PRATIK DOSHI

+1(831) 266-8773 ◇ Santa Cruz, CA
prdoshi@ucsc.edu ◇ linkedin.com/in/pratik-doshi-b2a493153/ ◇ github.com/Pratik-Doshi-99

EDUCATION

MS Computer Science, University of California, Santa Cruz — Expected: Apr 2025
Relevant Coursework: Neural Computation, Deep Learning, Compilers, Linear Algebra — GPA: 3.90/4.0

EXPERIENCE

Research Intern — 07/2024 - 09/2024
Data Care LLC, Utah, USA *(Kubernetes, PyTorch, Distributed Training/Inference, Docker)*

- Developed distributed training pipelines on a Kubernetes + Slurm setup using NVIDIA A6000 and A4000 GPUs.
- Ran experiments to integrate ZeRO memory optimizations and Flash Attention in the Company's GPUaaS.

Teaching Assistant — 09/2023 - 06/2024
University of California, Santa Cruz *(C, Low Level System Design, Multi-threading, Memory Management)*

- Led a section of students in developing a multi-threaded HTTP server in the C language.
- Delivered lectures on Multi-threading, Memory Management, and low-level IO in C-Programming.

Associate Software Engineer — 06/2021 - 03/2023
Rupeeseed Technology Ventures, Mumbai, India *(C#, System Design, Performance Profiling, AWS Cloud Services)*

- Reduced turnaround latency for a recommendation system from 15 minutes to 2 seconds using LINQ in C#.
- Designed a data processing pipeline in C# and improved its throughput by 50% using pipeline parallelism.
- Developed a statistical model to predict the profitability of customized trading strategies using probability distributions and derivative pricing models.

SKILLS

Machine Learning	LLM Inference, NLP, Deep Learning, Distributed Training, Transformers
Technical Skills	PyTorch, Python, Slurm, Grafana, Tensorboard, C, C#, REST APIs, MongoDB, SQL
Cloud	AWS EC2, Bedrock; GCP VMs, Kubernetes, Docker, Bash Scripting, CI/CD

PROJECTS

Image Captioning using VLMs. *Deep Learning, Multi-modal AI, Vision Language Models*
Trained a Vision-Language model on the image captioning task and achieved 25% accuracy improvement using dynamic attention from the paper "Show Attend and Tell". Developed a tokenizer for caption embedding and evaluated the model on the BLEU metric. (Github)

Predicting AI bias using SAEs *Synthetic Data, Deep Learning, Autoencoders, Classification*
Ran Inference on Llama 3-70B to generate a synthetic dataset with feature annotations. Used activations of Tiny-Stories-21M (base model) and its Sparse Autoencoder (SAE) to predict gender-bias in the model. (Github)

End-End Image Classification *Fine-tuning, Deep Learning, Google Cloud Platform (GCP)*
Fine-tuned ResNet150 and ViT models on an image classification task. Topped the competition through a custom training pipeline on GCP comprising data augmentation and ensembling. (Github)

Job Description Summarization using Llama 3 *LLM Inference, Continuous Batching, Ollama, Kubernetes*
Built a workflow using Llama 3-70B to fetch and summarize job descriptions. Optimized inference on a Kubernetes Cluster using continuous batching. (Github)

Small LM Training *LM Training, Kubernetes, PyTorch, Tensorboard, Grafana, Transformers*
Designed a small decoder-only LM and trained it on personal and group chats using a multi-GPU Kubernetes Cluster. Used Tensorboard and Grafana for real-time monitoring of training loss, GPU utilization, and Network IO.