Milestone 3

Learning Goal

In this milestone we will aim to turn our experimentation of achieving an optimal pipeline optimizer that will choose the best pipeline configuration, based on collected data. This milestone where one may design simple rules for optimization or use an ILP solver for c optimization problem is NP-hard so no point designing a provable approach—the best w efficient heuristics. Use of agents will be a bonus.

What is an Optimizer? The Optimizer is the auto-configuration module that will periodica load from the monitoring system, (2) predicts the next reference load based on the obse obtains the optimal configuration in terms of the variant used, batch size, and number o timestep, and finally (4) applies the new configuration using Kubernetes.

For example, if there are two available options per model in a analysis pipeline, the Opt choose more accurate models with small batch sizes to ensure low latency but will cho higher loads which have more replication and larger batch sizes to ensure high throughp

Tasks

T1. Define the accuracy of your pipeline. Note accuracy of a pipeline must be defined as accuracy of different stages of the pipelines. Define this combination. State any assump State any error correlations that you may identify. If there are model drifts, please define

T2. State the optimization problem you are solving. Each of the group looks at an optim paper assigned to them.

Understand the problem, map it to your pipeline setting and state it in your own words.

Video: Francisco Romero, Qian Li, Neeraja J Yadwadkar, and Christos Kozyrakis. INFa less inference serving. In 2021 USENIX Annual Technical Conference (USENIX ATC 21
ID Card: See paper Jeff Zhang, Sameh Elnikety, Shuayb Zarar, Atul Gupta, and Siddhar Dealing with fluctuating workloads in machine-learning-as-a-service systems. In 12th US Topics in Cloud Computing (HotCloud 20), 2020.
Audio: Jashwant Raj Gunasekaran, Cyan Subhra Mishra, Prashanth Thinakaran, Bikash Kandemir, and Chita R Das. Cocktail: A multidimensional optimization for model serving NSDI, pages 1041–1057, 2022.
Q&A: IPA: Inference Pipeline Adaptation to Achieve High Accuracy and Cost-Efficiency, Razavi, Mehran Salmani, Alireza Sanaee, Tania Lorido-Botran, Lin Wang, Joseph Doyle https://arxiv.org/abs/2308.12871v3

As part of the optimization problem, describe how different data values are collected in y
optimization problem.

T3. (Bonus) Show how you solved the optimization problem. Note if you use Gurobi solv
academic use. If you just used rules state those.

T4. (Bonus) Use an agentic framework to solve the optimization problem.

Deliverables:

1. A demonstration of your project. This should showcase the different pipeline variants
collect different metrics.
   In your demonstration also showcase a functional, documented Dockerfile to run your