Cuda Kernel Image Processing

Francesco Villi
francesco.villi@stud.unifi.it

Abstract

This project utilizes CUDA, a parallel computing platform and programming model developed by NVIDIA, to perform image processing tasks on a GPU. By implementing image processing algorithms as CUDA kernels, we can accelerate significantly the processing time compared to a traditional CPU-based implementation. The goal of the project is to demonstrate the effectiveness of using CUDA for image processing and to show the potential for processing high-resolution images.

1. Introduction

The developed application can be used to perform a convolution between kernel and image. Kernel image processing refers to a technique used in image processing where an image is convolved with a small matrix called a kernel. The kernel is a fixed matrix of coefficients, which is used to modify the pixels in the image. We apply the kernel to each element of the image by multiplying the pixel value with the corresponding value in the kernel, and then summing up the results. It can be written as:

$$g(x,y) = (f * g)(x,y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} k(i,j) f(x-i, y-j)$$

where f is the image, k the kernel, and g the output image. This process is applied to each pixel in the image, effectively transforming it. We can rotate by 180° the kernel k and compute the convolution using this formula that is easier to implement:

$$g(x,y) = (f * g)(x,y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} k'(i,j) f(x+i, y+j)$$

All kernels in the program are symmetric so we don't need rotation.

The output has a size smaller than the input due to we can't apply convolution in the borders; to solve that we have some techniques computationally expansive as increasing the kernel and image size:

- *zero-padding* where we add 0s around the border of the image
- *pixel mirroring* where we use the nearest border pixel as padding
- *pixel replication* where we copy border pixels.

The kernel matrix can be designed to perform different types of image processing operations such as image sharpening, blurring, edge detection.

Kernel-based image processing is a powerful technique that can be used to modify images in a variety of ways, but it has a high computational cost which can be time-consuming on a CPU. CUDA-based image processing allows to perform kernel-based image processing on a GPU, greatly reducing the time required for the operation. In particular this is an embarrassingly parallel problem because it can be easily split into smaller, independent tasks that can be run simultaneously on multiple processors. The tasks have no inter-dependencies and do not require communication between processors, making it simple to distribute the processing workload. This results in a high degree of parallelism and can lead to significant speedup in processing time.

1.1. CUDA

CUDA (Compute Unified Device Architecture) is a parallel computing platform and programming model developed by NVIDIA for general purpose computing on NVIDIA GPUs (Graphics Processing Units). CUDA enables developers to leverage the massive parallelism of NVIDIA GPUs to perform complex computations faster than on a CPU alone.

In CUDA, there are different types of memories that can be used for storing data and performing computation. These include:

- *Global memory*: This is the largest memory available on the device and it is shared among all threads. It's the slowest available on GPU

- *Shared memory*: This is a smaller memory that is shared among threads within a block.It's faster than global and constant memory.

- *Constant memory*: This is a read-only memory that is accessible by all threads and can be used to store constant data. It's faster then global.

In CUDA, there is also a distinction between host memory and device memory:

- *Host memory* refers to the memory that is available on the host (i.e. the CPU)

- *Device memory* refers to the memory that is available on the GPU

Data must be transferred between host and device memory in order to be processed on the GPU.

2. Structure

The structure of the code is:

- **ConvolutionCPU** implements static functions to perform a convolution in parallel and sequential way.

- **PaddingImage** implements static function to perform padding according to the type padding (zero, pixel mirroring, pixel replication)

- **Image** class that contains image size and image values. It doesn't read a real image but, given height and width, it creates a random matrix with values from 0 to 255.

- **Kernel** class that contains all possible kernels(with a size) that can be used such as edge detection (3x3), sharpen (3x3), constant blur (3x3) and gaussian blur (3x3, 5x5, 7x7).

- **ConvolutionGPU** implements static functions to perform a convolution using GPU. There are 3 possibilities: use gloabl memory, global and constant memory or shared memory and constant memory.

2.1. CPU version

In this version we have 2 main steps:

1. The function gets in input the image, kernel and type padding. At the begin it computes the padding according to the type and returning a new image padded.

2. The second step is about performing convolution between the image padded and kernel returning the edited image.

I implemented a parallel version and a sequential version in order to compare them. The parallel version uses directive yielded by OpenMP.

2.2. GPU version

In the GPU version we have 3 different types of implementation:

1. Using only *global memory*: I allocate space in the global memory for kernel, padded image and output image; then I compute the output image calculating the convolution between padded image and kernel reading all values from the global memory and storing the new values in the output image.

2. Using *constant memory*: I allocate space in the global memory for padded image and output image whereas the kernel has been stored in the constant memory. Now to compute the convolution I have to read image values from the global memory and kernel values from the constant memory that is faster.

