CS554 Project Ideas

DeepLearning: Deep Learning on GPUs with Limited Precision

Overview

Deep learning has revolutionized many domains from understanding complex images and videos, to speech recognition, to self-driving cars. NVIDIA saw an opportunity in deep neural network's ability to function correctly even with low precision data types (e.g. quarter precision and half precision data, 8-bit, 16-bit), and has produced optimized GPUs to support significantly faster computations when using these limited precision data types. This project will explore the performance of deep neural networks with varying precision (8-bit, 16-bit, 32-bit, and 64-bit), as well as the quality of the learning models (precision and recall).

Relevant Systems and Reading Material

- https://en.wikipedia.org/wiki/Deep_learning
- https://developer.nvidia.com/deep-learning
- http://proceedings.mlr.press/v37/gupta15.pdf
- https://arxiv.org/pdf/1410.0759.pdf
- https://arxiv.org/pdf/1412.7024.pdf
- http://ac.els-cdn.com/S0893608014002135/1-s2.0-S0893608014002135-main.pdf?_tid=aaa9f742-9809-11e7-8f8f-00000aacb362&acdnat=1505255474_1e63be99a40f9d5c311df0340d9debf1

Preferred/Required Skills

- Preferred: CUDA, OpenCL, neural networks

Evaluation

The evaluation should be done on real datasets (both training and testing) on GPUs in the Chameleon testbed (e.g. NVIDIA P100), on a single node across 1 or 2 GPUs.