Faculty of Engineering
Cairo University

Parallel Computing

#CMP4011

Lab 4 - Convolution

Report

Submitted to:

Eng. Mohamed Abdullah

Submitted By:

NAME	SEC	BN	ID
Yasmine Ashraf Ghanem	2	37	9203707
Yasmin Abdullah Nasser	2	38	9203717

Kernel 1 : basic implementation (no tiling)

a. *Small Images (5x 1000x500 images)*

GPU Activities : (Time) in tabular form:

Mask Size/ BlockSize	4x4	16x16	32x32	64x64
3x3	7.82 ms	2.55 ms	2.64 ms	-
9x9	38.3 ms	30.284 ms	29.50 ms	-

All GPU Activities:

i. 3x3 Mask

1. Block Size = 4x4

```
 GPU activities:   58.39%  12.259ms         4  3.0649ms     640ns  4.4039ms  [CUDA memcpy HtoD]
                   37.27%  7.8248ms         1  7.8248ms  7.8248ms  7.8248ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                    4.34%  911.20us         1  911.20us  911.20us  911.20us  [CUDA memcpy DtoH]
```

2. Block Size = 16x16

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   76.84%  11.504ms         4  2.8760ms     672ns  3.8434ms  [CUDA memcpy HtoD]
                   17.06%  2.5535ms         1  2.5535ms  2.5535ms  2.5535ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                    6.10%  913.92us         1  913.92us  913.92us  913.92us  [CUDA memcpy DtoH]
```

3. Block Size = 32x32

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   76.57%  11.635ms         4  2.9088ms     672ns  3.9030ms  [CUDA memcpy HtoD]
                   17.41%  2.6456ms         1  2.6456ms  2.6456ms  2.6456ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                    6.02%  914.43us         1  914.43us  914.43us  914.43us  [CUDA memcpy DtoH]
```

4. Block Size = 64x64 :
   Error => Exceeded the threads/block capacity

ii. *9x9 Mask*

1. Block Size = 4x4

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   77.42%  83.362ms         2  41.681ms  33.333ms  50.029ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                   21.17%  22.797ms         6  3.7995ms     673ns  7.4496ms  [CUDA memcpy HtoD]
                    1.41%  1.5171ms         2  758.53us  605.25us  911.81us  [CUDA memcpy DtoH]
```

2. Block Size = 16x16

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   58.05%  30.284ms         2  15.142ms  12.114ms  18.170ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                   39.04%  20.366ms         6  3.3944ms     640ns  4.7379ms  [CUDA memcpy HtoD]
                    2.91%  1.5186ms         2  759.31us  605.44us  913.18us  [CUDA memcpy DtoH]
```

3. Block Size = 32x32

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   55.63%  29.503ms         2  14.751ms  11.807ms  17.696ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                   41.49%  22.008ms         6  3.6680ms     671ns  6.6455ms  [CUDA memcpy HtoD]
                    2.88%  1.5270ms         2  763.49us  615.90us  911.07us  [CUDA memcpy DtoH]
```

b. Large Images (4000x3000)

GPU Activities Time in tabular form:

Mask Size/ BlockSize	4x4	16x16	32x32	64x64
3x3	493.6 ms	165.6 ms	171.4 ms	-
9x9	4.23 s	1.76 s	1.74 s	-

All GPU Activities:

i. 3x3 Mask

1. Block Size = 4x4

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   56.54%  713.48ms         6  118.91ms     640ns  237.74ms  [CUDA memcpy HtoD]
                   39.12%  493.59ms         2  246.79ms  190.49ms  303.10ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                    4.34%  54.736ms         2  27.368ms  25.669ms  29.067ms  [CUDA memcpy DtoH]
```

