

## CS4990 Spring 2024 Homework 3

Total points: 100

Due date: Friday, April 26, 2024

### Task Description:

In this assignment, you are tasked with developing a complete CUDA C/C++ program for an image blur application, also known as image smoothing that we learned in Module 3 “Multidimensional Grids and Data”.

**Convolution** serves as the fundamental operation for implementing the image blur process. Specially, one of the convolution kernels requested in this assignment should be optimized using the **tiled convolution technique** that we learned in Module 7 “Convolution”. This optimization involves leveraging GPU **shared memory and constant memory** to enhance performance.

Below are the specific requirements:

1. **Develop** a single CUDA program file named “**convolution.cu**” containing all the necessary code to blur an input image and generate its blurred image. For simplicity, we will use the average filter of size  $5 \times 5$  in this assignment, where each filter element holds the floating-point value  $1/25$ .

It is highly recommended to utilize the NCSA Delta GPUs for this assignment. However, if you're experienced in successfully building and installing OpenCV for C++ from sources, you may proceed with your own computers.

Below are the command-lines for compiling and executing your program using NCSA Delta GPUs:

- Log in to the NCSA supercomputer using your own NCSA account.
- Enter an interactive session with GPUs, for example:  

```
srunc --account=bchn-delta-gpu --partition=gpuA40x4-interactive --nodes=1 --gpus-per-node=1 --tasks=1 --tasks-per-node=16 --cpus-per-task=1 --mem=20g --pty bash
```
- Load the OpenCV module:  

```
module load opencv/4.9.0.x86_64
```
- To compile:  

```
nvcc -o convolution convolution.cu -I $OPENCV_HOME/include/opencv4/ -L $OPENCV_HOME/lib64 -lopencv_core -lopencv_imgcodecs -lopencv_imgproc
```
- To execute:  

```
./convolution inputImg.jpg
```

When running your program,

- Please make sure to replace “inputImg.jpg” with the name of your input image file, which should be located in the same directory as your program. For example, if

your input image is named "santa-grayscale.jpg", the execution command would be `./convolution santa-grayscale.jpg`

- No command argument for filter radius is necessary in this assignment, as we utilize the constant average filter of size  $5 \times 5$  where the filter radius is 2.
2. Within "convolution.cu", **implement** one host function and two CUDA kernels to perform convolution, respectively. Specifically,
    - a. A host function that performs the convolution operation using CPU-only.
    - b. A CUDA kernel that performs the convolution operation using GPU but without tiling. You may refer to the example provided on Slide 40 in Module 3, but **please modify it to incorporate the average filter matrix**. It is acceptable to load the average filter from either GPU global memory or constant memory.
    - c. **An optimized CUDA kernel that presents a "tiled" version of the convolution operation using GPU shared memory**. Specifically, in this optimized kernel, please **load the average filter from GPU constant memory**.
  3. **Ensure** that your code can handle images with **varying image sizes**. Please also **consider** boundary conditions to ensure proper handling in such cases.

For testing purposes, two input images of different sizes are provided in the zipped folder accompanying this assignment.

- "santa-grayscale.jpg": a grayscale image of size  $1,000 \times 1,000$
- "tree-grayscale.jpg": a grayscale image of size  $345 \times 346$

You may also choose to test your program with additional grayscale images if desired.

4. **Utilize** timing techniques such as CPU timers or CUDA events, **to measure the performance of** your implementation of the host function and two CUDA kernels as specified above.

We also suggest **structuring** the "convolution.cu" **by implementing the following macros, host functions, and CUDA kernels**. At the end of this assignment, we will provide screenshots of an example of the program's structure.

- **#define CHECK(call)**
  - A macro for error checking.
- **double myCPUTimer()**
  - A timer for measuring execution time.
- **void blurImage\_h(cv::Mat Pout\_Mat\_h, cv::Mat Pin\_Mat\_h, unsigned int nRows, unsigned int nCols)**
  - A host function for CPU-only convolution.
- **\_\_global\_\_ void blurImage\_Kernel(unsigned char \* Pout, unsigned char \* Pin, unsigned int width, unsigned int height)**
  - A CUDA kernel performs a simple convolution without using tiling.
- **void void blurImage\_d(cv::Mat Pout\_Mat\_h, cv::Mat Pin\_Mat\_h, unsigned int nRows, unsigned int nCols)**
  - A host function for handling device memory allocation and free, data copy, and calling the specific CUDA kernel, blurImage\_Kernel().

