Project Webpage

https://www.andrew.cmu.edu/user/qingzhen/15740/

Project Description

Machine learning tasks are typically computationally expensive and have high energy consumption. For example, in the era of big data, machine learning tasks need to process hundreds of Terabytes of data to learn the model. Those tasks need to make use of thousands of machines in a parallel manner. Many applications on mobile devices that use machine learning algorithms usually transmit the task to a cloud-based server and later retrieve their results. However, the computation power of mobile devices is getting to the point where we can do enough computation for some machine learning tasks. For example, some newest releases of mobile devices are equipped with GPUs that are powerful enough to speed up many numerical computations.

In this project, we are going to implement a specific machine learning algorithm on a given mobile device, so that we can verify whether or not computation can be done locally at mobile device more efficiently than going through the cloud. Specifically, we are going to implement an efficient kernel expansion algorithm named Fastfood\cite{1}. It can be used to classify images or recognize objects. We will implement this algorithm both on Google Nexus 9 and on a traditional local machine. We will then compare the latency and energy efficiency between using a cloud service to classify objects and using the mobile device to perform the same task.

Goals

- 100% goal: Implement the algorithm on some real device and the algorithm can do some simple object recognition tasks. The device can also turn to a local computer to do the query. We hope to compare their performance in terms of energy consumption, latency.

- 75% goal: We are able to use a simulator instead of a real device to do the experiments. We are not sure whether we can compare the performance in terms of energy consumption, latency with a simulator for now.
• 125% goal: The implementation of the algorithm is efficient such that using the algorithm on the mobile device is faster than querying the algorithm on local computer.

Schedule

• 10.27-10.31 Get familiar with the development kit. Learn how to develop an application on a mobile device with Android system using cuda.

• 11.1-11.7 Implement an algorithm that runs on local machines.

• 11.8-11.14 Implement an algorithm that runs on Simulator.

• 11.15-11.22 Using real device to test the algorithm and finish the communication between mobile device and local computer.

• 11.23-11.28 Organize the components to make them work together. Finish milestone report.

• 11.29-12.5 Further optimize the code on mobile device to improve performance.

Milestones

By Nov 26th, we hope to finish the main components: algorithm on local computer and algorithm on mobile devices and start testing the performance.

Literature search

[1] is the main algorithm we intend to implement. We hope to evaluate the algorithms using MNIST[2] dataset. Using mobile device to do machine learning task is quite new, we haven’t found exising literature about this so far.

Resource needed

• Tegra Android Development Pack

• Google Nexus 9

Getting starting

We are currently getting started by learning how to use the development pack to write application on Android.
References