2. Block Size = 16x16

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   78.73%  819.74ms         6  136.62ms     640ns  336.55ms  [CUDA memcpy HtoD]
                   15.91%  165.64ms         2  82.820ms  61.352ms  104.29ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                    5.36%  55.837ms         2  27.919ms  26.088ms  29.749ms  [CUDA memcpy DtoH]
```

3. Block Size = 32x32

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   76.08%  714.91ms         6  119.15ms     672ns  239.76ms  [CUDA memcpy HtoD]
                   18.24%  171.39ms         2  85.695ms  61.743ms  109.65ms  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                    5.68%  53.338ms         2  26.669ms  24.438ms  28.900ms  [CUDA memcpy DtoH]
```

ii. 9x9 Mask

1. Block Size = 4x4

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   79.72%  4.23300s         2  2.11650s  1.89746s  2.33554s  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                   18.88%  1.00235s         6  167.06ms     673ns  370.35ms  [CUDA memcpy HtoD]
                    1.40%  74.261ms         2  37.131ms  24.806ms  49.455ms  [CUDA memcpy DtoH]
```

2. Block Size = 16x16

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   71.23%  1.75923s         2  879.62ms  707.82ms  1.05141s  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                   26.62%  657.51ms         6  109.58ms     704ns  186.09ms  [CUDA memcpy HtoD]
                    2.14%  52.926ms         2  26.463ms  24.145ms  28.781ms  [CUDA memcpy DtoH]
```

