ECE408 Project Milestone3

Report

YiQing Xie(yiqing.17); Anbang Ye(anbang.17); **Zhouzhenyan Hong(zhouzhenyan.17)**

Team names: Life is short :)

ZJUI Campus students

Instructor: Professor Volodymr Kindratenko

Oct 22, 2020

1 Introduction

This is milestone3 report for ECE408 group project. In this project, we will get practical experience by using GPU framework. Also, we need to demonstrate command of CUDA and optimization approaches by designing and implementing an optimized neural-network convolution layer forward pass.

2 Deliverables

2.1 Show output of rai running your GPU implementation of convolution

We modify the file new-forward.cuh and realize the GPU convolution. Then, we run the m3.cc program with three different data size and the result is shown below.

```
Test batch size: 100
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
Op Time: 216.281 ms
Conv-GPU==
Op Time: 7.65913 ms
Test Accuracy: 0.86
```

Correctness and timing with size 100

```
Test batch size: 1000
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
Op Time: 269.531 ms
Conv-GPU==
Op Time: 58.7204 ms
Test Accuracy: 0.886
```

Correctness and timing with size 1000

```
Test batch size: 10000
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
Op Time: 789.162 ms
Conv-GPU==
Op Time: 8627.08 ms
Test Accuracy: 0.8714
```