3. Using *shared memory*: I allocate space in global memory for padded image and output image while the kernel is put in constant memory; then for each block I have to load a tile of the input image due to image size. To do that we have to copy for each block that represents a part of the total image also the elements "outside the border" (i.e. if a block is 16x16 and the kernel size is 3x3 the sub-image that we copy is 18x18 because we have to consider the pixel near the border to calculate correctly the output value).

2.3. Other implementation details

Every time before executing the convolution it creates the padded image and then it computes the convolution. All kernels are integers and only after the sum we apply the normalization factor to achieve the max precision.
In the context of GPU implementations, the size of the computational grid is determined differently based on the choice of memory type. In implementations utilizing global or constant memory, the grid size is determined by the original dimensions of the image. Conversely, in implementations utilizing shared memory, the grid size is dependent on the dimensions of the image after padding.

3. Environment settings

All tests have been run in a PC with intel core intel i7-12700H, 16GB of RAM DDR5, GPU NVIDIA GeForce RTX 3050 with 4GB of memory. The compiler used is Visual Studio Community version 17.4.3 and NVIDIA toolkit 12.0, OpenMP has been used to parallelize the CPU code.

4. Comparison

Firstly we'll compare the CPU version and GPU version; then we will analyze three different versions of a GPU implementations that utilize different types of memory.
We will be examining the performance, and overall effectiveness of each version to determine the best option for the given application. By comparing these different types of memory, we aim to provide insight into the advantages and disadvantages of each.
We will evaluate their performance under varying image sizes and kernel size. We will incrementally increase one parameter at a time and we measure the time it takes for each version to complete the task, i.e. load a matrix into memory, complete the convolution and return the new elaborated image. This will provide insight into how each type of memory handles larger workloads and how it may impact the overall performance of the system. The kernel used for tests is a gaussian filter.

5. Results and Discussion

5.1. CPU

The parallel version uses the max amount of threads (20) and to balance the workload I used a dynamic scheduling because the i7-12000H has 4 efficient cores and 8 high performance cores.

Size	Sequential	Parallel
1280x720	0.006941	0.002811
1920x1080	0.025705	0.003316
2560x1440	0.034740	0.005121
8120x4568	0.245991	0.036058

Table 1. 3x3 kernel processing speed using CPU(values in seconds)

Size	Sequential	Parallel
1280x720	0.022662	0.003011
1920x1080	0.051387	0.005131
2560x1440	0.034740	0.008211
8120x4568	0.577013	0.064652

Table 2. 5x5 kernel processing speed using CPU(values in seconds)

Size	Sequential	Parallel
1280x720	0.042273	0.004781
1920x1080	0.085983	0.009318
2560x1440	0.128911	0.013541
8120x4568	1.0025	0.128868

Table 3. 7x7 kernel processing speed using CPU(values in seconds)

We observe in the CPU case that the parallelization improves the general performance.
In particular:

1. Using a 3x3 kernel (tab. 1) the speed up is low (2) when the image and kernel size are small whereas speed up grows until 10 for the median sizes and then it stabilizes around 7 for high values.

2. Using a 5x5 kernel (tab. 2) the speed up is 6 for small image, then it reaches 10 for the median sizes and it stabilizes around 7 again.

3. Using a 7x7 kernel (tab. 3) the speed up is constantly around 9

We notice that in all cases the parallel version performs better than the sequential. The unique case takes in exam that performs with a low speed up is the first case in table 1 due to the additional time and resources required to coordinate and manage parallel operations.
In all other cases the improvements are significant but not enough to elaborate a millions of images in rational time. So we need to introduce GPU that has thousands of small processing cores that can work together to perform the same operation on multiple pieces of data at the same time.

5.2. GPU

We measured computational time for each implementation according to the memory used.

Size	Global	Constant	Shared
1280x720	$2.0e-05$	$1.82e-05$	$2.1e-05$
1920x1080	$1.98e-05$	$1.84e-05$	$1.90e-05$
2560x1440	$1.76e-05$	$1.8e-05$	$1.78e-05$
8120x4568	$2.05e-04$	$1.95e-04$	$1.82e-05$

Table 4. 3x3 kernel processing speed using GPU and various memories without data transfer(values in seconds)

We note that in general from table 4, 5 and 6 that the shared memory version has the best performance while the constant memory version is a little better than the global memory version. The difference in absolute terms isn't

