Design Space Exploration of Model Serving

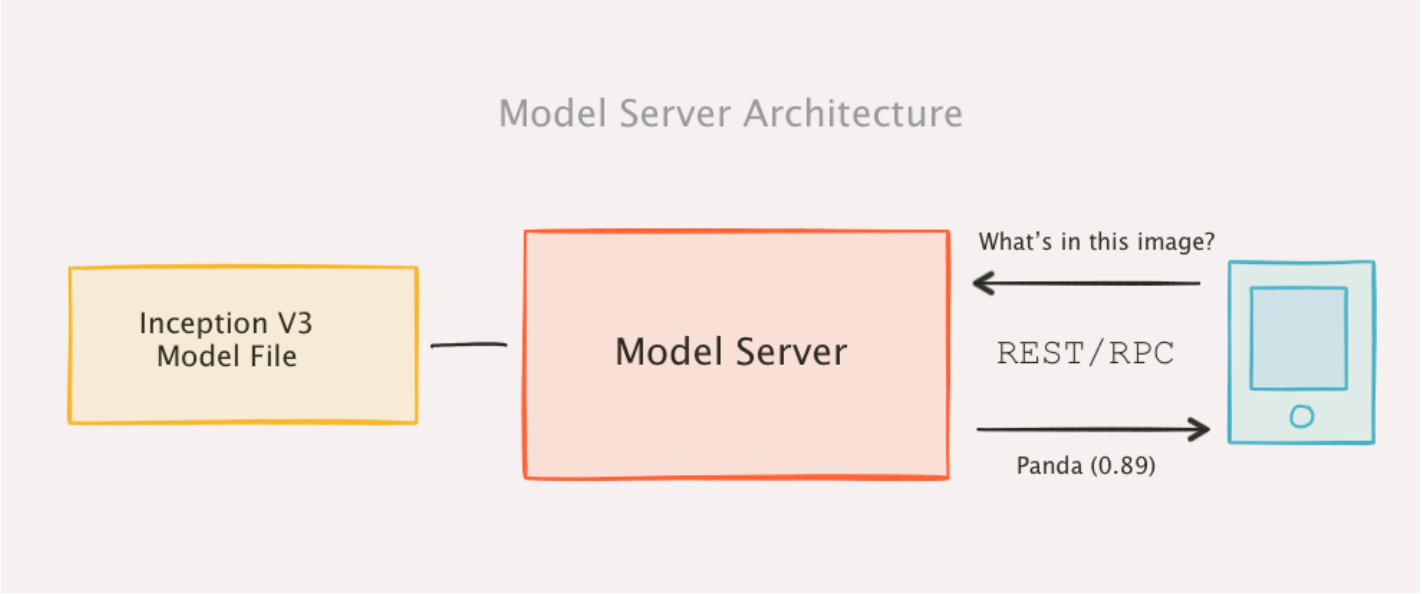
Course Project
CSCE 790
(Machine Learning Systems)

How projects will be evaluated?