3. Block Size = 32x32

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   68.12%  1.73807s         2  869.04ms  659.93ms  1.07814s  kernel1_batch(unsigned char*, float*, int, int, int, int, int)
                   29.84%  761.36ms         6  126.89ms     672ns  260.99ms  [CUDA memcpy HtoD]
                    2.04%  51.976ms         2  25.988ms  23.187ms  28.789ms  [CUDA memcpy DtoH]
```

Kernel 2 : tiling where each block matches the input tile size.

We change the size of the output tile and calculate the input tile:

INPUT_TILE_WIDTH = OUTPUT_TILE_WIDTH + MASK_WIDTH - 1

a. *Small Images (1000x500)*

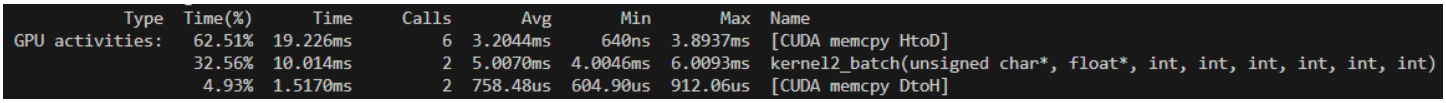
GPU Activities : (Time) in tabular form:

Mask Size/ OutputTile	4x4	16x16	32x32	64x64
3x3	10.01 ms	3.11 ms	-	-
9x9	45.39	16.87 ms	-	-

All GPU Activities:

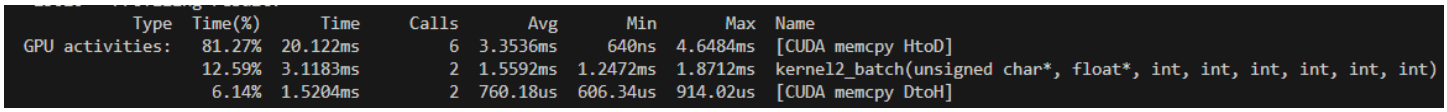
i. *3x3 Mask*

1. Output Tile Size = 4x4 | Input Tile Size (Block Size) = 6x6



Type	Time(%)	Time	Calls	Avg	Min	Max	Name
GPU activities:	62.51%	19.226ms	6	3.2044ms	640ns	3.8937ms	[CUDA memcpy HtoD]
	32.56%	10.014ms	2	5.0070ms	4.0046ms	6.0093ms	kernel2_batch(unsigned char*, float*, int, int, int, int, int, int)
	4.93%	1.5170ms	2	758.48us	604.90us	912.06us	[CUDA memcpy DtoH]

2. Output Tile Size = 16x16 | Input Tile Size (Block Size) = 18x18



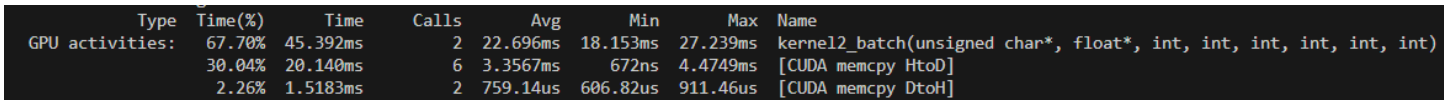
Type	Time(%)	Time	Calls	Avg	Min	Max	Name
GPU activities:	81.27%	20.122ms	6	3.3536ms	640ns	4.6484ms	[CUDA memcpy HtoD]
	12.59%	3.1183ms	2	1.5592ms	1.2472ms	1.8712ms	kernel2_batch(unsigned char*, float*, int, int, int, int, int, int)
	6.14%	1.5204ms	2	760.18us	606.34us	914.02us	[CUDA memcpy DtoH]

3. Output Tile Size = 32x32 | Input Tile Size (Block Size) = 34x34

Exceeded maximum number of threads/block allowed

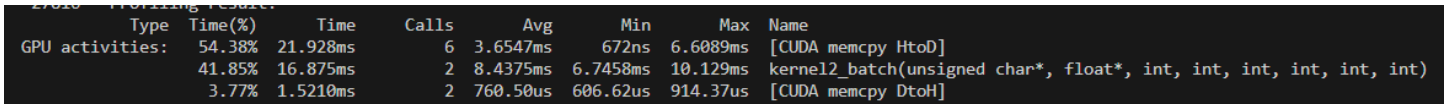
ii. 9x9 Mask

1. Output Tile Size = 4x4 | Input Tile Size (Block Size) = 12x12



Type	Time(%)	Time	Calls	Avg	Min	Max	Name
GPU activities:	67.70%	45.392ms	2	22.696ms	18.153ms	27.239ms	kernel2_batch(unsigned char*, float*, int, int, int, int, int, int)
	30.04%	20.140ms	6	3.3567ms	672ns	4.4749ms	[CUDA memcpy HtoD]
	2.26%	1.5183ms	2	759.14us	606.82us	911.46us	[CUDA memcpy DtoH]

2. Output Tile Size = 16x16 | Input Tile Size (Block Size) = 24x24



Type	Time(%)	Time	Calls	Avg	Min	Max	Name
GPU activities:	54.38%	21.928ms	6	3.6547ms	672ns	6.6089ms	[CUDA memcpy HtoD]
	41.85%	16.875ms	2	8.4375ms	6.7458ms	10.129ms	kernel2_batch(unsigned char*, float*, int, int, int, int, int, int)
	3.77%	1.5210ms	2	760.50us	606.62us	914.37us	[CUDA memcpy DtoH]

3. Output Tile Size = 32x32 | Input Tile Size (Block Size) = 40x40

Exceeded maximum number of threads/block allowed

b. *Large Images: (4000x3000)*

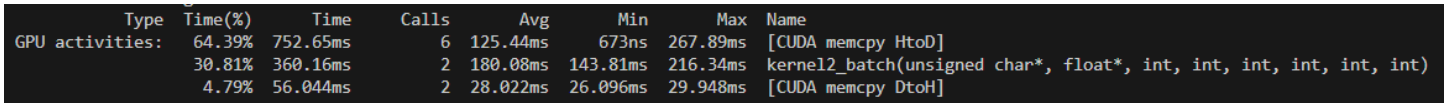
GPU Activities : (Time) in tabular form:

Mask Size/ OutputTile	4x4	16x16	32x32	64x64
3x3	360.16 ms	136.79 ms	-	-
9x9	1.68 s	589.96 ms	-	-

All GPU Activities:

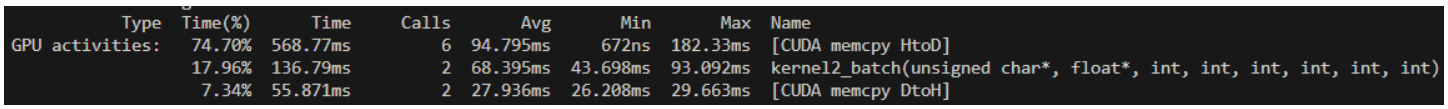
i. 3x3 Mask

1. Output Tile Size = 4x4 | Input Tile Size (Block Size) = 6x6



