|  |  |
| --- | --- |
| **Name:** | *Wentao Zhang* |
| **NetID:** | *wentao4* |
| **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*"). |
| *Test batch size: 1000*  *Loading fashion-mnist data...Done*  *Loading model...Done*  *Conv-GPU==*  *Layer Time: 65.9262 ms*  *Op Time: 1.62797 ms*  *Conv-GPU==*  *Layer Time: 54.2663 ms*  *Op Time: 6.27431 ms*  *Test Accuracy: 0.886* |
| 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 | *0.17677 ms* | | *0.63500 ms* | *1.163s* | *0.86* | | 1000 | *1.62797 ms* | | *6.27431 ms* | *9.691s* | *0.886* | | 10000 | *16.0546 ms* | | *62.7971 ms* | *1m34.838s* | *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). |
| *My nsys result shows*  *100.0 79100647 2 39550323.5 16187218 62913429 conv\_forward\_kernel*  *So, the only kernel is* ***conv\_forward\_kernel,***  *it account for 100% of time.* |
| 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). |
| *77.1 1081571065 8 135196383.1 18306 585526607 cudaMemcpy*  *15.6 218461729 8 27307716.1 82115 209579309 cudaMalloc*  *So there are two API:* ***cudaMemcpy and cudaMalloc*** *consume more than 90% of running time.* |
| 1. Explain the difference between kernels and CUDA API calls. Please give an example in your explanation for both. |
| *CUDA APIs are used for data transfer, control and resource management. For example,* ***cudaMemcpy, cudaMemset, cudaMalloc*** *are CUDA APIs.*  *CUDA kernels are mostly used for efficient computation in parallel , such as* ***convolution*** *and* ***matrix multiplication*** |
| 1. Show a screenshot of the GPU SOL utilization |
| *The GPU SOL utilization of two launch:* |