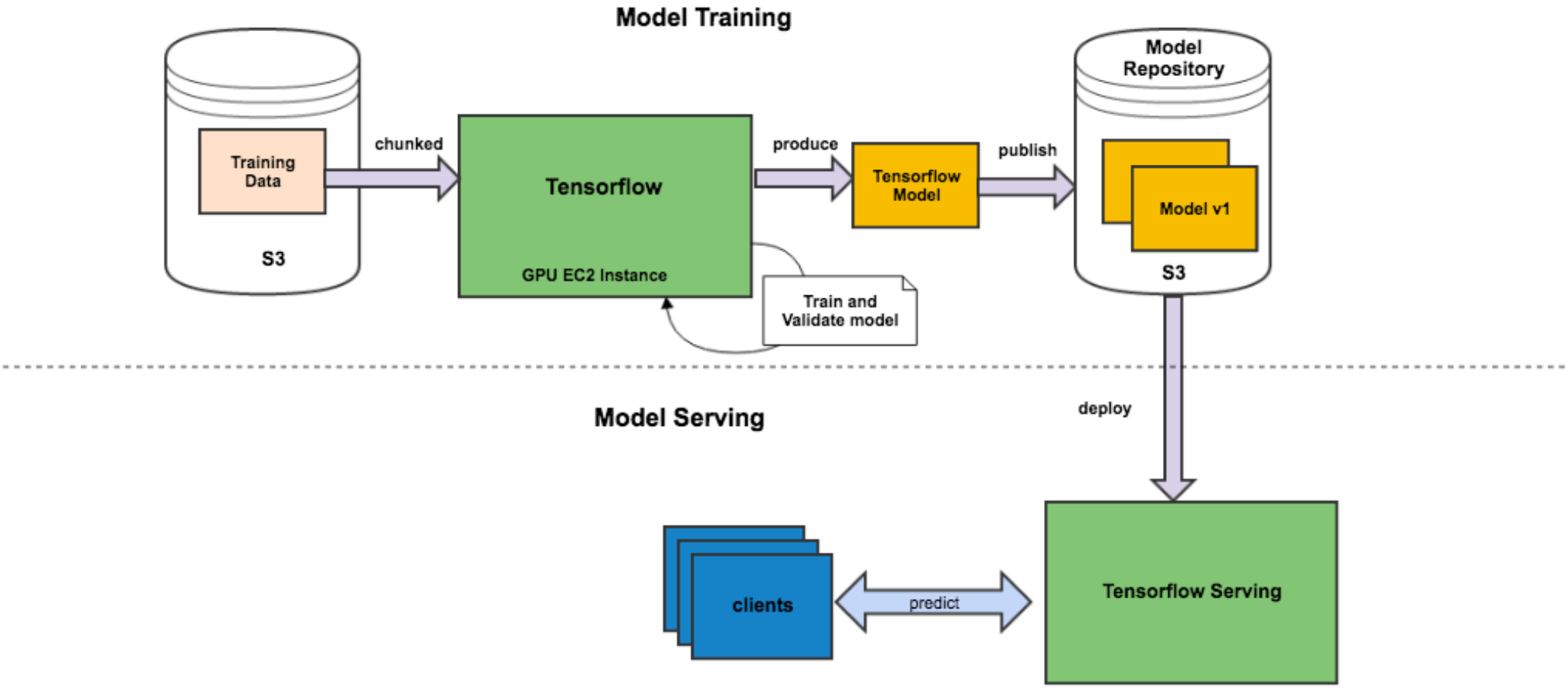
- Team up with other 2 students, so each team 3 persons
- Select one project
- No communications between the two teams
- Every teammate should be able to demonstrate her/his contribution
- The outcome will be evaluated based on the quality of the results, report, and final presentation.
- The final report is an iPython notebook that has documentation, results, comparisons, discussions, and related work.
- 60% of your final mark will be evaluated based on the course project.

Project description

- How you can decrease latency of model serving (TF serving) by changing some of parameters like caching, remote procedure call protocol, etc.



Model serving



Project goal

- The aim of the project is to perform design space exploration of model serving.
- The goal is to understand how the choice of configurations in the deployment environment of model serving can influence user perceived latency of model predictions of DNNs.

Selecting model server

- You first need to select a model server, e.g.:
  - TensorFlow Serving
  - Clipper
  - Model Server for Apache MXNet
  - DeepDetect
  - TensorRT
  - etc.

Deciding the configuration space

- You need to then choose the configuration space you would like to explore.
- For this, you need to select specific configuration options you can vary on the server side. E.g.:
  - CPU frequency
  - RAM
  - Batching
  - Number of models running concurrently
  - Number of threads
- References:
  - https://www.tensorflow.org/serving/serving_advanced
  - https://www.tensorflow.org/api_docs/serving/struct/tensorflow/serving/server-core/options

Selecting specific DNN architectures

- Select few pre-trained DNN architectures that fit onto your hardware platform, e.g.:
  - Any pre-trained CNN architecture
  - Use available implementations, e.g.,: https://github.com/tensorflow/benchmarks

Deciding about workload

- Choose 2 different workloads from existing datasets, e.g. UCI repository, or other available datasets
  - Image
  - Time-series
  - Text
  - etc.

Generating load

- One difference about this project comparing with other project is that you need to write/use code/script that generate different load patterns to the sever.
- For example you can generate loads with different stress level: e.g., light, medium, high
- This is part of the configuration space as you can imagine.

Start measurements

- Once you decided about the configuration space, you need to determine the configurations that you want to measure.
- At this stage you need to discretize the continuous variables.
- And think about using a sampling strategy, e.g., random sampling, or possibly Full factorial design
  - https://en.wikipedia.org/wiki/Design_of_experiments
- Do not forget that you need to measure both Inference time and energy consumption for each configuration

Analyzing data

- Once you measured configurations, you need to dig into data and find interesting trends.
  - You could look into Pareto-optimal configurations
  - You could find whether the optimal configurations in one DNN architecture is also optimal in other architectures, if not dig into and find out why.
  - You could look into correlation measures across different workloads
  - You may want to have a look at this to get some idea what kinds of analyses you may want to perform: https://arxiv.org/pdf/1709.02280.pdf

Final point

- Use your creativity when it comes to analyzing the results, try to surprise me!
- If you find a very interesting observations and dig into it by providing some insight, you will then get a good score!
- If you also produce very good results, you may also want to think about a potential paper, it’s optional, but I strongly recommend it.