Extending NetAdapt Neural Network Adaptation Analyses

Based on: Tien-Ju Yang et al's work in NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications.

Group Members: Nicodemus Mbwambo, Thomas Randall

Oct 25th 2019

Problem Statement:

A significant number of works in improving efficiency of design of Deep Neural Networks (DNN) has been focusing on optimizing *indirect metrics* such as the number of multiply-accumulate operations (MACs) or number of model parameters. These optimizations enable less powerful devices such as mobile phones to utilize the power of DNNs. However, performing optimization using *indirect metrics* does not guarantee a reduction in the desired *direct metrics* such as latency and energy consumption. Existing DNNs are often optimized for a single platform because manually optimizing a network for other platforms requires extensive knowledge of both network and platform architectures.

The work done by Yang et. al for NetAdapt automatically performs DNN adaptation for different platforms using a set of direct metrics (such as inference latency, memory footprint, and energy usage). The work claimed that NetAdapt could use energy metrics as part of its optimization constraints, but no energy analyses were presented in the paper. We plan to add an energy analysis of adapted networks as we verify the results reported in their work.

Planned Approach:

We will begin by gaining a basic understanding of the problem by reading the NetAdapt paper and other related works. We will review the implementation provided by Google as we work on replicating NetAdapt on the Palmetto Cluster. We are interested in using NetAdapt to adapt DNNs for a variety of GPUs and CPUs available to us on the Palmetto Cluster.

After establishing our work environment, we plan to develop new analyses to profile average power and energy consumption of adapted networks on each architecture we use. This will involve some amount of initial learning of how to perform these energy analyses on CPUs and GPUs.

Deliverables:

- Results from our analyses and steps to reproduce our results.
- A 15-minute class presentation of our work and analyses.
- A technical report regarding our work and analyses.

Scheduling and Coordinating Responsibilities:

We will begin work following our outlined approach starting October 25th. Our presentation will be ready by November 26th and our technical report will be ready by December 12th.

We plan to meet and share ideas as each of us reads related works and examines the Google implementation of NetAdapt. Both of us will use this implementation to adapt several DNNs for a variety of GPU and CPU architectures, and we will jointly develop analyses to profile power and energy consumption. To complete our deliverables, we will collaborate on the writing for the presentation and final report.