An AI Benchmarking Suite for Microservices-Based Applications

In the realm of AI applications, the deployment strategy significantly impacts performance metrics. This research internship aims to investigate and benchmark AI applications in two predominant deployment configurations: monolithic and microservices-based, specifically within Kubernetes environments.

The central question revolves around understanding how these deployment strategies affect various performance metrics and determining the more efficient configuration. This inquiry is crucial as the deployment strategy plays a pivotal role in the operational efficiency of AI applications.

Currently, the field lacks a comprehensive benchmarking suite that evaluates AI applications from an end-to-end deployment perspective. Our approach includes the development of a benchmarking suite tailored for microservice-based AI applications.

This suite will capture metrics such as CPU/GPU/Memory utilization, interservice communication, end-to-end and per-service latency, and cache misses.

Voraussetzungen

- Familiarity with Kubernetes
- Familiarity with Deep Learning frameworks (e.g., PyTorch or TensorFlow)
- Basics of computer networking

Kontakt

Email: navid.asadi@tum.de

Betreuer

Navidreza Asadi