RAKSHITH SHARMA SRINIVASA
Senior research scientist, Machine learning
Samsung Research America
Mountain View, CA
404.490.6520
rakshith.sharma.s@gmail.com
https://www.linkedin.com/in/rsrinivasa

SUMMARY

I am a Machine Learning Research Scientist with over four years of specialized experience in AI, focusing on computer vision, NLP, and speech processing. My research has led to multiple publications in top-tier machine learning conferences and journals. My interests are developing algorithms for multi-modal representation learning and enhancing model efficiency, both in training and inference. I am passionate about tackling complex technical challenges, developing innovative solutions and refining prototypes to meet specific application needs.

RESEARCH INTERESTS

Efficient inference for LLMs, LLMs for language translation, Multi-modal representation learning, vision-language pre-training, contrastive learning, speech representation learning, matrix factorization, low-rank matrix recovery, convex optimization

EDUCATION

Ph.D in Electrical and Computer Engineering Aug 2015 - December 2020
Georgia Institute of Technology, Atlanta, GA GPA:3.93/4.0
Advisor: Dr. Justin Romberg
Outstanding Research Award, 2020 (Center for Signal and Image Processing, Georgia Tech)
ITA Graduation Day Award, 2020

M.S in Electrical and Computer Engineering Aug 2014 - December 2020
Georgia Institute of Technology, Atlanta, GA GPA:4.0/4.0

B.Tech in Electronics and Communication Engineering July 2010 - May 2014
National Institute of Technology Karnataka, Surathkal, India GPA:9.36/10.0

EXPERIENCE

Senior Research Scientist (Machine learning), Samsung Research America (SRA) Dec 2021 – Present
Mountain View, CA

- Speculative decoding and multi-token prediction for efficient LLM inference
- Vision-language pre-training with SOTA zero-shot transfer accuracy
- Fine-tuning LLMs for improved language translation
- Instruction tuning of large multi-modal models (LMMs) for visual grounding and GUI understanding

Senior Machine learning Research Scientist - IQVIA Jan 2021 – Nov 2021
Cambridge, MA

- Developed ML solutions for clinical trial operations, health condition prediction and rare disease prediction
- Ranking medical providers for clinical trials, and fairness in patient selection for clinical trials

Machine Learning Research Intern – IQVIA Jan 2020 – May 2020
Cambridge, MA

- Developed an algorithm to improve computational efficiency of graph neural networks
- Graph neural network based models to study the spread of the COVID-19 using real world hospital data

Research Intern – Mitsubishi Electric Research Labs (MERL) May 2017 – Aug 2017
Cambridge, MA

- Convex optimization algorithms for array signal processing

Application Support Engineering Intern – MathWorks May 2015 – Aug 2015
Natick, MA

- Developed software in C++ for the signal processing toolbox, released as part of MATLAB R2016a

SELECTED PUBLICATIONS

- C. Lee, C. Yang, **R.S**, Y.M. Saidutta, J. Cho, Y. Shen, H. Jin, 'Leveraging self-supervised speech representations for domain adaptation in speech enhancement', **ICASSP**, Seoul, Korea, April 2024
- J.Cho, **R.S**, C. Lee, Y.M. Saidutta, C. Yang, **R.S**, Y. Shen, H. Jin, 'Zero-Shot Intent Classification Using a Semantic Similarity Aware Contrastive Loss and Large Language Model', **ICASSP**, Seoul, Korea, April 2024

- **R.S**, J. Cho, C. Yang, Y.M. Saidutta, C. Lee, Y. Shen, H. Jin, 'CWCL: Cross-Modal Transfer with Continuously Weighted Contrastive Loss', **NeurIPS**, New Orleans, Louisiana, December 2023
- C. Yang, Y.M. Saidutta, **R.S.**, C. Lee, Y. Shen, H. Jin, 'Robust Keyword Spotting for Noisy Environments by Leveraging Speech Enhancement and Speech Presence Probability', **INTERSPEECH Conference**, Dublin, Ireland, August 2023
- **R.S***, Y.M. Saidutta*, C. Lee, C. Yang, Y. Shen, H. Jin, 'To wake-up or not to wake-up: reducing keyword false alarm by successive refinement', **International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, Rhodes Island, Greece, June 2023 (* - equal contribution)
- **R.S**, S. Kim, K. Lee, 'Recovering sketched low-rank matrices with a shared factor by convex programming', **IEEE Journal on Special Areas in Information Theory (Special Issue: Sensing: Fundamental Limits and Modern Applications)**, 2023
- K. Lee, **R.S**, M. Junge, J. Romberg, 'Approximately low-rank recovery from noisy and local measurements by convex program', **Information and Inference: a journal of Institute of Mathematics and its Applications** (IMA), 2023
- **R.S**, C. Qian, B. Theodorou, J. Spaeder, C. Xiao, L.Glass, J. Sun, 'Clinical trial site matching with improved diversity using fair policy learning', **Preprint**, https://arxiv.org/abs/2204.06501
- J. Gao, **R.S**, C. Qian, L. Glass, J. Spaeder, J. Romberg, J. Sun, C. Xiao, 'STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence', **Journal of the American Medical Informatics Association** (JAMIA), November 2020
- **R.S**, C. Xiao, L. Glass, J. Romberg, J. Sun, 'FastGAT: Fast Graph Attention Networks Using Effective Resistance Based Graph Sparsification', **Preprint**, https://arxiv.org/abs/2006.08796
- **R.S**, K.Lee, J. Romberg, M. Junge, 'Tensor-norm-based convex program and performance guarantee for subspace-constrained blind deconvolution', **Invited paper, Asilomar conference on Signals, Systems, and Computers**, November 2020
- **R.S**, M. Davenport, J. Romberg, 'Sample complexity bounds for localized sketching' **AISTATS**, August 2020 , `https://arxiv.org/abs/2003.09097`
- **R.S**, M. Davenport, J. Romberg, 'Trading beams for bandwidth: imaging with randomized beamforming' **SIAM Journal on Imaging Sciences, 13:1, 317-350**, 2020, `https://doi.org/10.1137/19M1242045`
- **R.S**, K. Lee, M. Junge, J. Romberg, 'Decentralized sketching of low rank matrices' **Neural Information processing systems (NeurIPS)** , Vancouver, Canada, December 2019, `https://papers.nips.cc/paper/9200-decentralized-sketching-of-low-rank-matrices`

TALKS

- "Localized Sketching for matrix multiplication and regression", LightOn (Paris) summer seminar series, June 2020
- "Subspace learning and embedding with localized sketching" - Graduation day presentation, Workshop on Information theory and applications (ITA), San Diego, february 2020
- "Localized matrix sketching with applications to active array imaging", Spectrum Lab, Indian Institute of Science, Bangalore, India, Ferurary 2019

TECHNICAL SKILLS

- Python, Pytorch, C++, MATLAB, SQL, PySpark
- Linux, macOS, Git, LaTeX

SERVICE, TEACHING EXPERIENCE

- Area Chair, AISTATS 2023, 2024
- Reviewer, NeurIPS, ICLR, ICML
- Reviewer, Transactions on Signal Processing , Transactions on Information Theory
- Session Chair, Allerton Conference, 2018
- Teaching Assistant, **Math foundations of Machine learning, Statistical machine learning**