Ren Pang

☎ (484)747–2401 | ✉ ain-soph@live.com | ain-soph.github.io

My research focuses on developing safe, robust and resilient machine/deep learning applications. Experienced in addressing the security concerns in image classification, AutoML, etc.

EDUCATION

Ph.D.	*Information Sciences and Technology*	Pennsylvania State University	2019–2023
B.Sc.	*Mathematics*	Nankai University	2014–2018

WORK EXPERIENCE

Machine Learning Engineer (Intern), *Meta* 2022 Summer

Pages and Groups Integrity: Introduce new classification model for malicious page detection. It mitigates the impact of incorrect label annotation, and provides interpretable classification outputs for better user experience.

TorchVision: Provide the official TorchVision implementation of SwinTransformerV2.

Applied Research Scientist (Intern), *Amazon* 2023 Summer

Explore the vulnerabilities of LLMs to jail-breaking attacks, where Reinforcement Learning from Human Feedback (RLHF) is considered to enhance the attack and defense efficiency. During project development, I submitted several bug fixes and new features to Transformers, Peft and Trl libraries.

PUBLICATIONS

1. A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models,
R. Pang, H. Shen, X. Zhang, S. Ji, Y. Vorobeychik, X. Luo, A. Liu, and T. Wang,
Proceedings of *the ACM Conference on Computer and Communications Security* (**CCS**), 2020.
2. AdvMind: Inferring Adversary Intent of Black-Box Attacks,
R. Pang, X. Zhang, S. Ji, X. Luo, and T. Wang,
Proceedings of *the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (**KDD**), 2020.
3. i-Algebra: Towards Interactive Interpretability of Deep Neural Networks,
X. Zhang, **R. Pang**, S. Ji, F. Ma, and T. Wang,
Proceedings of *the AAAI Conference on Artificial Intelligence* (**AAAI**), 2021.
4. Graph Backdoor,
Z. Xi, **R. Pang**, S. Ji, and T. Wang,
Proceedings of *the USENIX Security Symposium* (**USENIX**), 2021.
5. On the Security Risks of AutoML,
R. Pang, Z. Xi, S. Ji, X. Luo, and T. Wang,
Proceedings of *the USENIX Security Symposium* (**USENIX**), 2022.
6. TrojanZoo: Towards Unified, Holistic, and Practical Evaluation of Neural Backdoors,
R. Pang, Z. Zhang, X. Gao, Z. Xi, S. Ji, P. Cheng, and T. Wang,
Proceedings of *the IEEE European Symposium on Security and Privacy* (**EuroS&P**), 2022.
7. The Dark Side of AutoML: Towards Architectural Backdoor Search,
R. Pang, C. Li, Z. Xi, S. Ji, and T. Wang,
Proceedings of *the International Conference on Learning Representations* (**ICLR**), 2023.
8. On the Security Risks of Knowledge Graph Reasoning,
Z. Xi, T. Du, C. Li, **R. Pang**, S. Ji, X. Luo, X. Xiao, F. Ma, and T. Wang,
Proceedings of *the USENIX Security Symposium* (**USENIX**), 2023.
9. An Embarrassingly Simple Backdoor Attack against Self-supervised Learning,
C. Li, **R. Pang**, Z. Xi, T. Du, S. Ji, Y. Yao, and T. Wang,
Proceedings of *the International Conference on Computer Vision* (**ICCV**), 2023.

10. Reasoning over Multi-view Knowledge Graphs,
Z. Xi, **R. Pang**, C. Li, T. Du, S. Ji, F. Ma, and T. Wang,
Arxiv Preprint, 2022.

OPEN-SOURCE CONTRIBUTION

1. **TrojanZoo** (*owner*) https://github.com/ain-soph/trojanzoo
   (70,000 Lines) Offer a universal, flexible PyTorch platform to conduct security analysis of attacks and defenses on deep neural network models.

2. **AlpsPlot** (*owner*) https://github.com/ain-soph/alpsplot
   (14,000 Lines) Offer a high-level python library to plot academic figures based on Matplotlib.

3. **TorchVision.SwinTransformerV2** https://github.com/pytorch/vision/pull/6246
   (400 Lines) Provide TorchVision official implementation of SwinTransformerV2.

4. **TorchVision.AutoAugmentation** https://github.com/pytorch/vision/pull/6609
   (400 Lines) Provide TorchVision official implementation of AutoAugmentation for object detection. This work is based on the next generation PyTorch APIs.

5. **Matplotlib.Text** https://github.com/matplotlib/matplotlib/pull/20101
   Fix Text class bug when font argument is provided without math_fontfamily.

TEACHING EXPERIENCE

CSE 017: Structured Programming and Data Structures, *Lehigh University* 2018 Fall

SELECTED PROJECTS

Mutual Reinforcement of Adversarial Inputs and Poisoned Models
The project presents a new unified attack model called “IMC” that jointly optimizes adversarial inputs and poisoned models. It shows that there are mutual reinforcement effects between the two attack vectors and enables a large design spectrum for the adversary to enhance existing attacks such as backdoor attacks. It also discusses potential countermeasures and technical challenges, pointing to promising research directions.

Inferring Malicious Intent of Adversarial Machine Learning
The project presents a new class of estimation models that infer the intent of black-box adversarial attacks in a robust and prompt manner by taking into account fake queries and proactively soliciting subsequent queries to maximize exposure of the adversary’s intent.

Exploring Vulnerabilities of AutoML Architectures
This project examines the potential security risks of using neural architecture search (NAS) in machine learning systems. The study finds that NAS-generated models are more vulnerable to malicious manipulations compared to manually designed models. The study also provides explanations for this vulnerability, such as early convergence during training, and suggests potential remedies such as increasing cell depth or suppressing skip connections.