FYP 2019-20 Detailed Project Plan
Building new AI applications for a distributed AI serving system

Waqas Ali
Supervisor: Dr. Heming Cui
Mentor: Shixiong Zhao

September 29, 2019

1 Background

Artificial intelligence is an area of computer science which focuses on granting machines the ability to act intelligently [1]. It's a vast field with limitless applications and each application has its own unique solution. Nowadays we see artificial intelligence everywhere from spam filters to movie recommendations to virtual assistants to self-driving. Moreover, customers are demanding more and more smart capabilities in their machines.

This leads to its own set of software development challenges; Nvidia summarises them with the PLASTER framework:

- Programmability
- Latency
- Accuracy
- Size of Model
- Throughput
- Energy Efficiency
- Rate of Learning

We could further divide these concerns into training and inference. Training is the phase of development where we train our model whereas inference is when we put the model in production ready for taking inputs and providing outputs.

According to research, customer satisfaction drops rapidly if it takes longer than half a second for a response. For a great customer experience, latency and throughput are really important during inference. There are many ways latency and throughput can be improved.

2 Objective

In this project, we aim to improve latency and throughput by implementing and deploying an AI application on a distributed system. As a proof of concept, we will choose an AI application which has a complex inference pipeline and distribute the tasks over a distributed system.

3 Methodology

First of all, we will choose an AI application. A few prospective ones are as follows:

- Smart Driving
- Stock Price Prediction

Second, we will develop a basic version of the application on our systems.

Third, we will deploy the model on cloud.

Fourth, when we are satisfied with the output we will convert the application in such a way that it can work on a distributed system. We will use python RPC.

Fifth, we will manually deploy the system onto cloud so we can figure out the details.

Sixth, we will devise a method to programmatically instantiate cloud resources and deploy the model. This way we can repeatedly deploy our application in the future without having to waste time.

Seventh, we will compare the latency and throughput of our application on monoliths and different sizes of distributed systems.

Our expectation is that the latency and throughput should vastly improve with a distributed system.

4 Schedule & Milestones

2019	
October	Choose & Design AI application to work on Develop a basic inference pipeline on a monolithic architecture
November	Convert pipeline to work on a distributed system using RPC
December	Manually deploy pipeline to distributed architecture on cloud
2020	
January	Programmatically instantiate cloud resources and deploy model First Presentation Detailed interim report
February	Vary deployments by cloud resources and measure latency/throughput on each
March	Enhance AI inference pipeline by adding more steps
April	Finalize implementation Final Presentation Final Report

References

[1] John McCarthy. *What Is Artificial Intelligence?* Tech. rep. Stanford University, 2007.