# ECE408 Project Report

Ayushi Patel Kartikeya Sharma Rishabh Goyal

## Team details

Team name: cudashouldawoulda
School affiliation: On campus students
Ayushi Patel (**ayuship2**)
Kartikeya Sharma (**ksharma**)
Rishabh Goyal (**rgoyal6**)

## MILESTONE 2

- **Include a list of all kernels that collectively consume more than 90% of the program time:**

| Time(%) | Name |
|---|---|
| 30.23% | [CUDA memcpy HtoD] |
| 18.00% | volta_scudnn_128x64_relu_interior_nn_v1 |
| 17.31% | volta_gcgemm_64x32_nt |
| 8.82% | fft2d_c2r_32x32 |
| 7.86% | volta_sgemm_128x128_tn |
| 6.62% | op_generic_tensor_kernel |
| 6.57% | fft2d_r2c_32x32 |
| 3.97% | cudnn::detail::pooling_fw_4d_kernel |
| 0.42% | mshadow::cuda::MapPlanLargeKernel |

- **Include a list of all CUDA API calls that collectively consume more than 90% of the program time.**

| Time(%) | Name |
|---|---|
| 42.85% | cudaStreamCreateWithFlags |
| 33.41% | cudaMemGetInfo |
| 20.90% | cudaFree |

- **Include an explanation of the difference between kernels and API calls**

  Kernels (GPU Activities) in the *nvprof* output represent actual usage of the GPU for any kind of task. The time taken for GPU Activities represents the difference between the times the task actually started executing on the GPU and finished executing on the GPU.

  API calls are made by the host code (or by other API calls made by the code) that access the CUDA runtime. A GPU Activity is performed by initiating it with some form of API call. However since API calls are asynchronous, their finishing time is not related to the GPU activity that it launches, it may even finish executing before the kernel code is done using the GPU.

- **Show output of rai running MXNet on the CPU**

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: ‘accuracy’: 0.8154
```

- **List program run time**
  19.70 seconds user
  6.46 seconds system

- **Show output of rai running MXNet on the GPU**
  Loading fashion-mnist data... done
  Loading model... done
  New Inference
  EvalMetric: ‘accuracy’: 0.8154

- **List program run time**
  5.05 seconds user
  3.40 seconds system

**CPU implementation**

- **List whole program execution time**
  87.40 seconds user
  10.34 seconds system

- **List Op Times**
  Op Time 1: 11.223992 seconds
  Op Time 2: 60.508100 seconds