RISHABH SRIVASTAVA

+1 (929) 358-5907 | rs4489@columbia.edu | linkedin.com/in/rishabhsrivastava6 | github.com/RishabhS66

Education

Columbia University New York, NY
MS in Computer Science (Machine Learning Track), GPA: 4.06/4.00 Dec 2024
Relevant Courses: NLP, LLM-based Generative AI Systems, Machine Learning, High-Performance ML, Databases
TA for: Projects in Advanced Machine Learning, Topics in Software Engineering, Advanced Software Engineering
Recipient of Data Science Institute Scholarship (Fall 2024)

Indian Institute of Technology Guwahati Assam, IN
BTech in Electronics and Electrical Engineering, Minor in Computer Science Jul 2021
Relevant Courses: Computer Vision, Probability, Data Structures and Algorithms
Recipient of Samsung Fellowship Award

Technical Skills

Languages: Python, CUDA, C++, Java, MySQL, MongoDB, MATLAB, React, NodeJS, TypeScript
Technologies: PyTorch, Scikit-learn, TensorFlow, OpenCV, vLLM, LangChain, Wandb, AWS, GCP, Kubernetes, Docker

Work Experience

Rubicon Robotics Inc. New York, NY
Software Engineer Intern May 2024 – Dec 2024

- Developed and implemented CV algorithms for swimmer detection by SwimBot, attaining a **90%** accuracy rate.
- Established a comprehensive training pipeline, including a GUI tool for dataset creation through video frame extraction and body part annotation, enabling efficient re-training of the OpenPose model for detailed posture analysis.
- Architected and deployed backend infrastructure using Django and AWS services (RDS, EC2, Load Balancers, Route53), and automated workflows like video thumbnail generation with AWS Lambda.
- Implemented CI/CD pipeline using GitHub Actions, boosting development efficiency and site reliability.

Adobe Inc. - Adobe Experience Manager (AEM) Assets Noida, IN
Software Development Engineer Level II Jul 2021 – Aug 2023

- Spearheaded enhancement of AEM Assets Search by utilizing Lucene indexing for efficient information retrieval, Hugging Face's BLIP APIs for asset auto-captioning and GPT-4 for query pre-processing.
- Led the end-to-end development of Smart Tags Block-list in AEM Assets Essentials, utilizing Micro-Frontends (MFEs) for the front-end interface, empowering users to block inappropriate tags and ensuring brand compliance.
- Volunteered as DevOps Champion, maintaining the CI/CD pipeline on Jenkins and managing Kubernetes cluster deployments.
- GenAI Hackathon - integrated Adobe Firefly to improve search experience for AEM Assets Essentials, allowing customers to generate custom images if search results are irrelevant; selected to be presented at Adobe EMEA Summit 2023.

Research Experience

Artificial Intelligence for Vision Science (AI4VS) Lab, CU Irving Medical Center New York, NY
Research Assistant under Prof. Kaveri Thakoor Sep 2024 – Dec 2024

- Developed AI-CNet3D, a 3D CNN with cross-attention mechanisms, enhancing glaucoma detection by analyzing 3D OCT volumes. Achieved **10x** parameter reduction and outperformed state-of-the-art models on key metrics.
- Implemented attention-alignment with expert gaze patterns in Vision Transformer to detect AMD from retina B-scans.

Advanced Research in Software Engineering (ARISE) Lab, Columbia University New York, NY
Research Assistant under Prof. Baisakhi Ray May 2024 – Aug 2024

- Fine-tuned DeepSeek-Coder-V2-Lite-base using custom-built PYX dataset to get SemCoder-S, a semantic-aware CodeLLM.
- Conducted experiments comparing SemCoder-S with other CodeLLMs , achieving superior performance with F1 score of **0.678** for code correctness and **62.4%** accuracy for execution prediction on HumanEval-based dataset.

