Brian Rodriguez
U0853593

Project Proposal

Memory Management in Deep Learning is an ongoing battle a developer has to face when developing deep convolutional neural networks. This problem has only grown in the recent years that these networks have gotten deeper and deeper. A virtualized Deep Neural Network (vDNN) aims to combat this problem by providing clean and efficient memory management that utilizes CUDA allocation/release system. The vDNN makes use of both the CPU and GPU memory banks, while minimizing its impact on performance. The vDNN's allocation, memory placement, movement and release of data is handled by the system architecture and the runtime system. This allows for the developer to focus more on the algorithms, rather than focusing on memory management.

In my project, I aim to implement the vDNN and do benchmark testing using the deep learning library Caffe2. I will do benchmark testing on GPU's and CPU's since I have the resources on the CHPC clucters to do so. The implemented vDNN will also aid me in my research with Rajeev; working with neural networks and GPU's are already on my agenda. The end goal of the project is also to help other developers use the vDNN that I aim to create with Caffe2. If there is a way to generalize the vDNN so that it can be used across multiple libraries, then I will try and implement it that way but for now I will implement it solely so that it can be used with the Caffe2 API.