Adversarial Robust Machine Learning*

Jia Fu
KTH Royal Institute of Technology
RISE Research Institutes of Sweden
jia.fu@ri.se

Sepideh Pashami
Halmstad University
RISE Research Institutes of Sweden
sepideh.pashami@ri.se

Fatemeh Rahimian
RISE Research Institutes of Sweden
fatemeh.rahimian@ri.se

Anders Holst
KTH Royal Institute of Technology
RISE Research Institutes of Sweden
anders.holst@ri.se

***Abstract*— We present an ongoing doctoral project aiming to improve the robustness of machine learning regarding adversarial attacks. First, we introduce the research background showing the significance of adversarial defense in reality. Then we specify the research questions which will be explored within this project. Finally, we discuss the methodology in the current implementation briefly.**

I. INTRODUCTION

Machine learning (ML) has been employed across a variety of domains and security-sensitive applications are no exception. Not only the opacity of many ML models raises concern, but they have also been proven to be vulnerable to adversarial attacks during both training and deployment. In particular, deep neural networks (DNNs) are not robust against specially designed perturbations of the input data. For example, autonomous cars can be misled by small stickers strategically placed on the stop signs and fail to recognize them [1]. Facial recognition systems can break down when people wear clothes with specially designed patterns [2]. It is also feasible for recommendation systems to push malicious content to the end-users, given fake user profiles [3]. The consequences of the collapse of these systems are unacceptable.

Adversarial ML refers to the technique where an attacker can exploit ML for malicious gains. In the most common form, an attacker tries to deceive an ML model with malicious inputs such that the model exhibits a behavior deviating from the expectation. Adversarial attacks intentionally supply crafted examples that are drawn from distributions different with the training data, where the IID (independent and identically distributed) assumption for most ML models is violated. Enhancing the adversarial robustness signifies filling the gap between the modeled decision boundary and the actual decision boundary.

Research so far has extensively focused on developing specific attacks and targeted defenses. However, there is a lack of robust representation capable of resisting multiple adversarial perturbations. In this project, we will investigate the common mathematical basis of various attacks to build a general defense framework. We seek to create an internal representation of input data that encodes task-related information and holds transformation invariance to subtle perturbations. We plan to benchmark the selected ML models for adversarial robustness improvement and evaluate them on different public datasets comprehensively.

*This work is within DataLEASH project under Digital Futures

II. RESEARCH QUESTION

To formulate, let $f(\boldsymbol{x};\boldsymbol{\theta})$ be the objective neural network with input $\boldsymbol{x}$, the optimizer will learn network parameters $\boldsymbol{\theta}$ to minimize the loss function $L(f(\boldsymbol{x};\boldsymbol{\theta}),y)$. The adversary's goal is to maximize L given fixed $\boldsymbol{\theta}$, and the generation of adversarial examples can be represented as

$$\arg\max_{\boldsymbol{\delta}} L\left(f\left(\boldsymbol{x}+\boldsymbol{\delta};\boldsymbol{\theta}\right),y\right) \quad s.t.\ \|\boldsymbol{\delta}\|_p \leq \epsilon \tag{1}$$

where $\boldsymbol{\delta}$ is the perturbation added to the input and ϵ is the bound on the l_p norm of $\boldsymbol{\delta}$. The typical strategy for adversarial defense is robust optimization, i.e. adversarial training expressed as

$$\arg\min_{\boldsymbol{\theta}} \mathbb{E}[\max_{\boldsymbol{\delta}} L\left(f\left(\boldsymbol{x}+\boldsymbol{\delta};\boldsymbol{\theta}\right),y\right)] \tag{2}$$

The gradient descent with respect to $\boldsymbol{\theta}$ is in essence a smoothing of the decision boundary. Another strategy is constructing a heuristic representation transformation g for input where the impact of $\boldsymbol{\delta}$ on results is masked during the forward propagation of information flow:

$$\min_{g} \mathbb{E}[\max_{\boldsymbol{\delta}} L\left(f\left(g(\boldsymbol{x}+\boldsymbol{\delta});\boldsymbol{\theta}\right),y\right)] \tag{3}$$

Our research focuses on both attack and defense processes in black-box settings. We assume the attacker cannot obtain the training data and the network parameters. To clarify, the project will focus on the following issues:

- For problem (1), how to capture the defect of neural networks concerning different adversarial attacks?
- For problem (2), what can be done during the model training to defend against adversarial attacks?
- For problem (3), how to encode the robustness for the structure of DNNs?
- Compare (2) and (3), which category of methods is more effective for evaluating several safety-concerned task domains?

