Hello,

I am a second year Master of Science in Computer Science student at University of Southern California Viterbi School of Engineering. I have taken CSCI 699 Special Topics - Computational Perspectives on the Frontiers of Machine Learning, CSCI 566 Deep Learning, CSCI 567 Machine Learning, and CSCI 570 Analysis of Algorithms courses so far. My research interests are chiefly Statistical Machine Learning, Edge-GPU Performance Characterization, Deep Learning Theory, Stochastic Optimization, Federated Learning, and Embedded Computing. I have been selected for the award of the prestigious J N Tata Endowment scholarship for the higher education of Indians, for the year 2023-24.

Currently, I work as a Machine Learning Software Intern at DeGirum Corp., Santa Clara, where I focus on model optimization for hardware. I am also a Research Assistant in the Department of Computer Science at USC Viterbi, advised by Prof. Vatsal Sharan, where I am contributing to an open source project on tensor decomposition methods, and working on a faster C++ implementation of a random forest based anomaly-detection algorithm. Previously, I was Project Associate - I at Distributed Research on Emerging Applications and Machines, DREAM:Lab, Indian Institute of Science Bangalore, advised by Prof. Yogesh Simmhan. I worked on performance characterization of Nvidia Jetson edge-accelerators in deep learning workloads, and also on a Federated Learning project during my tenure at the lab. Prior to that, I graduated with a Bachelor of Technology in Computer Science and Engineering from Indian Institute of Technology Dharwad. I worked on Federated Algorithms with Bayesian and Exponential Weighted Average approaches for my B.Tech. Project, as a member of LIaN research group, advised by Prof. B. N. Bharath.

Below is a list of publications where I had been an author/co-author —

[1] Prashanthi S. K, **Sai Anuroop Kesanapalli**, and Yogesh Simmhan. “Characterizing the Performance of Accelerated Jetson Edge Devices for Training Deep Learning Models”. In: *Proc. ACM Meas. Anal. Comput. Syst. 6, 3, Article 44 (December 2022)*. 2022. DOI: 10.1145/3570604.

[2] Prashanthi S. K, Aakash Khochare, **Sai Anuroop Kesanapalli**, Rahul Bhope, and Yogesh Simmhan. “Don’t Miss the Train: A Case for Systems Research into Training on the Edge”. In: *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 2022, pp. 985–986. DOI: 10.1109/IPDPSW55747.2022.00157.

[3] Prashanthi S. K, **Sai Anuroop Kesanapalli**, Aakash Khochare, and Yogesh Simmhan. “Characterizing the Performance of Deep Learning Workloads on Accelerated Edge Computing Devices”. In: *28th IEEE International Conference on High Performance Computing, Data & Analytics Student Research Symposium (HiPC SRS)*. 2021, [Poster].

[4] **Sai Anuroop Kesanapalli** and B. N. Bharath. “Federated Algorithm with Bayesian Approach: Omni-Fedge”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 3075–3079. DOI: 10.1109/ICASSP39728.2021.9413571.

I am interested to work on the areas of Deep Learning Theory, Embedded Computing, and Federated Learning (FL) in my academic and professional careers due to the following reasons:

From knowledge being gained through the graduate level courses on Deep Learning Theory, I am intrigued by open-ended questions such as the following – Why do deep neural networks generalize well on test data despite possessing the complexity to fit noise in train data?; How and why massive language models have the ability to learn something as difficult as and solve, a system of linear equations, just by taking the system of equations as a prompt?; Are there better ways to devise optimization objectives that could be efficiently solved and which implicitly capture the essence of regularization as well?

I am motivated to work on embedded computing from my experience at DREAM:Lab. In one of our works at the lab [2], we have explored how the accelerators at the edge are growing powerful by the day yet much work needs to be done into reconciling the training phase of the state-of-the-art neural networks with these edge-accelerators. Characterization of the computational bottlenecks and ways to address them is really important given that it is wasteful to let the hardware idle-out while training these massive neural networks [1], [3]. Since PyTorch is a popular open-source deep learning framework adopted by the research community and industry alike, how best one could fine-tune the features provided by PyTorch while implementing the models on edge is an interesting field of research in itself. As it is a common notion, we need to have models customized for the hardware and the hardware customized for models in order to achieve peak performance at both the ends.

Furthermore, FL is a Privacy-Preserving Machine Learning paradigm where the data stays where it is generated. It focuses on training ML models on a cluster of edge-devices in rounds, which could be the smartphones, say, and shares the model weights or gradients to the server where they are averaged and sent back to the edge-devices for the next round of training. However, FL comes with its own set of challenges such as system heterogeneity, non-I.I.D. data, stragglers, high communication costs and additional privacy concerns. Devising novel algorithms and architectures which address these challenges is a vast and open area of research which has the potential to impact billions of people and businesses. In [4], we have proposed one such algorithm based on a novel Bayesian approach. FL has a wide range of applications ranging from private personalized recommender systems on smartphones to private diagnosis of illnesses in a loop of hospitals. Further, since FL is a distributed ML paradigm at its core, solving theoretical challenges posed by FL involves the exploration of concepts in distributed and stochastic optimization, and more generally into statistical machine learning itself.

Thus, it is with these insights and enthusiasm that I wish to work on the aforementioned areas of research through which I believe I could impact billions of people and businesses positively. I wish to do so in the guidance and mentorship of engineers and researchers in the industry or academia. Towards achieving this goal, I am taking courses at the university through which I could enhance my knowledge of the the graduate-level concepts that are required towards the research work.

Thank you!

Best Regards

Sai Anuroop Kesanapalli

Correspondence Details –

Email ID: ksanu1998@gmail.com
Webpage: https://ksanu1998.github.io