KARTIK KANOTRA

+1 (201) 423-1192 kk5243@nyu.edu /kartik-kanotra /kartik0649 /kartik0649

Education

New York University (NYU), Courant Institute of Mathematical Sciences **September 2023 - May 2025**

Master of Science in Computer Science (Specialization in Data Science) ***GPA*** *3.8/4* *New York, USA*

Relevant Courses: Deep Learning , Predictive Analytics, Operating Systems, Cloud & Machine Learning (ML), Large Language Models

Birla Institute of Technology and Science (BITS), Pilani **August 2018 - July 2022**

*B.E. (Hons) in Electrical and Electronics Engineering (**First Division**)* *Hyderabad, India*

Relevant Courses: Object-Oriented Programming, Data Structures and Algorithms, Cryptography, Data Mining, Database Systems, Digital Design, Information Retrieval, Computer Vision, Microprocessors

Skills

Languages: Python, C/C++, Java, R, SQL, TypeScript, JavaScript, Rust, Ruby, C
Tools: AWS (EC2, S3, Lambda), GCP, Docker, Kubernetes, Apache Spark, Kafka, TensorFlow, PyTorch, JAX, Hugging Face, MLflow, Airflow
Technologies: Machine Learning, Deep Learning, LLMs, NLP, Computer Vision, Model Deployment (ONNX, TensorRT), Data Engineering, Azure, MLOps, Performance Optimization (CUDA, Nsight Compute)

Experience

Lindsay Lab, New York University (NYU) **June 2024 - Dec 2024**

Research Assistant *New York, USA*

- Designed a scalable, modular **Python** codebase integrating biologically plausible learning algorithms, replacing **backpropagation** for an n-layer MLP on MNIST, leveraging **PyTorch**, **NumPy**, and **CUDA**, under **Grace Lindsay**'s mentorship.
- Conducted extensive benchmarking and documentation on **Feedback Alignment**, **Predictive Coding**, **BrainProp**, and **Dendritic Error Backpropagation**, achieving a **3%** accuracy boost, optimizing compute with **TorchScript** and **ONNX**.

Visa **June 2022 - July 2023**

Software Engineer *Bangalore, India*

- Led the redesign of the Digital Configuration Platform with **Angular**, **TypeScript** for UI and **Spring Boot**, **Java** for backend, increasing website conversion by **30%**.
- Enhanced **API performance** through **Kafka**-based event-driven architecture, implementing robust **microservices** in **Java**, achieving a **17%** performance enhancement.
- Optimized **SQL** and **MongoDB** query efficiency, reducing **data retrieval times** by **70%** and improving system throughput.
- Built a **CI/CD pipeline** with **Jenkins**, **Docker**, and **Selenium** testing, accelerating deployment by **50%** and improving system reliability.

Samsung Research And Development (R&D) Institute **June 2021 - December 2021**

Machine Learning Intern *Bangalore, India*

- Developed and deployed an AI-powered **chatbot** for Samsung Finance Plus using **RASA**, **Transformers**, **BERT**, **Spacy**, and **FastText**, increasing user engagement by **50%** and automating **70%** of customer queries.
- Refined **NLP pipelines**, **intent classification**, and **named entity recognition (NER)** with **Hugging Face**, **PyTorch**, and **TensorFlow**, reducing query resolution time by **35%** and boosting response accuracy by **20%**.
- Integrated a Java-based application with the chatbot, utilizing **Docker** for containerization, reducing deployment time by **40%** and improving system scalability by **60%** across multiple environments.

Projects

Trajectory Learner - JEPA (Under Yann LeCun)

- Built a **Joint Embedding Predictive Architecture (JEPA)** using **PyTorch, CNN, LSTM, and MLP**.
- Modeled agent dynamics in a a two-room simulation with **2.5M** frames.
- Refined state embeddings using **VICReg**, boosting predictive accuracy and reducing loss from **220** to **3.4**. [Link]

Retrieval Augmented Generation (RAG)

- Developed a **RAG-based** academic research recommender using **ChromaDB, LangChain, and Llama/GPT-3.5 Turbo**.
- Deployed a dockerized **Flask app** on **Kubernetes** with hyperparameter tuning using **Kubeflow**.
- Generated top **3** article recommendations using **semantic search** with **MS MARCO**. [Link]

CNN Perf Estimation

- Optimized **ResNet-50, VGG-16, and MobileNetV2** on **Nvidia GPUs (A100, V100)** using **CUDA** and PyTorch.
- Used **Nsight Compute** to improve **FLOPS** and memory bandwidth by **23.65%**. [Link]

Kernext

- Wrote a Linux **Kernal** in **C++**, implemented process management, memory management and **IPC** using **Linux** system calls [Link]
- Optimized **virtual memory** with **paging, segmentation**, and **demand paging**, boosting **allocation efficiency** by **25%**.