Name: *Yize Chen*
NetID: *yizec2*
Section: *AL2*

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*").

```
Loading fashion-mnist data...Done
Loading model...Done
Conv-GPU==
Layer Time: 63.4949 ms
Op Time: 1.85346 ms
Conv-GPU==
Layer Time: 56.8583 ms
Op Time: 8.17948 ms

Test Accuracy: 0.886


real    0m9.678s
user    0m9.355s
sys     0m0.288s
```

2. 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	*0.199107 ms*	*0.831835 ms*	*0m1.213s*	*0.86*
1000	*1.85346 ms*	*8.17948 ms*	*0m9.678s*	*0.886*
10000	*18.2708 ms*	*81.5028 ms*	*1m35.538s*	*0.8714*

3. 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

4. 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).

cudaMemcpy
cudaMalloc

5. Explain the difference between kernels and CUDA API calls. Please give an example in your explanation for both.

CUDA API calls is current to the calling host thread, while kernel calls are asynchronous for CPU.
Example:
*cudaMalloc((void**)&device_x, ...); cudaMemset(device_x, ...); Allocate the device memory first, and then set them to zero.*
Kernel1<<<...>>>(); Kernel2<<<...>>>(); CPU won't wait for these 2 kernel calls.

6. Show a screenshot of the GPU SOL utilization