MANN PATEL

manncodes@gmail.com • linkedin.com/in/manncodes • https://github.com/manncodes
+1 213-913-8798 • Los Angeles, CA - 90007

EDUCATION

University of Southern California Aug 2023 - Present
Master's of Science, Computer Science

Devang Patel Institute of Advance Technology and Research May 2018 - Apr 2022
Bachelors of Technology, Computer Science & Engineering CGPA: 9.69

EXPERIENCE

SWE ML Intern May 2024 - August 2024
Google Search *Mountain View, CA*

- Spearheaded research to enhance agentic capabilities of **LLM-based web agents**, focusing on improving planning and navigation for targeted information extraction
- Developed and implemented a scalable pipeline for **Online Agent Optimization**, enabling automated creation and deployment of specialized agents across 100+ diverse web platforms aiding Google search
- Achieved performance gains through multi-pronged optimization approach through novel automatic prompt optimization, **LoRA-based Gemma fine-tuning**, Vizier's algorithmic (black box) optimizations
- Pioneered LLM to linear HMM **model distillation**, resulting in 70% reduction in TPU usage for training from prior setting of only TPU bulk inference while maintaining robust coverage for out-of-distribution scenarios

Student Researcher Aug 2023 - Present
GLAMOR LAB, USC *Los Angeles, CA*

- Investigating **scaling laws** of memorization and uncertainty in Large Language Models through the lens of **mechanistic interpretability** across continuous pretraining stages on controlled synthetic knowledge graph
- Researching the impact of tokenization patterns on factual knowledge acquisition during model pretraining, identifying three distinct phases of memorization
- Previously contributed to research on rationale generation in Vision-Language Models (LLava, Qwen, Deepseek)

Machine Learning Intern Jul 2022 - Sep 2022
Quantive *Sofia, Bulgaria*

- Contributed to development of Quantive Singularity, an AI platform enhancing strategic decision-making. Accomplished 30% growth across 0.5 million professionals and 2,000 organizations.
- Researched and built models for **Causal Discovery & Causal Inference.** Automated causal ground truth creation, discovery, and regression pipelines.
- In a team of four, developed a causal tuning algorithm for achieving state-of-the-art results in causal discovery for multivariate time series benchmarks using optimal Markov blankets.
- Implemented and modified DAGs with NO-TEARS for structure learning from 100+ concurrent time series data.

Machine Learning Researcher Feb 2022 - Jul 2022
Cliff.ai *Indore, India*

- Researched extensively on **Time series forecasting**, **Anomaly Detection**, Time Series Clustering, Root Cause Analysis, Network Analysis, and Event Similarity.
- Elevated Neural Basis Expansion & Neural Hierarchical Interpolation for Time Series Forecasting by 45% and 2x prediction length while seamlessly extending its application on joint distribution data.
- Orchestrated Multivariate forecasting using spectral-temporal graph neural networks. Devised N-permutation masking test.
- Engineered API for streamlined analysis of large-scale network graphs (1M+ nodes, 10M+ edges) to efficiently predict anomaly propagation.

Machine Learning Intern Jan 2022 - Jul 2022
Greendeck *Indore, India*

- Conducted research and in-depth exploration of the following challenges: **Exact Image retrieval**, **Hierarchical Contrastive learning**, and local image feature matching.

- Engineered and automated robust MLOps pipeline with Apache Airflow for online CLIP training, data ingestion, embedding generation & storage, and ANN search with post-search corrections utilizing *LoFTR*. Gained 92% Top-5 accuracy.
- Devised *NeRF-CLIP*, hierarchical joint embedding architecture for 3D object-text pair embedding generation.
- Formulated a scalable image downloading service, achieving a rate of 2,500 images per second on a free EC2 instance for downloading IP-restricted images.

SKILLS

Programming & Development: Python, C++, CUDA, Java, Rust, Javascript, Go, Protobuf, Git, Bazel, gRPC
ML & LLM Technologies: PyTorch, Hugging Face, vLLM, TensorFlow, JAX, LangChain, Distributed Training, Quantization
LLM Research: LLM Pretraining, Post training, Model Distillation, Architecture, Mechanistic Interpretability, Alignment
Data Engineering & Analysis: Apache Beam, Spark, NumPy, Pandas, Feature Engineering, NetworkX, SNAP
Infrastructure & Cloud: Docker, Kubernetes, AWS (EC2, S3, Lambda), GCP, MongoDB, PostgreSQL, Redis, Pinecone, FAISS, Airflow, Parquet

PROJECTS

Xeno [repo link] | *Python, CUDA*
Constructed a scalable deep learning framework from the ground up, utilizing solely NumPy. Implemented a wide range of layers, initializations, optimizers, activation functions, and loss functions commonly present in other renowned frameworks. Realized a notable 7-8x acceleration in performance on tasks such as MNIST and CIFAR10.

Domain Expansion: Modules of Efficiency in Measurements as Building Blocks for Composite Domains [repo link][Paper] | *Python, Huggingface, PyTorch*
We explore how the formation of arithmetic over the weight space can be used as a method of acquiring knowledge of the intrinsic domain of the intrinsic domain of different neural networks. We provide an addition, negation, and interpolation operation to form a composition which is a theoretically universal representation.

FontSearch: Semantic Font Discovery and Generation Engine [live demo] [repo link] | *Python, Flask, FAISS, Sentence Transformers, DiffVG, Stable Diffusion*
Award-winning solo hackathon project that revolutionizes font discovery through natural language queries. Built a comprehensive pipeline that scraped 13K fonts, generated rich semantic descriptions using Google Gemini, implemented dual-embedding search with FAISS and FontCLIP for multimodal matching, and integrated font generation capabilities using differentiable vector graphics. The application enables users to discover fonts through text descriptions and generate novel typography based on semantic concepts.

Quick3D [repo Link] | *Pytorch, Javascript, Flutter*
Quick3D takes a single portrait image as input and output are the full high-resolution 3D Human Digitized model, based on PiFuHD [paper link] research paper by Facebook research. Led a team of three individuals to form a synchronized open-source web app and Android app. Find the release link.

Offline Reinforcement Learning with Action Discretization [Github repo] [Paper] | *JAX, Python*
In this project, we tackled the challenges of offline reinforcement learning by developing an adaptive method for action quantization using VQ-VAE. This approach enabled state-conditioned action quantization, improving the performance of established offline RL methods like IQL and CQL. Joint training of VQ-VAE and offline RL techniques led to further enhancements. More here at: W&B Reports

Violence Detection [repo link] | *Pytorch, OpenCV, TFlite, Python*
Utilized CNN-LSTM and PoseNet to detect violent activities from incoming video streams, trained on various annotated video datasets. With an accuracy of 91.2 %(K10 fold Validation). Leading a group of 3, created a web app, Android and iOS App employing Flutter. Authored a paper that currently holds 26 citations.

PUBLICATIONS

Mann Patel - *Real-Time Violence Detection Using CNN-LSTM*, [arXiv].
Mann Patel - *FedGrad: Optimization in Decentralized Machine Learning*, [arXiv].
Mann Patel*, Divyajyoti Panda, Hilay Mehta, Parth Patel, Dhruv Parikh. - *Domain Expansion: Parameter-Efficient Modules as Building Blocks for Composite Domains*, [arXiv]

HONORS & AWARDS

- HackSC 2024 - Grand Prize Winner: Solo developer competing against 50 full teams.
- Won and was declared Intel AI DevMesh 2021 winner.
- Ranked globally 17th / 14,000 at Amazon AI Hackathon 2021.
- Secured AIR-19 and global rank 143rd in Google HashCode Extended Round 2021.