Privacy side-channel attacks against ML systems

Gregory Blanc, Houda Jmila, and Nesrine Kaaniche

Topics of interest — adversarial machine learning, countermeasures, privacy

Context

Artificial Intelligence (AI), often based on Machine Learning (ML) systems, is becoming pervasive with popular applications ranging from drug synthesis and tumor prediction to financial market and climate forecasts to chatbots and deep fakes. Indeed, industries are trying to push AI forward by marketing its ability to automate tasks, reduce human errors and prejudices. Nevertheless, policymakers and society remain cautious about the inherent risks associated with ML systems, especially those that may inadvertently contain biases introduced by their designers. Beyond these ethical considerations, ML systems as any other digital systems are the target of various attackers that are trying to manipulate the outputs of such systems for different purposes: misinformation, extortion or fraud, to name a few.

Adversarial machine learning [3] is the field of study of attacks against ML systems, as well as their defenses. A first type of attacks targets the privacy of data manipulated by ML systems and is known as *membership inference* [8], which aim at inferring the presence of a specific sample in the training set. Such attack can be generalized to the extraction of the whole training set. Common defenses include *differential privacy* [6] and *data deduplication* [7]. Another class of attacks tackles the integrity of the model: poisoning [2] aims at injecting enough forged samples in the training set so that it triggers an undesired decision. *Backdoors* are usually crafted to induce a specific malicious behaviour. *Data deduplication* has also been demonstrated to be efficient against poisoning. *Evasion* [1] attacks specifically perturb malicious samples to induce misclassification. Defending against evasion attacks is harder 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. There is thus a chance that attackers get caught by *stateful detectors* generating too many similar requests [4].

However, defenses might unintentionally become vulnerabilities if they were not originally designed to align with the specific conditions of their deployment environment. Indded, some of the defenses outlined above may actually consider the ML system to protect in a *vacuum*, while in reality, they are integral components of larger systems. [5]. It has been observed that other components within the ML pipeline that interact with these defenses can potentially give rise to side channels. These side channels represent unintentional pathways for information leakage, that may leak private information with respect to the model, the training data, or even test queries.

This project focuses on recreating selected attacks and developing strategies to mitigate them.

Activities

1. State of the art:

- study a subset of adversarial attacks (e.g., evasion),
- survey defenses that thwart the selected attacks.

2. Experimentation:
   - design a privacy attack as outlined by Debenedetti et al. [5]
   - evaluate the performance of the proposed privacy attack.
   - propose a contermeasure: either to thwart the attack or to strengthen the target system.

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, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

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

[4] S. Chen, N. Carlini, and D. Wagner. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pages 30–39, 2020.

[5] E. Debenedetti, G. Severi, N. Carlini, C. A. Choquette-Choo, M. Jagielski, M. Nasr, E. Wallace, and F. Tramèr. Privacy side channels in machine learning systems, 2023.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[7] N. Kandpal, E. Wallace, and C. Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR, 2022.

[8] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.