Towards problem-space adversarial traffic

Gregory Blanc, Houda Jmila, and Thomas Robert

Topics of interest — adversarial machine learning, problem space, network traffic

Context

Adversarial machine learning [2] is the field of study of attacks against ML systems, as well as their defenses. A typical approach to generate *adversarial examples* is to specifically perturb malicious samples to induce misclassification. This attack, known as *evasion* [1] is hard to defend against because the adversarial samples that are generated are usually close to legitimate samples. However, in a black-box setting, attackers need to generate many queries against the target model [6].

At the feature space, that is when dealing with the samples as they are represented in the machine learning model, many proposals were made to produce adversarial examples [2, 3, 5] against neural networks in domains as diverse as images, texts, sounds, programs or files. In order to counter them, the most effective measure seems to be *adversarial training* [4, 7]: simply put, it consists in retraining the model with a training dataset augmented with the evasive samples. Once retrained, models usually demonstrate increased robustness against adversarial examples.

However, perturbing malicious examples to make them evasive is not straightforward, if one wishes to avoid unnecessarily generating samples that would result in detection. This may due to insufficient perturbation of a sample. On the other hand, an over-perturbation may lead to an evasive yet invalid sample [8]. Invalid samples are samples that are inconsistent with the specifications or the semantics of its domain. In the network traffic domain, data are highly heterogeneous making any uncalibrated perturbation unlikely to be realistic. Such phenomenon, that we can call *feature-problem space duality* [9] is not specific to network traffic and has been witnessed in the other sample domains. Although some methods to adaptively perturb samples have been proposed [11], such transformations occur at the feature space and may not completely satisfy all requirements of the problem space [9], which include, besides realism (also known as *plausibility*), robustness to preprocessing, preservation of semantics, and restrictions to only available transformations. The latter would allow to describe a complete chain from a problem-space transformation of a (malicious) network traffic to its preprocessing and feature extraction, ultimately leading to a model evasion in the feature space. While in the problem space, the perturbed network traffic would eventually lead to a (successful) attack.

This project focuses on clarifying this chain for the specific problem of adversarial network traffic.

Activities

1. State of the art:
    - study existing perturbations for the network traffic domain,

- compare existing representations for network traffic.

2. Formalization:
   - formalize perturbations/transformations in the feature space,
   - adapt this model to the network traffic domain,
   - survey likely problem-space transformations and their impact on the feature space.
3. Experimentation:
   - design an adaptive approach that binds the problem and feature spaces,
   - evaluate its effectiveness on a well-known dataset [10],
   - survey or propose metrics to assess the quality of the adversarial examples.

References

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 387–402. Springer, 2013.

[2] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[3] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[4] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[5] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

[6] H. Jmila and M. I. Khedher. Adversarial machine learning for network intrusion detection: A comparative study. *Computer Networks*, 214:109073, 2022.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[8] M. A. Merzouk, F. Cuppens, N. Boulahia-Cuppens, and R. Yaich. Investigating the practicality of adversarial evasion attacks on network intrusion detection. *Annals of Telecommunications*, 77(11-12):763–775, 2022.

[9] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro. Intriguing properties of adversarial ml attacks in the problem space. In *2020 IEEE symposium on security and privacy (SP)*, pages 1332–1349. IEEE, 2020.

[10] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1:108–116, 2018.

[11] J. Vitorino, N. Oliveira, and I. Praça. Adaptative perturbation patterns: realistic adversarial learning for robust intrusion detection. *Future Internet*, 14(4):108, 2022.