# Project Milestone Report

## Reducing Latency in CPU-GPU Interactions

**Project URL: emmaloool.github.io/15400-Project**

Major Changes

There haven't been many changes to my project and its proposal since I submitted the proposal last month. Both I and my research advisor's PHD student, Pratik, who I will be working closely with on this effort, reviewed my proposal after I submitted it, and he was pleased with the way I presented the problem.

Current Progress

For the past few weeks, I've been familiarizing myself with the papers six-eight or so papers I briefly mentioned in my proposal writeup. Since I do not have a background in machine learning, but the workloads we're studying involve several different neural network implementations, I've also spent some time perusing different pages for courses offered at CMU that cover neural networks (like 10-315: Machine Learning and 11:485: Neural Networks) to get a basic understanding of the types of neural networks. I've downloaded PyTorch, the software we'll be using the construct and train the neural network workloads. I have yet to gain the necessary software with the PhD student, as communication has been slow between him and I in the past month.

Milestone Completion Assessment

I believe I've met the milestone for this semester, which was to read the set of papers I presented in detail and gain some background support for the workloads. What now needs to occur before the start of the semester is to fine-tune the scope of my project with Pratik and perhaps Professor Mowry.

For the scope of this research, since I will need to have a deep understanding of these workloads, after reading these papers, despite completing the milestone, I'm a little concerned about my understanding of these topics; I've never worked with machine learning before, much less neural networks. In regards to this topic, the PhD student I'm working with also has a similar level of unfamiliarity as I do, apparently. To address this uncertainty, I will try to meet with Pratik this week (I'm on campus for a week or two after finals complete) to discuss the gaps in our knowledge and we may come up with a plan (the same as I originally intended, or modified) to prepare for the beginning of the semester. This may include needing to add 10-315: Machine Learning or 11-485: Neural Networks to my schedule next semester, so that I can get a sufficient background to supplement my research efforts.

## Surprises

Though I knew I didn't know pretty much anything about machine learning, I'm surprised that I've spent most of my time learning about these workloads and not as much time on the systems that run them, or the motivating about why/how we could alter systems to accommodate these workloads. This isn't necessarily a bad thing, although it does point out that there's a lot of background knowledge I don't have about machine learning, specifically neural networks. I suppose this realization will be a reoccuring theme as the project matures, since it's known in the architectural sphere that research is driven by the workloads we tailor our projects towards. The best I can do is to continue studying the neural network workloads we have so far in order to understand them.

## Revisions to 15-400 Milestones

I will need to touch-base with Pratik and Professor Mowry based on my updates from the past month and the work that Pratik's been doing to start the research effort, in order to confirm whether the proposed milestones I set out for next semester are good as is. For now, I am operating under the assumption that they are fine, and supplementary work can be done over winter break. This includes independently studying GPGPU-Sim, Pytorch, and other software that will be given to me, as well as reviewing the neural network implementations we've selected on GitHub and how to use them based on the course notes I can find for the machine learning classes I mentioned earlier.

## Resources Needed

I have most of the resources I need, except for perhaps a GPU simulator (GPGPU-Sim), which I assume will be important for our efforts at least after we are able to run the neural network workloads on the GPU for the first time. I will acquire the software to simulate the workloads on the GPU in the near future when I meet with Pratik, and hopefully get a run-down from him or Professor Mowry about their procedure for conducting research on workloads.