III. RELATED WORK

Recent studies have shown that various learning systems are inescapably vulnerable to adversarial attacks. Szegedy et al. [4] first proved the existence of subtle perturbations to the images, where the perturbed images could fool neural networks into misclassification. Then the famous Fast Gradient Sign Method (FGSM) [5] was proposed to generate adversarial examples effectively by exerting perturbation in the gradient direction of the loss function. The universal adversarial perturbations [6] that are effective for all images and the one-pixel attack [7] that changes only one pixel in an image both pioneered the new research trends. Beyond convolutional neural networks (CNNs), attacks on other forms of networks are also well-developed. For example, Dai et al. [8] focused on the adversarial attacks that fool the graph neural networks (GNNs) by modifying the combinatorial structure of graphs. As for generative adversarial networks (GANs), attacks such as attackGAN [9] are also intractable to evade and defend.

Currently, the defenses against adversarial attacks are the primary strategy to boost the adversarial robustness of models. The most intuitive defense is to train neural networks with adversarial examples [4], namely adversarial training [10]. However, new adversarial examples can always be computed to deceive the classifiers [6]. Nie et al. [11] first forward diffused the adversarial examples with noise and then recovered the clean images using an inverse generative process. Instead of modifying the input, some methods that adjust the networks are of better generalization. Ross et al. [12] studied gradient regularisation that penalizes the degree of variation resulting in the output with respect to change in the input. Akhtar et al. [13] appended extra pre-input layers to the targeted networks and trained them to rectify a perturbed image. On the other hand, adversarial robustness is also investigated regarding various learning methods. For example, adversarial robust ML models have been proposed using transfer learning [14], few-shot learning [15], and continual learning [16] techniques.

IV. METHODOLOGY

The research project is currently in its early stage. We start in the context of multimodal ML, which involves data with more than one modality, such as image, text, and audio. Information from different sources correlated to the same task can be decomposed into a consistent part and a complementary part. In the multimodal fusion paradigm, when one of the fusion channels is under attack, the anomaly triggered by adversarial perturbations can be detected by differentiating the consistent component with that from other channels. We will try to reconstruct the attacked modality by modal translation or distill it by joint multimodal representation. The former means leveraging the generative models, and the latter means projecting all the unimodal signals to the same feature space. Both of them attempt to find the potential representation transformation g for problem (3) in a multimodal context. The research will be instantiated by the object recognition task using multi-sensor datasets KITTI [17] and/or Waymo [18]. In this way, we will achieve the goal of realizing multimodal adversarial defense. Further, quantifying the impact of attacks on the fusion taking place at different levels can help us determine inherently more robust representation transformation. In the future stages, we will land in a more general environment without limiting ourselves to the multimodal setting.

REFERENCES

[1] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

[2] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 665–681, Springer, 2020.

[3] K. Christakopoulou and A. Banerjee, "Adversarial attacks on an oblivious recommender," in *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 322–330, 2019.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[6] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

[7] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[8] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song, "Adversarial attack on graph structured data," in *International conference on machine learning*, pp. 1115–1124, PMLR, 2018.

[9] S. Zhao, J. Li, J. Wang, Z. Zhang, L. Zhu, and Y. Zhang, "attackgan: Adversarial attack against black-box ids using generative adversarial networks," *Procedia Computer Science*, vol. 187, pp. 128–133, 2021.

[10] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," *arXiv preprint arXiv:2102.01356*, 2021.

[11] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," *arXiv preprint arXiv:2205.07460*, 2022.

[12] A. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[13] N. Akhtar, J. Liu, and A. Mian, "Defense against universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3389–3398, 2018.

[14] A. Shafahi, P. Saadatpanah, C. Zhu, A. Ghiasi, C. Studer, D. Jacobs, and T. Goldstein, "Adversarially robust transfer learning," *arXiv preprint arXiv:1905.08232*, 2019.

[15] M. Goldblum, L. Fowl, and T. Goldstein, "Adversarially robust few-shot learning: A meta-learning approach," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17886–17895, 2020.

[16] H. Khan, N. C. Bouaynaya, and G. Rasool, "Adversarially robust continual learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2022.

[17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012.

[18] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.