	0.0157	0.0012
MIXED	0.4685	0.1162	0.3341	0.0066	0.0035	0.0011	0.0001



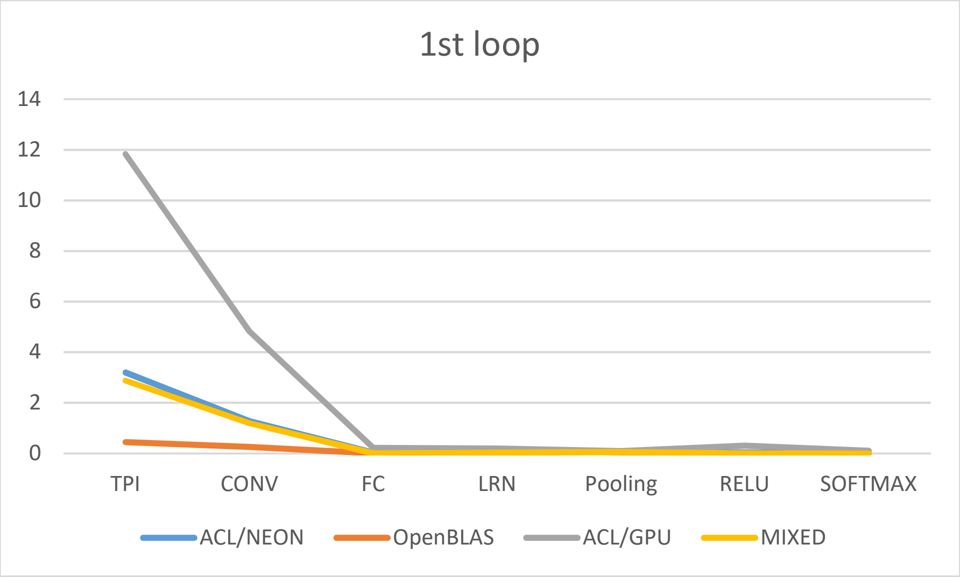


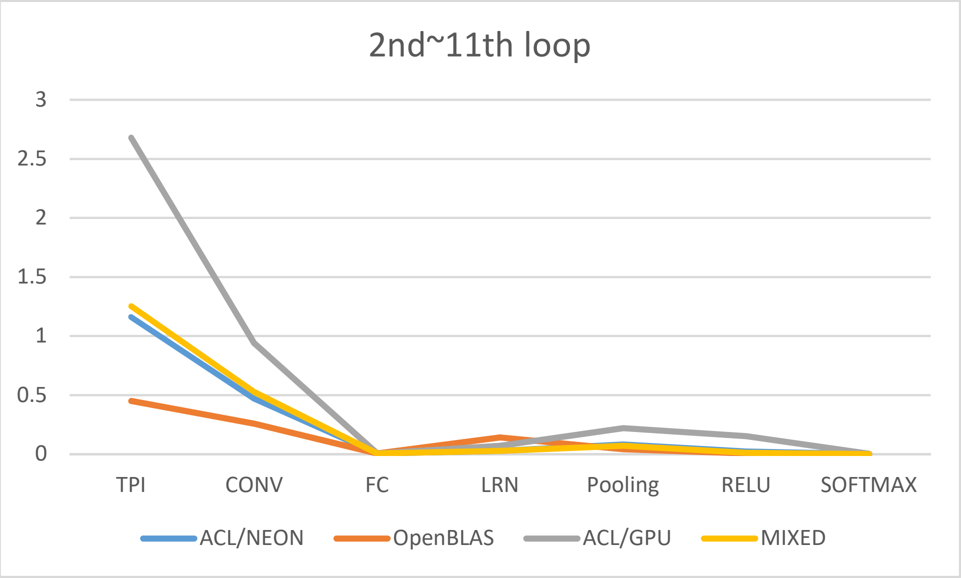
3.2 GoogleNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	3.2005	0.4244	0.7935	0.1407	0.3702
OpenBLAS	0.4519	0	0	0	0
ACL/GPU	11.8248	0.5431	1.1119	3.8728	0.5557

