Exploitation and Mitigation: Understanding Large-Scale Machine Learning Robustness under Paradigm Shift

Traditional Tutorial: 2 hours

Xi Li
University of Alabama at Birmingham
xli7@uab.edu

Ruixiang Tang
Rutgers University
ruixiang.tang@rutgers.edu

Muchao Ye
The University of Iowa
muchao-ye@uiowa.edu

Abstract

The advancement of machine learning (ML), particularly evident in large-scale models like Large Language Models (LLMs), has showcased impressive capabilities across multiple domains. However, the increased complexity and enhanced capabilities of these models present new challenges to existing robustness frameworks. In this tutorial, we aim to delve deeply into the evolving paradigm of ML robustness for both large-scale and conventional small models, and offer insights for future research. We will revisit the two primary types of attacks – test-time adversarial attacks and training-stage poisoning attacks – and discuss how the evolution of large-scale models has altered attack and defense mechanisms, which also impact smaller-scale models. More importantly, we will summarize the differences between traditional and novel approaches to highlight the paradigm shift in ML robustness. We believe this is an emerging and critical area in ML robustness and it is expected to garner significant attention from both researchers and practitioners in academia and industry.

1 TARGET AUDIENCE

The target audience includes researchers, students, and conference attendants with basic deep learning knowledge. In particular, researchers from the fields related to AI safety and adversarial machine learning are encouraged for further discussion and Q&A in this tutorial.

The audience from the SIAM data mining community will be interested in this topic because they mostly develop new data mining methods based on new machine learning techniques such as large language models and multi-modal large language models. Meanwhile, the robustness of such new types of ML models has not been systematically studied yet. Our tutorial will provide an opportunity for the audience to study the robustness of ML models under the current paradigm shift. After attending our tutorial, the participants will obtain a comprehensive understanding on ML robustness and will obtain the skill of developing robust and safe data mining methods based on large-scale machine learning models.

We believe our tutorial will attract the attendance of the audience because the data mining community is enthusiastic about attending tutorial related to AI safety and adversarial machine learning held in data mining conferences such as SDM, KDD, and CIKM.

2 TUTORS

Dr. Xi Li (presenter, xli7@uab.edu) is currently an Assistant Professor in the Department of Computer Science at the University of Alabama at Birmingham. She received her Ph.D. from the Department of Computer Science and Engineering at Pennsylvania State University. She earned her B.E. from Southeast University and her M.S. from Pennsylvania State University. Her research focuses on trustworthy AI and adversarial machine learning, with a specific emphasis on poisoning attacks and defenses against deep neural networks. Her work has contributed to several publications in conferences such as ICCV, AAAI, and ICASSP, as well as in journals like IEEE Transactions on Knowledge and Data Engineering (TKDE).

Dr. Ruixiang Tang (presenter, email: ruixiang.tang@rutgers.edu) is an Assistant Professor in Computer Science at the Rutgers University. He received his PhD from the Department of Computer Science at the Rice University. Before that, he obtained his Bachelor of Engineering degree in Automation at Tsinghua University. His research focus is in the realm of Trustworthy AI, and specialize in issues related to safety, privacy, and explainability. Additionally, he collaborate closely with health informaticians from Yale, UTHealth, and Baylor College of Medicine, leveraging AI to address critical challenges in healthcare. His work has been published on top venues, including NeurIPS,

ICLR, KDD, WWW, CIKM, ACL, EMNLP, NACCL, TKDD and CACM.

Dr. Muchao Ye (presenter, email: muchao-ye@uiowa.edu) is an Assistant Professor in Computer Science at the University of Iowa. He received his PhD from the College of Information Sciences and Technology at the Pennsylvania State University in 2024. Before that, he obtained his Bachelor of Engineering degree in Information Engineering at South China University of Technology. His research interests lie in the intersection of AI, security, and healthcare, with a focus on improving AI safety from the perspective of adversarial robustness. *His representative works in this direction include VLAttack [36], VQAttack [37], and ADR [35], which discuss the vulnerability of the new "pretrain-and-finetune" machine learning paradigm applied in large-scale models.* His research works have been published in top venues including NeurIPS, KDD, AAAI, ACL, and the Web Conference.

