Aounon Kumar

Profile

Researcher in **Trustworthy AI** focusing on the safety, security, and robustness of machine learning models. My research involves designing algorithms to certifiably defend models against adversarial inputs, for example, safeguarding large language models from prompts that circumvent safety guardrails. I have studied and contributed to model robustness in several machine learning domains, including **computer vision**, **reinforcement learning**, and **language modeling**. My work has been accepted at prominent machine conferences such as **ICML**, **ICLR**, and **NeurIPS**, and I am actively involved in collaborative projects within the academic community.

Contact details

@ aokumar@hbs.edu
🌍 **aounon.github.io**
✉ 150 Western Ave, Office #6220
Allston, MA 02134.

Skills

- Programming Languages: **Python**, MATLAB, C++.
- Deep Learning Frameworks: **PyTorch**, **Torchvision**, Hugging Face **Transformers**, Keras, TensorFlow.
- Other Tools: **Numpy**, Scipy, Pandas, Matplotlib, Linux, **Git**, LaTeX.

Experience

Harvard University, Research Associate. **2023–Present**
◇ Trustworthy Machine Learning, AI Security and Robustness, Language Modeling, Adversarial Machine Learning, Certified Robustness.

Amazon Go, Applied Scientist Intern. **Summer 2022**
◇ Human Action Recognition, Computer Vision, Uncertainty Estimation, Out-of-Distribution Detection, Anomaly Detection, Video Data.

Nokia Bell Labs, Research Intern. **Summer 2019**
◇ Machine Learning for Network Security, Anomaly Detection.
◇ Developing a firewall using network traffic to flag suspicious IP addresses.

Nokia Bell Labs, Research Intern. **Summer 2018**
◇ Neural Networks, Autoencoders.
◇ Analysing the class of problems that can be learned using a single-layer autoencoder with ReLU activation function.

University of Maryland, Graduate Research Assistant. **2021–2023**
◇ Adversarial Machine Learning, Certified Robustness, Distributional Robustness, Computer Vision, Image Classification, Adversarial RL, AI-content Detection.

Education

University of Maryland, Ph.D. in Computer Science. **2017–2023**
◇ Thesis title: *Extending the Scope of Provable Adversarial Robustness in Machine Learning.*
◇ Advisors: Soheil Feizi and Tom Goldstein.
◇ GPA: 3.84/4.0

Indian Institute of Technology Delhi, Master of Science (Research) in Computer Science and Engineering. **2015–2017**
◇ Thesis title: *The Capacitated k-Center Problem and its Variant with Vertex Weights.*
◇ Advisors: Naveen Garg and Amit Kumar.
◇ GPA: 9.69/10.

Indian Institute of Technology Mandi, B-Tech in Computer Science and Engineering. **2011–2015**
◇ Project title: *The Steiner Tree Problem.*
◇ GPA: 8.58/10.

Research Interests

Machine Learning, AI Security and Robustness, Language Models, Adversarial Robustness, Computer Vision, Distributional Robustness.

Media Coverage

The New York Times, Article link. **Aug. 2024**
◇ Featuring: *Manipulating Large Language Models to Increase Product Visibility,* Paper link.

Science News Magazine, Article link. **Feb. 2024**
◇ Featuring: *Certifying LLM Safety against Adversarial Prompting,* Paper link.

The Washington Post, Article link. **Jun. 2023**
◇ Featuring: *Can AI-Generated Text be Reliably Detected?* Paper link.

Other articles: Bloomberg, Wired, New Scientist, The Register, TechSpot.

Publications

For a complete and up-to-date list, please see Google Scholar.

Certifying LLM Safety against Adversarial Prompting [PDF] **COLM 2024**
◇ **Aounon Kumar**, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, Hima Lakkaraju.
◇ Paper: Conference version, ArXiv version.
◇ Code: `https://github.com/aounon/certified-llm-safety`
◇ Media Coverage: Science News Magazine, D^3 Institute at Harvard.

Manipulating large language models to increase product visibility [PDF] **Preprint 2024**
◇ **Aounon Kumar** and Himabindu Lakkaraju.
◇ ArXiv: `https://arxiv.org/abs/2404.07981`
◇ Code: `https://github.com/aounon/llm-rank-optimizer`
◇ Media Coverage: The New York Times.

Can AI-Generated Text be Reliably Detected? [PDF] **Preprint 2023**
◇ Vinu Sankar Sadasivan, **Aounon Kumar**, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi.
◇ ArXiv: `https://arxiv.org/abs/2303.11156`
◇ Code: `https://github.com/vinusankars/Reliability-of-AI-text-detectors`
◇ Media Coverage: The Washington Post, Bloomberg, Wired, New Scientist, The Register, TechSpot.

MedSafetyBench: Evaluating and Improving the Medical Safety of Large Language Models [PDF] **Preprint 2024**
◇ Tessa Han, **Aounon Kumar**, Chirag Agarwal, and Himabindu Lakkaraju.
◇ ArXiv: `https://arxiv.org/abs/2403.03744`

Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness [PDF] **ICML 2020**
◇ **Aounon Kumar**, Alexander Levine, Tom Goldstein, Soheil Feizi.
◇ Paper: Conference version, ArXiv version.