MIXED	2.8757	0.389	0.6906	0.1294	0.3542
Avg. Time					
ACL/NEON	1.1599	0.0335	0.4653	0	0.0545
OpenBLAS	0.4503	0	0	0	0
ACL/GPU	2.6791	0.1404	0.9176	0	0.2364
MIXED	1.2513	0.0654	0.4569	0.0139	0.0738

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	3.2005	1.2879	0.0267	0.0367	0.0726	0.0466	0.0015
OpenBLAS	0.4519	0.2582	0.0047	0.1425	0.0397	0.0067	0.0001
ACL/GPU	11.8248	4.8188	0.2239	0.1926	0.0949	0.3041	0.1068
MIXED	2.8757	1.198	0.0069	0.036	0.0607	0.0107	0.0002
Avg. Time							
ACL/NEON	1.1599	0.469	0.0064	0.0274	0.0814	0.0223	0.0001
OpenBLAS	0.4503	0.2581	0.0045	0.1404	0.0413	0.0059	0.0001
ACL/GPU	2.6791	0.9369	0.0045	0.0701	0.2186	0.1512	0.0035
MIXED	1.2513	0.5268	0.0058	0.0271	0.071	0.0106	0.0001

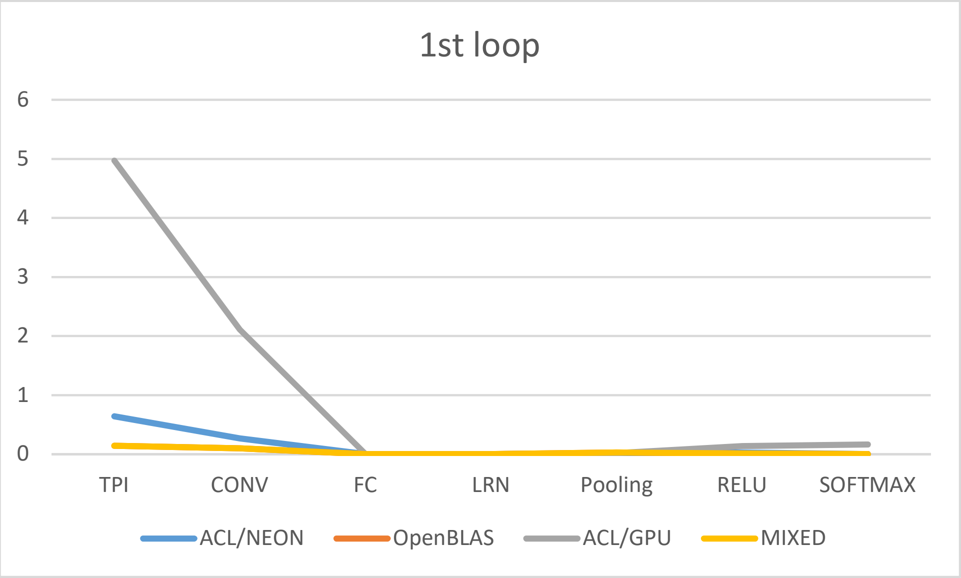


