

Suraj Srinivas

ssrinivas@seas.harvard.edu · suuraj.srinivas@gmail.com · [suraj-srinivas.github.io](https://github.com/suraj-srinivas)

research interests

Interpretable ML; Robust & Safe ML; Computationally Efficient ML;
Computer Vision; Large Language Models; Vision-Language Models

work experience

- 11/2024 - **Research Scientist 2**,
Robert Bosch LLC, Sunnyvale, CA, USA
- 01/2022 - **Postdoctoral Research Fellow**,
08/2024 Harvard University, MA, USA,
Faculty Advisor: Prof. Himabindu Lakkaraju

education

- 2017 - 2021 **Doctor of Philosophy**,
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland,
Faculty Advisor: Prof. François Fleuret
Thesis: Gradient-based Methods for Deep Model Interpretability
- 2014 - 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
- 2010 - 2014 **Bachelor of Engineering**,
PES University, Bangalore, India,
Major: Electronics and Communication Engineering

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

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

research articles

Total citations: 2500+ | **h-index:** 15

highlighted papers

- 2024 Usha Bhalla*, Alex Oesterling*, **Suraj Srinivas**, Flavio Calmon, Hima Lakkaraju.
Interpreting CLIP via Sparse Linear Concept Embeddings (SpLiCE).
Neural Information Processing Systems (NeurIPS)
- 2024 Aounon Kumar, Chirag Agarwal, **Suraj Srinivas**, Aaron Li, Soheil Feizi, Hima Lakkaraju.
Certifying LLM safety against adversarial prompting.
Conference on Language Modelling (CoLM)
- 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 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**
- 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**

additional peer-reviewed publications

- 2024 Tessa Han*, **Suraj Srinivas***, Hima Lakkaraju.
Characterizing Data Point Vulnerability as Average-Case Robustness.
Conference on Uncertainty in AI (UAI)
- 2023 Usha Bhalla*, **Suraj Srinivas***, Hima Lakkaraju. (*co-first-author)
Discriminative feature attributions: Bridging post hoc explainability and inherent interpretability.
Neural Information Processing Systems (NeurIPS)
- 2023 Anna Meyer*, Dan Ley*, **Suraj Srinivas**, Hima Lakkaraju.
On Minimizing the Impact of Dataset Shifts on Actionable Explanations.
Uncertainty in Artificial Intelligence (UAI) - **Oral (Top 5%)**
- 2022 Marwa El Halabi, **Suraj Srinivas**, Simon Lacoste-Julien.
Data-Efficient Structured Pruning via Submodular Optimization.
Neural Information Processing Systems (NeurIPS)
- 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 **Suraj Srinivas**, Andrey Kuzmin, Markus Nagel, Mart van Baalen, Andrii Skliar, Tijmen Blankevoort.
Cyclical Pruning for Sparse Neural Networks.
Computer Vision and Pattern Recognition Workshops (CVPRW) - **Oral**
- 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%)**

- 2019 **Suraj Srinivas**, François Fleuret.
Full-Gradient Representation for Neural Network Visualization.
Neural Information Processing Systems (NeurIPS)
- 2018 Akshayvarun Subramanya, **Suraj Srinivas**, R. Venkatesh Babu.
Estimating Confidence for Deep Neural Networks through Density Modelling.
IEEE Conference on Signal Processing and Communications (SPCOM)
- 2017 **Suraj Srinivas**, Akshayvarun Subramanya, R. Venkatesh Babu.
Training Sparse Neural Networks.
Computer Vision and Pattern Recognition Workshops (CVPRW) - **Oral**
- 2017 Lokesh Boominathan, **Suraj Srinivas**, R. Venkatesh Babu.
Compensating for Large In-plane Rotations in Natural Images.
Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)
- 2016 **Suraj Srinivas**, R. Venkatesh Babu.
Learning the Architecture of Deep Neural Networks.
British Computer Vision Conference (BMVC)
- 2015 **Suraj Srinivas**, R. Venkatesh Babu.
Data-free Parameter Pruning for Deep Neural Networks.
British Computer Vision Conference (BMVC)

book chapters

- 2017 **Suraj Srinivas**, Ravi Kiran Sarvadevabhatla, Konda Reddy Mopuri, Nikita Prabhu, Srinivas SS Kruthiventi, R. Venkatesh Babu.
A taxonomy of deep convolutional neural nets for computer vision.
Book chapter: *Deep Learning for Medical Image Analysis, Elsevier*
Journal version: *Frontiers in Robotics and AI*

talks

- 03/2024 *Introduction to Machine Learning and Contestability*
Tufts University
- 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
- 08/2021 *Pitfalls of Saliency Map Interpretation in Deep Neural Networks*
HES-SO, Sierre, Switzerland
- 04/2021 *Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability*
ICLR (virtual)
- 01/2020 *Neural Network Interpretability using Full-Gradient Representation*
Indian Institute of Science, Bangalore

- 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
- 05/2019 *Complete Saliency Maps using Full-Jacobians*
Valais / Wallis AI workshop, Martigny
- 07/2018 *Knowledge Transfer with Jacobian Matching*
ICML, Stockholm
- 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

service

- 2023 Co-organizer of "XAI in Action: Past, Present, and Future Applications"
NeurIPS 2023 workshop
- 2024 Co-organizer of "Interpretable AI: Past, Present, and Future"
NeurIPS 2024 workshop