Bhooyas Kapadia

Mumbai | +91 7045361098 | bak05102000@gmail.com | in | |

PROFESSIONAL EXPERIENCE

Colaberry Inc. **Work From Home**
Machine Learning Engineer *January 2025 - Present*

- Developed an automated benchmarking script for evaluating model performance across **FastAPI, MLServer, and Triton**, supporting multiple frameworks such as **PyTorch, TensorFlow, and scikit-learn**.
- **Developing Argo workflow template** for automating batch model deployment, streamlining the deployment process.

QUANTIPHI, **Mumbai**
MLOps Engineer *August 2022– January 2025*

- Containerized ml applications and deployed them on Kubernetes using helm charts on different cloud providers as well as on prem ecosystem.
- Deployed a LLM model on Triton using Nemo Inference Microservice resulting in reducing latency by approximately **20%**. Additionally, collaborated closely on deploying a RAG pipeline for suggesting drinks from the menu.
- Worked closely in multiple workshops to optimize training time using technologies like Distributed Data Parallelism, Model Parallelism, Slurm and Enroot. The efforts resulted in reducing the training time from **30%-50%** based on the use case.
- Led the setup of a **DGX Cluster** comprising multiple DGX nodes and a SuperMicro as the head node using **Base Command Manager**.
- Converted models to **TensorRT engines** and deployed them on **Triton Inference Server** to server 10K users per day with latency of 5 secs per query.
- Collaborated closely to mitigate VM vulnerabilities and keep the cost under track for GCP.
- Additionally, I have utilized Prometheus and Grafana to monitor Docker utilization. I have also written a Bash script, resulting in a **30%** reduction in effort.
- Set up and managed multiple Kubernetes clusters with **NGINX Ingress Controller** to streamline external access to services, enhancing scalability and load balancing. Implemented Horizontal Pod Autoscaling (HPA) to support applications serving up to 100K users per day.
- Additionally, responsible for scoping new projects, defining deliverables, and establishing timelines to ensure alignment with business objectives and resource availability.

SFITY INDIA, **Work From Home**
Machine Learning Intern *April 2021-June 2021*

- Developed a chatbot by fine-tuning the **GPT-2 345M** parameter model for the company.
- Integrated this developed bot into the companies website.

BEING DIGITAL, **Work From Home**
Full Stack Machine Learning Engineer, Internship *December 2020–March 2021*

- Developed a fully functional **mobile application** utilizing Convolution Neural Networks trained with **Transfer Learning** for fruit and vegetable classification.
- Implemented functionality to suggest recipes based on the classified fruit and vegetable.

SKILLS

Cloud: GCP, AWS, AZURE, OCI, On Prem (DGX).
MLOps: Docker, Kubernetes, Helm, Slurm, Base Command Manager, Python, Keras, Pytorch, ONNX, LoRA, GenAI, Distributed Data Parallelism, Triton, TensorRT, Large Language Models, Tensort-LLM, Github Workflow, Azure Devops, Netron, Prometheus, Grafana, Terraform, MLFlow.
Other: Bash, HTML, CSS, JavaScript, Flask, MySQL, Pandas.

PROJECTS

tinyllama Fine-Tunning Successfully fine-tuned **TinyLlama (1.1B parameters)**, a lightweight language model, using **LoRA** adaptation technique on the **Databricks Dolly 15k dataset**. The implementation focused on enhancing the model's ability to understand and respond to context-based instructions while maintaining computational efficiency. This optimization resulted in improved performance for instruction-following tasks while keeping resource requirements minimal.

StoryNet Designed and implemented a custom transformer architecture from scratch using PyTorch, resulting in StoryNet—a specialized neural network for creative text generation. Successfully trained the model to generate coherent short stories, demonstrating expertise in transformer architecture, sequence modeling, and natural language generation.

ConditionalVAE Implemented a **Conditional Variational Autoencoder (ConditionalVAE)** that generates handwritten digits based on specified class inputs, trained on the **MNIST dataset**. Demonstrated the ability to control the generation process through conditional parameters, highlighting expertise in deep generative modeling, latent space manipulation, and complex neural architecture design for controlled image synthesis.

ACHIEVEMENTS

Think Tank Award Recived Think tank Award in Quantiphi.

Research Paper on Text Summariser Published a Paper in INDIACom - 2020

CERTIFICATIONS

NVIDIA-Certified Professional: AI Infrastructure **Dec 2024 - Dec 2025**

NVIDIA-Certified Professional: AI Operations **Dec 2024 - Dec 2025**

Google: Associate Cloud Engineer **Dec 2022 - 2025**

Oracle: Oracle Cloud Infrastructure 2023 Foundations Associate **July 2023 - 2025**

Oracle: Oracle Cloud Infrastructure 2023 AI Certified Foundations Associate **December 2023 - 2025**

EDUCATION

BTech Information Technology, **MPSTME, Mumbai— NMIMS**
CGPA: 3.87/4.00 *2019-2022*

Diploma in Computer Technology, **Vartak Polytechnic, Mumbai— MSBTE**
Percentage: 83.06% *2016-2019*

X **M.K.V.V.I.V, Mumbai— CBSE**
CGPA: 8.8/10 *2016*