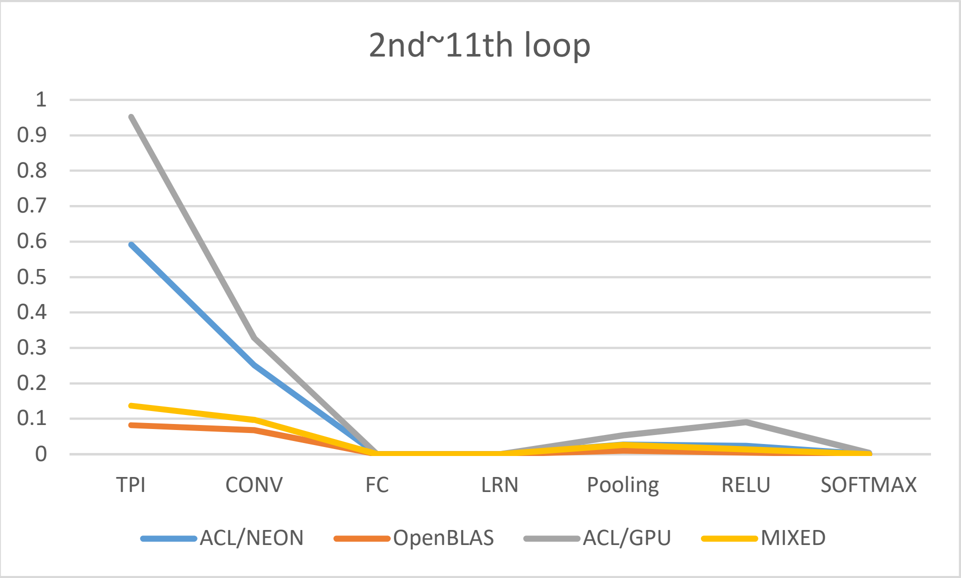


3.3 SqueezeNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.639	0.0757	0.1678	0.0345	0.0574
OpenBLAS	0.1395	0	0	0	0
ACL/GPU	4.9702	0.1578	0.3355	1.8661	0.1819
MIXED	0.1403	0	0	0	0
Avg. Time					
ACL/NEON	0.5913	0.01734	0.2331	0	0.0386
OpenBLAS	0.0818	0	0	0	0
ACL/GPU	0.952	0.0605	0.2961	0	0.1204
MIXED	0.1368	0	0	0	0

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	0.639	0.2646	0	0	0.0103	0.0285	0.0002
OpenBLAS	0.1395	0.0992	0	0	0.0264	0.0138	0.0002
ACL/GPU	4.9702	2.1096	0	0	0.0165	0.1378	0.1651
MIXED	0.1403	0.1	0	0	0.0263	0.0138	0.0002
Avg. Time							
ACL/NEON	0.5913	0.2516	0	0	0.0269	0.0237	0.0001
OpenBLAS	0.0818	0.0679	0	0	0.0096	0.0042	0.0001
ACL/GPU	0.952	0.3275	0	0	0.053	0.0908	0.0038
MIXED	0.1368	0.097	0	0	0.0258	0.0139	0.0001