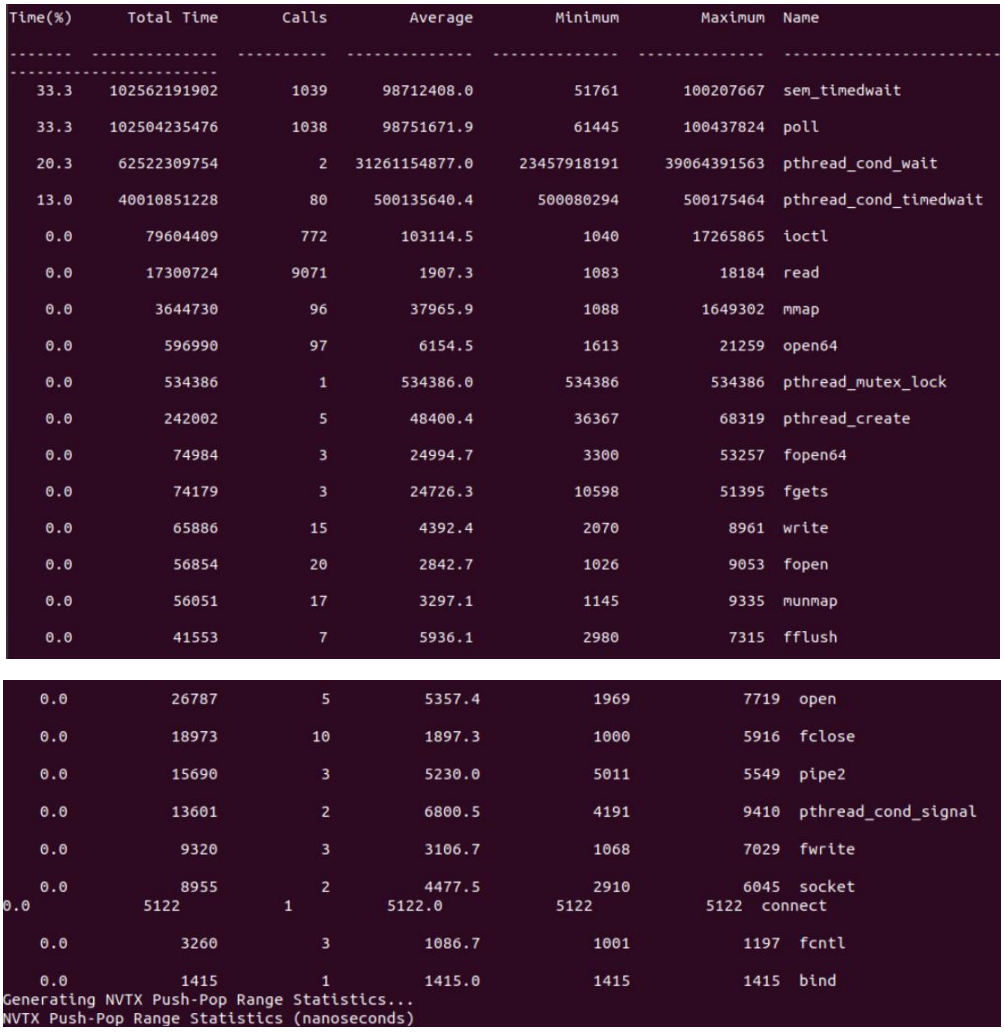
Correctness and timing with size 1000

2.2 Demonstrate nsys profiling the GPU execution

```
Running nv-nsight-cu-cli --section ".*" -o analysis_file ./m3 10000
Test batch size: 10000
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
==PROF== Connected to process 513 (/build/m3)
==PROF== Profiling "conv_forward_kernel" - 1: 0%....50%....100% - 74 passes
Op Time: 12261.9 ms
Conv-GPU==
==PROF== Profiling "conv_forward_kernel" - 2: 0%....50%....100% - 74 passes
Op Time: 37147.6 ms
Test Accuracy: 0.8714
==PROF== Disconnected from process 513
==PROF== Report: /build/analysis_file.ncu-rep
```

```
Exported successfully to
/build/report1.sqlite
Generating CUDA API Statistics...
CUDA API Statistics (nanoseconds)

Time(%)      Total Time       Calls         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ----------------------
   98.3      9229204977           6    1538200829.5           84059      8590131913  cudaMemcpy
    1.7       159211218           6      26535203.0           65732       157179446  cudaMalloc
    0.0         2886897           6        481149.5           53037         1335176  cudaFree
    0.0          252666           2        126333.0           27137          225529  cudaLaunchKernel
Generating CUDA Kernel Statistics...

Generating CUDA Memory Operation Statistics...
CUDA Kernel Statistics (nanoseconds)

Time(%)      Total Time   Instances         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ----------------------
  100.0      8249263968           2    4124631984.0        48257918      8201006050  conv_forward_kernel
```

```
CUDA Memory Operation Statistics (nanoseconds)

Time(%)      Total Time  Operations         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ----------------------
   92.7       905888219           2     452944109.5       388353061       517535158  [CUDA memcpy DtoH]
    7.3        71471909           4      17867977.2            1216        38443629  [CUDA memcpy HtoD]


CUDA Memory Operation Statistics (KiB)

           Total      Operations            Average            Minimum            Maximum  Name
---------------  --------------  -----------------  -----------------  -----------------  ----------------------
       1722500.0               2           861250.0          722500.000          1000000.0  [CUDA memcpy DtoH]
        538919.0               4           134729.0               0.766           288906.0  [CUDA memcpy HtoD]
Generating Operating System Runtime API Statistics...
Operating System Runtime API Statistics (nanoseconds)
```



Time(%)	Total Time	Calls	Average	Minimum	Maximum	Name
33.3	102562191902	1039	98712408.0	51761	100207667	sem_timedwait
33.3	102504235476	1038	98751671.9	61445	100437824	poll
20.3	62522309754	2	31261154877.0	23457918191	39064391563	pthread_cond_wait
13.0	40010851228	80	500135640.4	500080294	500175464	pthread_cond_timedwait
0.0	79604409	772	103114.5	1040	17265865	ioctl
0.0	17300724	9071	1907.3	1083	18184	read
0.0	3644730	96	37965.9	1088	1649302	mmap
0.0	596990	97	6154.5	1613	21259	open64
0.0	534386	1	534386.0	534386	534386	pthread_mutex_lock
0.0	242002	5	48400.4	36367	68319	pthread_create
0.0	74984	3	24994.7	3300	53257	fopen64
0.0	74179	3	24726.3	10598	51395	fgets
0.0	65886	15	4392.4	2070	8961	write
0.0	56854	20	2842.7	1026	9053	fopen
0.0	56051	17	3297.1	1145	9335	munmap
0.0	41553	7	5936.1	2980	7315	fflush
0.0	26787	5	5357.4	1969	7719	open
0.0	18973	10	1897.3	1000	5916	fclose
0.0	15690	3	5230.0	5011	5549	pipe2
0.0	13601	2	6800.5	4191	9410	pthread_cond_signal
0.0	9320	3	3106.7	1068	7029	fwrite
0.0	8955	2	4477.5	2910	6045	socket
0.0	5122	1	5122.0	5122	5122	connect
0.0	3260	3	1086.7	1001	1197	fcntl
0.0	1415	1	1415.0	1415	1415	bind