Policy Smoothing for Provably Robust Reinforcement Learning [PDF] **ICLR 2022**
◇ **Aounon Kumar**, Alexander Levine, Soheil Feizi.
◇ Paper: Conference version, ArXiv version.

Provable Robustness against Wasserstein Distribution Shifts via Input Randomization [PDF] **ICLR 2023**
◇ **Aounon Kumar**, Alexander Levine, Tom Goldstein, Soheil Feizi.
◇ Paper: Conference version, ArXiv version.
◇ Code: `https://github.com/aounon/distributional-robustness`

Center Smoothing: Provable Robustness for Networks with Structured Outputs [PDF] **NeurIPS 2021**
◇ **Aounon Kumar** and Tom Goldstein.
◇ Paper: Conference version, ArXiv version.
◇ Code: `https://github.com/aounon/center-smoothing`

Certifying Confidence via Randomized Smoothing [PDF] **NeurIPS 2020**
◇ **Aounon Kumar**, Alexander Levine, Soheil Feizi, Tom Goldstein
◇ Paper: Conference version, ArXiv version.
◇ Code: `https://github.com/aounon/cdf-smoothing`

Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks [PDF] **ICLR 2024**
◇ Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, **Aounon Kumar**, Atoosa Chegini, Wenxiao Wang, Soheil Feizi.
◇ Paper: Conference version, ArXiv version.

Detection as Regression: Certified Object Detection by Median Smoothing [PDF] **NeurIPS 2020**
◇ Ping-yeh Chiang, Michael J. Curry, Ahmed Abdelkader, **Aounon Kumar**, John Dickerson, Tom Goldstein.
◇ Paper: Conference version, ArXiv version.

Provable Robustness for Streaming Models with a Sliding Window [PDF] **Preprint 2023**
◇ **Aounon Kumar**, Vinu Sankar Sadasivan, Soheil Feizi.
◇ ArXiv: `https://arxiv.org/abs/2303.16308`

Tight Second-Order Certificates for Randomized Smoothing [PDF] **Preprint 2020**
◇ Alexander Levine, **Aounon Kumar**, Thomas Goldstein, Soheil Feizi.
◇ ArXiv: `https://arxiv.org/abs/2010.10549`

Publications (contd.)

On the cost of essentially fair clusterings [PDF] **APPROX 2019**

◇ Ioana O. Bercea, Martin Groß, Samir Khuller, **Aounon Kumar**, Clemens Rösner, Daniel R. Schmidt and Melanie Schmidt (ordered alphabetically by last names).
◇ Paper: Conference version, ArXiv version.

Capacitated k-Center Problem with Vertex Weights [PDF] **FSTTCS 2016**

◇ **Aounon Kumar**
◇ Paper link

Select Projects

Defending LLMs against Adversarial Prompting (AI Security, Language Modeling) **2023–2024**

◇ Designed a class of methods to defend language models against adversarial attacks that add malicious tokens to an input prompt to bypass safety guardrails.
◇ Paper to appear at COLM 2024.
◇ Code on GitHub.
◇ Media Coverage: Science News Magazine, D^3 Institute at Harvard.

Reliability of AI-Content Detection (Trustworthy Machine Learning, Language Modeling) **2023–2024**

◇ Designed and evaluated a theoretical framework to understand the fundamental challenges of AI generated image and text detection methods.
◇ Paper on AI-text detection, Code on GitHub.
◇ Paper on AI-image detection.
◇ Media Coverage: The Washington Post, Bloomberg, Wired, New Scientist, The Register, TechSpot.

Provable Robustness to Distribution Shifts (Computer Vision, Distributional Robustness) **2021–2023**

◇ Developed a robustness certificate for image classification models under input distribution shifts such as RGB shifts, hue shifts, and brightness/saturation changes.
◇ Paper at ICLR 2023.
◇ Code on GitHub.

Certifiable Robustness for Reinforcement Learning (Adversarial RL, Certified Robustness) **2021–2022**

◇ Developed a robustness certificate for an RL agent that guarantees that the total reward obtained by the agent under an adversarial attack remains above a certain threshold.
◇ Paper at ICLR 2022.

Academic Service

Served as a reviewer for prominent machine learning conferences:
◇ **NeurIPS** in 2021, 2022, and 2024.
◇ **ICLR** in 2024.

Relevant Coursework

◇ **Ph.D.:** Introduction to Quantum Information Processing, Scientific Computing, Advanced Numerical Optimization.
◇ **M.S.(R):** Advanced Algorithms, Theory of Computation and Complexity Theory, Cryptography and Computer Security, Machine Learning.
◇ **B-Tech:** Advanced Algorithms, Modern Techniques in Theory of Computation, Advanced Theory of Computation, Advanced Complexity Theory, Mathematical Concepts in Computer Science, Algorithm Design and Analysis, Advanced Data Structures and Algorithms, Formal Languages and Automata Theory, Artificial Intelligence, Pattern Recognition, Machine Learning.

Teaching Experience

University of Maryland, Teaching Assistant. **2017–2020**

◇ CMSC250: Discrete structures
◇ CMSC351: Algorithms
◇ CMSC451: Design and analysis of computer algorithms

Indian Institute of Technology Delhi, Teaching Assistant. **2015-2017**

◇ Discrete mathematics
◇ Introduction to Automata and Theory of Computation
◇ Analysis and Design of Algorithms