2D Convolution using CUDA

1. 2D Convolution

Input Matrix

1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8

Kernel

1	1	1
1	1	1
1	1	1

Input Matrix 1st Element * Kernel

0^{*1}	0^{*1}	0^{*1}	0	0	0	0	0	0	0
0^{*1}	1^{*1}	1^{*1}	1	1	1	1	1	1	0
0^{*1}	2^{*1}	2^{*1}	2	2	2	2	2	2	0
0	3	3	3	3	3	3	3	3	0
0	4	4	4	4	4	4	4	4	0
0	5	5	5	5	5	5	5	5	0
0	6	6	6	6	6	6	6	6	0
0	7	7	7	7	7	7	7	7	0
0	8	8	8	8	8	8	8	8	0
0	0	0	0	0	0	0	0	0	0

Output

6							

Input Matrix 2nd Element * Kernel

0	0^{*1}	0^{*1}	0^{*1}	0	0	0	0	0	0
0	1^{*1}	1^{*1}	1^{*1}	1	1	1	1	1	0
0	2^{*1}	2^{*1}	2^{*1}	2	2	2	2	2	0
0	3	3	3	3	3	3	3	3	0
0	4	4	4	4	4	4	4	4	0
0	5	5	5	5	5	5	5	5	0
0	6	6	6	6	6	6	6	6	0
0	7	7	7	7	7	7	7	7	0
0	8	8	8	8	8	8	8	8	0
0	0	0	0	0	0	0	0	0	0

Output

	9						

Final Output

6	**9**	**9**	**9**	**9**	**9**	**9**	**6**
12	**18**	**18**	**18**	**18**	**18**	**18**	**12**
18	**27**	**27**	**27**	**27**	**27**	**27**	**18**
24	**36**	**36**	**36**	**36**	**36**	**36**	**24**
30	**45**	**45**	**45**	**45**	**45**	**45**	**30**
36	**54**	**54**	**54**	**54**	**54**	**54**	**36**
42	**63**	**63**	**63**	**63**	**63**	**63**	**42**
30	**45**	**45**	**45**	**45**	**45**	**45**	**30**

The final output and input size are same due to zero padding.

Details of data used for profiling:

Size of input matrix	: (16000, 16000)
Size of kernel	: (5,5)

2. CPU

CPU	: Intel® Core™ i7-4610M @ 3.00 GHz
Cores	: 2
Logical Processors	: 4

Average time taken to run 2D convolution on CPU : **7.5 seconds**

3. GPU

GPU	: Nvidia Quadro K1100M
CUDA Capability Version	: 3.0
Number of Streaming Multiprocessors(SM)	: 2
Number of CUDA Cores/SM	: 192
Shared Memory per block	: 48KBytes
Warp Size	: 32
Max number of threads per SM	: 2048
Max number of threads per block	: 1024
Max dimension size of a thread block (x,y,z)	: (1024, 1024, 64)
Max dimension size of a grid size (x,y,z)	: (2147483647, 65535, 65535)

With the above limitations, the grid dimension used is (500x500) blocks and each block containing (32x32) treads. This implies, each SM has to run 125000 blocks. Shared memory and thread synchronization was used within a block to avoid redundant copy of data from global memory and thereby reducing memory fetching time too.

Average time taken to run 2D convolution on GPU : **0.348 seconds**

4. Performance Comparison

Timing(s)

CPU	GPU
7.5	**0.348**

The mentioned GPU ran the convolution code ~21 times faster than CPU.