Generating NVTX Push-Pop Range Statistics...
NVTX Push-Pop Range Statistics (nanoseconds)

Output with size 10000

```
✱ Running nv-nsight-cu-cli --section ".*" -o analysis_file ./m3 1000
Test batch size: 1000
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
==PROF== Connected to process 514 (/build/m3)
==PROF== Profiling "conv_forward_kernel" - 1: 0%....50%....100% - 74 passes
Op Time: 2528.14 ms
Conv-GPU==
==PROF== Profiling "conv_forward_kernel" - 2: 0%....50%....100% - 74 passes
Op Time: 4497.73 ms
Test Accuracy: 0.886
==PROF== Disconnected from process 514
==PROF== Report: /build/analysis_file.ncu-rep
```

```
Exported successfully to
/build/report1.sqlite
Generating CUDA API Statistics...
CUDA API Statistics (nanoseconds)

Time(%)      Total Time       Calls         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  --------------------
   60.2       181122634           6      30187105.7           67942       180318465  cudaMalloc
   39.5       118950107           6      19825017.8           77292        57188126  cudaMemcpy
    0.2          725978           6        120996.3           51446          257111  cudaFree
    0.0           51397           2         25698.5           25544           25853  cudaLaunchKernel
Generating CUDA Kernel Statistics...

Generating CUDA Memory Operation Statistics...
CUDA Kernel Statistics (nanoseconds)

Time(%)      Total Time   Instances         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  --------------------
  100.0        21354331           2      10677165.5         4829147        16525184  conv_forward_kernel
```

```
CUDA Memory Operation Statistics (nanoseconds)

Time(%)      Total Time  Operations         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ------------------------
   92.5        88504561           2      44252280.5        36964193        51540368  [CUDA memcpy DtoH]
    7.5         7138537           4       1784634.2            1216         3814019  [CUDA memcpy HtoD]


CUDA Memory Operation Statistics (KiB)

            Total      Operations            Average            Minimum            Maximum  Name
----------------  --------------  -----------------  -----------------  -----------------  ----------------------------
        172250.0               2            86125.0          72250.000           100000.0  [CUDA memcpy DtoH]
         53903.0               4            13475.0              0.766            28890.0  [CUDA memcpy HtoD]
Generating Operating System Runtime API Statistics...
Operating System Runtime API Statistics (nanoseconds)
```

Time(%)	Total Time	Calls	Average	Minimum	Maximum	Name
33.9	9513811752	109	87282676.6	50661	100155611	sem_timedwait
33.6	9417988021	109	86403559.8	57924	100262508	poll
32.1	9002533930	18	500140773.9	500103376	500149439	pthread_cond_timedwait
0.3	89916595	760	118311.3	1068	17227727	ioctl
0.0	3663950	97	37772.7	1006	1670630	mmap
0.0	1843059	944	1952.4	1313	12592	read
0.0	621476	97	6407.0	2571	21812	open64
0.0	234615	5	46923.0	32761	62611	pthread_create
0.0	72497	3	24165.7	11453	49394	fgets
0.0	69327	15	4621.8	2143	10100	write
0.0	67453	21	3212.0	1030	9809	fopen
0.0	59087	19	3109.8	1283	8227	munmap
0.0	46066	3	15355.3	2896	31619	fopen64
0.0	39020	7	5574.3	3133	8294	fflush
0.0	28317	5	5663.4	4328	7616	open
0.0	14793	3	4931.0	4598	5230	pipe2

