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 [3]. It's a vast field with limitless applications and each application has its own unique solution. Machine learning, specifically, is a subset of artificial intelligence which learns from data. [4] Nowadays we see artificial intelligence everywhere from spam filters [1] to movie recommendations [2] to virtual assistants to self-driving.

Customers are demanding smarter and smarter capabilities in their machines which leads to its own set of software development challenges; Nvidia summarises them with the PLASTER [5] framework :

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

The development lifecycle of any machine learning application can be summarized as follows:

1. Training (Learning from data)
2. Inference (Returning an output given a single input)

End-users of machine learning applications are only concerned with inference. For example, for virtual assistants training may take several days going through tonnes of voice recordings and figuring out what sounds correspond to which words. However, the customer is only concerned with sending their own voice instruction and expecting the assistant to understand their instruction. In this scenario, the latter is inference and real-time response (low latency) is expected. Moreover, the virtual assistant should be able to cope up with large traffic as well (high throughput).

Therefore, as also mentioned in the PLASTER framework [5], latency and throughput are highly important for any artificial intelligence application.

It's typical of such applications to have pipelines made out of several compute-intensive tasks. Some of them could even be run in parallel. To make such applications faster, it makes sense to distribute these tasks across several machines and that is exactly where distributed architecture could shine.

2 Objective

The project's aim is to improve latency and throughput of AI applications by implementing and deploying them on a distributed system. As a proof of concept, an AI application will be chosen which has a complex inference pipeline and it will be developed on a distributed system.

To prove that a distributed architecture can indeed improve latency and throughput, performance will be compared across systems of various specifications:

1. 1 node for n tasks (baseline)
2. less than n nodes for n tasks
3. n nodes for n tasks (optimum)

3 Methodology

1. Choose AI Application i.e. Smart Driving, Stock Price Prediction, etc.
2. Develop and run basic inference pipeline on a monolith architecture.
3. Convert the pipeline to run using RPC (remote procedure call) on one server.
4. Deploy the above on multiple servers.
5. Devise a method to programmatically instantiate cloud resources and deploy model.
6. Compare latency and throughput of model on distributed systems of various specifications.

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] Ion Androutsopoulos et al. “Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach”. In: *CoRR* cs.CL/0009009 (2000). URL: `http://arxiv.org/abs/cs.CL/0009009`.

[2] George Lekakos and Petros Caravelas. “A hybrid approach for movie recommendation”. In: *Multimedia tools and applications* 36.1-2 (2008), pp. 55–70.

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

[4] Thomas Mitchell. *Machine Learning (McGraw-Hill Series in Computer Science).* McGraw-Hill Education, 1997.

[5] David A. Teich and Paul R. Teich. *PLASTER: A Framework for Deep Learning Performance*. Tech. rep. TIRIAS Research, 2018.