```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   64.39%  752.65ms         6  125.44ms     673ns  267.89ms  [CUDA memcpy HtoD]
                   30.81%  360.16ms         2  180.08ms  143.81ms  216.34ms  kernel2_batch(unsigned char*, float*, int, int, int, int, int, int)
                    4.79%  56.044ms         2  28.022ms  26.096ms  29.948ms  [CUDA memcpy DtoH]
```

2. Output Tile Size = 16x16 | Input Tile Size (Block Size) = 18x18



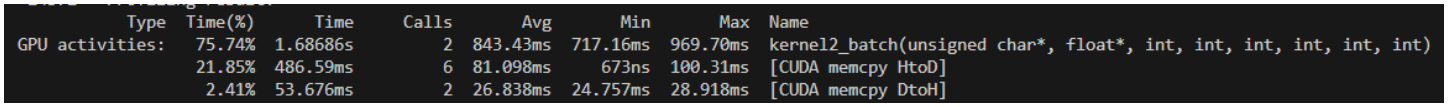
```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   74.70%  568.77ms         6  94.795ms     672ns  182.33ms  [CUDA memcpy HtoD]
                   17.96%  136.79ms         2  68.395ms  43.698ms  93.092ms  kernel2_batch(unsigned char*, float*, int, int, int, int, int, int)
                    7.34%  55.871ms         2  27.936ms  26.208ms  29.663ms  [CUDA memcpy DtoH]
```

3. Output Tile Size = 32x32 | Input Tile Size (Block Size) = 34x34

<u>Exceeded maximum number of threads/block allowed</u>

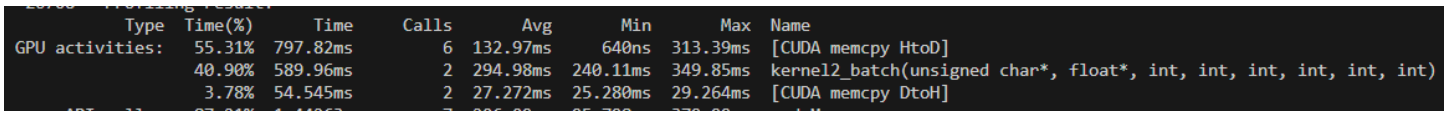
ii. 9x9 Mask

1. Output Tile Size = 4x4 | Input Tile Size (Block Size) = 12x12



```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   75.74%  1.68686s         2  843.43ms  717.16ms  969.70ms  kernel2_batch(unsigned char*, float*, int, int, int, int, int, int)
                   21.85%  486.59ms         6  81.098ms     673ns  100.31ms  [CUDA memcpy HtoD]
                    2.41%  53.676ms         2  26.838ms  24.757ms  28.918ms  [CUDA memcpy DtoH]
```

2. Output Tile Size = 16x16 | Input Tile Size (Block Size) = 24x24



