Sunit Singh

Columbus, OH, USA 43201
614-812-9608
sunitsingh.bitsg@gmail.com

OBJECTIVE

Data Scientist with 6 years of full-time work experience in building and deploying reliable machine learning systems. Proficient in data analytics, cloud-based ML systems, big data frameworks like Apache Spark and MLOps. Experienced in training / fine-tuning LLMs for generative AI tasks.

Seeking full-time opportunity from May 2024. Open to relocation.

EXPERIENCE

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

- Built learning-to-rank based DNN scorer for 'Similar Homes', a content-based recommendation system for users prospecting real-estate properties, that achieved over a 20% improvement in customer clickthrough rate.
- Created data processing and warehousing pipeline on Databricks with Spark SQL, creating a near real-time ETL workflow.
- Built a scalable MLflow pipeline for reproducible model training and experimentation, and integrated inference workflow with Apache Kafka for real time generation of property recommendations.
- Created document intelligence system using OCR and BM25 ranker, to create a textual semantic based document ranker for property seller-buyer contracts.

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, generating an average ROI of 8% by providing real-time cues and lead prioritization recommendations to sales agents.

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 on customer conversations, that raised customer satisfaction rating from 3.2 to 4.3, and improved issue resolution rate by over 25%.
- Deployed cascaded-CNN based face detection (MTCNN) and recognition (Facenet) models using on-premise NVIDIA Triton Inference Servers, ensuring fault tolerant failovers and integrated with 'Amelia' using GRPC framework.

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

- 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:** Data Science, Large Language Models, Conversational AI, Task Oriented Dialog Agents, Recommender Systems, MLOps, Distributed system design, Software Development
- **Software and frameworks:** AWS, PyTorch, Tensorflow, Trax, Scikit-learn, Numpy, Pandas, JAX, Apache Spark, Apache Kafka, Apache Airflow, Databricks, MS Excel, MS Access, Airflow, RDBMS, SQL, DynamoDb, MongoDb, Apache Cassandra, Docker, GIT, Atlassian JIRA, ElasticSearch
- **Operating Systems:** Windows, UNIX, Mac

- **Programming Languages:** Python, Java, C, C++
- **Personal:** Presentation and public speaking skills, collaborative research, highly organized, project leadership and team management

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 PEFT techniques, knowledge distillation and model sharding with DeepSpeed's Zero Redundancy Optimizer (ZeRO Stage 3 Offload) for training LLMs.

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

PROJECTS

Amazon – Alexa Taskbot Challenge 2023, Apr 2023 - Sep 2023
- Led the development of Tacobot, an AWS native task-oriented dialog system for the Amazon Alexa Taskbot Challenge - 2023.
- Built question-answering module with instruction-finetuned LLMs using PEFT techniques (LoRA) that helped increase customer engagement and an increase in task success rate by 26% compared to legacy system.

Knowledge Distillation to train Question Answering Models, Mar 2024 – present
- Applied knowledge distillation to train a small language model on domain specific Question Answering task, using gpt3.5-turbo as teacher model.
- Leveraged LangChain Expression Language(LCEL) for retrieval augmented generation(RAG) to generate informative answers to user queries.

Reinforcement Learning from Human Feedback(RLHF) Summary Detoxifier, Jan 2024 – Apr 2024
- Built an instruction-finetuned text summarizer using Proximal Policy Optimization (PPO) to remove toxic(offensive/bias) content from summaries.

Cross Task Collaboration in Large Language Models, 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.