

Harshit Kumar

Boston, MA | [kHarshit.github.io](https://github.com/kHarshit) | kumar.hars@northeastern.edu | +1 (857) 693-9361

 [github/kHarshit](https://github.com/kHarshit) |  [linkedin/kHarshit](https://www.linkedin.com/in/kHarshit) | [Top-15 PyTorch on](#)  [stackoverflow/6210807](https://stackoverflow.com/questions/tagged/pytorch)

Education

Northeastern University, Khoury College of Computer Sciences, Boston, MA

Sep 2022 - Present

MS in Artificial Intelligence

GPA: 3.83/4.0

Graduate Teaching Assistant: Data Mining Techniques, Programming for Data Science

Courses: Foundations of AI, Algorithms, Machine Learning, Pattern Recognition Computer Vision, Large Language Models

Guru Gobind Singh Indraprastha University, Delhi, India

Aug 2016 - Sep 2020

BTech in Computer Science and Engineering, top-4 in class

GPA: 8.72/10.0

Experience

Deep Learning Research Intern - The Jackson Laboratory, Bar Harbor, Maine

Jul 2023 - Dec 2023

- Researched 5+ **Explainable AI** methods: Saliency Maps, Integrated Gradients, SHAP, GNN Explainer for DNA sequence data and **Graph Neural Networks** using **PyTorch Geometric**, leveraging Slurm on High Performance Computing (HPC) clusters.
- Conducted quantitative assessments with cross-entropy and AUC-ROC to evaluate attribution scores against ground truth, optimizing 160+ model architecture configurations (GCNConv, GraphConv, dropout, L2 regularization) through grid search.

Machine Learning Engineer - Vehant Technologies, Noida, India

Jun 2019 - Aug 2022

- Implemented **10+ People and Traffic Analytics** solutions - object detection, tracking, line crossing, action recognition, crowd counting, pose estimation in Python and C++, leading to acquisition of **4 new Smart City contracts**.
- **Optimized** multi-GPU end-to-end pipeline 1.5x for real-time surveillance w/ low precision fp16 Mixed Precision, INT8 Quantization.
- Mentored 2 fellow teammates and gave technical sessions on **Edge AI** topics w.r.t. **Video Analytics** for Smart Cities.
- Executed 30+ camera **DeepStream pipeline profiling** to detect bottlenecks, optimizing for maximum **throughput**, minimizing **latency**, and devising efficient hardware distribution strategy for video stream processing.
- **Deployed** and trained **8+ deep learning models** (YOLO, Faster R-CNN, Mask R CNN, Vision Transformer, ResNet) using PyTorch, Tensorflow, Nvidia TensorRT, DeepStream SDK, TAO, Intel OpenVINO, ONNX, leveraging **MLOps**.
- Integrated MLFlow, DVC (Data Version Control), Kafka for ML data handling, leveraging **CI/CD** and **Docker** for containerization.

Computer Vision R&D Engineer - Vehant Technologies, Noida, India

Jun 2019 - Jul 2020

- Applied **image processing, classification, clustering** for 15+ multi-label Pedestrian Attribute Recognition e.g. clothing, gender.
- Explored monocular **depth estimation** methods for under-vehicle object detection with 90% accuracy with RGB-D data.
- Utilized **Semantic, Instance Segmentation** for Indian road scene understanding with U-Net, DeepLab, and Mask R-CNN.

NLP Intern - Arbunize Digital Media Pvt Ltd, Delhi, India

Jun 2018 - Aug 2018

- Leveraged **text-processing** techniques, including **Named Entity Recognition (NER)**, to parse resumes with nltk, scikit-learn.
- Extracted resume skills and applied supervised learning **random forest, gradient boosting** with 0.89 F1-score to predict job.
- Developed **Multinomial Naive Bayes, Support Vector Machine (SVM)** classifiers with 0.85 R2 across 4 MBTI personalities, utilizing natural language processing techniques like word embeddings, TF-IDF, **dimensionality reduction (PCA)**.

Projects

Visual Question Answering with Generative AI [\[github\]](#)

- Integrated Hugging Face pre-trained tokenizers, Vision Transformer for images, and LLMs for generating answers.
- Achieved 0.3 WUPS with RoBERTa and BEiT outperforming all 4 model combinations viz. ViT, DEiT with BERT, GPT.
- Optimized model with Parameter-Efficient Fine-Tuning (PEFT), LoRA (Low-Rank Adaptation), and 4-bit quantization.
- Deployed multimodal VQA in Docker for containerization, Kubernetes for orchestration, ensuring scalable, efficient service.

RAG for Financial Document Summarization [\[github\]](#)

- Developed Chatbot leveraging Retrieval Augmented Generation (RAG) with GPT, LLama 2, Gemma models to extract, summarize key performance indicators (KPIs) from 10-Q financial docs with LangChain, HuggingFace for LLM, Chroma for vector databases.

Sentiment Analysis on Amazon SageMaker [\[github\]](#)

- Deployed sentiment analysis model on AWS Sagemaker using PyTorch, with data processing, training on EC2 and S3 data.
- Integrated custom inference code into web app using AWS Lambda, API Gateway, and IAM roles for seamless functionality.

Image colorization of historical paintings with GAN [\[github\]](#)

- Leveraged U-Net and pix2pix Convolutional Generative Adversarial Network, to colorize grayscale historical paintings.
- Utilized CIELAB color space conversion and Patch discriminator for enhanced image-to-image translation.

Comparative Analysis of Local and Global Temperature Trends with SQL [\[github\]](#)

- Analyzed local and global temperature trends using SQL for data extraction revealing a consistent global temperature rise.

Skills

Programming Languages: Python, C++, C, SQL, Java, R, JavaScript

Machine Learning: PyTorch, Tensorflow, OpenCV, scikit-learn, xgboost, numpy, pandas, matplotlib, plotly, nltk, Dask, LangChain

Tools & Frameworks: Django, Flask, GStreamer, Git, DVC, Docker, Kubernetes, Kafka, Slurm, ONNX, Linux, HuggingFace

MLOps: AWS SageMaker, Azure, GCP, Nvidia DeepStream, TensorRT, MLFlow, Intel OpenVINO, GitHub Actions, CI/CD