Rhitvik Sinha

+1 (760) 759-3197 | rhitviksinha@nyu.edu | linkedin.com/in/rhitvik-sinha | rhitviksinha.github.io/

Education

Courant Institute of Mathematical Sciences, New York University — New York, NY
Master of Science in Computer Science — *2022 - 2024*
Courses: *Computer Vision, Machine Learning, Reinforcement Learning, Large Language & Vision Models, DL Systems*

Indian Institute of Technology, Kharagpur — Kharagpur, India
Bachelor of Technology (Honours) in Electrical Engineering — *2018 - 2022*
Courses: *Deep Learning, Data Analytics, Transform Calculus, Probability & Stochastic Processes*

Technical Skills

Languages: *Python, C/C++, CUDA, MATLAB, R, JavaScript, Scala, Java*
Relevant Libraries / Frameworks: *NumPy, PyTorch, OpenAI, TensorFlow 2.x, OpenCV, Faiss, Langchain, Gym/Gymnasium/PettingZoo, multiprocessing, Z3, Pandas, GeoPandas, Capstone*

Experience

Machine Learning Research Intern — Summer, Fall 2024
Axiado Corp., San Jose, CA — *Supervisor: Mr. Zhichao Zhang*
Work with the research team to develop the following POC(s):
- Design end-to-end vulnerability detection pipeline for BIOS images using Retrieval Augmented Generation (RAG) with state of the art LLMs (Faiss, Langchain & OpenAI API).
- Boot-time anomaly detection: Use LLMs to parse through & reduce Boot-time logs, followed by RAG-assisted anomaly detection.
- Miscellanous Scripting & Automation: Compile & disassemble linux kernels to collect CVE-listed vulnerabilities; Trigger persistent BSoD events from a python script & recover; Scrape the web to build a dataset of BIOS images.

Systems and Applications Engineering Intern — Summer 2023
Cirrus360 Corp., Richardson, TX — *Supervisor: Dr. Alan Gatherer*
- Improved compile time of Gabriel™(a platform for RAN deployment described using a Domain Specific Language) by 20-30% (logarithmic speedup as CPU cores are increased) through parallelization strategies using Python's multiprocessing module.
- Developed a Flask-hosted Automatic Speech Recognition application utilizing OpenAI's Whisper and PyTorch's Torchaudio, improving transcription efficiency with 8-bit quantization.
- Created multiple Python utilities to automate testing, increasing productivity and reducing development time.
- Modified configuration files to simulate constrained hardware environments, ensuring robustness and adaptability of Gabriel™.

Projects

Image Encoding Schemes for Vision Transformers | **Paper** — Fall 2023
Course Project (CSCI-GA 3033 Special Topics: Large Language & Vision Models), NYU — *Guide: Prof. Saining Xie*
- Improved image classification performance by 7% on CIFAR-10 and 8% on CIFAR-100 using Mixed-Resolution Tokenization, leveraging a saliency scorer for adaptive patch sizes, outperforming baseline Vision Transformers.
- Achieved 85% accuracy with patch embeddings, demonstrating their robustness and efficiency despite low computational cost, validated through extensive experiments with varying patch sizes.
- Explored advanced tokenization techniques, including VQ-VAE and initial convolutions, identifying limitations in their application for classification tasks, and highlighting mixed-resolution tokenization as a promising direction for further research.

Adaptive SphereFormer: Dynamic Radial Windows for Better Sparse Learning | **Paper** — Fall 2023
Course Project (CSCI-GA 2271 Computer Vision), NYU — *Guide: Prof. Rob Fergus*
- Developed the Adaptive SphereFormer, improving the existing state-of-the-art SphereFormer's radial window self-attention with dynamic adjustments based on point distance, increasing overall mean Intersection over Union (mIoU) by 1%.
- Achieved efficient memory management and better segmentation accuracy by progressively expanding window sizes with distance from the origin, validated on the SemanticKITTI dataset.

Optimizing Diffusion Models for Image De-Noising | **Paper** — Fall 2022
Course Project (CSCI-GA 2565 Machine Learning), NYU — *Guide: Prof. Rajesh Ranganath*
- Modified and trained diffusion models for noisy image inputs, analyzing the impact of noise levels, diffusion steps, and cycles; implemented and trained a class-conditioned diffusion model.
- Developed a diffusion model to regenerate images with missing pixels, effectively functioning as a Masked Auto-Encoder.
- Conducted literature review on VAEs, GANs, and diffusion models, and reproduced benchmarks for Denoising Diffusion Probabilistic Models (DDPM).

Barenet: INT8 Quantized Transformers from Scratch | **Code** — Spring 2024
Course Project (CSCI-GA 3033 Special Topics: Big Data & Machine Learning Systems), NYU — *Guide: Prof. Jinyang Li*
Developed a fully functional Multi-Layer Neural Net with int8 quantization and attention mechanisms, with CUDA kernels for matrix operations to achieve efficient transformer inference and performance optimization, creating a limited-functionality PyTorch clone.

Deep Learning for Extreme Weather Forecasting — Spring, Fall 2021
Undergraduate Research, IIT Kharagpur — *Guide: Prof. Adway Mitra*
Conducted research on extreme weather forecasting using Capsule Neural Networks and Analog Weather Forecasting techniques, demonstrating superior performance over CNNs and logistic regression in predicting heat/cold waves 1-5 days ahead using T2m and Z500 data from the NCAR CESM-LENS dataset.