

Suraj Srinivas

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

research interests

Robustness, Interpretability, Computational Efficiency of Deep models;
Generative modelling; Representation learning

work experience

01/2022 **Postdoctoral Research Fellow**,
- Present Harvard University, USA,
Advisor: Prof. Hima Lakkaraju.

education

2017 **Doctor of Philosophy**,
- 2021 École Polytechnique Fédérale de Lausanne (EPFL), Switzerland,
Advisor: Prof. François Fleuret.

2014 **Master of Science (Engineering)**,
- 2017 Indian Institute of Science, Bangalore, India,
Advisor: Prof. R. Venkatesh Babu.

2010 **Bachelor of Engineering**,
- 2014 PES University, Bangalore, India.

internships

08/2020 **Research Intern**, *Qualcomm AI Research, Netherlands*,
- 01/2021 Research on algorithms to sparsify neural networks.

06/2016 **Research Intern**, *DataGrokr, India / Verisk Analytics, USA*,
- 08/2016 Speeding up inference on deep neural networks using tensor factorization.

01/2014 **Engineering Intern**, *Tonbo Imaging, Bangalore*,
- 06/2014 Implemented image processing algorithms on FPGA for a thermal imaging camera.

06/2013 **Research Intern**, *Indian Institute of Science, Bangalore*,
- 08/2013 Research on computational photography to perform camera jitter compensation.

publications

- 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**
- 2023 Usha Bhalla*, **Suraj Srinivas***, Hima Lakkaraju. (*co-first-author)
“Verifiable feature attributions: A bridge between 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**

- 2022 **Suraj Srinivas***, Kyle Matoba*, Hima Lakkaraju, François Fleuret. (*co-first-author)
 “Efficient Training of Low-Curvature Neural Networks”
Neural Information Processing Systems (NeurIPS)
 Code: github.com/kylematoba/lcnn (Jointly authored)
- 2022 Tessa Han, **Suraj Srinivas†**, Hima Lakkaraju. (†advising role)
 “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**
- 2022 Marwa El Halabi, **Suraj Srinivas**, Simon Lacoste-Julien.
 “Data-Efficient Structured Pruning via Submodular Optimization”
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**
 Code: github.com/idiap/rethinking-saliency
- 2019 **Suraj Srinivas**, François Fleuret.
 “Full-Gradient Representation for Neural Network Visualization”
Neural Information Processing Systems (NeurIPS)
 Code: github.com/idiap/fullgrad-saliency
- 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**
- 2017 **Suraj Srinivas**, Akshayvarun Subramanya, R. Venkatesh Babu.
 “Training Sparse Neural Networks”
Computer Vision and Pattern Recognition Workshops (CVPRW) - **Oral**
- 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

- 03/2023 *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
- 05/2021 *Pitfalls of Saliency Map Interpretation in Deep Neural Networks*
Harvard University, USA
- 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

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 Karnataka Common Entrance Test (CET) for entrance to undergraduate engineering programmes.

reviewing

Conferences AACL, 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