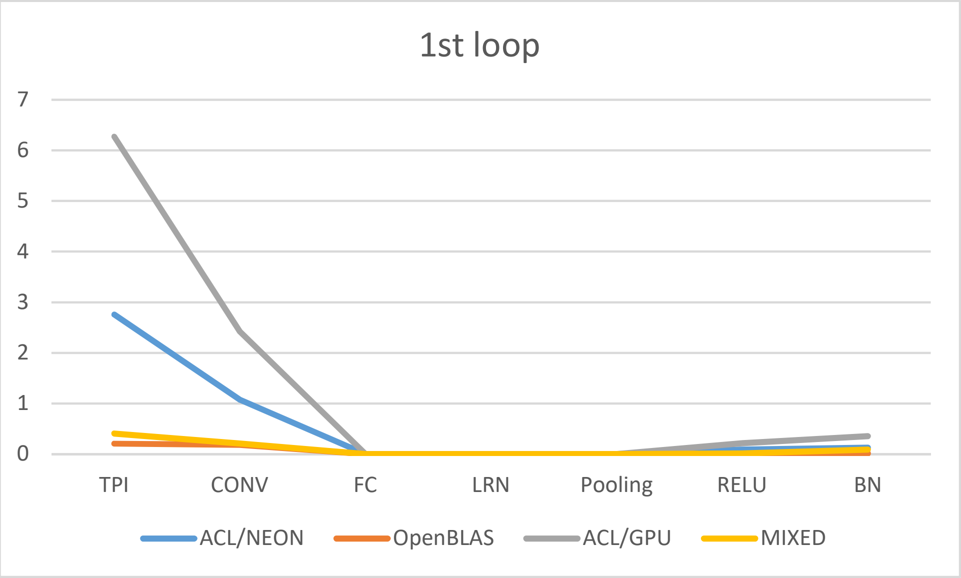


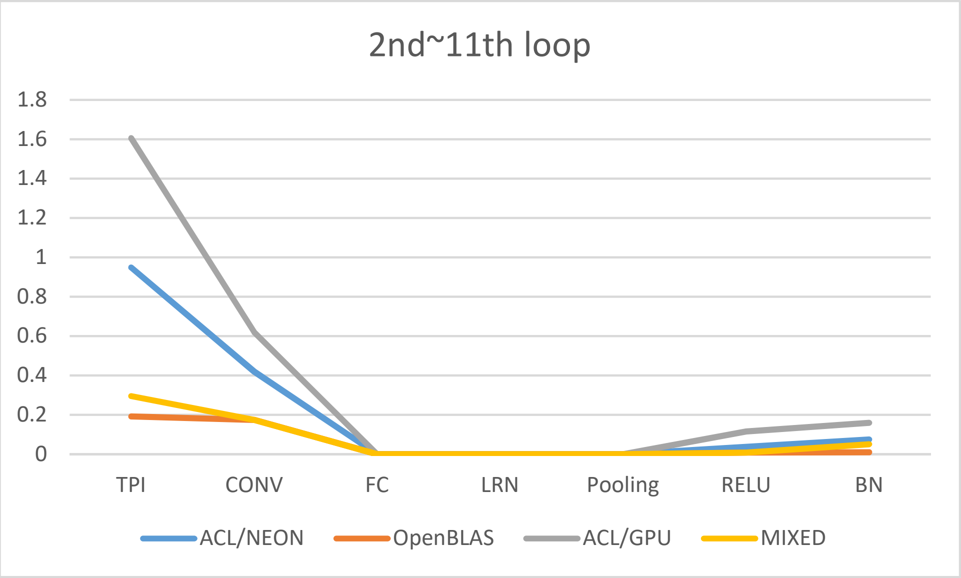


3.4 MobileNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	2.7558	0.4723	0.5227	0.0693	0.4013
OpenBLAS	0.206	0	0	0	0
ACL/GPU	6.2691	0.5955	0.5025	1.6128	0.5653
MIXED	0.4067	0.0419	0.0256	0.0004	0.0297
Avg. Time					
ACL/NEON	0.9484	0.0411	0.2905	0	0.0848
OpenBLAS	0.1913	0	0	0	0
ACL/GPU	1.6058	0.1105	0.3889	0	0.2113
MIXED	0.2955	0.0135	0.023	0	0.0267

	TPI	CONV	FC	LRN	Pooling	RELU	BN
1^{st}							
ACL/NEON	2.7558	1.0747	0	0	0.0004	0.0866	0.1286
OpenBLAS	0.206	0.1859	0	0	0.0001	0.0082	0.0119
ACL/GPU	6.2691	2.4199	0	0	0.0003	0.2174	0.3555
MIXED	0.4067	0.2111	0	0	0.0001	0.0098	0.0881
Avg. Time							
ACL/NEON	0.9484	0.4194	0	0	0.0002	0.0373	0.0752
OpenBLAS	0.1913	0.1733	0	0	0.0001	0.0082	0.0097
ACL/GPU	1.6058	0.6192	0	0	0.0006	0.1155	0.1597
MIXED	0.2955	0.1734	0	0	0.0001	0.0082	0.0507





4. 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.

	AlexNet(s)	GoogleNet(s)	SquezzeNet(s)	MobileNet(s)
FC/ACL/GPU	0.077	0.0045	0	0
FC/ACL/NEON	0.2559	0.0064	0	0
FC/OpenBLAS	0.3381	0.0045	0	0