3 DESCRIPTIONS OF THE TOPICS TO BE COVERED

3.1 Introduction (20 mins)

Machine learning (ML) is undergoing a significant transformation from traditional small-scale models to large-scale frameworks such as Large Language Models (LLMs, e.g., GPT series [1], LLaMA [26], Gemini[9]), large vision models (e.g., ViT[6]), and large multi-modal models (e.g., CLIP[21] and Stable Diffusion [22]). These foundation models (FMs) have shown remarkable capabilities across various domains, fundamentally changing paradigms of data processing and analysis. However, their complexity and advanced features introduce challenges to existing robustness frameworks, requiring innovative approaches to ensure model reliability and security. We will begin the tutorial as follows:

- We will briefly introduce the applications of popular foundation models, highlighting their advancements across several key tasks, including natural language processing, image recognition and generation, visual question answering, and image captioning.

- We will briefly review traditional ML robustness, focusing on classic attacks and defenses, specifically adversarial attacks at inference time and poisoning attacks during the training stage. We will introduce the threat models and assumptions for these attacks and defenses, using examples to illustrate each. Detailed discussions on the new paradigm of ML robustness can be found in Sections 3.2 and 3.3.

3.2 Recent Paradigm Shift in Attacking and Defending ML in Test Time (30 mins)

Test-Time Attacks. In this section, we will introduce the recent paradigm shift in attacking ML models. We will discuss test-time attacks for large-scale ML models obtained through two popular learning paradigms: "pretrain-and-finetune" [19, 5] paradigm and test-time adaptation [29]. First, we will emphasize the treatment of transferring adversarial examples constructed from pretrained models to fine-tuned ones if the large-scale ML models are learned in the fashion of "pretraining-and-finetuning". We will highlight a novel type of ML threat in test time: because of the homogeneity of structures applied in downstream tasks and the pre-trained ones, attackers can construct adversarial examples against the open-sourced pre-trained models and transfer them directly to the black-box fine-tuned models. Specifically, we will discuss how attackers can construct transferable adversarial examples by utilizing the accessibility of open-source pretrained model. We will take PDCL-Attack [34] and ADR [35] as examples to illustrate that structure-wise homogeneity leads to a threat on ML models in uni-modal tasks. Besides, we will further take representative multi-modal attack methods including Co-Attack [38], VLAttack [36], and VQAttack [37] such an idea also works for attacking multi-modal ones.[1] Second, we will discuss the security threat brought by the use of test-time adaptation [29], which becomes popular under ML paradigm shift because it avoids the necessity of collecting new data and fine-tuning the model. We will take Distribution Invading Attack [32] as an example to discuss how maliciously constructed adversarial examples injected in the test-time adaption will prevent the model from learning good results from testing data, which will make ML models misclassify in test time. We will also discuss gradient-based attacks [4] and on test-time adaptation.

Test-Time Defenses. After discussing the works above regarding new threats to ML models, we will then introduce the recently proposed works in text adversarial defense. First, we will discuss techniques [30, 17, 2] investigating the adversarial robustness specifically for "pretrain-and-finetune" models like CLIP [20]. We will detail the specific design of "pretraining" or "fine-tuning" that can amend the adversarial robustness of large-scale ML models. We will also discuss methods [23] using non-transferable learning and gradient masking to prevent adversarial perturbations from being transferred during the "pretrain-and-finetune" pro-

[1] ADR, VLAttack, and VQAttack come directly from the tutors' own research.

cess. Moreover, we will reveal the solutions [18, 8] to defending against malicious test samples during test-time adaptation by taking Sotta [8] and Medbn [18] as instances.

3.3 Rethinking Robustness at the Training Stage (50 mins)

Training-Time Attacks. We begin with a brief introduction to how FMs enhance the performance of small models. While FMs offer vast knowledge and versatility, they have limitations in scenarios requiring fast responses, limited computation, or data resources. Small models remain crucial in these contexts and can benefit from FMs through techniques such as knowledge distillation [10, 33, 31], synthetic data generation [16, 27], and model compression [3, 40, 15]. This allows small models to leverage the generality of large-scale models while adapting to specific downstream tasks.

However, the interaction between FMs and small models introduces new poisoning attack strategies. The attacker's approach evolves significantly due to the advanced capabilities of FMs. For instance, a backdoor attack can be triggered in a large language model (LLM) using a simple natural language prompt [28, 11, 24]. These vulnerabilities can easily transfer to small models that interact with compromised large-scale models. Backdoors planted in teacher models can be passed to student models during knowledge distillation [7]. Similarly, synthetic data generated by FMs may carry hidden backdoors from malicious prompts, compromising small models that rely on the synthetic dataset [13].

