Explainability-Guided Adversarial Examples: Generation, Evaluation, and Defense Mechanisms

Gregory Blanc, Majd Shalak

1 Context

Deep neural networks have achieved remarkable success across various domains, yet their vulnerability to adversarial examples poses significant security concerns for real-world deployments [1]. Adversarial examples are carefully crafted inputs, often imperceptible to human observers, that exploit the high-dimensional decision boundaries learned by DNNs, revealing fundamental limitations in their robustness and generalization capabilities.
Recent advances in explainable AI (XAI) have been used in security and introduced sophisticated methods for interpreting model decisions through attribution maps, saliency visualizations, and feature importance rankings [2]. Paradoxically, these transparency mechanisms, originally developed to enhance trust and accountability in AI systems, proved to be powerful tools for crafting effective adversarial attacks [3][4][5]. This dual-use nature of explainability methods presents both challenges and opportunities for the security of machine learning systems.
This research proposal aims to systematically investigate the intricate relationship between model explainability and adversarial robustness. We propose to develop a unified framework for understanding how explainability methods can be strategically leveraged for both offensive (attack generation) and defensive (robustness enhancement) purposes.
The primary objective is to characterize and exploit the bidirectional relationship between explainability and adversarial robustness to advance both attack and defense capabilities. Specific research goals include:

- Conducting a systematic analysis of existing explainability methods and their application to adversarial example generation
- Developing a comprehensive evaluation framework for comparing adversarial generation approaches
- Proposing defense mechanisms.

2 Activities

2.1 State-of-the-Art Analysis

The research will begin with literature review covering two main areas:

- **Explainability methods:** Comprehensive analysis of current explainability methods including white box methods (Integrated Gradients [9], DeepLift [10]), and black box methods (LIME [6], SHAP [7], LEMNA [8]).

- **Adversarial attack methodologies** Systematic study of classical adversarial generation methods (FGSM [11], BIM [12], DeepFool [14], C&W [13]) to recent explainability-driven approaches.

2.2 Comparative Analysis of Adversarial Generation Methods

Building upon existing adversarial robustness benchmarks, we will develop and implement a comprehensive evaluation framework to systematically compare explainability-driven attacks with traditional approaches. The framework will encompass:

- **Attack effectiveness**: Measuring evasion rates under different attack scenarios.
- **Adversarial examples quality**: Analyzing perturbation characteristics using Euclidean distance $\|x - x'\|_2$, Mahalanobis distance $\sqrt{(x - x')^T \Sigma^{-1} (x - x')}$ and semantic preservation measures
- **Novel metrics**: Eventually developing new evaluation criteria.

2.3 Defense Mechanism Development

The final phase focuses on designing, implementing, and rigorously evaluating defense strategies to counter explainability-guided attacks:

- **Explainability-aware adversarial training:** : Incorporating explainability-driven adversarial examples into the training process.
- **Explanation manipulation through fine-tuning**: Investigating controlled modification of model explanations to reduce attack surface while preserving model accuracy.
- **Uncertainty-based detection mechanisms**: Development of probabilistic frameworks for adversarial detection, this can include approaches like Bayesian deep learning approaches, Ensemble-based methods, Statistical hypothesis testing and conformal prediction.

3 Expected Outcomes

This research is expected to provide:

1. A comprehensive taxonomy of explainability-guided adversarial attacks
2. A standardized evaluation framework for assessing adversarial example quality
3. New defense mechanisms that explicitly consider the role of explainability in adversarial robustness

References

[1] M. M. Hassan, M. R. Hassan, S. Huda and V. H. C. de Albuquerque, ”A Robust Deep-Learning-Enabled Trust-Boundary Protection for Adversarial Industrial IoT Environment,” in *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9611-9621

[2] Charmet, F., Tanuwidjaja, H.C., Ayoubi, S. et al. Explainable artificial intelligence for cybersecurity: a literature survey. *Ann. Telecommun.* **77**, 789–812 (2022). `https://doi.org/10.1007/s12243-022-00926-7`

[3] Satoshi Okada, Houda Jmila, Kunio Akashi, Takuho Mitsunaga, Yuji Sekiya, et al.. XAI-driven adversarial attacks on network intrusion detectors. *European Interdisciplinary Cybersecurity Conference (EICC)*, Jun 2024, Xanthi, Greece. pp.65-73

[4] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016) The limitations of deep learning in adversarial settings. In: *IEEE European symposium on security and privacy (EuroS&P)*

[5] Okada, S., Jmila, H., Akashi, K. et al. Xai-driven black-box adversarial attacks on network intrusion detectors. *Int. J. Inf. Secur.* 24, 103 (2025). https://doi.org/10.1007/s10207-025-01016-0

[6] Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp 1135–1144

[7] Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: *Advances in neural information processing systems*, p 30

[8] Guo W et al (2018) Lemna: Explaining deep learning based secu- rity applications. In: *proceedings of the 2018 ACM SIGSAC con- ference on computer and communications security*, pp 364–379

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML* 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70), Doina Precup and Yee Whye Teh (Eds.). PMLR, 3319–3328

[10] Shrikumar, Avanti et al. “Learning Important Features Through Propagating Activation Differences.” *International Conference on Machine Learning* (2017).

[11] Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and har- nessing adversarial examples. In: *3rd International Conference on Learning Representations, ICLR*

[12] Kurakin A, Goodfellow I, Bengio S (2017) Adversarial examples in the physical world. In: *5th International Conference on Learn- ing Representations, ICLR*

[13] Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: *IEEE symposium on security and privacy. IEEE*

[14] Moosavi-Dezfooli SM, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: *IEEE conference on computer vision and pattern recognition*