Project Update 2

07-400

Justin Zhang
https://www.andrew.cmu.edu/user/justinz/

1 Progress Report

Rashmi Vinayak and I have set up regular weekly meetings with our google collaborators. We had our first meeting with each other this week, where we discussed my literature review and where to begin in our collaboration. Specifically, we decided that we will no longer be working with the DLRM model and instead, we will be using the DCN model (Google claims that this model is better than Facebook's DLRM, but I remain skeptical). Apparently more simple, my new objective is to now implement the DCN model and run some beginning benchmarks on it using the kriteo dataset by the end of next week.

2 Major Changes

The only major change is the shift in studied architecture. Rather than using multilayer perceptron stacks, the DCN model uses cross and deep networks. The goal of the project is still the same however; we want to run the model on TPUs and gain insight to how we can utilize fast TPU bandwidth for theoretical and practical speed gains.

3 Meeting your Milestone

I had met my milestone which was to finish the DLRM model implementation in Tensorflow as well as run some baseline tests. Sadly, since we are switching over to the DCN model, I will have a little more work to implement the DCN model. However, google has pretty good documentation for it, and I have already done the hardwork of figuring out how to run tests, so I don't believe it should be too difficult.

4 Surprises

My main surprise was the switch in model. Other than that, nothing too surprising; the google people were pretty nice!

5 Revisions to your 07-400 Milestones

I have no revisions to make at this time.

6 Resources Needed

The current resources I need are Google Cloud TPU credits which the google people will apparently supply to me. I am currently running on a free account which is good enough so far.

References