We then specialize the novel backdoor poisoning attack mechanism in FM-integrated federated learning (FL) scenarios [13, 14, 12].[2] A compromised FM integrated into FL introduces a new one-access external poisoning mechanism, fundamentally different from classic backdoor attacks against FL. We will detail this novel poisoning strategy, highlight changes in the threat model, provide practical attack scenarios, and present quantitative results demonstrating the effectiveness and feasibility of this new attack approach.

Finally, we will compare traditional and novel poisoning attack strategies, highlighting key differences in attacker abilities and methods, and summarize the "new paradigm" of poisoning attacks, along with the limitations of existing defenses.

Training-Time Defenses. In response to these novel poisoning strategies, training-time defenses must be reconsidered and enhanced to protect against vulnerabilities arising from the interaction between FMs and small models. Standard training-time defenses, such as gradient masking, model pruning, and robust training, were designed for classic backdoor scenarios but may fall short in FM-enhanced environments.

Recent research has identified various strategies to defend against backdoor attacks during the training phase of pre-trained language models (PLMs). One notable defense mechanism is Moderate Fitting [39], which suggests restricting the training dynamics of PLMs to a moderate-fitting stage to prevent backdoor triggers from being learned effectively. The approach is based on the observation that PLMs undergo two distinct learning phases: an initial stage where they focus on general features and a subsequent overfitting stage where they learn minor features, including backdoor triggers. To prevent the model from entering the overfitting stage, this defense proposes reducing the model's capacity, training epochs, and learning rate, thereby maintaining the PLM's focus on benign data while ignoring backdoor patterns.

Another defense strategy, Honeypot-Based Defense [25], introduces a novel module called the honeypot classifier, which traps backdoor triggers using lower-layer representations of the PLM.[3] This module absorbs the backdoor features early in the training process, allowing the main task classifier to focus on legitimate data. By isolating backdoor features, the honeypot effectively neutralizes the adversary's ability to embed malicious behaviors. Experimental results indicate that the honeypot defense significantly reduces the attack success rate across various backdoor attacks, including word-level, syntactic, and style-based attacks, while maintaining high performance on benign samples.

We will explore how existing training-time defense frameworks can be extended and propose new defense mechanisms specifically designed for FM-integrated systems. Our evaluation will compare the effectiveness of traditional defenses versus novel strategies, providing a comprehensive overview of protection techniques in FM-enhanced models.

3.4 Conclusion and Future Work (20 mins)

In the final section, we will summarize the aforementioned works and discuss possible future research directions such as designing attack and defense mechanisms for large-scale machine learning models applied in real-world applications such as healthcare and education. We will also discuss how to provide a comprehensive evaluation benchmark for evaluating the effects of novel evasion attacks and poisoning attacks.

[2] Works discussed here come directly from the tutors' own research.

[3] This work comes directly from the tutors' own research.

4 Potential Societal Impacts

Since recently developed AI systems such as OpenAI o1 and GPT-4o adopt new learning paradigms for training, which have attracted great attention from the general public and their concerns on large-scale models, our tutorial will let the audience and the general public understand what the vulnerabilities are for those large-scale models and how to make them safe and secure. Our tutorial can provide the audience with a basic understanding of robustness in recent large-scale AI models and a direction for developing more reliable AI systems in the future.

5 Similar Tutorials

The following list includes the previous similar tutorials and we differentiate ours from them as follows. We find that there is a lack of data mining tutorials on the vulnerability of large-scale machine learning robustness trained with new paradigms.

- "Vulnerabilities of Large Language Models to Adversarial Attacks" at ACL 2024 in Bangkok, Thailand. **Similarities and differences**: This tutorial provides a comprehensive overview of vulnerabilities in Large Language Models (LLMs) revealed through adversarial attacks. We will explore poisoning attacks on LLMs alongside adversarial attacks, focusing on the new paradigm in machine learning robustness. We'll discuss differences between traditional and novel approaches, with insights to enhance large-scale model robustness.

- "Robustness at Inference: Towards Explainability, Uncertainty, and Intervenability" at CVPR 2024 in Seattle. **Similarities and differences**: This tutorial focuses on a human-centric approach to robust image understanding, making neural networks explainable, uncertainty-aware, and open to human intervention. Our tutorial will also discuss the robustness of LLMs, and provide a deeper exploration of ML robustness, thoroughly addressing both test-time and training-stage attacks, along with advanced defense strategies.

