Jinam Shah

Raleigh, NC | (919)922-8481| jbshah@ncsu.edu | github.com/jinamshah | linkedin.com/in/jinamshah

I am a Machine learning engineer with over three years of experience in the NLP industry. I have designed and developed applications in the domain of ML, NLP, and Big Data. I have experience of developing for all three major cloud providers - AWS, GCP, Azure. I have designed systems that scale to handle multi-million requests and parallelize cost-effectively.

WORK EXPERIENCE

Machine Learning Intern, Cactus Communications *May 2022 – Present*

- Handling large scale **pattern recognition** on **250 TB** of raw data.
- Leading the **Machine learning track** for a project in the domain of **disambiguation of records** in a data lake, coordinating between 3 teams, 7 team members and 2 external vendors.
- Created a serverless highly available API that orchestrates **~100K long-running requests** with **sub-second SLA**.

Senior Software Engineer, Cactus Communications *June 2020 – July 2021*

- Guided the **architecture planning** and implementation of various products in ML/NLP and BigData, **bridging the gap** between the business and tech teams.
- Lead and Co-ordinated the efforts to have an open channel of communication with the Machine Learning teams at AWS and Azure.
- Designed a **data processing pipeline** for a Machine Learning product with **cumulative 24K CPU cores, 48Tb RAM**, generating over ~4.5Tb of data in under 2.5 hours. This was executed at **1/5th of the proposed cost from AWS Big Data Team**.
- Designed and implemented a **BigData platform** that ingests over 1.5TB every day and generates over 8TB every week. It manages over **~900TB** of data in the data lake.
- Setup the best practices and operational runbooks for the team to operate on Cloud across AWS, GCP and Azure.

Python Developer, Cactus Communications *June 2018 – June 2020*

- Reported and worked with the AWS **S3 team** on **fixing a bug** on prefix throughput.
- Designed and implemented products in the **image recognition** domain, leading the efforts in creating a **new business vertical** in the company.
- Designed and implemented an **ML product** from scratch in under a week that scales with **zero downtime**.

PROFESSIONAL PROJECTS

Transformer-based Document Classification

- Ensemble of DL and ML models (based on **BERT**) performing document classification for over an unprecedented **1500 classes** deployed using **serverless** architecture.
- Saved the organization around **$1M per annum** and reduced the TAT for the service from **8 hours to under 2 minutes** (down by 99.6%).

Serverless Image Recognition

- Image recognition software for determining **ethical compliance of images** added in research papers.
- Achieved state-of-the-art performance (**99.8% accuracy**) and deployed using **serverless** architecture.

Automated Language Correction

- Worked on the "**explainable AI**" portion behind an NLP software that focused on automated grammar correction.
- Led the efforts of building **scalable infrastructure and API** for the product.

Bias detection in text

- Worked on using NLP along with statistical machine learning to identify stereotypical biases in text.
- The project acts as a tool to assess the data quality for all future NLP tasks and help integrate bias detection as the part of the workflow for any text-based decision-making company.

Image caption generator

- Worked on the adversarial neural network for text generation and for image recognition.
- Achieved 96% accuracy on the flickr-30 dataset.

EDUCATION

Master of Science in Computer Science, North Carolina State University *August 2021 – May 2023*
GPA: 4.0
Thesis: Academic author name disambiguation using research topics
Relevant Courses: High-performance Machine learning, Neural Networks, Natural Language Processing, Artificial Intelligence-1, Automated Learning and Data Analysis, Design and Analysis of Algorithms, Software Engineering.

TECHNICAL SKILLS

Programming tools: Python, Pytorch, Tensorflow, Keras, Pandas, Django, Flask, Spark, SQL, Git, C, C++

Domain Expertise: Machine Learning, Deep Learning, Natural Language Processing, Image Recognition, Distributed Training, Big Data

Cloud technologies: AWS, GCP, AWS EC2, AWS S3, AWS Lambda, AWS RedShift, AWS Kinesis, AWS API Gateway, AWS CloudFormation, AWS ECS

Design Principles: Cost-effective, scalable, secure, reproducible, and reliable architecture, serverless architecture.

ACTIVITIES

Courses: Machine learning from Stanford University, AI programming with python nanodegree by Udacity.
Hackathons: Top 25 percentile in annual Reply Code hackathon with over 10000 teams worldwide, twice in a row.
Volunteer: AI4Good foundation, working on solving United Nations' Sustainable Development Goals.
Open source: Contributor of AllenAI's S2AND and Specter repositories, consistent contributor through HacktoberFest.