0.0	69327	15	4621.8	2143	10100	write
0.0	67453	21	3212.0	1030	9809	fopen
0.0	59087	19	3109.8	1283	8227	munmap
0.0	46066	3	15355.3	2896	31619	fopen64
0.0	39020	7	5574.3	3133	8294	fflush
0.0	28317	5	5663.4	4328	7616	open
0.0	14793	3	4931.0	4598	5230	pipe2
0.0	14475	8	1809.4	1202	3635	fclose
0.0	9731	2	4865.5	3782	5949	socket
0.0	6385	2	3192.5	1042	5343	fwrite
0.0	5879	1	5879.0	5879	5879	connect
0.0	2340	2	1170.0	1084	1256	fcntl
0.0	1458	1	1458.0	1458	1458	bind

Generating NVTX Push-Pop Range Statistics...
NVTX Push-Pop Range Statistics (nanoseconds)

Size 1000

```
Running nv-nsight-cu-cli --section ".*" -o analysis_file ./m3 100
Test batch size: 100
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
==PROF== Connected to process 513 (/build/m3)
==PROF== Profiling "conv_forward_kernel" - 1: 0%....50%....100% - 74 passes
Op Time: 1271.68 ms
Conv-GPU==
==PROF== Profiling "conv_forward_kernel" - 2: 0%....50%....100% - 74 passes
Op Time: 1008.52 ms
Test Accuracy: 0.86
==PROF== Disconnected from process 513
==PROF== Report: /build/analysis_file.ncu-rep
```

```
Exported successfully to
/build/report1.sqlite
Generating CUDA API Statistics...
CUDA API Statistics (nanoseconds)

Time(%)      Total Time       Calls         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ----------------------
   92.6       183326121           6      30554353.5           69011       182690384  cudaMalloc
    7.1        14069887           6       2344981.2           67480         6532096  cudaMemcpy
    0.3          575101           6         95850.2           55822          173829  cudaFree
    0.0           62665           2         31332.5           29465           33200  cudaLaunchKernel
Generating CUDA Kernel Statistics...

Generating CUDA Memory Operation Statistics...
CUDA Kernel Statistics (nanoseconds)

Time(%)      Total Time   Instances         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ----------------------
  100.0         2032337           2       1016168.5          505244         1527093  conv_forward_kernel
```

```
CUDA Memory Operation Statistics (nanoseconds)

Time(%)      Total Time  Operations         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ------------------------
   92.9         9242524           2       4621262.0         4222081         5020443  [CUDA memcpy DtoH]
    7.1          710747           4        177686.7            2400          428189  [CUDA memcpy HtoD]


CUDA Memory Operation Statistics (KiB)

           Total      Operations            Average            Minimum            Maximum  Name
---------------  --------------  -----------------  -----------------  -----------------  ----------------------
         17225.0               2             8612.0           7225.000            10000.0  [CUDA memcpy DtoH]
          5402.0               4             1350.0              0.766             2889.0  [CUDA memcpy HtoD]
Generating Operating System Runtime API Statistics...
Operating System Runtime API Statistics (nanoseconds)
```

```
Time(%)      Total Time       Calls         Average         Minimum         Maximum  Name
-------  --------------  ----------  --------------  --------------  --------------  ----------------------
   35.1      1186052353          26      45617398.2           49827       100155131  sem_timedwait
   32.6      1101538463          26      42366864.0           52180       100215905  poll
   29.7      1001572524           2     500786262.0       500094722       501477802  pthread_cond_timedwait
    2.5        82898233         760        109076.6            1127        20215064  ioctl
    0.1         4095511          94         43569.3            1013         2016488  mmap
    0.0          653532          97          6737.4            1687           23251  open64
    0.0          356770         131          2723.4            1401           18849  read
    0.0          239386           5         47877.2           33752           59759  pthread_create
    0.0           78452          15          5230.1            2029           12206  write
    0.0           75030           3         25010.0           11623           50097  fgets
    0.0           65552          21          3121.5            1047           12323  fopen
    0.0           55955           3         18651.7            3123           36824  fopen64
    0.0           36519           7          5217.0            3040            7484  fflush
    0.0           34563          10          3456.3            1390            7489  munmap
    0.0           29596           5          5919.2            1983            8629  open
    0.0           17757           9          1973.0            1021            4995  fclose
    0.0           15992           3          5330.7            3920            7363  pipe2
```

```
    0.0           8288          2         4144.0           3493           4795  socket
    0.0           6331          3         2110.3           1066           3460  fwrite
    0.0           6236          1         6236.0           6236           6236  connect
    0.0           1322          1         1322.0           1322           1322  bind
    0.0           1081          1         1081.0           1081           1081  putc
0.0           1066          1         1066.0           1066           1066  fcntl
Generating NVTX Push-Pop Range Statistics...
NVTX Push-Pop Range Statistics (nanoseconds)
```

