CMSC5743 Lab 03

CUDA Programming Language

1 Sample Code:

- Install the CUDA environment
  - Use `nvcc --version` to check whether it is successful or not
  - Run `nvidia-smi` to check the status of your GPUs.
- Run the vector_add example:
  - Go to the `./Lab03-CUDA/code/vector_add`
  - Run `./compile.sh` script to compile the CUDA kernel
  - Run `./vector_add` script to get the final result

2 Assignments:

Q1 Learn the code in `./Lab03-CUDA/code` and it contains three folders (vector_add, gemm, wmma)

- Learn the code style and components of `vector_add.cu` file
- Complete all of the code in `gemm` folder
- Try to make your gemm kernel more efficient
  - shared memory
  - tiling size
  - block and thread size

Q2 Learn the `wmma.cu` from the `/Lab03-CUDA/code/wmma` to run it successfully by `compile.sh` script

- Learn the different data type in CUDA programming language such as Float16, Int8
- Learn the basic knowledge of Tensor Core and WMMA in CUDA programming language
- Learn the difference between FLOPs and FLOPS
- Change the tiling size in `wmma.cu` to get the different TFLOPS

Useful Materials:

- Performance Metrics
- CS 179 GPU Programming
- Tensor Core
- CUTLASS
- High Performance Computer Architecture
- CUDA C++ Programming Guide

Tips: You should learn the code style from the sample code to build your project.