Suraj Srinivas

ssrinivas@seas.harvard.edu · suuraj.srinivas@gmail.com · Ph: (+1)857-995-7812

about me

I'm a machine learning researcher with 7+ years of academic experience in ML research. My experience includes areas such as interpretable & explainable machine learning, robust deep learning, computer vision, large language models (LLMs), representation learning and computationally efficient deep learning. My research has been published at machine learning conferences such as ICML, NeurIPS, ICLR, UAI, BMVC and CVPR Workshops.

For more information about my work, check out my webpage: suraj-srinivas.github.io

work experience

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

summer 2016 **Research Intern**, *DataGrokr, India / Verisk Analytics, USA*,
Speeding up inference on deep neural networks using tensor factorization.

fall 2014 **Engineering Intern**, *Tonbo Imaging, Bangalore*,
Implemented image processing algorithms on FPGA for a thermal imaging camera.

summer 2013 **Research Intern**, *Indian Institute of Science, Bangalore*,
Research on computational photography to perform camera jitter compensation.

skills

Python, Pytorch, Bash, Git, Slurm

Github stars: 200+

selected research publications

Total citations: 1900+ | **h-index:** 12

2024 Usha Bhalla*, Alex Oesterling*, **Suraj Srinivas**, Flavio Calmon, Hima Lakkaraju.
Interpreting CLIP via Sparse Linear Concept Embeddings (SpLiCE).

2023 Aounon Kumar, Chirag Agarwal, **Suraj Srinivas**, Soheil Feizi, Hima Lakkaraju.
Certifying LLM safety against adversarial prompting.

2023 **Suraj Srinivas***, Sebastian Bordt*, Hima Lakkaraju. (*co-first-author)
Which Models have Perceptually-Aligned Gradients? An Explanation via Off-Manifold Robustness.
Neural Information Processing Systems (NeurIPS) - **Spotlight (Top 3%)**

2022 **Suraj Srinivas***, Kyle Matoba*, Hima Lakkaraju, François Fleuret. (*co-first-author)
Efficient Training of Low-Curvature Neural Networks.
Neural Information Processing Systems (NeurIPS)

2022 Tessa Han, **Suraj Srinivas**, Hima Lakkaraju.
Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post hoc Explanations.
Neural Information Processing Systems (NeurIPS)
ICML Interpretable ML for Healthcare Workshop - **Best Paper Award**

2021 **Suraj Srinivas**, François Fleuret.
Rethinking the Role of Gradient-based Attribution Methods in Model Interpretability.
International Conference on Learning Representations (ICLR) - **Oral (Top 1%)**

2018 **Suraj Srinivas**, François Fleuret.
Knowledge Transfer with Jacobian Matching.
International Conference on Machine Learning (ICML)
NeurIPS Learning with Limited Data (LLD) Workshop - **Best Paper Award**

selected 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

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

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

selected talks

11/2023 *On the Missing Conceptual 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

06/2022 *Pitfalls and Opportunities with Feature Attribution Methods*
Vanderbilt University, USA

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

07/2016 *Making Deep Neural Networks Smaller and Faster*
Deep Learning Conf, Bangalore

reviewing

Conferences AAAI, CVPR, ECCV, NeurIPS (2020) ; WACV, ICML, ICCV, NeurIPS (2021); ICLR, ICML, NeurIPS (2022); ICLR, AISTATS (2023)

Journals IEEE SP-Letters, Elsevier Neural Networks, IEEE T-PAMI, 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

2021 **Guest Lecturer** on Interpretability for *Deep Learning for Computer Vision*
Instructor: Prof. R. Venkatesh Babu
Indian Institute of Science, Bangalore

research supervision

2023 Usha Bhalla & Alex Oesterling (PhD students, Harvard)
Concept Decompositions with CLIP, ongoing

2022-23 Tessa Han (PhD candidate, Harvard)
Local Function Approximation to Characterize Explanations, NeurIPS 2022
Efficient Estimation of Local Robustness, ICML Workshops, 2023

2023 Usha Bhalla (PhD student, Harvard)
Verifiable Feature Attributions, NeurIPS 2023

2023 Daniel Ley (PhD student, Harvard)
On Minimizing the Impact of Dataset Shifts on Actionable Explanations, UAI 2023

service

2023 Co-organized "XAI in Action: Past, Present, and Future Applications"
NeurIPS 2023 workshop