Aaron Pham

aaron.pham.2k21@gmail.com
linkedin.com/in/aarnphm
aarnphm.xyz
+1(415)-535-8841

Experience

BentoML

Machine Learning Engineer
05/2022-04/2023
On-site - San Francisco, CA
05/2023 - Present
Remote - Toronto, CA

Authored and lead *OpenLLM*, an open-source platform for operating large language models (LLMs) in production. Implemented state-of-the-art inference optimisation, such as continuous batching, quantisation, sliding context windows using *PyTorch*, optimised inference kernels with *OpenAI Triton*.
Maintained and contributed features improvement to HuggingFace's ecosystem, including *optimum, transformers, accelerate, PEFT*. Added NVIDIA Triton Inference Server support for BentoML Runners. Added and designed general improvement for BentoML.
Technologies: Python, PyTorch, OpenAI Triton, BentoML, Transformers,.

Software Engineer
01-08/2021
Remote - Vietnam
08/2021-04/2022
Remote - Hamilton, ON

Added a Docker image releases management, provides supports for multiple Linux OS. Implemented and added supports for every major ML frameworks.
Proposed and designed a easy-to-use CI system. Added robust testing and structured typing to BentoML's codebase.
Improved and enhanced internal batching algorithms that increase general throughput by **2.2x**.
Provided supports for label creation implementation in Golang for BentoCloud, mentored Fall 2021 Batch MLH Fellows.
Technologies: Python, Pyright, Docker, GitHub Action, Bash, Golang, AWS EC2

MLH Fellowship Fellow
09-12/2020
Remote - Hamilton,ON

Wrote a gRPC metrics exporter and added Prometheus support for YataiService, BentoCloud's predecessor. Added GPU supports for BentoML. Selected as one of **80** fellows to be part of MLH Fellowship Open Source Batch 1.
Technologies: Python, PyTest, Prometheus, Docker, TypeScript

Project

OpenLLM

github
mle @bentoml
06/2023 - Present

OpenLLM is an open-source platform designed to facilitate the deployment and operation of large language models (LLMs) in real-world applications. With OpenLLM, you can run inference on any open-source LLM, deploy them on the cloud or on-premises, and build powerful AI applications.

Key features include support for a wide range of open-source LLMs and model *runtimes*, Flexible serving APIs over *HTTP/gRPC* endpoint, first class support for *LangChain*, *LlamaIndex*, *Hugging Face*, allowing you to easily create your own AI applications by composing LLMs with other models and services, **Fine tune** your custom LLMs, with optimisation built-in, such as *quantisation*, *continuous batching*, *optimised inference kernels* based on latest research in LLM inference.

Streamlined deployment by automatically generate Docker images or **serverless** deployment via BentoCloud., supports token streaming via **server-sent events** (SSE).

Technologies: PyTorch, OpenAI Triton, CUDA, LangChain

onw

github
hack the north 2021
remote

A real-time navigation tool that harnesses the power of community to make community safer.

Implemented **Gaussian Mixture Model** to find the safest path between different locations, trained on past assault data provided by Toronto Police Department from 2015 onwards.

Integrated with Google Maps API to show heatmaps of avoided zone. Implemented designs from *Figma* in **React Native** and **Expo**, shipped with *AWS Fargate*

Awarded: Finalists at Hack the North 2021.

Technologies: AWS, React Native, TypeScript, GraphQL, Apache Spark MLlib

dha-ps

github
contractor @gpayvn
06/2020-12/2020

Designed and created a product-based price recommender API.

Implemented and trained a custom *BERT* model for sentence similarity in PyTorch, served behind a *FastAPI* app.

Implemented a leaky-bucket ingress controllers in *Golang* and deployed with **Kubernetes**, hosted on Google Cloud Platform (GKE).

Added unit and end-to-end testing for both Go and Python backends.

Technologies: Python, Go, Kubernetes, Docker, PyTorch, Google Cloud Platform

Education

McMaster University
Expected Grad: 12/2025

Bachelor of Engineering and Management Software Engineering Co-op.