Online Image Recognition Service

Qiancheng Fu, Lina Qiu, Cheng Zhang, Zichen Zhu

GitHub link: https://github.com/qcfu-bu/CS655-Project
Geni link: ??

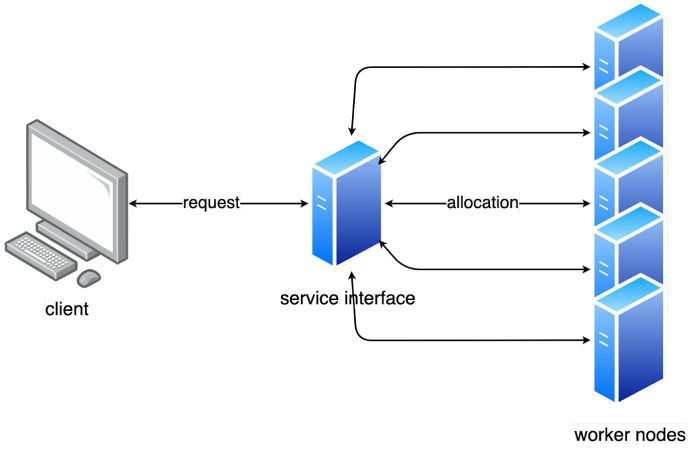
Introduction

Computational tasks in the visual domain have grown at an incredible rate. This is largely due to the breakthroughs in both hardware and algorithm design. The ubiquity of online compute services gives developers and researchers access to hardware resources that may otherwise be too expensive to purchase or maintain. We develop such an online service for image recognition tasks. Our service abstracts away the details of image recognition, allowing clients to perform inference tasks by submitting requests to our website. In order to accomodate a possibly large number of concurrent requests, we adopt a distributed architecture. Our design can provide responses to each request in a reasonable amount of time.

Experimental Methodology

Design

The following figure presents the design of our service. Instead of performing computationally expensive image recognition tasks directly, a client can submit a request through our website. Our server will break up the task, allocate relatively independent subtasks to workers, and send the results back to the client after collecting them from the workers. Since there could be a large number of clients submitting multiple requests simultaneously, we adopt a distributed architecture to resolve this problem.



Workload allocation algorithm: We assume that the time required to classify an image (one job) is constant. Our web server, which is also responsible for job allocation, maintains a table that records the number of incomplete jobs of each worker. Upon receiving a request,

the server parses and breaks the request into jobs, and then assigns each job to the worker with the shortest waiting queue. The server would update the table accordingly when it receives a response indicating that a job is complete.

Experiments

To demonstrate the scalability of our design, we plan to measure the average response delays against various number of simultaneous requests, and against various number of workers. We will also compare the performance of our service to the performance of local image recognition on a comparable hardware.

Division of Labor

Zicen Zhu: The front-end website (jQuery+ajax) , experimental shell scripts and figure scripts;

Lina Qiu: Client requests processing (Flask+REST) and worker computational resources management.

Cheng Zhang: Socket programming between worker machine and web server. Including data (images and results) transmission and load balance control (co-design with Lina)

Qiancheng Fu: Image recognition algorithms and their corresponding network interfaces