Vaibhav Sahu

2679280709 | vaibhavs@seas.upenn.edu | linkedin.com/in/vaibhav | Portfolio

Education

University of Pennsylvania Philadelphia, PA
Master of Science in Scientific Computing - GPA: 3.83/4 *Aug. 2022 – May 2024*

Indian Institute of Science Bangalore, India
Bachelor of Science in Physics - GPA: 8.2/10 *Aug. 2016 – June 2020*

Technical Skills

Languages: Python, C/C++, MATLAB
ML Frameworks: PyTorch, TensorFlow, SQL, Spark, PyTorch, Scikit-learn, Hugging Face, OpenAI-API
Other Frameworks: pandas, NumPy, Matplotlib, Seaborn, OpenMP, AWS, Azure
Courses: Deep Learning, Computer Vision, Big Data Analytics, Applied Machine Learning, Natural Language Processing
Conceptual Skills: Machine Learning, Large Language Models, PEFT, RAG, LoRA, Natural Language Processing, Parallel Computing, Data Science

Experience

AI Engineer November 2024 – Present
IgniteIQ.ai *Remote, United States*

- Made an auto-tagging endpoint utilizing LLMs to convert correspondence letters between Contractors for a construction project to convert them into structured data with fields such as summary, key points, clauses cited from contract, etc.
- Designed a clause detection model using GPT to match detected clauses to agreement documents with 95% recall.
- Implemented an algorithm to match letters with other related letters in different conversation chains using semantic search, Regex, and Breadth-First Search.
- Added a new agentic chunking method using LLMs that can chunk documents and preserve contexts across a page.
- Skills: PostgreSQL, GPT-4o, Semantic Search, Faiss, RAG, Agentic Workflows, Python, Pydantic, FastAPI

High Performance Computing Engineer June 2021 – July 2022
Simyog Technology *Bangalore, India*

- Performed a hot-spot analysis of the **C++** Finite Element Analysis computational solver code using performance analysis tools such as **Intel Vtune**
- Achieved a 22% speedup by optimizing the Matrix-Vector Product in **Intel MKL** (detected bottleneck) by parallelizing using **OpenMP** to speed up computational solvers
- Implemented experiments on achieving the best way to parallelize 1000s of Matrix-Vector operations in the numerical solver
- Implemented concurrent **GMRES** algorithm for the computational solver utilizing contiguous memory for optimal memory fetch, resulting in a 40% speed improvement.

AI Research Intern April 2021 – June 2021
Simyog Technology Pvt. Ltd. *Bangalore, India*

- Setup the Pipeline for a Black-box measurement-based model to predict the output of ICs using **Machine Learning** and **Neural Networks**
- Implemented models in **Python** and trained for various different ICs using **TensorFlow**
- Translated existing **MATLAB** code to reconstruct waveforms for ICs using trained models to that in **Python** and packed the whole pipeline into a standalone executable

Research and Part-time experience

Graduate Teaching Assistant January 2023 – Present
University of Pennsylvania *Philadelphia, US*

- Teaching Assistant for courses 'Big Data Analytics', 'Computer Vision', 'Natural Language Processing'
- QA Tested Automated Homework Coding Notebooks
- Hosted Recitation covering concepts such as Deep Learning and Web Data Scraping using XML
- Provided support and guidance to students for understanding course material and solving homework
- Guided several teams across different semesters on various data science projects
- Skills: PyTorch

Graduate Research Assistant May 2023 – June 2024
Prof. Talid Sinno, University of Pennsylvania *Philadelphia, PA*

- Measured the Performance of Neural Network Potentials at predicting properties of materials
- Trained Symmetry Preserving Neural Network Potentials on the DeePMD framework for Copper systems
- Ran Molecular Dynamics simulations on LAMMPS using high capacity H100 and A100 GPUs on the cloud for analyzing the Performance of Neural Network Potentials

Projects

Playing Hangman with Transformers | *PyTorch, Hugging Face, NumPy*

- Devised a pre-training task in the form of a multiclass classification for the transformer to learn guessing letters in hangman
- Pre-trained Google CANINE on a corpus of 380k words to achieve optimal performance of 56% game winning accuracy
- Devised a self-play fine-tuning task for better performance
- Fine-tuned the model to achieve 63% winning accuracy within 6 wrong guesses and 87% accuracy within 10 wrong guesses

Masked Face identification using One-Shot Learning | *Python, PyTorch, NumPy, OpenCV*

- Applied transfer learning on Inception-ResnetV1 deployed as a Siamese Network for one-shot face identification to achieve 91% accuracy of the LFW Dataset
- Generated database of masked faces using LFW Dataset and Image Editing using OpenCV
- Retrained the models to achieve 82% accuracy on masked faces

Extracting Lexical Stylistic Notions From Words Using LLMs | *PyTorch, Hugging Face, NumPy, Scikit-learn*

- Performed Literature Review on Extracting Directions attributing to features, such as Complexity and Formality of Text
- Improved the Performance of LLM-based Contextual Word Embeddings on extracting lexical features and using them to classify phrases using cluster-based Anisotropy removing - accuracy improved from 64% to 83%
- Fine-tuned LLMs to do document-level classification for these features
- Used ML models to make new similarity measures that performed better than cosine similarity

Synthetic Data Classification with GPT | *Python, OpenAI-API, Pandas*

- Synthetic Data is a neat way to avoid privacy issues and cheap to generate
- Generated data by prompting GPT to generate queries by tuning the temperature setting
- Engineered prompts to classify these disputes into 3 categories
- Used Few-shot In-context learning to achieve $\approx$ 95 % accuracy

Analyzing Crime across LA | *Python, Pandas, NumPy*

- Plotted Geo-spatial correlation of crime and gun violence across LA
- Analyzed the relationships between time taken for crime reports with victim profiles and nature of crime
- Conducted t-tests to prove the significance of the insights from the previous analysis statistically - E.g. Women take longer to report crime than men in LA!

Certifications

Generative AI with Large Language Models	Certificate
Introduction to Machine Learning in Production	Certificate
DeepLearning.AI: Deep Learning Specialisation	Certificate
Fundamentals of Parallelism on Intel Architecture	Certificate