Jaydeep Chauhan

LinkedIn : linkedin.com/in/jaydeep-chauhan-57ab3171/ Mobile: +1 (630) 943-9664
Github: github.com/Jd8111997 Email: jaydeept126@gmail.com
Kaggle: kaggle.com/jd8111997
Portfolio: jd8111997.github.io/

Education

- **Indiana University Bloomington, USA** Aug 2021 - May 2023
  Masters in Computer Science *Overall GPA:3.7/4*
  Courses: *Applied Algorithms, Elements of AI, Machine Learning, Computer Vision, Advance OS, Applied ML for Computational Linguistics, Masters Thesis, Music Data mining, Advance NLP*

- **Dharmsinh Desai University, Nadiad, India** Aug 2015 - May 2019
  B.Tech in Computer Science Engineering *Overall GPA:7.57/10*
  Courses: *Operating Systems, Data Structures, Analysis Of Algorithms, Artificial Intelligence, Machine Learning, Networking, Databases*

Skills Summary

- **Languages**: Python, C/C++, JavaScript, Java SE, R, SQL, Bash, PHP
- **Libraries**: PyTorch, Scikit, NLTK, Matplotlib, SpaCy, Gensim, TensorFlow, Keras, JAX, Pandas, Numpy, Transformers, OpenCV, LangChain, AutoGen, LlamaIndex, PineCone, PyTorch Distributed, Megatron, Lightning, FSDP
- **Tools**: GIT, MySQL, GCP, AWS, Microsoft Azure, VS Code, SQLite, Docker, Wandb, Kubernetes, Latex, Terraform, MLFlow, Sagemaker

Experience

- **Machine Learning Engineer**
  Indiana University School of Public Health-Bloomington, Remote, USA *Jan 2025 - Present*
  - **Scalable Data Pipeline Development**: Designed and implemented a scalable data pipeline to download and process 8 million PubMed documents, leveraging techniques such as TF-IDF and LLM-based semantic matching.
  - **PubMed Document Classification**: Developed a classifier to identify evidence-based nudge articles using in-context learning and few-shot learning with open-source LLMs.
  - **LLM Fine-Tuning & Quantization**: Fine-tuned and quantized an instruction-tuned LLaMA 3.1-8B model for structured information extraction, enabling domain-specific filtering in JSON format.
  - **Retrieval-Augmented Generation (RAG) System Development** : Building a RAG-based retrieval pipeline to recommend targeted nudge behaviors for patient problems, utilizing BGE and LLaMA-based embeddings with FAISS for semantic matching.
  - **Reminiscence Therapy AI Agent**: Designed and developed a fully speech-to-speech, persona-grounded AI agent for reminiscence therapy targeting dementia and Alzheimer's patients. Built using automatic speech recognition (ASR), large language models (LLMs), text-to-speech (TTS), and real-time avatar interaction in Unity. Developed in collaboration with researchers at Indiana University and the University of Notre Dame. The prototype secured an NIH R21 grant and is being prepared for clinical trials.

- **Machine Learning Software Engineer**
  Causify, Remote, USA *Feb 2024 - Dec 2024*
  - **Cryptocurrency Pipeline Development**: Designed and maintained Apache Airflow DAGs to handle downloading, resampling, transforming, and forecasting cryptocurrency prices using real-time and historical data from multiple vendors, exchanges, and trade types.
  - **Exchange Data Integration**: Built a robust pipeline for processing exchange data from the Crypto.com vendor, enhancing the system's capability to handle diverse trade data efficiently.
  - **API Development and Codebase Optimization**: Contributed to the Datapull team by introducing API enhancements and revising the codebase, leading to improved performance, maintainability, and ease of use.
  - **Custom LLM Research**: Collaborated on research projects to develop custom Large Language Models (LLMs) and intelligent agents to streamline and optimize the internal codebase and workflows.

- **Computer Vision Research Intern**
  Vimaan Robotics, Santa Clara, USA *June 2022 - Sept 2022*
  - **Image Quality Assessment (IQA)**: Developed a cutting-edge 'No Reference Image Quality Assessment method utilizing deep learning exclusively trained on synthetic data. This innovation accurately assesses image quality and pinpoints distorted regions in multitask settings, and exhibits robust generalization capabilities when applied to real-world data.
  - **Data valuation**: Worked on data valuation methods by implementing a meta-reinforcement learning approach, quantifying sample quality. This significantly contributed to enhancing the robustness of the data utilized in various projects.
  - **Image enhancement**: Played a key role in developing a pipeline for image enhancement using state-of-the-art deep learning methods for image deblurring.