Size 100

2.3 Include a list of all kernels that collectively consume more than 90% of the program time.

conv_forward_kernel

2.4 Include a list of all CUDA API calls that collectively consume more than 90% of the program time.

cudaMemcpy, cudaMalloc, cudaFree

2.5 Include an explanation of the difference between kernels and API calls

The API calls part is the section contains the time of CPU using, while the profiling result is the time that the GPUs taking. The total time of the API call is from the moment it is launched to the moment it completes, so will overlap with executing kernels.

2.6 Screenshot of the GPU SOL utilization in Nsight-Compute GUI for your kernel profiling data

Speed Of Light [%]

SOL SM Breakdown		SOL Memory Breakdown	
SOL SM: Issue Active [%]	27.66	SOL L1: Data Pipe Lsu Wavefronts [%]	30.63
SOL SM: Inst Executed [%]	27.65	SOL L1: Lsu Writeback Active [%]	24.17
SOL SM: Inst Executed Pipe Lsu [%]	21.83	SOL L1: Lsuin Requests [%]	21.83
SOL SM: Pipe Fma Cycles Active [%]	19.91	SOL L1: Data Bank Reads [%]	3.73
SOL SM: Pipe Alu Cycles Active [%]	17.75	SOL L2: T Sectors [%]	2.63
SOL SM: Mio2rf Writeback Active [%]	11.39	SOL L2: Xbar2lts Cycles Active [%]	2.31
SOL SM: Mio Inst Issued [%]	10.91	SOL L2: Lts2xbar Cycles Active [%]	2.24
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	7.17	SOL L2: T Tag Requests [%]	1.95
SOL SM: Mio Pq Read Cycles Active [%]	0.06	SOL L1: M L1tex2xbar Req Cycles Active [%]	1.32
SOL SM: Mio Pq Write Cycles Active [%]	0.05	SOL L1: M Xbar2l1tex Read Sectors [%]	1.28
SOL SM: Inst Executed Pipe Adu [%]	0.00	SOL GPU: Dram Throughput [%]	0.91
SOL SM: Inst Executed Pipe Xu [%]	0	SOL L2: D Sectors [%]	0.74
SOL IDC: Request Cycles Active [%]	0	SOL L1: Data Bank Writes [%]	0.18
SOL SM: Inst Executed Pipe Tex [%]	0	SOL L2: D Sectors Fill Device [%]	0.09
SOL SM: Inst Executed Pipe Ipa [%]	0	SOL L1: F Wavefronts [%]	0.00
SOL SM: Inst Executed Pipe Fp16 [%]	0	SOL L1: Texin Sm2tex Req Cycles Active [%]	0.00
SOL SM: Pipe Fp64 Cycles Active [%]	0	SOL L1: Data Pipe Tex Wavefronts [%]	0
SOL SM: Pipe Shared Cycles Active [%]	0	SOL L1: Tex Writeback Active [%]	0
SOL SM: Pipe Tensor Cycles Active [%]	0	SOL L2: D Atomic Input Cycles Active [%]	0
		SOL L2: D Sectors Fill Sysmem [%]	0

Speed Of Light [%]

