Zoaib Sihorwala

704-805-0669 | zoaibsihorwala@gmail.com | linkedin.com/in/zoaibsihorwala | github.com/zoaibs | U.S. Citizen

Education

Georgia Institute of Technology Expected May 2027

Bachelor of Science in Computer Science | *GPA: 4.0/4.0* *Atlanta, GA*

***Organizations**: I2P Startup, Big Data Big Impact, AI @ GT*

***Coursework:** Data Structures, Algorithms, Object Oriented Design, Linear Algebra, Artificial Intelligence*

Experience

Intuit Credit Karma May 2025 – August 2025

Data Science Intern *Charlotte, NC*

- **Built and processed a 22GB dataset** of user features, logins, notification history, and time-domain data using **BigQuery, SQL** and 6 DAGs in **Apache Airflow.**
- Used **Scikit-learn** to train a random forest model on the dataset to **generate 210000+ synthetic counterfactual labels**, estimating user logins had no notification been sent.
- Designed, built, and trained a **Tensorflow-based deep twin network** on the synthetic data and implemented an importance sampling based inference algorithm for accurate predictions.
- **Improved attribution rate by 52%** by scoring model on notification-attributed logins and identifying that less than half were validly attributed, potentially **saving costs for over 15 million notifications** daily.

Datacurve (YC W24) June 2024 – August 2024

Machine Learning Engineer Intern *Remote*

- **Developed and solved 150+ algorithmic coding problems** to train **4** large language models (LLMs) for clients using Python, Java, C++, and C#.
- Worked on in-house GenAI tools to improve the generation of training problems for client-specific models, **leading to 13% more effective model training.**
- **Identified and reported 6 platform bugs** to improve model training system stability and performance.

Michigan State University June 2023 – August 2023

AI Research Intern *Lansing, MI*

- **Created a 40GB dataset of over 7500 high-resolution satellite images** along with meteorological and air quality data using Python-based geospatial analysis software and 3 weather APIs.
- Led the development of an ensemble computer vision model with Pytorch using transformers, convolutional neural networks, and several image processing techniques to predict Air Quality Index in polluted areas.
- Trained model using parallel processing on MSU's **HPCC supercomputer and eight GPU servers.**
- **Achieved testing accuracy of 94%**, superior to existing state-of-the-art computer vision models.

Projects

Fairplay | *Python, AWS, Apache Spark* February 2025 – Present

- **Selected as a winner (top 4%) out of 750+** participants at GT Hacklytics for a CV web app that distinguishes between fouls and dives in soccer; earned entrance into Georgia Tech's I2P startup incubator.
- Currently architecting a scalable data pipeline using **AWS S3 and Apache Spark** to build a large scale training dataset from SoccerNet, an existing database of soccer clips.
- Designing the microservice architecture for the inference system to enable real time foul predictions, ensuring the app is **scalable** and **fault tolerant** for future production deployment.

MemoirAI | *Python, LangChain, OpenAI CLIP, VectorDB* August 2024 – December 2024

- Engineered a query optimization agent using **LangChain and GPT-4o**, which leveraged NLP to convert user inputs into vector search queries, achieving an average **processing time of 0.6 seconds.**
- Spearheaded the development of the analysis pipeline, **creating two AI agents** that established the app's primary text-to-image retrieval system.
- Implemented an image analysis agent using the **OpenAI CLIP model** to automate the conversion of **1500+** images into vector embeddings, building the foundational semantic index stored in a Pinecone vector database.

Technical Skills

Languages: Java, Python, SQL, C, C++, C#, HTML, CSS, Javascript
Libraries/Tools: PyTorch, TensorFlow, Apache Airflow, BigQuery, CircleCI, Pandas, Git (Version Control), Pytest