Apoorve Kalot

MLOps Engineer

Tran5wert Apoorve-Kalot ✉ apoorve.kalot.ep17.iith@gmail.com ☎ 8949437977

Experience

MLOPS ENGINEER | Saarthi.AI

August 2022 – August 2023 | Bangalore, India

Collaborating with Research Scientists/Research Engineers, responsible for maintaining existing model deployments on Cloud platforms (AWS/Azure) and managing the end-to-end pipeline (CI/CD). Additionally, integrating open-source tools to enhance the current infrastructure in the following manner:

- → Developing **'Conversational SDK'** offering Speech-to-Text (STT) and Text-to-Speech (TTS) services in **11 Indian Languages**, leveraging **gRPC**, **Kubernetes**, **PostgreSQL**, web-sockets, and **Prometheus/Grafana**. Achieved **20 % latency reduction**.
- → Creating a central **LLM server** using **Text-generation-Inference Engine** and Docker-swarm for internal automation, utilizing **Lang-chain**.
- → Implemented an end-to-end training pipeline for multiple **NLU** models using **MLFlow**, **DVC**, **Azure Blob storage**, and **GitHub Actions**. Improved training efficiency by **30 %**.
- → Configured **CI** for an ASR model repository using **Jenkins**, **SonarQube**, and **ACR**. Reduced build failures by **15 %**.
- → Utilized Nvidia's **Triton Inference Server** for staging model deployment with **TensorRT/openVINO** acceleration, achieving a **15 ms latency drop**.
- → Conducted infrastructure stress testing using Nvidia's **Model Analyzer**, deploying regular model updates on **Azure AKS** and **AWS EKS**.

ASSOCIATE SOFTWARE ENGINEER | Legato Health Technologies

June 2021 – May 2022 | Bangalore, India

- → Being Recruited as member of Data science team, which is responsible for generating business value from the data from different services offered by Anthem.
- → Worked on creating business insights for current ”Provider's portal”, which will be used to optimize latter one, using AI and ML techniques

DEEP LEARNING INTERN | Tech IndusOne Services

June 2020 – August 2020 | Chandigarh, India

- → As an intern, my work consisted of completing projects on AWS virtual EC2 instance along with mentoring Junior interns for their respective Projects
- → Drowsiness Detection. : Detecting a person's drowsiness on live video stream using OpenCV and , and deploying on web-page using Flask.
- → Invisibility of People using Pre-trained Mask-RCNN model. : Similar to Harry Potter's Cloak of Invisibility.

Courses

Data Science/ ML/ IR

Data Science Analysis • Artificial Intelligence, Data Mining • DBMS • MOOCs [Coursera Deep Learning Specialization • Coursera TensorFlow Developer Specialization]

Education

IIT HYDERABAD

B.Tech. in Engineering Physics, with minor in Entrepreneurship

May 2017 - Aug 2021

Cum. GPA: 7.56 / 10.00

Skills

PROGRAMMING

Experienced:

C • Python • $\LaTeX$

LIBRARIES/FRAMEWORKS

Tensorflow • Joblib • ONNX • locust • OpenVINO • TensorRT • Numpy • Pandas

TOOLS/PLATFORMS

Azure (Blob-storage, ML Studio, ACR, AKS) • AWS (EC2, EKS, S3) • Docker-swarm • Kubernetes (Minikube, K8) • Git • DVC • AirFlow • Inference Server (Triton by Nvidia, Text-Generation-Inference by Hugging Face) • MLFlow • Prometheus/Grafana • Jenkins • SonarQube

Position Of Responsibilities

Placement Coordinator @ Office of Career Services| (2020-2021) Club

Coordinator @ EPOCH, Machine Learning Club of IIT Hyderabad | (2020-2021) Club

Coordinator @ Rang De Manch, Dramatics Club of IIT Hyderabad | (2018-2019)