Devam Sheth

devamsheth21 | devamsheth21 | devamsheth20@gmail.com | 602.768.7274

Experience

Software Development Engineer (SDE)-AI/ML Mar. 2025 – Present
Amazon Web Services (AWS) – Bedrock Data Automation *Bellevue, Washington*

- Led cross-functional coordination across **3+** teams to deliver customer-facing AI features on schedule, managing dependencies across distributed services and deployment pipelines.
- Optimized AI inference workloads through concurrency tuning, batching, and access pattern analysis, achieving **20% latency improvement and 30% cost reduction** while processing millions of monthly requests.
- Led load testing and optimization of ECS-hosted Triton-based Responsible AI service across 7 regions, identifying thread-bound bottlenecks and resource underutilization to achieve **$28K/month** in cost savings through fleet rightsizing.
- Improved end-to-end request latency by **30%** by redesigning client connection pooling and request concurrency mechanisms.
- Eliminated **45+ high-severity incidents** by resolving systemic concurrency and race-condition issues in DynamoDB-backed workflows improving system stability under peak loads.
- Built comprehensive monitoring infrastructure with CloudWatch dashboards, custom metrics, and automated fault analysis scripts, reducing MTTD for production issues by **60%**.

Data Science Analyst June. 2024 – Mar. 2025
Mayo Clinic (AI & Informatics) *Phoenix, Arizona*

- Fine-tuned M3D multi-modal LLM (Llama-7B) on Mayo CT datasets using distributed training (FSDP), improving report generation quality through custom loss functions and architectural enhancements.
- Developed clinical retrieval system using ALBED-SS embeddings for natural language-based patient case matching, enabling physicians to find similar reports with configurable filters (BIRADS, similarity scores) and top-k retrieval.
- Built custom neural architectures using domain-specific pre-trained models (MammoCLIP) and improved interpretability with 3D Grad-CAM and attention visualization
- Designed and implemented distributed training pipelines using PyTorch DDP for large-scale medical imaging datasets, optimizing data loading, preprocessing, and GPU utilization.

Machine Learning Research Assistant Aug. 2023 – May 2024
Arizona State University, Wu Lab *Tempe, Arizona*

- Improved Brain MRI white matter segmentation with nn-Unet, achieving **0.65** mean dice score across 5 folds.
- Engineered distributed training pipelines and custom data preprocessing for seamless multi-modal training.
- Integrated LLM embeddings (GPT-2, BERT) with clinical notes and fine-tuned CLIP models, improving AUC by **8%** for Headache classification and achieving **10%** increase in classification performance on medical data.

Machine Learning Engineer May 2023 – Aug 2023
Arrow Electronics Inc. (e-Infochips) *Ahmedabad, India*

- Developed and deployed YOLO-based computer vision models for object detection, improving product quality by **15%**.
- Optimized low-light image enhancement models using Neural Architecture Search, reducing computational cost by **82%** while maintaining image quality metrics (PSNR, SSIM).
- Optimized deep learning models for real-time inference on edge devices using distillation, quantization, and pruning.

Machine Learning Research Assistant May 2021 – May 2022
Nirma University *Ahmedabad, India*

- Developed bearing fault classification pipeline achieving **98%** accuracy using EfficientNet CNNs with novel vibration-to-image feature engineering, FFT analysis, and transfer learning techniques.

Technical Skills

Languages: Java, Python, C/C++, SQL (Postgres), JavaScript, TypeScript, HTML/CSS, R, Bash
ML Frameworks: PyTorch, HuggingFace, Llama-Index, LangChain, ONNX, Keras, TensorFlow, TensorRT, vLLM
Developer Tools: Git, AWS (Lambda, ECS, S3, DynamoDB), Docker, Kubernetes, Linux, Jira, Confluence
Libraries: Transformers, Diffusers, FAISS, OpenCV, Scikit-learn, Pillow, Seaborn, Pandas, NumPy, Matplotlib, OpenAI

Education

Arizona State University Aug. 2022 – May 2024
Master of Science in Computer Science, ***GPA: 4.0/4.0*** *Tempe, Arizona*
Courses : Machine Learning, Natural Language Processing, Data Processing at Scale, Data Visualization

Nirma University July. 2018 – May 2022
Bachelor of Technology in Computer Engineering, Mechanical Engineering, ***GPA: 7.95/10.0*** *Ahmedabad, India*

Projects

Hallucination Evaluation | *LLMs, Hugging-Face, NLP, HPC Clusters, Git* Oct 2023 – Sep 2023

- Led team evaluating MLLMs (InstructBLIP, Open-Flamingo) for hallucination detection using HPC clusters and developed end-to-end pipeline with CLIP-score filtering and Llama-based Q&A generation.

Achievements

Invited to join IEEE Eta Kappa Nu's ASU Chapter (Honors Society of IEEE) in Fall 2023
Graduate Fellowship Scholarship awarded in Fall 2023 by Arizona State University.