|  |  |
| --- | --- |
| **Name:** | *John Li* |
| **NetID:** | *johnwl2* |
| **Section:** | *AB* |

**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*"). |
|  |
| 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 | *0.175693 ms* | *0.63736 ms* | *7.14276 ms* | *0.86* | | 1000 | *1.63478 ms* | *6.29276 ms* | *72.3058 ms* | *0.886* | | 10000 | *16.161 ms* | *137.53 ms* | *614.038 ms* | *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). |
| *conv\_forward\_kernel(): 100%* |
| 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). |
| *CUDA memcpy D to H: 92.9%*    *cudaMempcy()*  *cudaMalloc()* |
| 1. Explain the difference between kernels and CUDA API calls. Please give an example in your explanation for both. |
| *Kernels are functions/computer programs that when called, they are executed N times in parallel by N different CUDA threads. Kernels are also core components of operating systems, while an API is like an extension or library that you could interact when you are implementing kernel level code. It includes extra functions, definitions, protocols, or tools. An example of a kernel would be defined as something like \_\_global void MatrixMultiply(){<code>} from MP2. Some examples of CUDA API calls/functions are cudaMalloc(), cudaMemcpy(), cudaFree(), etc.* |
| 1. Show a screenshot of the GPU SOL utilization |
|  |