- **\_\_global\_\_ void blurImage\_tiled\_Kernel(unsigned char \* Pout, unsigned char \* Pin, unsigned int width, unsigned int height)**
  - A CUDA kernel that performs a “tiled” version of convolution using GPU shared memory and loading the filter elements from GPU constant memory.
- **void blurImage\_tiled\_d(cv::Mat Pout\_Mat\_h, cv::Mat Pin\_Mat\_h, unsigned int nRows, unsigned int nCols)**
  - A host function for handling device memory allocation and free, data copy, and calling the specific CUDA kernel, blurImage\_tiled\_Kernel().
- **int main(int argc, char\*\* argv)**
  - The main entry point of the program
- **bool verify(cv::Mat answer1, cv::Mat answer2, unsigned int nRows, unsigned int nCols)**
  - A function to validate if the blurred image result using a CUDA kernel matches that of the OpenCV’s blur function, cv::blur().
  - Note that you may call this verification function **three** times in your main function, in order to
    - Compare the blurred image result of blurImage\_h() with that of cv::blur().
    - Compare the blurred image result of blurImage\_d() with that of cv::blur().
    - Compare the blurred image result of blurImage\_tiled\_d() with that of cv::blur().

Please note that if needed, you may review how to use NCSA Delta GPUs and how to use OpenCV with CUDA programs by referring to the video recording of the lecture session on 02/23/2024, as well as the CUDA&&OpenCV program examples we covered during lectures on Module 3.

### **What to Submit?**

Please **compress** the following **six** required files into a zip file, named following the format **"yourname\_p3.zip"**, and **submit** the zipped file on Canvas.

1. **"convolution.cu"**: a single CUDA program file containing all required code.
2. **"blurredImg\_opencv.jpg"**: the blurred image generated by OpenCV’s cv::blur(), given an input image of your choice.
3. **"blurredImg\_cpu.jpg"**: the blurred image generated by blurImage\_h() using CPU-only, given the same input image.
4. **"blurredImg\_gpu.jpg"**: the blurred image generated by blurImage\_d() using GPU but without tiling, given the same input image.
5. **"blurredImg\_tiled\_gpu.jpg"**: the blurred image generated by blurImage\_tiled\_d() using GPU and the tiling, given the same input image.
6. **"output.jpg"**: a screenshot presenting the verification results of the "verify" function and the timing results.

To demonstrate the necessary information for your screenshot, here's an example of a screenshot displaying the output of my vector-addition program. It includes the verification results and the timing results. However, please note that this example is not for convolution.

```
[vccjihao@gpub001 CS4990]$ ./main
Vector size 16777216
vecAdd on CPU:                                0.035908 s

      cudaMalloc:                             0.000372 s
      cudaMemcpy:                             0.010065 s
      vecAddKernel<<<(32768,1,1),(512,1,1)>>>: 0.025956 s
      cudaMemcpy:                             0.009967 s
vecAdd on GPU:                                0.048771 s
```

Verifying results...TEST PASSED

**Below please find screenshots of an example for the program's structure for this assignment.** Please note that the CUDA program shown in the screenshots is incomplete. You also need to modify the part of the program responsible for loading an input image specified by the user. For example,

