SAMYAK JAIN

(+91)9179144039 ⋄ samyakjain.cse18@itbhu.ac.in ⋄ DOB: 1st December, 1999
LinkedIn ⋄ Github ⋄ Webpage ⋄ Google Scholar ⋄ Twitter

EDUCATION

Indian Institute of Technology (BHU) Varanasi *August 2018 - May 2023*
Integrated Dual Degree (B.Tech + M.Tech) in Computer Science - CGPA : 9.55/10.0 *Master's Thesis*

AREAS OF INTEREST

AI Alignment, Science of Deep Learning, Interpretability, Understanding Learning Dynamics

EXPERIENCE

Microsoft Research India July 2024 - Present
Research Fellow *Mentor Navin Goyal*

Five AI and Torr Vision Group, Oxford October 2023 - June-2024
Research Intern *Mentor Puneet Dokania*

Krueger AI Safety Lab, Cambridge University May 2023 - October-2023
Research Intern *Mentor David Krueger*

Vision and AI Lab, Indian Institute of Science, Bangalore May 2020 - May-2023
Research Intern *Mentor Venkatesh Babu*

Theoretical Foundations of AI, Technical University of Munich May 2021 - August-2021
Research Intern *Mentor Debarghya Ghoshdastidar*

PUBLICATIONS

1. **What Makes Safety Fine-tuning Methods Safe? A Mechanistic Study**
   Samyak Jain, Ekdeep Singh, Kemal Oksuz, Tom Joy, Phil Torr, Amartya Sanyal, Puneet Dokania
   ICML workshop on Mechanistic Interpretability, 2024 **(Spotlight)**
   NeurIPS 2024 **[main]**

2. **Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks**
   Samyak Jain*, Robert Kirk*, Ekdeep Singh*, Hidenori Tanaka, Robert Dick, Tim Rocktaschel, Edward Grefenstette, David Krueger
   ICLR 2024 **[main][code]**

3. **DART: Diversify-Aggregate-Repeat Training Improves Generalization of Neural Networks**
   Samyak Jain*, Sravanti Addepalli*, Pawan Sahu, Priyam Dey, RV. Babu
   CVPR-2023 **[main][code]**

4. **Efficient and Effective Augmentation Strategy for Adversarial Training**
   Sravanti Addepalli*, Samyak Jain*, RV. Babu
   NeurIPS 2022 **[main][code]**

5. **Scaling Adversarial Training to Large Perturbation Bounds**
   Sravanti Addepalli*, Samyak Jain*, Gaurang Sriramanan, RV. Babu
   ECCV 2022 **[main][code]**

6. **Boosting Adversarial Robustness using Feature Level Stochastic Smoothing**
   Sravanti Addepalli*, Samyak Jain*, Gaurang Sriramanan*, RV. Babu
   SAIAD Workshop CVPR 2021 **[main][code]**

7. **Towards Understanding and Improving Adversarial Robustness of Vision Transformers**
   Samyak Jain, Tanima Dutta
   CVPR 2024 **[main]**

ACADEMIC PROJECTS AND COLLABORATIONS

Understanding the role of inductive biases of loss landscape in convergence *Navin Goyal*

- Analyzing the connections between Bayesian posterior of high likelihood models and lottery ticket hypothesis using mechanistic interpretability and their role in convergence of the model.

Mechanistic understanding of safety fine-tuning and jailbreaking attacks *Puneet Dokania, Ekdeep Singh, Amartya Sanyal, Phil Torr*

- Safety fine-tuning projects unsafe samples into model's (low rank) null space, resulting in safety.
- Model is unsuccessful in projecting jailbreaks into its null space, thus circumventing the safety.
- **Gemma Scope** highlighted value of using sparse autoencoders in LLMs based on insights in this work.

Mechanistic understanding of fine-tuning *Robert Kirk, Ekdeep Singh, David Krueger*

- Demonstrated that fine-tuning is unable to alter the model mechanistically, giving pretense of change.
- Reverse fine-tuning has become the staple method for evaluating unlearning.
- This work is often used to counter use of safety-finetuning as an assurance protocol. Some works [1], [2] have used this work to submit comments to **RFI related to NIST's** executive order concerning AI.

Exploring loss basin to find generalized solutions *RV. Babu, Sravanti Addepalli*

- Showed that using weight averaging of diverse models during training increases the convergence time for learning spurious features and aids the learning of robust features.
- Proposed method DART shows improvements on both in-domain and out of domain settings.

Using data augmentations effectively in adversarial training *RV. Babu, Sravanti Addepalli*

- Demonstrated for the first time that it is possible to use augmentations effectively in adversarial training irrespective of the type of augmentation and adversarial training (AT) method used.
- Demonstrated that weight space smoothing can help in preventing catastrophic overfitting.

Aligning adversarial training with Ideal training objectives *RV. Babu, Sravanti Addepalli*

- Observed that standard AT cannot generalize to larger perturbation bounds due to conflict in training.
- Developed Oracle-Aligned Adversarial Training (OAAT), which aims to align the model's predictions with the oracle labels of adversarial images.

Calibrating robust models to allow rejection of adversarial samples *RV. Babu, Sravanti Addepalli*

- Inspired by variational inference, proposed a stochastic classifier which aims to learn smoother class boundaries by sampling noise multiple times in its latent space during inference.
- Proposed method demonstrated improved robustness along with improved calibration.

Understanding gradient masking in vision transformers *Tanima Dutta*

- Past works have demonstrated gradient masking in vision transformers, but failed to analyze the cause.
- Demonstrated that softmax in attention causes floating point errors leading to gradient masking in VITs.

SCHOLASTIC ACHIEVEMENTS

- Recipient of **DAAD-WISE**, a research oriented scholarship program by German Government.
- Fellow of Berkeley Existential Risk Initiative **(BERI)**, which supported my research at Cambridge.
- Recipient of Summer Research Fellowship 2020 **(SRFP)**, a research program by Indian Government.
- All India rank 922 in JEE Advanced 2018 and 346 in JEE Mains 2018 out of 1 million+ candidates.
- Selected for the KVPY 2018 Fellowship (IISc, Bangalore) by the Govt. of India.
- Ranked in amongst **Top 300** students in India for Maths, Physics and Astronomy Olympiads at national level – INMO, INPhO, INAO 2018 and city topper in NTSE 2016.

FEATURED POSITIONS

Reviewer: NeurIPS 2024, ICLR 2024, ICML 2023, NeurIPS 2023, CVPR 2023 **(outstanding reviewer)**, CVPR 2022 **(outstanding reviewer)**, ICLR 2022 **(highlighted reviewer)**, ECCV 2022, NeurIPS 2022.

Teaching Assistant: Introduction to Database Management, Introduction to Machine Learning

- Conducted lab classes of undergraduate students with a batch size of over 80 students and managed lab evaluations along with assignments.