SOL SM Breakdown	
SOL SM: Issue Active [%]	23.84
SOL SM: Inst Executed [%]	23.84
SOL SM: Inst Executed Pipe Lsu [%]	18.69
SOL SM: Pipe Fma Cycles Active [%]	17.13
SOL SM: Pipe Alu Cycles Active [%]	15.32
SOL SM: Mio2rf Writeback Active [%]	10.47
SOL SM: Mio Inst Issued [%]	9.35
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	6.06
SOL SM: Mio Pq Read Cycles Active [%]	0.20
SOL SM: Mio Pq Write Cycles Active [%]	0.17
SOL SM: Inst Executed Pipe Adu [%]	0.00
SOL SM: Inst Executed Pipe Xu [%]	0
SOL IDC: Request Cycles Active [%]	0
SOL SM: Inst Executed Pipe Tex [%]	0
SOL SM: Inst Executed Pipe Ipa [%]	0
SOL SM: Inst Executed Pipe Fp16 [%]	0
SOL SM: Pipe Fp64 Cycles Active [%]	0
SOL SM: Pipe Shared Cycles Active [%]	0
SOL SM: Pipe Tensor Cycles Active [%]	0

SOL Memory Breakdown	
SOL L1: Data Pipe Lsu Wavefronts [%]	30.92
SOL L1: Lsu Writeback Active [%]	25.85
SOL L1: Lsuin Requests [%]	18.69
SOL L1: Data Bank Reads [%]	4.46
SOL GPU: Dram Throughput [%]	4.16
SOL L2: T Sectors [%]	3.98
SOL L2: Xbar2lts Cycles Active [%]	2.94
SOL L2: Lts2xbar Cycles Active [%]	2.76
SOL L2: T Tag Requests [%]	2.33
SOL L1: M L1tex2xbar Req Cycles Active [%]	1.68
SOL L1: M Xbar2l1tex Read Sectors [%]	1.58
SOL L2: D Sectors [%]	1.38
SOL L2: D Sectors Fill Device [%]	0.35
SOL L1: Data Bank Writes [%]	0.28
SOL L1: F Wavefronts [%]	0.00
SOL L1: Texin Sm2tex Req Cycles Active [%]	0.00
SOL L1: Tex Writeback Active [%]	0
SOL L2: D Atomic Input Cycles Active [%]	0
SOL L2: D Sectors Fill Sysmem [%]	0
SOL L1: Data Pipe Tex Wavefronts [%]	0

ECE408 Project Milestone2

Report

YiQing Xie(yiqing.17); Anbang Ye(anbang.17); **Zhouzhenyan Hong(zhouzhenyan.17)**

Team names: Life is short :)

ZJUI Campus students

Instructor: Professor Volodymr Kindratenko

Oct 14, 2020

1. Optimization Approach and Results

1. Output of rai running Mini-DNN on the CPU:

Size 100:

✱ Running /bin/bash -c "time ./m2 100"

Test batch size: 100

Loading fashion-mnist data...Done

Loading model...Done

Conv-CPU==

Op Time: 829.391 ms

Conv-CPU==

Op Time: 2410.14 ms

Test Accuracy: 0.86

real

0m4.207s

user

0m4.195s

sys

0m0.012s

Size: 1000

✱ Running /bin/bash -c "time ./m2 1000"

Test batch size: 1000

Loading fashion-mnist data...Done

Loading model...Done

Conv-CPU==

Op Time: 8319.1 ms

Conv-CPU==

Op Time: 24160.7 ms

Test Accuracy: 0.886

real

0m41.883s

user

0m41.775s

sys

0m0.108s

.

Size: 10000

✱ Running /bin/bash -c "time ./m2 10000"Test batch size: 10000

Loading fashion-mnist data...Done

Loading model...Done

Conv-CPU==

Op Time: 91014.9 ms

Conv-CPU==

Op Time: 263174 ms

Test Accuracy: 0.8714

real

7m29.176s

user

7m28.252s

sys

0m0.924s

2. List Op Times

Size 100:

Conv-CPU==

Op Time: 829.391 ms

Conv-CPU==

Op Time: 2410.14 ms

Size 1000:

Conv-CPU==

Op Time: 8319.1 ms

Conv-CPU==

Op Time: 24160.7 ms

Size 10000:

Conv-CPU==

Op Time: 91014.9 ms

Conv-CPU==

Op Time: 263174 ms

3. List whole program execution time

Size 100: 0m4.207s

Size 1000: 0m41.883s

Size 10000: 7m29.176s