Size	Global	Constant	Shared
1280x720	$2.7e-05$	$2.1e-05$	$1.9e-05$
1920x1080	$1.98e-05$	$1.88e-05$	$1.74e-05$
2560x1440	$1.85e-05$	$1.76e-05$	$1.78e-05$
8120x4568	$2.20e-05$	$1.80e-05$	$1.70e-05$

Table 5. 5x5 kernel processing speed using GPU and various memories without data transfer(values in seconds)

Size	Global	Constant	Shared
1280x720	$2.1e-05$	$1.95e-05$	$1.92e-05$
1920x1080	$1.8e-05$	$1.6e-05$	$1.5e-05$
2560x1440	$2.1-05$	$2.2e-05$	$2.01e-05$
8120x4568	$2.4e-05$	$1.75e-05$	$1.73e-05$

Table 6. 7x7 kernel processing speed using GPU and various memories without data transfer(values in seconds)

much between all three versions but we have to remember that we will apply convolution to millions of images and so a small speed improvement for a single image can mean a big speed improvement for elaborating millions of images.
Comparing these results with the results obtained using a CPU we notice a speed up ranging from 100 to 7449; bigger is the image and kernel, and greater is the speed up.
Using a GPUs, one of the key challenges is transferring data between the host (i.e. the CPU) and the device (i.e. the GPU). This process, known as data transfer or data movement, can have a significant impact on the performance of the system. Considering also the time needed for the data transfer the results in terms of time become:

Size	Global	Constant	Shared
1280x720	0.001107	0.001072	0.001242
1920x1080	0.002462	0.002822	0.002231
2560x1440	0.003020	0.002667	0.003089
8120x4568	0.015869	0.017503	0.016138

Table 7. 3x3 kernel processing speed using GPU and various memories with data transfer(values in seconds)

Size	Global	Constant	Shared
1280x720	0.001248	0.001175	0.0011532
1920x1080	0.003057	0.002209	0.002365
2560x1440	0.002680	0.0029334	0.0026348
8120x4568	0.017908	0.01884	0.015280

Table 8. 5x5 kernel processing speed using GPU and various memories with data transfer(values in seconds)

We note that the issue is the data transfer, in particular in shared memory we have to allocate also the memory in each block for the partial image and the kernel.

Size	Global	Constant	Shared
1280x720	0.001274	0.0012639	0.0011666
1920x1080	0.0025554	0.002370	0.0021321
2560x1440	0.0034905	0.003237	0.002869
8120x4568	0.0233028	0.0233129	0.0172099

Table 9. 7x7 kernel processing speed using GPU and various memories with data transfer(values in seconds)

So, the bottleneck in this process is often the transfer of data from the CPU to the GPU, as the CPU must send the data to the GPU before computation can begin and retrieve the results once computation is complete. This transfer can take a significant amount of time, limiting the overall speedup gained by using a GPU. Despite this bottleneck, the use of GPUs can still result in significant performance gains in certain types of computations like this one.
So, considering the data transfer we confirm the previous result.
Analyzing tab 7, 8 and 9 we observe quite similar computation times for all GPU implementations; the best outcome has been reached by shared memory in most cases.
A comparison of the data presented in Table 11 and Table 12 reveals a significant impact imposed by the data transfer from and to the device and host. This is evidenced by the re-calculation of the speed up metric, which reveals that the maximum speed up achieved is 7.5, while the lowest value obtained is a 1.5 speed up when utilizing a 3x3 kernel as we can see in table 10.
Due to the short amount of time needed to perform convolution operations and transferring the data, the results may be affected by GPU throttling or task management by the operating system; so that in the code there is a function to measure the computation time with and without considering data transfer that confirm the previous deductions.

Size	Speed-up
1280x720	2.2
1920x1080	1.5
2560x1440	1.9
8120x4568	2.2

Table 10. 3x3 kernel processing speed up between parallel CPU and GPU shared version considering data transfer

6. Conclusion

In conclusion, this project has demonstrated that GPU is a superior choice for kernel image processing compared to CPU. The parallel processing capabilities of GPU allows for faster processing times, even when faced with the added overhead of transferring data from the host to the device and viceversa. The results of this project show a significant

Size	Speed-up
1280x720	249
1920x1080	621
2560x1440	673
8120x4568	7449

Table 11. 7x7 kernel processing speed up between parallel CPU and GPU shared version without considering data transfer

Size	Speed-up
1280x720	4.1
1920x1080	4.3
2560x1440	4.7
8120x4568	7.5

Table 12. 7x7 kernel processing speed up between parallel CPU and GPU shared version considering data transfer

performance boost when using GPU for kernel image processing, making it a valuable tool for tasks such as image filtering and feature extraction.

Overall, it is clear that GPU are a powerful and efficient choice for image processing tasks.