```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   55.31%  797.82ms         6  132.97ms     640ns  313.39ms  [CUDA memcpy HtoD]
                   40.90%  589.96ms         2  294.98ms  240.11ms  349.85ms  kernel2_batch(unsigned char*, float*, int, int, int, int, int, int)
                    3.78%  54.545ms         2  27.272ms  25.280ms  29.264ms  [CUDA memcpy DtoH]
```

3. Output Tile Size = 32x32 | Input Tile Size (Block Size) = 40x40

<u>Exceeded maximum number of threads/block allowed</u>

Kernel 2 : tiling where each block matches the input tile size.

We change the size of the output tile and calculate the input tile:

INPUT_TILE_WIDTH = OUTPUT_TILE_WIDTH + MASK_WIDTH - 1

a. *Small Images (1000x500)*

GPU Activities : (Time) in tabular form:

Mask Size/ OutputTile	4x4	16x16	32x32	64x64
3x3	7.93 ms	3.07 ms	3.32 ms	-
9x9	28.04 ms	11.22 ms	11.20 ms	-

All GPU Activities:

i. 3x3 Mask

1. Output Tile Size (Block Size) = 4x4 | Input Tile Size = 6x6

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   67.31%  19.468ms         6  3.2447ms     672ns  4.0542ms  [CUDA memcpy HtoD]
                   27.43%  7.9334ms         2  3.9667ms  3.1755ms  4.7579ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    5.26%  1.5207ms         2  760.35us  606.50us  914.21us  [CUDA memcpy DtoH]
```

2. Output Tile Size (Block Size) = 16x16 | Input Tile Size = 18x18

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   80.67%  19.148ms         6  3.1913ms     672ns  3.8301ms  [CUDA memcpy HtoD]
                   12.92%  3.0668ms         2  1.5334ms  1.2264ms  1.8404ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    6.41%  1.5207ms         2  760.37us  606.56us  914.18us  [CUDA memcpy DtoH]
```

3. Output Tile Size (Block Size) = 32x32 | Input Tile Size = 34x34

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   81.58%  21.323ms         6  3.5539ms     672ns  5.7970ms  [CUDA memcpy HtoD]
                   12.69%  3.3166ms         2  1.6583ms  1.3291ms  1.9876ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    5.74%  1.4995ms         2  749.73us  598.46us  900.99us  [CUDA memcpy DtoH]
```

ii. 9x9 Mask

1. Output Tile Size (Block Size) = 4x4 | Input Tile Size = 12x12

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   57.01%  28.039ms         2  14.019ms  11.215ms  16.823ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                   39.90%  19.622ms         6  3.2703ms     640ns  4.2413ms  [CUDA memcpy HtoD]
                    3.09%  1.5179ms         2  758.96us  605.38us  912.54us  [CUDA memcpy DtoH]
```

2. Output Tile Size (Block Size) = 16x16 | Input Tile Size = 24x24

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   61.14%  20.062ms         6  3.3437ms     672ns  4.7346ms  [CUDA memcpy HtoD]
                   34.19%  11.221ms         2  5.6103ms  4.4866ms  6.7341ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    4.67%  1.5310ms         2  765.49us  605.54us  925.44us  [CUDA memcpy DtoH]
```

3. Output Tile Size (Block Size) = 32x32 | Input Tile Size = 40x40

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   60.10%  19.161ms         6  3.1936ms     672ns  3.8371ms  [CUDA memcpy HtoD]
                   35.14%  11.202ms         2  5.6010ms  4.4820ms  6.7201ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    4.76%  1.5184ms         2  759.18us  605.06us  913.31us  [CUDA memcpy DtoH]
```

b. Large Images (4000x3000)

GPU Activities : (Time) in tabular form:

Mask Size/ OutputTile	4x4	16x16	32x32	64x64
3x3	312.3 ms	122.52 ms	126.63 ms	-
9x9	1.33 s	397.13 ms	400.99 ms	-

All GPU Activities:

i. 3x3 Mask

1. Output Tile Size (Block Size) = 4x4 | Input Tile Size = 6x6

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   57.10%  498.35ms         6  83.058ms     640ns  103.55ms  [CUDA memcpy HtoD]
                   35.78%  312.31ms         2  156.15ms  115.08ms  197.22ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    7.12%  62.176ms         2  31.088ms  28.354ms  33.822ms  [CUDA memcpy DtoH]
```

