Vatsal Baherwani

Allendale, NJ | vatsalbaherwani@gmail.com | linkedin.com/in/vatsal-baherwani | github.com/vatsal0

RESEARCH INTERESTS

Mixture-of-experts, Compute-efficient language models, AI Interpretability, AI Safety/Oversight

EDUCATION

University of Maryland – College Park, MD *Aug 2021 – May 2025*
B.S., Computer Science • Computational Finance Minor • Cumulative GPA: 3.98/4.0
Relevant Coursework: Computer Systems, Design & Analysis of Algorithms, Applied Probability & Statistics, Probability Theory, Data Science, Advanced Data Structures, Machine Learning, Intro to Deep Learning, Deep Learning (PhD level), AI/ML at Scale (PhD level)

RESEARCH EXPERIENCE

University of Maryland – College Park, MD

Efficient Gradient Estimation for Sparse Mixture-of-Experts Language Models *Jul 2024 – Present*
Supervisor: Dr. Tom Goldstein - Department of Computer Science

- Implemented logging to track similarity of sparse gradients to true dense gradient in mixture-of-experts models
- Proposed novel method for estimating dense gradient with 99% accuracy using nearest neighbor activations
- Developed efficient Triton kernel implementation for gradient estimation, minimizing FLOPs increase to 0.4%
- Trained 1B parameter LLMs with gradient estimator on 10B tokens of Fineweb dataset on Frontier supercomputer
- Improved validation perplexity by 3% over baseline method, and decreased imbalance in routing inputs to experts

Improving One-Shot Motion Customization in Video Generative Models *Feb 2024 – Present*
Supervisor: Dr. Abhinav Shrivastava - Department of Computer Science

- Implemented LoRA fine-tuning for reproducing preliminary work on customizing motion in diffusion models
- Traced motion learning in diffusion models by applying DDIM inversion and resampling at various timestep ranges
- Improved motion customization fidelity by restricting fine-tuning of temporal attention layers to early timesteps
- Conducted ablation study to localize motion information in attention layers, improving training speed by >100%

Understanding Historical Determinants of Market Liquidity and Depth *Sep 2023 – Dec 2023*
Supervisor: Dr. Pete Kyle - Department of Finance

- Developed Python library for loading and merging high-frequency trading data and calculating financial statistics
- Modeled stock market liquidity and market depth with time series regression on 1 terabyte of historical market data
- Analyzed time series autocorrelation, reduced noise with outlier filters, and implemented parallel data processing
- Aggregated monthly liquidity and depth for >5,000 companies, revealing a logarithmic correlation (r^2=0.95)

Interpreting Lack of Compositional Understanding in Image Generative Models *Sep 2023 – Dec 2023*
Supervisor: Dr. Soheil Feizi - Department of Computer Science

- Reproduced prior work on failure cases of text-to-image models understanding compositionality of multiple objects
- Conducted experiments generating images of compositions for each combination of 25 objects and 10 attributes
- Traced information flow across 70 model layers, revealing mixing of representations between objects and attributes

Predicting Human Responses Using Electrical Brain Signals *Sep 2022 – Dec 2022*
Supervisor: Dr. Alec Solway - Brain & Behavior Institute

- Preprocessed and cleaned EEG data of 14 patients to study signals influencing human actions and decision making
- Predicted patient decisions and response times from brain activity using a time series convolutional neural network

Natural Language Processing Techniques for Invasive Species Threat Detection *Jan 2022 – May 2022*
Supervisor: Dr. Lars Olson - Department of Agricultural and Resource Economics

- Fine-tuned natural language processing model to identify economic threat indicators among 1000+ invasive species
- Achieved >90% accuracy in classifying phrases describing negative economic impacts in the CABI database

PUBLICATIONS

Video Diffusion Models Encode Motion in Early Timesteps, *Under Review at CVPR 2025*
V. Baherwani, Y. Ren, A. Shrivastava

Dense Backpropagation Improves Routing for Sparsely-Gated Mixture-of-Experts, *NeurIPS OPT 2024, Under Review at ICLR 2025*
A. Panda*, **V. Baherwani***, Z. Sarwar, B. Thérien, S. Sahu, S. Rawls, S. Chakraborty, T. Goldstein

Racial and Gender Stereotypes Encoded Into CLIP Representations, *ICLR 2024 Tiny Papers Track*
V. Baherwani, J. Vincent

EMPLOYMENT

University of Maryland – College Park, MD
Teaching Assistant – BUFN400: Introduction to Financial Markets and Financial Datasets *Aug 2023 – Dec 2023*

- Held weekly discussions demonstrating financial quantitative analysis techniques using Pandas, NumPy, and SciPy
- Assisted students in office hours with debugging Python code for homework assignments involving data modeling

Wolverine Trading – Chicago, IL
Software Engineer Intern *May 2023 – Aug 2023*

- Designed snapshot service to deliver daily historical price and market data for 50,000+ traded option contracts
- Improved average request speed by >75% and reduced database request load by >80% with server-side caching
- Built trader-facing client with .NET framework to query and cache market snapshots and listen to real-time updates
- Implemented client-server RPC communication and publisher/subscriber API for sending intraday market updates

Bloomberg L.P. – New York, NY
Software Engineer Intern *May 2022 – Aug 2022*

- Optimized calculation of 72 daily indicators for 3 million fixed-income securities with Apache Airflow workflows
- Parallelized execution and implemented autonomous error handling, reducing average calculation runtime by 35%

Software Engineer Intern *Sep 2020 – Aug 2021*

- Developed Java Spring REST API to create, read, update, and delete 900+ Bloomberg Law account permissions
- Reduced service login time by >200ms through identifying and removing 300+ unused or expired permissions

PERSONAL PROJECTS

UnityPack

- Extracts assets including meshes, textures, and animations from compiled Unity games into glTF format JSON files
- Decompresses bit-packed mesh data, calculates animation transform matrices, and generates scene graphs

NBA Shot Selection Analysis - vatsal0.github.io/cmsc320final/

- Data cleaning, visualization, and regression analysis on 30 NBA teams predicting win rate with 82% accuracy
- SVM classification and k-means clustering to evaluate and rank shot efficiency from 16 areas on the NBA court

MNIST Digit Classifier - github.com/vatsal0/mnist-digit-classifier

- Neural network written from scratch in C with implementations of regularization and mini batch gradient descent
- Classifies 10,000 handwritten digits from the MNIST image database test set with 97.1% accuracy

ADDITIONAL SKILLS

Technical Skills: Python (Flask, Pandas, NumPy, SciPy, TensorFlow, Airflow), Java (Spring), JavaScript (Node, React, Vue), Ruby (Rails), SQL, GraphQL, Swift, C, Matlab, Docker, AWS, Linux
Certifications: Stanford Machine Learning, Kaggle Deep Learning, Bloomberg Market Concepts
Personal Blog: medium.com/@vatsalbaherwani