Tyler Chang, Pablo Ordorica Wiener, Athitheya Gobinathan, Tim Chang

A Study of Optimization Techniques for Scalable Image Classification Inference on the Cloud

Applied Machine Learning for Cloud Computing - Final Project Proposal

Overview

For this project, we aim to analyze various latency reduction techniques for a deployed image model. At its core, we will be using a pretrained CNN for either OCR or image classification and running it on a GCP instance. We plan to build a simple interface around this model in the form of an API, allowing users to interact with the model by passing in an image as input and getting the classification from the inference as output. We will deploy the model as-is (i.e, without any attempts to reduce latency) and 2 more instances of this model, each with a new type of optimization to improve inference latency. Users will be able to interact with each deployment to highlight the increases in speed between each one. This project can be broken down into two main components: model optimization and inference interfacing.

Model Optimization

Deploying 3 Instances of the Model with Different Optimizations:

1. Standard model with no optimizations (control)
2. Model with optimization at the cloud architecture level (use of inference-optimization services, GPU utilization, etc.)
3. Model with optimization at the model level (quantization, pruning, etc.)

Inference Interfacing

1. API development using a service of choice (Streamlit, Flask, etc.)
2. Preprocessing of input image data (dimension resizing, etc.) before inference
3. Call to inference service
4. Return the classification to the user

Measuring Performance

Our goal is to minimize the latency of our image application from the perspective of the user. We will not only measure the total amount of time it takes to go from input to output, but we will also track the time each part takes (preprocessing, inference, network, etc.) to inform our optimization decision.

Course Topics

This project touches on all three core topics from the course: cloud computing, deep neural networks, and performance analysis. The image classification model itself, inference usage, and model-level optimization dive deep into deep neural networks. Developing the API interface for inference and model optimizations at the cloud architecture level allows for exposure to work with cloud computing technologies. Comparing the performance and scalability of each optimization technique will also allow us to apply the measurement techniques from the course.