Rajan Vivek

+1 602 463 6531 ◊ rvivek@alumni.stanford.edu ◊ rvivek3.github.io ◊ /in/rajanvivek717

Education

Stanford University *2022 –2024*
MS Computer Science: Artificial Intelligence Specialization GPA: 4.06
Georgia Institute of Technology *2018 –2022*
BS Electrical Engineering: Signal Processing and Data Science Coursework Focus GPA: 3.96

Technical Skills

Languages Python, C++, Java, MATLAB, SQL, C, HTML, VHDL
Machine Learning PyTorch, Tensorflow, scikit-learn (sklearn), numpy, pandas, Deep Graph Library, ROS
Cloud Computing AWS Lambda, S3, EC2, API Gateway, DynamoDB, Domino, MapReduce
Relevant Courses Machine Learning, NLP, Foundation Models, Speech Processing, Statistics, Decision Making

Experience

Contextual AI Fall 2023
Research Intern
Ran 200+ experiments for LM (draft) model distillation for speculative decoding, achieving 2-2.6X inference speed up. Implemented semantic entropy uncertainty estimation for LLMs and productionized with LLM serving platform.

Stanford NLP Group Fall 2022 – Winter 2023
NLP Researcher
Designed representative data selection technique for efficient language model evaluation, leading to EACL publication. Studied benchmark hill-climbing during LM pretraining, data-relatedness metrics, zero-shot vs. fine-tuned performance.

Scale AI Summer 2023
Machine Learning Research Engineering Intern
Ran 100+ experiments to design novel spacio-temporal Q-former for video foundation model, surpassing InstructBLIP, VideoLLaMA, & MPlugOwl at causal and temporal video question answering (on NextQA). Used AWS Sagemaker.

JPMorgan Chase Asset Management Summer 2022
Deep Learning Intern
Developed and productionized a transformer-based entity recognition model for invoice processing. Designed custom data augmentation techniques, improving model performance by 11% (F1-score). Used AWS Lambda, S3, DynamoDB.

Lockheed Martin Missiles and Fire Control: Applied Research Summer 2021 & 2020
Deep Learning Intern
2020: Wrote image processing scripts for end-to-end classification pipeline and experimented with CNN architectures.
2021: Performed 60+ experiments with embedding algorithms for large graph data and segmentation for 3D point clouds.

Publications

Anchor Points: Benchmarking Models with Much Fewer Examples EACL 2024 Main (Long Paper)
Rajan Vivek, Kawin Ethayarajh, Diyi Yang, Douwe Kiela
Explainable Activity Recognition for Smart Home Systems ACM Transactions on Interactive Intelligence 2023
Devleena Das, Yasutaka Nishimura, **Rajan Vivek**, Naoto Takeda, Sean T. Fish, Thomas Ploetz, Sonia Chernova

Projects and Awards

CS 330 Best Project Award (2023): Top in ~100 student teams for "Synthetic Data Generation for Few-Shot Learning"
NSF Graduate Research Fellowship Honorable Mention (2022): National recognition for achievements in undergraduate research in explainable AI, as well as K-12 STEM outreach and robotics curriculum development.
HackGT 2019: Received two 1st place awards and 2nd overall (of 250 teams) for *Smooth.io*: A voice assistant-connected food scale and iOS app that calculates and visualizes nutritional content of ingredients in real-time. Used REST API, Arduino, C.
Opportunity Research Scholars 2019: 2nd overall (of 25 teams) for "Robust Deep Learning-Based Motion Planner"