RELATED WORK - NON-EVASION ATTACKS

Similar to evasion attacks, data poisoning and backdoor attacks aim to compromise model accuracy. However, they achieve it by tampering the training data to create deceptive model decision boundaries. In addition, backdoor attacks also require perturbing the test instance to result in a misclassification. This is achieved by introducing manipulated training data with triggers that can be activated during the testing phase.

Goldblum et al. [2] and Cinà et al. [1] review recent literature on attack methodologies and countermeasures for both poisoning and backdoor attacks. Both of these surveys found that existing research made overly-optimistic assumptions when designing / validating attack techniques, e.g., assuming the knowledge of a large portion of training data. They advocate for researchers to test proposed methods in more realistic situations to better assess the potential threats. Furthermore, they encourage exploration of the relationship between poisoning attacks and evasion attacks. This could lead to the creation of attacks that produce less noticeable poisoning examples, or defensive strategies that can safeguard models against both backdoor and evasion attacks.

In addition to undermining model accuracy, adversarial attacks also aim at breaching the privacy and confidentiality of training data. In particular, membership inference attacks [4] attempt to determine whether a specific data point was part of the training set used to train the model. Hu et al. [3] present a comprehensive survey of existing research efforts on membership inference attacks. They find that, similar to evasion attacks, the membership inference attack success rate decreases as the number of training samples increases. However, all these attacks are orthogonal to our survey, as we focus on adversarial evasion attacks.

REFERENCES

[1] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. 2023. Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *Comput. Surveys* (2023).

[2] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. 2022. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[3] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *Comput. Surveys* 54, 235 (2022), 1–37.

[4] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (SP)*. 3–18.