

Towards Realistic Adversarial Examples against Network Intrusion Detection Systems

Gregory Blanc, Majd Shalak

September 2026

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 [SPW⁺22]. 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.

Yet, many mathematically valid perturbations yield no real traffic [CQA⁺25]. Indeed, legacy adversarial machine learning methods may optimize the similarity of adversarial examples with benign samples in the feature space, at the expense of the validity or the inherent relationships of the features [MCBCY22]. Recent approaches have thus focus on the constraints brought by the problem space, that is the real environment in which network traffic is generated and collected to produce more realistic and compliant adversarial examples [CPRV24, VDV⁺26].

2 Proposed work plan

- Review of the literature on problem-space adversarial examples.
- Study the effects of problem-space perturbations on the feature space.
- Implement a feature-space constrained perturbation method to evade ML-based network intrusion detection systems.
- Implement a problem-space perturbation method that is transferable to the feature space within the bounds revealed by the previous studies.

References

- [CPRV24] Marta Catillo, Antonio Pecchia, Antonio Repola, and Umberto Villano. Towards realistic problem-space adversarial attacks against machine learning in network intrusion detection. In *Proceedings of the 19th international conference on availability, reliability and security*, pages 1–8, 2024.
- [CQA⁺25] Jacopo Cortellazzi, Erwin Quiring, Daniel Arp, Feargus Pendlebury, Fabio Pierazzi, and Lorenzo Cavallaro. Intriguing properties of adversarial ml attacks in the problem space [extended version]. *ACM Transactions on Privacy and Security*, 28(4):1–37, 2025.
- [MCBCY22] Mohamed Amine Merzouk, Frédéric Cuppens, Nora Boulahia-Cuppens, and Reda Yaich. Investigating the practicality of adversarial evasion attacks on network intrusion detection. *Annals of Telecommunications*, 77(11):763–775, 2022.
- [SPW⁺22] Ryan Sheatsley, Nicolas Papernot, Michael J Weisman, Gunjan Verma, and Patrick McDaniel. Adversarial examples for network intrusion detection systems. *Journal of Computer Security*, 30(5):727–752, 2022.

- [VDV⁺26] Miel Verkerken, Laurens D’hooge, Bruno Volckaert, Filip De Turck, and Giovanni Apruzzese. “what is the problem space?” defining host-space adversarial perturbations against network intrusion detection systems. In *Proceedings of the ACM Asia Conference on Computer and Communications Security*, pages 1043–1059, 2026.