in/suhasraja (US Citizen)
github.com/Suhas7

Suhas Raja

Experience

Apple | *Systems Engineer, AirPods Applied ML* | C, Obj-C, Swift, PyTorch *02/2024 – Present*

- Incubating 4 multimodal feature **prototypes from concept to functional demos**, architecting inference stacks optimized for latency-bound operation, efficient model pipelining, and cross-device execution.
- Developed 2 composition models, collaborating across 6 ML & firmware teams to **orchestrate inference of audiovisual embedding models** across devices.
- Drove 2-year product roadmap proposals through internal scoping, competitive benchmarking, and literature reviews; 2 prototypes nominated for tentpole-level programs following executive reviews.
- Mentored ML research intern for 4 months, culminating with a **novel multimodal clustering pipeline**.

Qualcomm | *Research Intern, Multimedia Speech ML R&D* | Pytorch *05/2023 – 08/2023*

- Applied transformer & GRU-based architectures for on-device speech processing in various wireless channel conditions, **reducing reconstruction loss by over 25% against SOTA** methods.
- Developed pipeline and tools for experiment management, visualization, and analysis.
- Finetuned compression model using **RLHF on internal MOS dataset**.
- Results identified as one of **top-5 organization accomplishments** at SVP-level quarterly all-hands.

Amazon | *Software Engineer, Alexa Speech ML* | C++, FreeRTOS, SQL, Python *06/2021 – 06/2022*

- Led development of forced speech alignment for edge devices to support announcements.
- Supported on-call operations for on-device speech processing across over 100M devices.
- Refactored 5K+ LOC firmware codebase for custom SoC integration.

CMU Robotics Institute — Dr. Katia Sycara | *RISS Research Intern* *05/2020 – 09/2020*

The Walt Disney Company | *Graph ML Research Intern* *05/2019 – 08/2019*

Education

Carnegie Mellon University | *M.S. Computer Science* *3.9/4.0*
Coursework: On-Device ML (PyTorch), Operating Systems (C, x86), Adv. ML & Game Theory (PyTorch), Formal Verification (WhyML), Deep RL (PyTorch), Modern Computer Architecture (SystemVerilog)

The University of Texas at Austin | *B.S. Electrical & Computer Engineering, B.S.A. Mathematics* *3.9/4.0*
Scholarship: Gail and Howard Neal Endowed Scholarship in Electrical Engineering ($4700, awarded 2019 & 2020)
Teaching Assistant: Intro to Computing Systems, Embedded Systems, Algorithms *Minors:* Economics

Whitefish Bay High School | *High School Diploma – Milwaukee, WI*

Publications

Individualized Mutual Adaptation in Human-Agent Teams *IEEE Transactions on Human Machine Systems Journal, 2021*
Dynamic Programming Method to Optimally Select Power Distribution System Reliability Upgrades
IEEE Open Access Power and Energy Journal, 2021
Adaptive Agent Architectures for Realtime Human Agent Teaming *AAAI PAIR Workshop, 2020*

Projects

Low-Level CUDA Kernels for Transformer Primitives | *Personal Project* *03/2025 – Present*

- Writing and benchmarking kernels for model inference, ranging from softmax to multihead & flash attention.
- Bound kernels into minimal PyTorch pipeline for empirical comparison with standard operator performance.

Reputation-Aware Gossip Learning | *Adv. ML & Game Theory Research Project, CMU* *09/2022 – 12/2022*

- Led team of 3 to explore robust trust mechanisms for fully decentralized gossip-learning architectures.
- Applied imitation learning to determine and validate optimal heuristic, mitigating adversarial vulnerability.

Federated Learning Under Resource Constraints | *Senior Capstone Project* *01/2020 – 12/2020*

- Researched bandwidth optimization protocols for federated training architectures on edge devices.

Skills

Technologies: PyTorch, C/C++, CUDA, ARM & x86 Assembly, Swift & Obj-C, SystemVerilog, Python, Git, Blender.
Interests: FPV Drones, 3D Printing, Indie Games & Animation, Weightlifting, Home Automation.