Spandan Das

(571) 446 8105 • spandand515@gmail.com • sd325.github.io
spandand • SD325

Education

Carnegie Mellon University **August 2021 – May 2025**

B.S. Computer Science

Relevant Coursework: (PhD) Intro to Deep Learning [Python], Deep Reinforcement Learning [Python], (PhD) Advanced NLP [Python], Algorithm Design and Analysis, Machine Learning with Large Datasets [Python], (PhD) Convex Optimization, Intro to ML [Python], Intro to Computer Systems [C], Probability and Computing, Statistics and Computing

Experience

NVIDIA **May – August 2024**

Artificial Intelligence Engineer Intern

- Developed an anomaly detection system for NVIDIA's CI/CD pipeline for TEGRA chip production environment - increased efficiency of issue resolution and site reliability engineer team efficiency
- Implemented monitoring system to autonomously report issues in build, packaging, and testing processes via email and Slack
- Designed and built a real-time harmless error filter using ElasticSearch database to perform LLM search over vector database of log embeddings
- Technologies: **Python**, **ElasticSearch**, **Flask**

Apple **May – August 2023**

Machine Learning Engineer Intern

- Designed and implemented Golang backend service to automate labeling queries with LLM-based natural language understanding results
- Created LangChain-inspired LLM integration library to filter and annotate semantic search results over 2.3 billion datapoints across various Siri domains
- Improved a number of Siri functionalities used by millions of users around the world including web video ("show me how to bake a cake"), app launch ("open Facebook"), and sentence usage ("use autonomous in a sentence")
- Technologies: **Golang**, **Amazon Web Services (AWS)**, **Docker**

NASA Goddard Space Flight Center **June - August 2020; June – August 2021**

Research Intern

- Trained machine learning models (**TensorFlow**, **Scikit-learn**, **XGBoost**) on data from NASA's Global Precipitation Measurement mission's Core Observatory Satellite to reduce satellite costs
- Utilized NASA Center for Climate Simulation (NCCS) supercomputing cluster to work with large data (2016 and 2017 annual satellite data) and optimize training of bagging models using multithreading
- Presented research to GSFC Climate and Radiation Lab and at international conference (AGU Fall Meeting)
- `https://github.com/SD325/NASA_Internship_2020`

Research

CMU Language Technologies Institute (CX Group) **February – May 2024**

- Developed an active learning based approach for data-efficient instruction tuning for LLMs by utilizing data impact models
- Improved pretraining efficiency and effectiveness by continuously adapting to models' evolving data preferences
- Submitted to NeurIPS 2024
- Technologies: **HuggingFace**, **PyTorch**

Visual Question Answering with LLMs **May – August 2023**

- Redesigned the Winoground dataset as a visual question answering (VQA) problem
- Designed and evaluated modified data with various multi-modal LMs including MiniGPT4, Salesforce BLIP2, PromptCap, ViperGPT, LLaVA, and GPT4
- Submitted paper to EMNLP 2023
- Technologies: **HuggingFace**, **OpenAI API**, **SceneXplain**

CMU Robotics Institute (AirLab) **May – August 2022**

Research Assistant

- Developed an online camera calibration algorithm for a multi-view stereo setup (6 cameras; Double Sphere model) on drones used to determine real-time depth maps
- Technologies: **PyTorch**, **CUDA**, **OpenCV**, **Docker**

Spandan Das

☎ (571) 446 8105 • ✉ spandand515@gmail.com • 🌐 sd325.github.io
in spandand • SD325

Publications

Yu, Z.; **Das, S.**; Xiong, C. MATES: Model-Aware Data Selection for Efficient Pretraining with Data Influence Models. 2024. *Submitted to NeurIPS 2024*. [`https://arxiv.org/abs/2406.06046`]

Das, S.; Wang, Y.; Gong, J.; Ding, L.; Munchak, S.J.; Wang, C.; Wu, D.L.; Liao, L.; Olson, W.S.; Barahona, D.O. A Comprehensive Machine Learning Study to Classify Precipitation Type over Land from Global Precipitation Measurement Microwave Imager (GPM-GMI) Measurements. Remote Sens. 2022, 14, 3631. [`https://doi.org/10.3390/rs14153631`]

Pandey, R.; **Das, S.**; Thrush, T.; Liang, P.P.; Salakhutdinov, R.; Morency, L.-P. WinogroundVQA: Zero-shot Reasoning with Large Language Models for Compositional Visual Question Answering. 2023. [Link to Paper]

Das, S.; Samuel, V.; Noroozizadeh, S. TLDR at SemEval-2024 Task 2: T5-generated Clinical-Language Summaries for DeBERTa Report Analysis. Carnegie Mellon University, 2024. [`https://arxiv.org/abs/2404.09136`]

Achievements

- 2021 USA Math Olympiad – Top 2% (Top 550 out of 30,000+ contestants; 232.5 USAMO Index)
- American Invitational Math Exam - Score: 12/15 (2021), 11/15 (2020)
- USA Computing Olympiad (USACO) – Top 600 in nation (Gold Division)
- 2022 Goldman Sachs Quantathon – Honorable Mention
- 2021 CMU Math and Informatics Competition – 8th place (out of 220+ teams)
- 2021 PurpleComet Math Meet – Honorable Mention (3000+ teams), 1st place in state
- Carnegie Mellon University Dean's List - Fall 2021, Fall 2023

Extracurriculars

Clubs: Undergraduate Entrepreneurship Association (Executive Board), Scottie Ventures - Venture Capital Club (Analyst), CMU Sahara - Bollywood Fusion Dance Team, Quant Club, Intramural Tennis, Intramural Basketball
Activities & Interests: Tennis, Hindustani Classical Music, Basketball, Card/Board Games, Running (Marathon)