Subject: Server VS. Serverless Resources and Cost Analysis of Machine Learning/Deep Learning Model Inference on Cloud Computing Instances
– Individual Project

Project Member: Zixuan Zhou 59421870

Nature: Technical

Project Summary:

Introduction

With the emerging development of Machine Learning and Deep Learning, different kinds of ML algorithms and DL models are utilized in various areas, from credit card fraud detection to AI chatbot. However, the inference process of these models can be extremely computational intensive, which requires strong data and computing infrastructure. Thus, a lot of start-up AI companies choose to make use of the cloud computing resources and cooperate with large providers like Amazon to support their services. Between server and serverless choices with different combinations of resources (CPU, GPU, Memory, Bandwidth etc.), it is significant for companies to find the most cost-effective choice for their services. Therefore, within this project, I want to make an experiment on server vs. serverless inference on the resources and cost analysis of ML/DL models with various resources.

Server VS. Serverless Methods

Server-based model inference follows the traditional Infrastructure as a Service (IaaS) model. The company needs to rent a virtual machine and deploy their model and inference service on the virtual machine. Since the company takes control of the virtual machine, they can adjust their usage of OS and network accordingly. In contrast, serverless-based model inference follows the Function as a Service (FaaS) model. The cloud service provider only gives a runtime environment to the company, the company does not have the control over the OS and network usage.

Objective

The broad objective of this project is to suggest the best model inference choice and find the tradeoff relationship between resources and amount of requests received under server and serverless circumstance. The detailed objective of this project (models, metric etc.) will be further discussed later.

Experiment Workflow

The whole workflow of the experiment would likely be the following steps:

1. Preparing test dataset and ML/DL models offline
2. Writing scripts for inferencing on AWS EC2 instances[1] (server) and AWS Lambda[2] (serverless)
3. Collecting resource usage and cost information from AWS
4. Discussing server vs. serverless choices under different conditions.
5. Analyzing the tradeoff relationships under different conditions.

Website Links
[1] Amazon EC2: https://aws.amazon.com/cn/ec2/
[2] Amazon Lambda: https://aws.amazon.com/cn/lambda/