- **Associate Data Scientist - DL Research**
  ParallelDots, Gurugram, India *Jun 2019 - Jul 2021*
  - **Object Detection Benchmarking**: Led the benchmarking of Domain Invariant Object Detectors across diverse client datasets, contributing to the enhancement of the ShelfWatch product. Evaluated and implemented state-of-the-art object detection models to replace the generalized object detector.
  - **Managed and optimized Computer Vision pipelines:**: Effectively oversaw and enhanced end to end Computer Vision pipelines for leading FMCG companies, including BAT, RB and Mondelez. Led client communications, addressing their requirements, and continuously improved pipelines to align with client expectations.
  - **Object Detection for Mobile Device**: Trained and optimized object detection models using ONNX, TensorRT and Tfjs for mobile devices.
  - **Research Work**: Made significant contributions to ongoing research in semi-supervised learning for dense object detection, resulting in publication at the CVPR 2021 RetailVision Workshop.

- **Software Engineering Intern**
  Bhaskaracharya Institute For Space Applications and Geo-Informatics(BISAG) *Dec 2018 - Mar 2019*
  - **Project**: Developed a decentralized chatting application for BISAG scientists using the open-source framework 'Matrix Synapse,' showcasing proficiency in Python, Angular, and SQLite.

Projects

- **Training diffusion models with RL**: Experiment with various RL techniques to improve the generation capabilities of stable diffusion models by fine-tuning it using LoRA adapters. Trained novel GFlowNet architecture to predict high reward trajectories during sampling an image with text prompt (Nov '24)
- **Kaggle LLM Science Exam challenge**: Experiment with various open book models such as Playtypus, Flan-T5, Mistral using LoRA for fine-tuning and also designed retriveal strategies (RAG) from Wikipedia to enhance the MAP on test dataset. Achieved 80.69% MAP on leaderboard. (Oct '23)
- **Google AI4Code – Understand Code in Python Notebooks**: Finetuned a CodeBert language model using mixed-precision training to predict the ranking of markdown cells in a code segment. Achieved a 75.36% Kendall Tau correlation with the ground truth test dataset. (Nov '22)
- **Volumetric rendering using Neural Radiance Fields**: Implementation of NeRF(https://www.matthewtancik.com/nerf) to render a 3D view of a scene. (May '22)
- **Supervised Contrastive Learning for pretrained language models**: Implementation of supervised contrastive learning for pretrained language models paper and analyzed the embeddings(https://arxiv.org/abs/2011.01403). (April '22)
- **Feedback Prize - Evaluating Student Writing and predicting effective arguments**: Evaluated student writing through Name Entity Recognition (NER), employing SOTA transformer models such as BigBird and Longformer. Used back-translation strategy for data augmentation and deepspeed to optimize training time. Achieved a 70% leaderboard score. (Mar '22)
- **Jigsaw Multilingual Toxic Comment Classification Challenge and Severity prediction**: Attained 94% Leader board accuracy by ensembling multiple XLM-Roberta-large language models which are trained using distributed training for classification challenge. Achieved 90% Leader board accuracy for predicting severity by combining Ridge Regression on top of embeddings of Roberta large language model. (Feb '22).
- **Speech Enhancer**: This project is about enhancing a audio files using end to end deep learning system using SEGAN(Speech Enhancement Generative Adversarial Network). (Aug '20)
- **Autoregressive Generative Models**: Implementation of Autoregressive Generative Models such as PixelCNN, PixelCNN++, Diffusion and Score based models on Standford Dog Dataset. (Aug '20)
- **Twitter sentiment extraction challenge**: Experimented with various Bert models, employing techniques such as K-Fold Cross Validation, differential learning rate, and gradient accumulation. Achieved a 71.5% Jacquard similarity score with ground truth test data. (June '20)
- **Music Recommendation Engine**: A music hosting site and music recommendation engine built in C# using collaborative filtering. (Oct '17)

Publications

- **Semi-supervised Learning for Dense Object Detection in Retail Scenes**: CVPR 2021 RetailVision Workshop, arXiv preprint:2107.02114(Jul '21)
- **Comparative Study of GAN and VAE**: International Journal of Computer Applications (0975 - 8887) Volume 182 - No.22, October 2018

Honors and Awards

- Won a silver medal for Cassava Leaf Disease Classification challenge on kaggle (ranked top 4% worldwide). - Jan 21
- Won a silver medal for SIIM-ISIC Melanoma Classification challenge on kaggle (ranked top 2% worldwide) - Aug 20
- 'Expert' in Kaggle Competitions(Current rank 851 out of 223, 635)
- Participated in and competed in 55 Kaggle competitions across various domains and modalities, including Computer Vision, NLP, Healthcare, Biotech, Speech, Time Series, Regression, GenAI etc.
- Successfully defended Master's thesis on Out-of-Distribution Generalization in Deep Learning.