- "Towards Adversarial Learning: from Evasion Attacks to Poisoning Attacks" at KDD 2022 in Washington, D.C. **Similarities and differences**: This tutorial focuses on the vulnerabilities of conventional deep neural networks (DNNs) to attacks at both the test and training stages. By contrast, we revisit machine learning robustness from the perspective of large-scale models, examining the vulnerabilities of foundation models and traditional small-scale models in light of the current paradigm shift.

- "Machine Learning Robustness, Fairness, and their Convergence" in KDD 2021 held virtually. **Similarities and differences**: This tutorial includes a discussion of robust training for machine learning in image data besides discussing fairness and training convergence. Nonetheless, this tutorial is still based on image data and lacks a discussion of adversarial attack methods.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 699–708, 2020.

[3] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 2018.

[4] Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. In *International Conference on Machine Learning*, pages 4421–4435. PMLR, 2022.

[5] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Yunjie Ge, Qian Wang, Baolin Zheng, Xinlu Zhuang, Qi Li, Chao Shen, and Cong Wang. Anti-distillation backdoor attacks: Backdoors can really survive in knowledge distillation. In *ACM MM*, 2021.

[8] Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Google. Gemini. https://gemini.google.com/, 2024.
[10] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.
[11] Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Backdoor attacks for in-context learning with language models. *CoRR*, abs/2307.14692, 2023.
[12] Xi Li and Jiaqi Wang. Position paper: Assessing robustness, privacy, and fairness in federated learning integrated with foundation models. *CoRR*, abs/2402.01857, 2024.
[13] Xi Li, Songhe Wang, Chen Wu, Hao Zhou, and Jiaqi Wang. Backdoor threats from compromised foundation models to federated learning. 2023.
[14] Xi Li, Chen Wu, and Jiaqi Wang. Unveiling backdoor risks brought by foundation models in heterogeneous federated learning. In *PAKDD*, 2024.
[15] Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *ICML*, 2020.
[16] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *EMNLP*, 2023.
[17] Changdae Oh, Mijoo Kim, Hyesu Lim, Junhyeok Park, Euiseog Jeong, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. *arXiv preprint arXiv:2311.01723*, 2023.
[18] Hyejin Park, Jeongyeon Hwang, Sunung Mun, Sangdon Park, and Jungseul Ok. Medbn: Robust test-time adaptation against malicious test samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5997–6007, 2024.
[19] Alec Radford. Improving language understanding by generative pre-training. 2018.
[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
[23] Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. *arXiv preprint arXiv:2402.16382*, 2024.
[24] Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. *CoRR*, abs/2304.12298, 2023.
[25] Ruixiang Ryan Tang, Jiayi Yuan, Yiming Li, Zirui Liu, Rui Chen, and Xia Hu. Setting the trap: Capturing and defeating backdoors in pretrained language models through honeypots. *Advances in Neural Information Processing Systems*, 36:73191–73210, 2023.
[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
[27] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. Generating faithful synthetic data with large language models: A case study in computational social science. *CoRR*, abs/2305.15041, 2023.
[28] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *NeurIPS*, 2023.
[29] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.
[30] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24502–24511, 2024.
[31] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, 2022.
[32] Tong Wu, Feiran Jia, Xiangyu Qi, Jiachen T Wang, Vikash Sehwag, Saeed Mahloujifar, and Prateek Mittal. Uncovering adversarial risks of test-time adaptation. *ICML 2023*, 2023.
[33] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *CoRR*, abs/2402.13116, 2024.
[34] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Prompt-driven contrastive learning for transferable adversarial attacks. *ECCV 2024*, 2024.
[35] Muchao Ye, Xiang Xu, Qin Zhang, and Jonathan Wu. Sharpness-aware optimization for real-world adversarial attacks for diverse compute platforms with enhanced transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-*

tern Recognition, pages 2937–2946, 2024.

[36] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Ziyi Yin, Muchao Ye, Tianrong Zhang, Jiaqi Wang, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vqattack: Transferable adversarial attacks on visual question answering via pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6755–6763, 2024.

[38] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022.

[39] Biru Zhu, Yujia Qin, Ganqu Cui, Yangyi Chen, Weilin Zhao, Chong Fu, Yangdong Deng, Zhiyuan Liu, Jingang Wang, Wei Wu, et al. Moderate-fitting as a natural backdoor defender for pre-trained language models. *Advances in Neural Information Processing Systems*, 35:1086–1099, 2022.

[40] Michael Zhu and Suyog Gupta. To prune, or not to prune: Exploring the efficacy of pruning for model compression. In *ICLR Workshop Track Proceedings*, 2018.