Shreyas S K (LinkedIn) (GitHub) (Youtube) Contact - +91 7019667673

Machine Learning Engineer skshreyas714@gmail.com

SKILLS

Libraries/Frameworks	TRL, Diffusers, PyTorch, Numpy, Pandas, Sklearn, Numba
Generative AI	Mistral 7B, Llama2, Stable Diffusion (1.5, 2.1), Kandinsky, DeepFloyd
Cloud Expertise	AWS : EKS, EC2, S3, Sagemaker, Bedrock
LLM Tools	LangChain, LlamaIndex, Nemo Guardrails, Haystack
MLOps	MLFlow, Github Actions, CML, Docker, Seldon
Distributed Computing Tools/ Accelerators/ Inference Engines	vLLM, accelerate, DeepSpeed, Pytorch DDP, Nvidia Triton, TensorflowLite, LoRAX

EXPERIENCE

Khoros India R&D Private Ltd, Bangalore - *Machine Learning Engineer*

Feb 2022 - Present

Products - Khoros Care, Khoros Communities & Khoros Marketing

- Developed **Content Genie** application, which handles the complete cycle of creating a social media post. It's features include **editing images** such as **fill, replace or remove** specific objects in the image by using natural language prompts with **Stable Diffusion** Inpainting and **Kandinsky** Inpainting, generating social media **captions** and **hashtags** in multiple languages with **Mistral 7B LLM**
- Enhanced search performance of **RAG pipeline** in **Helix,** An AI powered Knowledge system used to provide detailed response to queries from Khoros Bot. Used **Langchain, LlamaIndex, Google Vertex**, **Open AI** to build this pipeline.
- Developed and deployed a **Stylized Image Captioning** model which generates captions based on personality traits by considering multi-modal features (image, text). Improved the model performance (CIDEr, BLEU) further by **10-15%** using **Reinforcement Learning** techniques (**Adaptive SCST, REINFORCE, PPO**)
- Designed an **LLMOps** pipeline for **Agent Auto Assist,** which helps agents to write responses quickly by utilizing AI based autocomplete suggestions based on context of conversations. **Mistral-7B LLM** model was fine tuned for a large corpus of agent-customer conversations.

Facebook, Hyderabad - *Machine Learning Engineer (Contract)*

Jun 2020 - Jan 2022

AI based Pipeline Automation for Facebook VA/VR devices (Portal and Oculus)

Assistant Platform Engineering Team, Facebook

- Designed the pipeline with open source projects for Object detection (Detectron 2), Optical Character Recognition (OCR), Textual semantic similarity, Text-toSpeech (TTS), Template matching (OpenCV)
- Implemented image binarization technique based on research paper as a pre-processing step for OCR, optimized the performance using Scipy's Low Level Callables and Numba Jit compiler which reduced the computation time by 97%
- Involved in data collection, data pre-processing, model selection and training, deployment to production of an Object detection model (Detectron 2) to detect AR Effects in WhatsApp and Messenger Video calling
- Implemented K-means Clustering to segregate VA commands according to their respective domains. Collected VA responses from database, fine tuned and quantized ALBERT transformer for Named Entity Recognition (NER) to extract entities from VA response for advanced processing
- Analyzed various Transformer models fine tuned for STS task (DistilBERT, BERT Large NLI, Laser). DistilBERT was selected and deployed based on similarity score and latency
- Built and deployed a CLI client which was responsible in reducing execution time by 50%, reduced test case failure rate by 25% due to improved quality of speech, image transcriptions and advanced processing (Unsupervised Clustering, Token Classification)

