Rajath Rao

+1 408-300-4564
rajath.rao@stonybrook.edu
https://www.linkedin.com/in/rajath-rao/
https://github.com/Jarhatz
https://jarhatz.github.io/rajath-rao/

EDUCATION

M.S. Data Science Stony Brook University, *New York* *Aug 2023 - Dec 2024*
B.S. Computer Science & Engineering University of California - Irvine, *California* *Sep 2019 - Mar 2023*

PAPERS

[1] **Rao, R.**, Ryant, N., Schwartz, HA. (2025). WhiSPA: Semantically-Psychologically Aligned Whisper with Self-Supervised Contrastive Learning. ***Under Review at ACL 2025***.

[2] Varadarajan, V., Lahnala, A., Ganesan, A., Dey, G., Mangalik, S., Bucur, A.M., Soni, N., **Rao, R.**, Lanning, K., Vallejo, I., Flek, L., Schwartz, H., Welch, C., & Boyd, R. (2024). Archetypes and Entropy: Theory-Driven Extraction of Evidence for Suicide Risk. *In Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024) (pp. 278–291).* ***ACL 2024***.

[3] **Rao, R.**, Zhou, L., Samaras, D. (2024). BrainDiffusion: Reconstructing Visual Semantics from Non-Invasive Neural Activity Readings. *Unpublished.*

[4] **Rao, R.**, Chitale, P., Tiwari, A. (2024). Thought2Text: Semantic Language Generation from Non-Invasive Neural Activity Readings. *Unpublished.*

RESEARCH EXPERIENCE

Graduate Research Assistant, Stony Brook University, *Stony Brook, NY* Aug *2023 - Present*
Human Language Analysis Beings (HLAB), *Advised by Dr. H. Andrew Schwartz*

- Developed WhiSPA [1], a powerful audio encoder which captures language-based semantics and affect directly from audio achieving high accuracy in emotion/personality recognition and semantic similarity tasks
- Contributed to a novel multimodal approach for mental illness prediction integrating audio/video data with transformer embeddings [1,2]
- Researched and developed deep/shallow-fusion techniques for speech processing models with language models
- Integrated psycholinguistic and neuroscientific insights with ML architectures for enhanced emotional assessments [2]
- **Utilized:** *PyTorch, NLP/CV, HuggingFace, CUDA, NCCL, NLTK/DLATK, SQL, Linux, Academic Writing*

PROFESSIONAL EXPERIENCE

AI/Deep Learning Engineer Intern, Intel, *Hillsboro, OR* May *2024 - Aug 2024*

- Profiled and enabled Microsoft's state-of-the-art SLM, Phi3, to run on Intel Habana's Gaudi2 hardware accelerators
  - Implemented graphs compilation (lazy execution) for 40% increase in inference throughput
  - Developed memory efficient optimization of reusing Key/Value cache matrices for 80% HPU utilization
- Conducted experiments and discovered performance regression on Intel Habana's HF transformers with DeepSpeed
- Benchmarked and optimized *Llama, Mixtral, StarCoder*, etc. for customer tickets with 15% higher GPU utilization on Gaudi2 using profiler trace analysis on kernel operations
- Developed model quantization, mixed precision, and collate functions for dynamic batching on NVIDIA's Triton Inference
- **Utilized:** *Python, PyTorch, HuggingFace, TensorBoard, DeepSpeed, CUDA, NVIDIA, SYCL, HCCL, Linux*

HPC Software Engineer Intern, Intel, *Santa Clara, CA* Jun *2021 - Apr 2023*

- Proposed and developed models for failure prevention, pattern recognition, data automation for HPC data centers
- Spearheaded development of an ensemble of neural networks for regressing hard drive failures up to 4 months prior
  - Trained ensemble notifies of imminent failures with a 92% accuracy on test environment servers
  - Saved terabytes of potential data loss alerting maintenance teams to backup hard drives
- **Utilized:** *Python, Machine Learning, Feature Engineering, Jupyter, Sci-Kit Learn, Flask, MongoDB, Docker*

SKILLS

Languages: *Python, Java/JavaScript/TypeScript, C/C++/C#, HTML/CSS, R, SQL, LISP, Perl, Verilog, Linux*
Skills: *Deep Learning, AI/ML, Optimization, Computer Vision, NLP, PyTorch, CUDA, HuggingFace, Data Analysis, AWS, Azure, Google Cloud, Apache/Hadoop, MPI, React/Node.js, Docker, Kubernetes, Git, JIRA*
Courses: *Reinforcement Learning, Computer Vision, NLP, ML/Statistical Computing, Advanced Algorithms & Data Structures, Probability, Time Series Analysis, Big Data Systems, Operating Systems, Embedded Systems Design*

PROJECTS

BrainDiffusion: Generate Images with Your Mind, Stony Brook University, *Research Paper* *Feb 2024 - Oct 2024*

- ❖ Proposed and developed a novel self-supervised contrastive learning framework to align EEG embeddings with CLIP image embeddings
- ❖ Developed a Masked Autoencoder (MAe) to extract latent feature representations for EEG signals
- ❖ Validated and fine-tuned Stable Diffusion on CLIP aligned EEG embeddings to generate images corresponding to respective EEG signals
- ❖ **Utilized:** Python, *PyTorch, HuggingFace, Computer Vision, CUDA, NCCL, Git*

Thought-to-Text: Generate Text With Your Mind, *Stony Brook University, Research Paper* *Jan 2024 - Jun 2024*

- ❖ Developed an Encoder-Decoder framework using fMRI stimulus responses from 6 subjects as they perceived speech from audio books
- ❖ Spearheaded development of predicting fMRI voxel-wise BOLD features using stimulus matrix from GPT2 logits
- ❖ Implemented a candidate sequence token prediction algorithm with Beam Search and nucleus sampling
- ❖ **Utilized:** Python, *PyTorch, HuggingFace, NLP, CUDA, NCCL, Git*

Stocker: Agentic RAG Stock Sentiment and Price Forecasting, *HackAI Dell & NVIDIA Challenge* *Sep 2024 - Oct 2024*

- ❖ Implemented a robust Agentic-RAG framework for portfolio asset management and financial question answering, utilizing knowledge graphs to effectively source and analyze popular stock news and articles
- ❖ Enhanced TimeSeriesTransformer performance for S&P500 stocks forecasting through strategic implementation of automatic mixed precision, optimizing training efficiency without compromising model accuracy
  - ➢ Developed an efficient sliding window algorithm for batching the data during tensor preprocessing
- ❖ Developed a full-stack web application for concurrent users with React (front-end) & Flask REST API (back-end)
- ❖ **Utilized:** *Python, NVIDIA AI Workbench, CUDA, HuggingFace, Langchain, React/Node.js, Docker, Git, Devpost, Video*

Autonomous IoT Shopping Cart with Intelligent Tracking, *UC - Irvine* *Oct 2022 - Mar 2023*

- ❖ Spearheaded the development of an IoT based autonomous shopping cart with user-following, lane-correction, object-collision, and product-search features
- ❖ Constructed a Mealy Finite State Machine for the autonomous drive-state decision making process of the cart
- ❖ Devised a WiFi triangulation algorithm using 3-point trilateration of Received-Signal-Strength-Indicator (RSSI)
- ❖ **Utilized:** *Arduino, ESP8266, C/C++/C#, Python, Scikit-Learn, Firebase, MQTT Broker, Mealy Machine (FSM), Git*