Suraj Srinivas

ssrinivas@seas.harvard.edu · suuraj.srinivas@gmail.com · suraj-srinivas.github.io

research interests

Trustworthy Machine Learning (Interpretability, Explainability and Robustness);
Computationally Efficient Machine Learning; Large Language Models; Generative Models

work experience

2022 - **Postdoctoral Research Fellow**,
Harvard University, USA,
Faculty Advisor: Prof. Himabindu Lakkaraju
Research Topics: Interpretable Machine Learning, Robustness.

education

2021 **Doctor of Philosophy**,
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland,
Faculty Advisor: Prof. François Fleuret
Thesis: Gradient-based Methods for Deep Model Interpretability.

2017 **Master of Science (Engineering)**,
Indian Institute of Science, Bangalore, India,
Faculty Advisor: Prof. R. Venkatesh Babu
Thesis: Learning Compact Architectures for Deep Neural Networks.

internships

winter 2020 **Research Intern**, *Qualcomm AI Research, Netherlands*,
Research on algorithms to sparsify neural networks.

awards and honors

2022 **Best paper award** at ICML *Interpretable ML for Healthcare* Workshop

2022 **Highlighted reviewer** at *International Conference on Learning Representations (ICLR)*

2021 EPFL EDEE **PhD thesis distinction award** for top 8% thesis in EE

2019 ICML travel grant for ICML 2019

2017 **Best paper award** at NeurIPS *Learning with Limited Data* Workshop

2015 Xerox Research India travel grant for BMVC 2015

2014 **All India Rank 399** (99.8%ile) in the Graduate Aptitude Test in Engineering (GATE) for entrance to graduate school in electronics and communications engineering

2010 **State Rank 191** (99.8%ile) in the Statewide Common Entrance Test (CET) for entrance to undergraduate engineering programmes.

research summary

Total citations: 1750+ | **h-index:** 10

Long Papers (Long Conference + Journal papers): 12
Short Papers (Short Conference + Workshop papers): 7

highlighted talks

11/2023 *On the (Missing) Foundations of Interpretable Machine Learning*
Indian Institute of Technology, Hyderabad

03/2023 *Pitfalls and Opportunities with Feature Importance Methods*
MERL seminar series, Boston

07/2022 *Pitfalls and Opportunities with Feature Attribution Methods*
Simons Institute, UC Berkeley

03/2022 *Cyclical Pruning for Neural Network Sparsity*
Google Sparsity Reading Group

01/2020 *Full-Gradient Representation for Neural Network Visualization*
ML for Astrophysicists Club

11/2019 *Full-Gradient Representation for Neural Network Visualization*
Swiss Machine Learning Day, Lausanne

reviewing

Conferences
AAAI Conference on Artificial Intelligence (AAAI) - 2020
IEEE Conference on Computer Vision & Pattern Recognition (CVPR) - 2020
European Conference on Computer Vision (ECCV) - 2020
International Conference on Computer Vision (ICCV) - 2021
Workshop on Applications of Computer Vision (WACV) - 2021
Neural Information Processing Systems (NeurIPS) - 2020, 2021, 2022
International Conference on Machine Learning (ICML) - 2021, 2022
International Conference on Learning Representations (ICLR) - 2022, 2023
International Conference on Artificial Intelligence and Statistics (AISTATS) - 2023

Journals
IEEE Signal Processing Letters
Elsevier Neural Networks
IEEE Transactions in Pattern Analysis and Machine Intelligence
Nature Communications

teaching

2023 **Co-instructor** for *Interpretability and Explainability in ML*
Instructors: Prof. Hima Lakkaraju, Jiaqi Ma, Suraj Srinivas
Harvard University, USA
Webpage: `https://interpretable-ml-class.github.io/`

2018, '19, '21 **Teaching Assistant** for *Deep Learning*
Instructor: Prof. François Fleuret
EPFL, Switzerland

research mentoring

2023 Usha Bhalla & Alex Oesterling (PhD students, Harvard)
2022-23 Tessa Han (PhD candidate, Harvard)
2023 Usha Bhalla (PhD student, Harvard)
2023 Daniel Ley (PhD student, Harvard)
2022 Vincent Micheli & Karthigan Sinnathamby (MSc students, EPFL)
2017 Akshayvarun Subramanya (Research Assistant, IISc)
2016 Lokesh Boominathan (Research Assistant, IISc)

service

2023 **Co-organizer** of "XAI in Action: Past, Present, and Future Applications"
Workshop @ Neural Information Processing Systems (NeurIPS) 2023, New Orleans, USA