HIMANSHU GOYAL

+1 773-231–3457, goyalhimanshu414@gmail.com , Google Scholar

Work Experience (7+ years)

JCI (Acquired FogHorn.io) **Data-Scientist** *Oct'20 - Aug'24*

- **OB-Ask Me Anything ChatBot:** Natural language Query & insight generation using LLMs for technical dashboards.
  - Developed novel data insight generation tool using Langchain's MRKL agent with **Re-Act** framework & **Zero & Few-shot Chain of Thoughts** prompting techniques.
  - Implemented Auto-CoT for process self-improvement & developed **prompt-injection detection** framework.
  - Patented the process & showcased in *JCI's Tech-Challenge'24* finals as one of the top eight, out of 3,000+ submissions.
- **Points to Asset Mapping:** Fine-tuned LLMs to predict standard names of points based on short-names & units.
  - Fine-tuned Pythia-410m on custom dataset & implemented solution to perform **RLHF** using **DPO & PPO** techniques.
  - Developed features for **model explainability** for mapped classes & online training of deployed models to address *data-drift.*
  - Expanded solution for **multiple languages** enhancing overall sales & reducing onboarding time from 6 months to 1 month.
- **On-Devise ML for CV problems:** Developed computer vision solutions optimized to run in constrained edge environment for realtime monitoring of Flaring, BOP and WorkerSafety-norms-violation in Oil & Gas industry.
  - Performed Transfer-Learning experiments for hyper-parameter optimization to fine-tune computer vision model like **EffcientDet**, YOLOx, Faster R-CNN for object Detection tasks; **Mask R-CNN**, DeepLab for Mask-Segmentation task.
  - Optimized CV models for edge deployment using **Knowledge Distillation**, Quantization & Pruning leading to 60% speed improvements while maintaining baseline accuracy.
  - Implemented **CI/CD** pipeline for edge deployement & Training pipeline for experiment tracking, seamlessly integrated with MLflow and Databricks; Contributed **MLflow** integration to open source YOLOx repo. having **9k+ stars.**
  - Productionized at 100+ sites having annual revenue of $0.5mil with BOP receiving *IoT Edge Computing Excellence Award 2021*, WorkerSafety receiving *Nascom AI GameChanger Award 2024* & Flare *granted patent* and *published in AdConIP.*
- **Exploratory Projects:** Various POC projects for process improvement, Benchmarking & feasibility testing
  - Developed efficient **GPU Utilization Strategy** by using batch-inferencing, optimizing concurrent usage of GPU by multiple models in Edgeml® docker container, **doubling** the number of **solution** deployed on same hardware.
  - Implemented SoTA **Copy-Paste Augmentation Technique** (ranked #1 in **OD(Object Detection)** task in 2021) to improve accuracy on rare classes. Created **Rule Based Accuracy Matrices** reporting Framework.

Citigroup **Technical Analyst** *Aug'17 - Oct'19*

- Developed a low-cost, s3 object storage application using Java SpringBoot/MVC for microservices API & Angular for UI.
- **Trade Document OCR:** Developed n-gram based **recommendation engine** for assistance in data entry from trade documents. Predicted the value for missing fields with **mAP of 0.9 using precision@k** metrics.

Projects

- **High Performance Computing:** Implemented and optimized ANN in C/CUDA with forward and backpropagation steps.
  - Optimized ANN performance by parallelizing code with OpenMP, enabling multi-node distributed processing using MPI.
  - Developed a CUDA-based version for a multi-GPU setup, leveraging **cuBLAS** for efficient matrix multiplications.
  - Benchmarked various implementations, accelerating image processing from **4K to 300K images/sec**, a 75x improvement.

Technical Proficiency & Publications

- **Technology:** Python, Pytorch, C with CUDA, GoLang, Pandas, Numpy, Java, Docker, SQL, Angular, Azure/AWS
- **Theoretical:** Machine-Learning, Computer Vision, NLP, Deep-Learning, GenAI, MLops, GPU Programming
- **Papers:** *"Deep Learning based Flare Image Analytics at the Edge",* 2022 IEEE Int'l symp. on AdConIP, Vancouver, Canada.
- **Patents:** i.) Granted for developing Computer Vision based *"Flare Monitoring System and Method"* (US2023/0272910) ii.) Filed for *"Gen-AI Based Building System with Analysis & Contextual-Insight Generation"(Pending)*

Education

- **M.S. in Computer Science, The University of Chicago, IL (4.0/4.0)** *May' 25 (expected)*
  Coursework: High Performance Computing, Parallel Programming, Computer Systems, Cloud Computing
- **B.Tech. in Textiles, IIT-Delhi, India — GPA: 8.0/10** *Aug'13 - July'17*
  Relevant Coursework: Data Structures & Algorithms, Linear Algebra, Machine Learning, System Designing

Certifications

- Coursera: Deep Learning Specialization, ML Engineering for Production (MLOps) Specialization, CitiGroup: CFA-Level-2 Passed, Engineering Excellence level-2.