2. Output Tile Size (Block Size) = 16x16 | Input Tile Size = 18x18

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   76.82%  580.12ms         6  96.687ms     704ns  190.71ms  [CUDA memcpy HtoD]
                   16.23%  122.52ms         2  61.261ms  43.294ms  79.229ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    6.95%  52.494ms         2  26.247ms  23.384ms  29.109ms  [CUDA memcpy DtoH]
```

3. Output Tile Size (Block Size) = 32x32 | Input Tile Size = 34x34

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   80.32%  762.70ms         6  127.12ms     640ns  288.82ms  [CUDA memcpy HtoD]
                   13.33%  126.63ms         2  63.313ms  46.737ms  79.890ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    6.35%  60.282ms         2  30.141ms  29.183ms  31.099ms  [CUDA memcpy DtoH]
```

ii. 9x9 Mask

1. Output Tile Size (Block Size) = 4x4 | Input Tile Size = 12x12

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   58.95%  1.33703s         2  668.52ms  628.32ms  708.72ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                   38.35%  869.79ms         6  144.97ms     671ns  383.96ms  [CUDA memcpy HtoD]
                    2.70%  61.223ms         2  30.611ms  30.341ms  30.882ms  [CUDA memcpy DtoH]
```

2. Output Tile Size (Block Size) = 16x16 | Input Tile Size = 24x24

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   55.97%  571.18ms         6  95.196ms  1.7280us  182.49ms  [CUDA memcpy HtoD]
                   38.92%  397.13ms         2  198.57ms  160.57ms  236.56ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    5.12%  52.200ms         2  26.100ms  22.879ms  29.321ms  [CUDA memcpy DtoH]
```

3. Output Tile Size (Block Size) = 32x32 | Input Tile Size = 40x40

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   55.56%  566.03ms         6  94.339ms  1.6000us  183.63ms  [CUDA memcpy HtoD]
                   39.36%  400.99ms         2  200.50ms  156.84ms  244.15ms  kernel3_batch(unsigned char*, float*, int, int, int, int, int, int)
                    5.08%  51.734ms         2  25.867ms  22.568ms  29.167ms  [CUDA memcpy DtoH]
```

Different Declarations of the Mask

The previous profiling for all kernels was done where the mask was declared as a constant in the constant memory, we will see the difference between this implementation and accessing the global memory each time where each thread has its own mask. This was done for kernel 1 only for the same filter, block size, and the image sizes:

Constant Memory:

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   77.00%  1.71099s         2  855.50ms  669.20ms  1.04179s  kernel1_batch_cm(unsigned char*, float*, int, int, int, int, int)
                   20.71%  460.16ms         6  76.694ms     641ns  92.239ms  [CUDA memcpy HtoD]
                    2.30%  51.028ms         2  25.514ms  22.284ms  28.744ms  [CUDA memcpy DtoH]
```

Without Constant Memory:

```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   80.53%  2.11846s         2  1.05923s  760.82ms  1.35764s  kernel1_batch(unsigned char*, float*, float*, int, int, int, int, int)
                   17.54%  461.32ms         5  92.263ms  91.990ms  93.015ms  [CUDA memcpy HtoD]
                    1.93%  50.772ms         2  25.386ms  22.001ms  28.771ms  [CUDA memcpy DtoH]
```

The difference between the time when using the constant memory and the one without is evident; using the constant memory decreases the access time significantly.

Comments

- Increasing the block size from 4x4 to 16 decreases the time significantly however increasing it from 16x16 to 32x32 the time is almost the same. Having a block size of 4x4 is inefficient because the number of threads is 16/block while the minimum working block should have at least a single warp which is 32 threads.
- When increasing the block size to 64x64 this results in an error as it exceeds the maximum number of threads per block allowed.
- Also as the dimensions of the images increase the time increases the time to convolute the image increases significantly.
- Kernel 3 where the block size matches the output tile size each thread is responsible for loading multiple input elements and computing a single output element. The numbers show that this kernel is the fastest.
- Kernel 2 comes in second place where the block size matches the input tile, however this is more risky since the input tile size depends on the size of the filter and can easily exceed the maximum number of threads per block.
- Kernel 1 where there is no tiling involved comes in last since each thread accesses all the elements needed for each cell even though some of these cells are used by all the threads.