HCL Technologies, Bengaluru — *Lead Engineer*

July 2020 - Jan 2022

Orchestrating ML Model Life cycle with MLOps Open-source pipeline

AI COE Team, Next.ai Lab, HCL Technologies

- Expertise in ideation and design of MLOps Architecture using open-source tools and frameworks
- Crafted CI, CD, CT pipelines with MLOps best practices using MLFlow, DVC, Github Actions, CML, Airflow, Kafka, Docker, Kubeflow and Kubernetes
- Achieved Continuous Training by collecting data from Kafka, scheduling jobs in Airflow pipelines and tracking the experiments using MLFlow
- Implemented continuous monitoring with Elasticsearch, Kibana and Grafana
- Integrated a feature store (Feast) to operationalize analytics data for model training and online inference
- Trained and deployed Kernel Shap explainability model from Alibi Explainer module
- Expertise in deploying ML models using Seldon deploy, Openshift's Source to Image (S2I)

HCL Technologies, Bengaluru — *Senior Software Engineer*

July 2019 - June 2020

Quantization and Benchmarking COVID-19 Image Classification on ARM Mali GPU

AI COE Team, Next.ai Lab, HCL Technologies

- Collected Lung X-Ray images for 14 different chest abnormalities, applied spatial augmentation techniques to increase size of the training set.
- Trained an Image classification based model using modified Densenet-121 architecture, achieved 96% classification accuracy due to spatial augmentation
- Applied weight clustering and quantization techniques to make the model suitable for Edge deployment
- Model size reduced by 50% to 75% with 5% and 15% compromise in accuracy due to quantization for FP16 and INT8 precisions respectively
- Developed an android application to benchmark inference time on ARM's Mali GPU
- Hardware acceleration improved by 30% and 80% when run on FP-32 and FP-16 TFLITE models on GPU respectively

Cellstrat, Bengaluru - *Deep Learning Intern*

Sep 2019 - March 2020

Webinars at Cellstrat Meetups (YouTube Playlist)

- Orchestrating ML Model lifecycle with MLOps open source pipelines
- NLP with ELMo and FLAIR Embeddings
- Bayesian Optimization - Reinforcement Learning
- Stylish Image Caption Generator with Reinforcement Learning Techniques
- Reformer - The Efficient Transformer

PUBLICATIONS

Shreyas and Kameshwar (2021). "**Diagnostic Decision Support for Medical Imaging and COVID-19 Image Classification on ARM Mali GPU**". **IEEE Global Communications Conference 2021 - Edge-AI-IoT** - Published on 7th December 2021

Shreyas and Dey, (2019). "**Application of soft computing techniques in Tunneling and Underground excavations: State of the art and future prospects"**. **Innovative Infrastructure Solutions**. Springer Publications. Published on 25th September 2019

Shreyas and Dey, (2019). "**Comparative Efficacies of Ensemble methods and Hybrid Intelligent model in Predicting Penetration Rate of Tunnel Boring Machine". International IACMAG Symposium, IIT Gandhinagar 2019**. Published on 7th March 2019

EDUCATION

Indian Institute of Technology, Guwahati — *M.Tech in Civil Engineering* **(9.79/10)**

August 2017 - May 2019

PES University, Bengaluru — *B.Tech in Civil Engineering* **(9.14/10)**

August 2013 - May 2017

ACHIEVEMENTS AND AWARDS

- Secured **3rd** position in **SEBI Hackathon** out of 1200 teams on the theme "**Detecting misleading claims made by Finfluencers in Securities market with Gen AI**", organized as part of Global Fintech Fest 2023 (GFF), Mumbai.
- Named one among **top 25 teams** out of 400 teams in the **Autonomous Agents Hackathon** organized by Lablab.ai, Weaviate, LamaIndex, BabyAGI.
- Secured **4th position** out of 60 teams in **Generative AI** Hackathon on theme **"AI for India"** organized by **WhatFix, AWS, Entrepreneur First**
- Secured **A+** grade in **AI for Medical Imaging Analysis**, 6 month semester course offered by **IISc CCE 2022**.
- Secured **A** grade in **Deep Reinforcement Learning**, 6 month semester course offered by **IISc CCE 2023**
- **ERS Innovation Champion** and **ERS Business Champion** for outstanding performance in Facebook Client Project
- Secured **first place** in the merit list of **M.Tech Civil Engineering** department from IIT Guwahati
- One among **top Ten merit students in B. Tech Civil Engineering** from PES University