Sunit Singh

1450B Worthington Street
Columbus, OH, USA 43201
614-812-9608
sunitsingh.bitsg@gmail.com

OBJECTIVE

Machine Learning Engineer with 7 years of full time work experience in building and deploying reliable machine learning systems. Proficient in cloud-based ML systems and building scalable event-based microservice architectures. I am experienced in training / fine-tuning LLMs for downstream tasks in conversational AI and recommendation systems.

EXPERIENCE

Orbit Systems, Columbus, OH
Machine Learning Engineer, Aug 2024 - Present

- Created the NER-Engine to classify and annotate news articles with entities indicating useful leads for JobsOhio to approach businesses looking to invest and expand operations.
- Designed and developed asynchronous architecture using Apache Kafka to prompt large language model(GPT-4o) in a parallelized event micro-batching approach to tag continuous stream of scraped articles and persist the annotated output to MongoDb.
- Used LangChain's `ChatPromptTemplate` and `StructuredOutputParser` modules to enforce schema for unstructured I/O during LLM prompting.
- Deployed a pull-based monitoring system using Prometheus client to scrape metrics from services and Grafana to create monitoring dashboards and alerts.
- Deployed the end-to-end system on Azure cloud in a microservice architecture using FastAPI to communicate between services.

Amazon – Alexa Taskbot Challenge 2023, Columbus, OH
Graduate Research Assistant, Professor H. Sun, Apr 2023 - Sep 2023

- Co-led the development of Tacobot, a task-oriented dialog system for the Amazon Alexa Taskbot Challenge - 2023.
- Built comprehensive question answering and dialogue management by leveraging low rank adaptive (LoRA) LLM fine-tuning on Amazon's Wizard-of Tasks customer interaction datasets.
- Observed 37% increase in customer engagement rate with new dialogue strategy, and an overall 24% increase in task success rate.
- Trained and deployed end-to-end dialog system on Amazon AWS infrastructure.

Compass Real Estate, Bangalore, India
Senior Machine Learning Engineer, Jul 2021 - Jul 2022

- Built 'Similar Homes', a content-based recommendation system for users prospecting real-estate properties powered by deep neural network (DNN) trained on customer events on Compass website (www.compass.com). Achieved a 20% improvement in customer clickthrough rate on recommended properties, compared to the legacy system.
- Created ETL pipeline to automate log data extraction, validation, preprocessing and data warehousing, which led to reduction in data ingestion and analytics time by over a week.
- Leveraged Databricks to build and deploy cloud based MLflow pipeline to ensure model versioning and model serving and CI/CD of prediction models with real time streaming of property listing recommendations using Apache Kafka.

- Created OCR and NLU based document intelligence system, for vetting of Seller-Buyer Contracts for the state of California.

Salesken, Bangalore, India
Senior Data Scientist, May 2020 – Jun 2021

- Developed a conversational autopilot AI tool for sales conversations by fine tuning GPT2 model and ensemble learners to track user intents and sales pitch score during sales conversations.
- Generated an average ROI of 0.8 by providing real-time cues and lead prioritization recommendations to sales agents.
- Assigned sales agent to customers based on agent's pitch score and customer's intent to buy. Agent re-assignment achieved annual preventable losses of $14 million for EdTech vendor.

IPsoft, Bangalore, India
Research and Development Engineer, Mar 2018 – Apr 2020

- Developed an automated customer service response system for IPsoft's digital employee 'Amelia' by training hierarchical bidirectional LSTM with customer urgency index modeling.
- Raised customer satisfaction rating to 4.3 from 3.2, and improved issue resolution rate by over 25%.
- Deployed the model using on-premise NVIDIA Triton Inference Servers, ensuring fault tolerant failovers and integrated with 'Amelia' using GRPC framework.
- Developed cascaded-CNN based face detection and face recognition customer authentication system.

Tonbo Imaging, Bangalore, India
Software Development Engineer, Jul 2015 – Feb 2018

- Developed a vehicle monitoring system using motion segmentation, Kalman tracking and CNNs applied to IR images from toll booth cameras.
- Built a scalable and modular network communication library based on UDP and TCP protocols for optimized and secure communication of tracker annotated video feed.
- Developed infrared image processing pipeline comprising of image smoothing, non-uniformity correction(NUC) and histogram equalization using CUDA C++ programming.

QUALIFICATIONS & SKILLS

- **Interests:** Large Language Models, Conversational AI, Task Oriented Dialog Agents, Recommender Systems, MLOps, Distributed system design, Software Development
- **Software and frameworks:** AWS, PyTorch, Tensorflow, Scikit-learn, Numpy, Pandas, Google JAX, Apache Spark, Apache Kafka, Apache Airflow, Databricks, MS Excel, MS Access, Airflow, RDBMS, SQL, DynamoDb, MongoDb, Apache Cassandra, Docker, GIT, Atlassian JIRA, ElasticSearch, FASTAPI
- **Operating Systems:** Windows, UNIX
- **Programming Languages:** Python, Java, C, C++
- **Personal:** Presentation and public speaking skills, collaborative research, highly organized, project leadership skills

PUBLICATIONS

"Roll Up Your Sleeves: Working with a Collaborative and Engaging Task Oriented Dialogue System" - ACL/ISCA Special Interest Group on Discourse and Dialogue [SIGDIAL 2023]

"Optimal Route Searching in Networks with Dynamic Weights Using Flow Algorithms" - 2015 International Conference on Computational Intelligence and Communication Networks (CICN)

EDUCATION

The Ohio State University, Columbus, OH GPA: 3.85 (4.0 scale)
M.S., Computer Science and Engineering, Apr 2024
- **Thesis:** "Task Oriented Dialogue Systems"
Divided research into two phases. **Phase 1:** Exploring in-context learning and parameter efficient instruction fine-tuning of LLMs to develop domain specific and open-ended Question and Answering module for task-oriented dialogue agent. **Phase 2:** Utilizing DeepSpeed's Zero Redundancy Optimizer (ZeRO Stage 3 Offload) for memory efficient large model sharding during training.

Birla Institute of Technology and Science, (BITS), Pilani, India
B.E., Electrical and Electronics Engineering + M.Sc., Mathematics, Jul 2015

PROJECTS

Question Prompted Passage Summarization, Aug 2022 – Apr 2023
- Developed a novel approach for summary generation by using questions as cues.
- Combined T5 encoder, trained on question answering task, with T5, trained on passage summarization, using multi-headed self-attention.

Dialog Agent for Home Recommendation, Apr 2023 – Aug 2023
- Developed a chat-based requirement gathering tool for home recommendations.
- Created detailed dialog flow and NLU module to capture useful information from user responses.
- Used ElasticSearch to retrieve properties aligned with user preferences.

PROFESSIONAL ACTIVITIES AND EXTRA CURRICULARS

- "Best Projects" award for Advanced Database Systems, The Ohio State University, Columbus, OH
- "Budding Data Science Leadership Award" – Salesken, Bangalore, India
- Pursued Minors in Music (Singing) – The Ohio State University – School of Music
- Vocalist in Music Society at BITS Pilani, Goa