Kartik Gupta

kartikgt23@gmail.com kartikgt.github.io
linkedin.com/in/kartikgupta23 (413) 468-0353

EDUCATION

University of Massachusetts Amherst *Sep 2022 – May 2024*
MS in Computer Science

Indian Institute of Technology Kharagpur *Jul 2013 – Apr 2018*
BS + MS in Electrical and Computer Engineering

EXPERIENCE

Data Science Intern | ***Mathematica*** *Jun 2023 – Aug 2023*

- Developed ML models for multi-variate causal analysis, propensity scores, and random forest regression predictions
- Employed statistical techniques, Sparse PCA, and Lasso Regularization for feature selection, and SHAP for explainability
- Built a Python framework for a 100TB+ database creation in Redshift using Pandas, Botocore, and AWS Step Functions
- Identified use cases of large language models for keyword extraction and summarizing online medical records

Graduate Deep Learning Researcher | ***Adobe*** *Jan 2023 – Jun 2023*

- Developed neural image compression algorithms for optimizing latency in multi-task (segmentation, object detection) pipelines.
- Benchmarked texture recognition task performance for the baseline model and our custom neural architecture and demonstrated improvements in top-5 accuracy for 1 bit (43% to 90%) and 8 bits (89% to 92%) per pixel quantization setups
- Employed PyTorch Lightning for efficient training in a multi-GPU setup, and Weights and Biases for workflow monitoring

Software Engineer 2 | ***Uber*** *May 2021 – Aug 2022*

- Saved $3.5M in compute costs by developing a Spark and Presto query optimization tool for finding and fixing anti-patterns
- Lead engineer for managing the central facts tables (10PB+) and batch pipelines, and modeling downstream datasets using advanced SQL concepts like max structs, exploding lateral views, self joins, window functions, and nested grouping
- Reduced disk space usage by >40% via converting all datasets to Parquet format and using ZSTD compression
- Performance tuned and optimized 50+ Apache Spark pipelines correcting for out-of-memory, skew, and small file issues
- Created multiple real-time event analytics pipelines using Apache Flink on Kafka streams and a real time K-V datasore
- Oversaw the Hadoop data lake (100PB+) and handled resources (disk space and namenodes) and YARN queue assignments

Senior Data Engineer | ***Envestnet*** *Oct 2020 – Apr 2021*

- Lead the design and development of a Neo4j, a NoSQL graph database, based Master Data Management System on AWS
- End-to-end application development on AWS with EC2, S3, EMR, IAM, and Lambda on the web, CLI, and boto3 interfaces

Software Engineer | ***SAP*** *Jul 2018 – Sep 2020*

- Developed a Java and Selenium based automation test suite for feature testing and logging errors on Chrome and Firefox
- Developed an organization-wide data discovery tool for end-to-end analytics. Built a Hadoop HDFS datalake using Spark and Hive, and engineered a Java application for interactive API based querying of the data warehouse in Elastic Search
- Developed ARIMA time series forecasting models to predict issues based on historical seasonality and periodicity trends

SELECTED PROJECTS

- **Mechanistic interpretation of code generating large language models** - Novel research on mechanistically identifying the activations and circuits inside language models for code generation tasks, including the ability of the model to execute given code for a variety of programming languages, and to understand code as a set of basic arithmatic operations.
- **Evaluating in-context learning for finetuned LLM agents** - Used Langchain to create agents from LLama2 7b and simulated inter-agent debate. Analyzed conversational drift, topic modeling, and perplexity from the agent responses, and quantified incontext learning and influence via measuring delta across iterations with aspect based sentiment analysis.
- **Transfer Learning and Few Shot improvements for medical deep learning** - Utilized pretrained RESNET18 and VGG16 as backbones and partially finetuned them on Covid-19 classification dataset. Simulated class imbalances leading to increase in AuC (0.64 to 0.86) and implemented 4-way 5-shot learning schemes to further enhance accuracy on novel test datasets.

CERTIFICATIONS

Microsoft Certified: Azure Data Scientist Associate Credential ID: H446-0997
Neo4j Certified Professional Credential ID: 17127043

SKILLS

Languages and Tools	Python, Java, Golang, R, SQL, Gradle, Cypher, Vercel, Gorilla, Neovim, Flask, Docker, Linux, Shell
Data Science	Pandas, Numpy, Sklearn, Statistical Modeling, Scipy, NLTK, Spacy, CausalML, Bayesian Statistics
Data Engineering	Spark, PySpark, Flink, Kafka, Hadoop, Hive, MapReduce, Databricks, K-V databases, Turso
AI and ML	Pytorch, HuggingFace, Langchain, Lightning, OpenCV, FSDP, RAG, PEFT, Unsloth, Azure AI Studio