Soumyadeep Pal

✉ palsoum1@msu.edu | Google Scholar | GitHub

RESEARCH INTERESTS

Trustworthy AI : Machine Unlearning, Robustness (Backdoor, Evasion Attacks), Interpretability
Optimization for Machine Learning : Bi-Level Optimization, Zeroth-Order Optimization

EDUCATION

PhD in Computer Science — Sept. 2024 -
Michigan State University — Advisor: : Dr. Sijia Liu

Master of Science (Thesis) in Computing Science — Sept. 2019 - Jan 2023
University of Alberta — Advisor: Dr. Nilanjan Ray
GPA : 3.68 / 4.0

Bachelor of Electrical Engineering — August 2015 - August 2019
Jadavpur University, Kolkata, India
CGPA: 8.78 / 10

SELECTED RESEARCH EXPERIENCE

Graduate Research Assistant, Michigan State University — Sept. 2024 -
Topic : Machine Unlearning in LLMs — Advisor: : Dr. Sijia Liu

– Exploring the opportunities and pitfalls of applying Zeroth-Order Optimization in machine unlearning for large language models.
– Developing novel LLM unlearning algorithms using modern preference optimization methods, resulting in convergent unlearning methods.
– Leveraging data creation through a self-training framework to enhance the robustness of LLM unlearning against jailbreak, relearning attacks (Manuscript in preparation for ICML 2025)

Research Assistant, Alberta Machine Intelligence Institute — Jan. 2023 - Aug. 2024
Topic: Enhancing Classification Robustness using neural data — Advisor: Dr. Alona Fyshe

– Regularized the training of vision classification models using human fMRI data of visual cortices, obtained from the Natural Scenes Dataset
– Explored the robustness properties of such models in terms evasion attacks and common corruptions.

Research Collaboration, Michigan State University — May 2022 - Feb. 2023
Topic: Defense Against Backdoor Attacks — Advisor: Dr. Sijia Liu

– Developed a novel self-training approach using strong data augmentations to defend against backdoor attacks, with further exploration into self-supervised learning. (**Best Paper Finalist in Safe AI '23 @ AAAI 2023**)
– Developed a novel method for automatic identification of backdoor data in poisoned datasets using bi-level optimization, leveraging the prediction invariance of poisoned data to an input scaling factor. Achieved 4%-36% improvement in AUROC over baselines across diverse backdoor attack scenarios. (Publication in **ICLR 2024**)

Graduate Research Assistant, University of Alberta — May 2020 - Dec 2022
Topic : Diffeomorphic Image Registration — Advisor: : Dr. Nilanjan Ray

– Developed a postprocessing layer for deformable image registration that integrates seamlessly with deep learning pipelines for deformable registration.
– Ensured diffeomorphism by exponentiating Jacobians and reconstructing registration fields through Poisson reconstruction.
– Demonstrated effectiveness in 3D brain MRI registration. (Publication in **ICPR 2022**)

SELECTED PUBLICATIONS

Backdoor Secrets Unveiled: Identifying Backdoor Data with Optimized Scaled Prediction Consistency
Soumyadeep Pal, Yuguang Yao, Ren Wang, Bingquan Shen, Sijia Liu
The Twelfth International Conference on Learning Representations (ICLR 2024)

Towards Understanding How Self-training Tolerates Data Backdoor Poisoning
Soumyadeep Pal, Ren Wang, Yuguang Yao, Sijia Liu
The AAAI’s Workshop on Artificial Intelligence Safety, 2023. (Best Paper Award Candidate)

Towards Positive Jacobian: Learn to Postprocess Diffeomorphic Image Registration with Matrix Exponential
Soumyadeep Pal, Matthew Tennant, Nilanjan Ray
26th International Conference on Pattern Recognition 2022

SERVICES

Conference Reviewer: ICLR ’25, CPAL ’25, AISTATS ’25, ’24, ICASSP ’23, ’24
Journal Reviewer: IEEE TSP
Workshop PC: AdvML: New Frontiers in Adversarial Machine Learning @ ICML’ 22, ICML ’23, NeurIPS ’24

SELECTED PROJECTS / REPORTS

Automated Paper Review Generation using LLMs [Link]	Fall 2024
Brief note on PAC Learning [Link]	Winter 2021
Investigation of Action Imbalance in Experience Replay Buffers [Link]	Fall 2020
Survival Prediction using Probabilistic Graphical Models [Link]	Fall 2019

HONORS

Best Paper Finalist Award: The AAAI’s Workshop on Artificial Intelligence Safety, 2023
SURGE 2018 : Among 9 students in India, selected for the prestigious Students-Undergraduate Research Graduate Excellence program @ IIT, Kanpur.

SKILLS

Hugging Face, DeepSpeed, Pytorch, Python, C, Matlab, R, Git, LaTeX