Zachary Coalson

Corvallis, OR | coalsonz@oregonstate.edu

Summary

I am a prospective PhD student working on trustworthy and socially-responsible AI under the supervision of Prof. Sanghyun Hong.

Education

Oregon State University, Corvallis, OR — Sept 2020 – Present

B.S. in Computer Science, Minor in Mathematics (GPA: 4.0/4.0)

Honors Thesis: Auditing the Robustness of Neural Architecture Search to Data Distribution Shifts

Academic advisor: Prof. Sanghyun Hong

Honors and Awards

Oregon State University **Honor Roll** — 2020 – 2024

Drucilla Shepard Smith Award for maintaining a cumulative 4.0 GPA — 2022, 2024

Finnley Academic Excellence Scholarship — 2020

Publications

Conference Publications

– **Zachary Coalson**, Gabriel Ritter, Rakesh Bobba, Sanghyun Hong, "BERT Lost Patience Won't Be Robust to Adversarial Slowdown", In the *37th Conference on Neural Information Processing Systems* **(NeurIPS)**, 2023, https://openreview.net/forum?id=TcG8jhOPdv. **(acceptance rate: 26.1%)**

Preprints

– **Zachary Coalson**, Huazheng Wang, Qingyun Wu, Sanghyun Hong, "Hard Work Does Not Always Pay Off: Poisoning Attacks on Neural Architecture Search", *arXiv preprint*, 2024, https://arxiv.org/abs/2405.06073.

– **Zachary Coalson**, Jeonghyun Woo, Shiyang Chen, Yu Sun, Lishan Yang, Prashant Nair, Bo Fang, Sanghyun Hong, "PrisonBreak: Jailbreaking Large Language Models with Fewer Than Twenty-Five Targeted Bit-flips", *arXiv preprint*, 2024, https://arxiv.org/abs/2412.07192.

Research Experience

Bit Flip Attacks to Jailbreak Large Language Models — April 2024 – Nov 2024

– Created a comprehensive bit flip attack pipeline.

– Evaluated the pipeline on eight open-source large language chat models across two harmful tasks.

– Demonstrated state-of-the-art attack success while flipping minimal bits.

Data Poisoning on Neural Architecture Search — Dec 2023 – May 2024

– Developed a gradient-based clean-label poisoning attack to audit the robustness of NAS algorithms.

– Evaluated the attack on two representative NAS algorithms and one computer vision dataset.

– Discovered that such algorithms are surprisingly robust to practical poisoning attacks.

Slowdown Attacks on Input-Adaptive NLP Models — Aug 2022 – Dec 2023

– Designed an objective function for gradient-based slowdown attacks.

– Developed two slowdown attacks based on the state-of-the-art adversarial text attacks on NLP models.

– Performed an evaluation of the attacks on three input-adaptive NLP models across seven datasets.

– Demonstrated 100% attack success and proposed potential countermeasures such as input sanitization.

Professional Academic Activities

Conference Presentations

- BERT Lost Patience Won't Be Robust to Adversarial Slowdown, Poster Presentation, NeurIPS '23 — Dec 2023