Srimanth Agastyaraju

Bloomington, Indiana (Willing to relocate) | **Mobile**: (812) 803-5167 |
Email: srimanthagastyaraju.98@gmail.com **| LinkedIn**: linkedin.com/in/asrimanth/ **| GitHub:** github.com/asrimanth
Availability: https://calendly.com/asrimanth/30minmeeting

AI Engineer | Deep Learning Research Engineer | Full Stack Development Analyst | Software Engineer

PROFILE OVERVIEW

- Result-oriented and experienced AI engineer specializing in Computer Vision, NLP, Deep Learning, and Software Engineering.
- Full-stack developer proficient in **Python**, **Rust**, and JavaScript, with expertise in building scalable web applications and RESTful APIs.
- Proficient with Gen AI systems such as **GAN**s, **Stable Diffusion** models (SDXL, SDXL Turbo), and LLMs (Mixtral, LLaMA3, Phi3, GPT4).
- Proven track record of innovation in deep learning demonstrated through contributions to **open-source** projects and research.
- Skilled in containerization and orchestration technologies like **Docker** and **Kubernetes** for efficient application deployment
- Proficient in **CI/CD** pipeline setup using Azure DevOps and GitHub Actions for automated builds, tests, and deployments.
- Excellent **team player** with good analytical, strategic planning interpersonal, and communication skills along with highly motivated, enthusiastic, and a self-starter attitude.

TECHNICAL EXPERTISE

Python | FastAPI | LlamaIndex | LangChain | RAG | NLTK | SpaCy | Wandb (MLOps) | Qdrant | Amazon Web Services (AWS) | Git | Github | Linux (bash, zsh) | Jupyter | Jira (Agile Workflow) | Scikit-Learn (sklearn) | PyTorch | OpenCV

WORK EXPERIENCE

Software Engineer | eBS Minds IT Inc. – Remote, USA **Mar 2024 - Present**

- Designed RESTful APIs for a web application using **FastAPI**, SQLModel (Pydantic), and Python using Strategy and Factory patterns.
- Facilitated secure database operations using **ORM** with SQLModel and **Postgres**, emphasizing data-driven processing.
- Optimized **Apache Airflow DAGs** for 10+ workflows, reducing data errors and saving 20 hours monthly. Added 5 data sources in a quarter.
- Built a CI/CD pipeline using **Azure DevOps** and **GitHub Actions**. Configured auto builds, tests, and deployments for smoother delivery.

AI Engineer | Easel AI – Remote, USA **Aug 2023 - Feb 2024**

- Re-engineered existing image-generative AI to an **SDXL**-based pipeline, greatly improving image-generation quality.
- Developed and deployed a **RAG pipeline in Python** using LangChain with internal text prompts to improve recommendations by 20%.
- Finetuned various foundational models such as **SDXL**, **SDXL Turbo**, and **BLIP** for user image generation and captioning tasks.
- Initiated a **Rust-based** recommendation system for prompt recommendation using FastEmbed, Qdrant, and Unsupervised ML systems.
- Developed and deployed a profanity detector to moderate user input, resulting in a 15% reduction in redundant generations.
- Researched and implemented **AI-based filters** using SDXL ControlNet, ensuring the preservation of crucial details like facial features.
- Integrated a **Mobile Image segmentation** pipeline to segment facial data from an image for 30% faster inference.
- Curated a user safety and content moderation dashboard in **Python (Streamlit)** to prevent misuse and deployed it using **Docker** and **GKS**.

Deep Learning Research Engineer | Indiana University Bloomington – Bloomington, Indiana **May 2022 - Jul 2023**

- Orchestrated a 3D reconstruction deep learning model with 10 million parameters on multiple compute nodes.
- Conducted extensive experiments to augment the performance of existing networks by 10%, with experiment tracking using wandb.
- Augmented the coco dataset for the multi-focus imaging task and trained a Super-Resolution **GAN** model for image enhancement.
- Trained multiple **image super-resolution** models as part of the 3D reconstruction network to enhance the model performance.
- Researched and developed Computer Vision models on HPC using Slurm and Linux from 10+ research papers.

Full Stack Development Analyst | Accenture – Hyderabad, India **Jul 2020 - Jul 2021**

- Tech Lead/developer involved in the technical architecture for microservices application, also leading the analysis, design, and coding.
- Worked on data migration of Product Feed into GCP storage using dataflow, DAGs using Apache beam, and Airflow.
- Led the team under tight deadlines and delivered internal design and code for a single-page checkout application.
- Migrated legacy Search JSP pages to micro-frontend using React/Redux and Node JS.
- Responsible for production support, writing test scripts using Gatling IO, and integrating them with the Jenkins CI/CD pipeline.

Software Engineer Intern | WebileApps Pvt Ltd – Hyderabad, India **Jan 2020 - Jun 2020**

- Revamped an Android application in MVVM architecture following scrum methodologies and reduced bugs by 30%.
- Expedited parsing of pages by 40% with the help of OCR, Natural Language Processing, and Machine Learning lite models.
- Spearheaded a Proof-of-Concept app for 3D model rendering based on keyword retrieval from the OCR output.

ACADEMICS & ACCREDITATIONS

- **MS in Data Science** - Indiana University Bloomington - Bloomington, Indiana **Aug 2021 - May 2023**
- **B. Tech in Computer Science** - Sreenidhi Institute of Science and Technology - Hyderabad, India **Aug 2016 - Jun 2020**

PROJECTS HANDLED

HuggingFace Open-Source Contributions | PyTorch, HuggingFace, ONNX **Mar 2024**

- Fixed training resume problem for SDXL Consistency Distillation on FP16 – PR #6840, for diffusers library on HuggingFace.
- Added ONNX runtime support to RegNet – PR #833 - Improved the pre-trained model's availability and ease of use, for optimizers.
- Fixed gradient checkpointing and use_cache bug for BLOOM - PR #21956, a large language model, for transformers.

Image Captioning using Pretrained feature extractors | PyTorch, NumPy, OpenCV, wandb **Dec 2022**

- Assembled an LSTM model and a pre-trained ResNet and ResNeXt model to transcribe images.
- Tracked experiments and visualized intermediate results using the Weights and Biases (wandb) platform.

Homicide Data Analysis using PySpark| PySpark, NumPy, Pandas, Plotly, Data Analysis **Nov 2022**

- Established a data pipeline to analyze more than 600,000 rows in PySpark to reduce processing time by 60%.
- Employed IU's Jetstream 2 cloud infrastructure to run the PySpark SQL framework on tiny hardware.
- Visualized key findings using Plotly, including geographic, to communicate insights and support decision-making.