Adobe Inc. Noida, IN
Media and Data Science Research Intern Apr 2020 – Jul 2020

- Implemented Deep Q-Network to extract top relevant patterns from temporal, sequential datasets.
- Proposed algorithm allowed monitoring and improving user-targeting based on certain Key Performance Indicators.

Hanyang University Ansan, KR

Research Intern under Prof. Frank Rhee May 2019 – Jul 2019

- Designed a new algorithm Adaptive Shadowed C-Means (ASCM), clustering data using fuzzy and shadowed sets to reduce impact of noise.
- Implemented algorithm on Iris and Breast Cancer Wisconsin data sets, and demonstrated its use for image segmentation.

Publication

- Kenia, R., Li, A., **Srivastava, R.**, Thakoor, K. A., "AI-CNet3D: An Anatomically-Informed Cross-Attention Network for Enhanced Glaucoma Detection and Interpretability in 3D OCT Volumes," in review at IEEE Transactions on Medical Imaging

Projects

ReAct Agent for Search, Compare and Analysis | *LangChain, Gemini, SerpAPI* | Nov 2024

GitHub: RishabhS66/ReAct-Agent-for-Search-Compare-and-Analysis

- Designed and implemented a ReAct (Reasoning and Acting) agent, integrating search, compare and analyze tools.
- Developed a Gradio UI for seamless query input and concise result display with robust error handling and tool transitions.

FOMC Statement Hawkish-Dovish Analysis Using LLMs | *Transformers, BeautifulSoup* | Jun 2024 - Aug 2024

Supervisor: Prof Ali Hirsa, and associated with Morgan Stanley

- Used CentralBankRoBERTa to predict market dovishness/hawkishness from FOMC statements and meeting minutes from Jan 2019 to July 2024.
- Web scraped, cleaned, and extended the dataset, then prompt-engineered GPT-4 to classify text into pre-defined categorical labels, benchmarking results against the CentralBankRoBERTa model and MacroMicro AI Hawkish-Dovish index.
- Stress tested models by modifying prompts, data size, and order of inputs, and analyzed label consistency across models using Kendall's W.

Inference Acceleration of Stable Diffusion | *Quantization, Pruning* | Apr 2024 - May 2024

GitHub: RishabhS66/Inference-Acceleration-of-Stable-Diffusion

- Devised Time-step calibrated quantization for Stable Diffusion, achieving the lowest FID score and highest CLIP score compared to other quantization techniques.
- Conducted L1-unstructured pruning and combined quantization, compressing the model by **20%** and reducing inference time by **5%** without significant performance loss.

Abstract Art Interpretation Using ControlNet | *Stable Diffusion, ControlNet, BLIP* | Apr 2024

GitHub: RishabhS66/Abstract-Art-Interpretation-Using-ControlNet

- Leveraged ControlNet and Stable Diffusion to enhance spatial control over image composition and enable interpretation of abstract art through detailed geometric conditions.
- Developed a custom dataset of **14,279** image pairs to train model, achieving high-quality image generation with innovative artistic representations.

CUDA-Accelerated Image Convolution | *CUDA, cuDNN* | Mar 2024

GitHub: RishabhS66/CUDA-Accelerated-Image-Convolution

- Implemented custom CUDA-based image convolution with shared memory and tiling, optimizing performance and reducing execution time from **47.087 ms** (simple convolution) to **33.246 ms**.
- Performed convolution using the cuDNN library to compare performance of optimized libraries with custom implementation, and achieved an execution time of **38.338 ms**.

Lexical Substitution Task with WordNet, Word2Vec Embeddings, and BERT | *NLTK, Transformers* | Nov 2023

GitHub: RishabhS66/Lexical-Substitution-using-BERT

- Devised a novel fusion strategy, combining BERT's contextual understanding with Word2Vec's semantic similarity and WordNet's semantic relations, to improve lexical substitution accuracy and suggest contextually fitting word replacements.
- Attained a precision of **0.189** and recall of **0.189** on 206 attempted instances with mode-specific scoring.