|  |  |
| --- | --- |
| **Name:** | *Jiahan Chang* |
| **NetID:** | *Jiahanc2* |
| **Section:** | *AL1* |

**ECE 408/CS483 Milestone 2 Report**

|  |
| --- |
| 1. Show output of rai running Mini-DNN on the basic GPU convolution implementation for batch size of 1k images. This can either be a screen capture or a text copy of the running output. Please do not show the build output. (The running output should be everything including and after the line "*Loading fashion-mnist data...Done*"). |
| *<output here>* |
| 1. For the basic GPU implementation, list Op Times, whole program execution time, and accuracy for batch size of 100, 1k, and 10k images. |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Batch Size | Op Time 1 | Op Time 2 | Total Execution Time | Accuracy | | 100 | *0.494676 ms* | *6.95472 ms* | *0m1.190s* | *0.86* | | 1000 | *5.02055 ms* | *17.3097 ms* | *0m9.729s* | *0.886* | | 10000 | *49.7845 ms* | *167.509 ms* | *1m37.002s* | *0.8714* | |
| 1. List all the kernels that collectively consumed more than 90% of the kernel time and what percentage of the kernel time each kernel did consume (start with the kernel that consumed the most time, then list the next kernel, until you reach 90% or more). |
| *<answer here>* |
| 1. List all the CUDA API calls that collectively consumed more than 90% of the API time and what percentage of the API time each call did consume (start with the API call that consumed the most time, then list the next call, until you reach 90% or more). |
| *<answer here>* |
| 1. Explain the difference between kernels and CUDA API calls. Please give an example in your explanation for both. |
| *<answer here>*  *The kernel is the code run on the device. It is run in parallel and is executed by different CUDA threads. Each CUDA thread that executes the kernel has a different ID.*  *The API calls are all on the host. An example of a API call is setting up the memory for the kernel. They are an extension to the language the developer is using, and calls the lower level C API.* |
| 1. Show a screenshot of the GPU SOL utilization |
| *<nsight output here>*  *1-122 conv\_forward\_kernel*    *4-143 conv\_forward\_kernel* |