|  |  |
| --- | --- |
| **Name:** | *Ge Yuhao* |
| **NetID:** | [*Yuhaoge2@illinois.edu*](mailto:Yuhaoge2@illinois.edu) |
| **Section:** | *ZJ1/ZJ2* |

**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*"). |
| *文本  描述已自动生成* |
| 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 | *0.174333ms* | *0.630749ms* | *1.216s* | *0.86* | | 1000 | *1.63886ms* | *6.25089ms* | *9.673s* | *0.886* | | 10000 | *16.0759ms* | *62.9824ms* | *1m34.672s* | *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). |
| *99.9% conv\_forward\_kernel* |
| 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). |
| Batch size = 1000:  68.8% cudaMalloc  28.8% cudaMemcpy  Batch size = 10000:  76.2% cudaMemcpy  16.7% cudaMalloc |
| 1. Explain the difference between kernels and CUDA API calls. Please give an example in your explanation for both. |
| The CUDA API is the build in function such as cudaMalloc and cudaMemcpy which are called by the main function.  The Kernel is the cuda function we defined in our code which are also called by the main function. |
| 1. Show a screenshot of the GPU SOL utilization |
| *Batch size = 1000:*    *Batch size = 10000* |