- Parse the command-line arguments to extract the filename of the input image specified by the user.
- Load the specified input image dynamically at runtime, instead of using a hardcoded filename.
- Ensure error handling for cases where the specified image file cannot be found or loaded properly.

```
1  #include <opencv2/opencv.hpp>
2  #include <sys/time.h>
3
4  #define FILTER_RADIUS 2
5
6  // for simplicity, we use the constant average filter only in this assignment
7  const float F_h[2*FILTER_RADIUS+1][2*FILTER_RADIUS+1] = {
8      {1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25},
9      {1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25},
10     {1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25},
11     {1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25},
12     {1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25, 1.0f / 25}
13 };
14
15 __constant__ float F_d[2*FILTER_RADIUS+1][2*FILTER_RADIUS+1];
16
17 // check CUDA error if exists
18 #define CHECK(call) { \
19     const cudaError_t cuda_ret = call; \
20     if(cuda_ret != cudaSuccess) { \
21         printf("Error: %s:%d, ", __FILE__, __LINE__); \
22         printf("code: %d, reason:%s\n", cuda_ret, cudaGetErrorString(cuda_ret)); \
23         exit(-1); \
24     } \
25 }
26
27 // check if the difference of two cv::Mat images is small
28 bool verify(cv::Mat answer1, cv::Mat answer2, unsigned int nRows, unsigned int nCols) {
29     const float relativeTolerance = 1e-2;
30     for(int i=0; i<nRows; i++){
31         for(int j=0; j<nCols; j++){
32             float relativeError = ((float)answer1.at<unsigned char>(i,j) - (float)answer2.at<unsigned char>(i,j))/255;
33             if (relativeError > relativeTolerance || relativeError < -relativeTolerance) {
34                 printf("TEST FAILED at (%d, %d) with relativeError: %f\n", i, j, relativeError);
35                 printf("    answer1.at<unsigned char>(%d, %d): %u\n", i, j, answer1.at<unsigned char>(i,j));
36                 printf("    answer2.at<unsigned char>(%d, %d): %u\n", i, j, answer2.at<unsigned char>(i,j));
37                 return false;
38             }
39         }
40     }
41     printf("TEST PASSED\n\n");
42     return true;
43 }
```

```

44 // CPU timer
45 double myCPUTimer(){
46     struct timeval tp;
47     gettimeofday(&tp, NULL);
48     return ( (double)tp.tv_sec + (double)tp.tv_usec/1.0e6);
49 }
50
51 // A CPU-implementation of image blur using the average box filter
52 void blurImage_h(cv::Mat Pout_Mat_h, cv::Mat Pin_Mat_h, unsigned int nRows, unsigned int nCols)
53 {
54 }
55
56 // A CUDA kernel of image blur using the average box filter
57 global void blurImage_Kernel(unsigned char * Pout, unsigned char * Pin, unsigned int width, unsigned int height)
58 {
59 }
60
61 // A GPU-implementation of image blur using the average box filter
62 void blurImage_d(cv::Mat Pout_Mat_h, cv::Mat Pin_Mat_h, unsigned int nRows, unsigned int nCols)
63 {
64 }
65
66 // An optimized CUDA kernel of image blur using the average box filter from constant memory
67 global void blurImage_tiled_Kernel(unsigned char * Pout, unsigned char * Pin, unsigned int width, unsigned int height)
68 {
69 }
70
71 // A GPU-implementation of image blur, where the kernel performs shared memory tiled convolution using the average box filter from constant memory
72 void blurImage_tiled_d(cv::Mat Pout_Mat_h, cv::Mat Pin_Mat_h, unsigned int nRows, unsigned int nCols)
73 {
74 }
75
76 int main(int argc, char** argv){
77     cudaDeviceSynchronize();
78
79     double startTime, endTime;
80
81     // use OpenCV to load a grayscale image.
82     cv::Mat grayImg = cv::imread("santa-grayscale.jpg", cv::IMREAD_GRAYSCALE);
83     if(grayImg.empty()) return -1;
84
85     // obtain image's height, width, and number of channels
86     unsigned int nRows = grayImg.rows, nCols = grayImg.cols, nChannels = grayImg.channels();
87
88     // for comparison purpose, here uses OpenCV's blur function which uses average filter in convolution
89     cv::Mat blurredImg_opencv(nRows, nCols, CV_8UC1, cv::Scalar(0));
90     startTime = myCPUTimer();
91     cv::blur(grayImg, blurredImg_opencv, cv::Size( 2*FILTER_RADIUS+1, 2*FILTER_RADIUS+1 ), cv::Point(-1, -1), cv::BORDER_CONSTANT);
92     endTime = myCPUTimer();
93     printf("opencv's blur (CPU): %f s\n", endTime - startTime); fflush(stdout);
94
95     // for comparison purpose, implement a cpu version
96     cv::Mat blurredImg_cpu(nRows, nCols, CV_8UC1, cv::Scalar(0)); // cv::Mat constructor to create and initialize an cv::Mat object; note that CV_8UC1 implies 8-bit
97     unsigned, single channel
98     startTime = myCPUTimer();
99     blurImage_h(blurredImg_cpu, grayImg, nRows, nCols);
100     endTime = myCPUTimer();
101     printf("blurImage on CPU: %f s\n", endTime - startTime); fflush(stdout);
102
103     // implement a gpu version that calls a CUDA kernel
104     cv::Mat blurredImg_gpu(nRows, nCols, CV_8UC1, cv::Scalar(0));
105     startTime = myCPUTimer();
106     blurImage_d(blurredImg_gpu, grayImg, nRows, nCols);
107     endTime = myCPUTimer();
108     printf("blurImage on GPU: %f s\n", endTime - startTime); fflush(stdout);
109
110     // implement a gpu version that calls a CUDA kernel which performs a shared-memory tiled convolution, and filter elements are loaded from constant memory
111     cv::Mat blurredImg_tiled_gpu(nRows, nCols, CV_8UC1, cv::Scalar(0));
112     startTime = myCPUTimer();
113     blurImage_tiled_d(blurredImg_tiled_gpu, grayImg, nRows, nCols);
114     endTime = myCPUTimer();
115     printf("(tiled)blurImage on GPU: %f s\n", endTime - startTime); fflush(stdout);
116
117     // save the result blurred images to disk
118     bool check = cv::imwrite("./blurredImg_opencv.jpg", blurredImg_opencv);
119     if(check == false){ printf("error!\n"); return -1;}
120
121     check = cv::imwrite("./blurredImg_cpu.jpg", blurredImg_cpu);
122     if(check == false){ printf("error!\n"); return -1;}
123
124     check = cv::imwrite("./blurredImg_gpu.jpg", blurredImg_gpu);
125     if(check == false){ printf("error!\n"); return -1;}
126
127     check = cv::imwrite("./blurredImg_tiled_gpu.jpg", blurredImg_tiled_gpu);
128     if(check == false){ printf("error!\n"); return -1;}
129
130     // check if the result blurred images are similar to that of OpenCV's
131     verify(blurredImg_opencv, blurredImg_cpu, nRows, nCols);
132     verify(blurredImg_opencv, blurredImg_gpu, nRows, nCols);
133     verify(blurredImg_opencv, blurredImg_tiled_gpu, nRows, nCols);
134
135     return 0;
136 }

```