SAND2022-1493C

Position Paper: Counter-Adversarial Machine Learning is a Critical Concern

Jeremy D. Wendt
Sandia National Laboratories

Machine learning (ML) techniques are used widely to enable a wide variety of modern capabilities – including on-line search, self-driving cars, cybersecurity, and national security. As ML becomes more ubiquitous, and is used in more critical systems, ML security increases in importance. However, a rapidly growing body of research indicates that – while it is critical to enabling many current techniques – ML exposes new security concerns. While there is still little evidence of open attacks exploiting most of these security concerns,[1] we expect exploits (so-called "Adversarial ML") to become common soon. To ensure national security, nuclear missions, and maintain our many other capabilities, all-of-government efforts must unite with academic and industry partners to develop defenses and best practices – establishing defenses around ML systems so we can defend them quickly as attacks arise.

Machine Learning Attack Categories

As Adversarial ML is still relatively new, terminology, categorization, etc. are still in in flux.[2] Therefore, we present the following four attack categories as our best-effort categorization – recognizing that other researchers may use different language and categories.

Evasion: Evasion attacks seek to avoid proper classification of an input item by slightly altering the input item. In the literature, these are generally called "adversarial examples". The first widely referenced, very successful attack is the Fast Gradient Sign Method (FGSM) (Goodfellow, Schlens, & Szegedy, 2015). The FGSM requires the attacker can measure the targeted model's gradient at the location of a to-be-submitted input. Taking a down-hill step from this position is often sufficient to ensure the model misclassifies the input. Many follow-on techniques have been proposed that can target a specific false class (Yuan, He, Zhu, & Li, 2018), iterate over many smaller steps for less noticeable alterations (Kurakin, Goodfellow, & Bengio, 2016), and can generate misclassifications without gradient access (Papernot, et al., 2017).

Subversion: Subversion attacks seek to avoid proper classification of an input item by altering the model while it is being created. In the literature, many of these attacks use "data poisoning". A foundational analysis in this space placed small additional features on input images (e.g., a small, yellow rectangle on stop signs for driving sign classification) and gave them different labels (e.g., the modified stop signs are labeled as speed limit signs) (Gu, Dolan-Gavitt, & Garg, 2019). They showed that these altered input images resulted in high likelihood of misclassification on altered test data. Many further attacks have replaced some training data with different labels (Kegelmeyer, et al., 2015), moved values of specific input data (Biggio, Nelson, & Laskov, 2013), and other subversions showing that there are many possible ways to poison models to create backdoors.

Theft: Theft attacks seek to steal data that was not intended to be visible from a model. How these attacks succeed varies widely – from similar to those that require subversion-like poisoning, to only evasion-like query-level access. Thus, theft attacks are identified by their goal only (not the technique as well). Model inversions are a particularly frightening theft attack that enable an attacker to extract example images for each class from a model (Fredrikson, Jha, & Ristenpart, 2015). In brief, a model inversion attack starts with an uninformed input (e.g., gray image), and march up the gradient direction to find a specific maximal value for a class. These results can often be quite informative of what data was used to train the model. Other theft attacks furthered model inversions (Yang, Chang, & Liang, 2019) (Yin, et al., 2019), enabled querying for if a specific value was in the training data (Shokri, Stronati, Song, & Shmatikov, 2016), or encoded any data the attacker chose into the model (Song, Ristenpart, & Shmatikov, 2017).

---

[1] Although not widespread, there is continuing evidence of some on-going Adversarial ML attacks: https://kb.cert.org/vuls/id/425163, https://nvd.nist.gov/vuln/detail/CVE-2019-20634, https://interestingengineering.com/deepfaked-voice-of-ceo-used-to-steal-almost-250000-from-company.

[2] Notably, NIST has begun to unify terminology: https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf.

Misuse: Misuse attacks seek to leverage a benign ML technology to nefarious ends. In the literature, the most common of these attacks are "deep fakes". In a deep fake attack, a model can be trained that can transfer training data's features to a secondary data's space (Mirsky & Lee, 2020). While used for a variety of other input types, the most common examples transfer one person's facial expressions to another person – seamlessly replacing one actor's performance with another's. There are many examples of very convincing deep fakes being used in real-world attacks causing results that can be very expensive to the targets.[3]

Counter-Adversarial Machine Learning

These many novel attacks have led to a wide variety of new defenses being proposed. These defenses range from those that require models be retrained to be more robust,[4] to those that seek to identify attack inputs (Zhang, Chen, & Koushanfar, 2021). However, thus far most of these defenses have been quickly overcome by new attacks (Carlini & Wagner, 2017) (Tramer, Carlini, Brendel, & Madry, 2020). Some recent research also indicates that some of these attacks may be exploiting fundamental issues with modern ML (Ilyas, et al., 2019) (Javanmard & Soltanolkotabi, 2020).

However, it is often the case that in new competitive spaces, attacks outpace defenses at first. For instance, cybersecurity attacks and spam email attacks frequently overcame all proposed defenses during the early days of both technologies. However, in both cybersecurity and email there are now many techniques that prevent or rapidly detect a wide variety of attacks before they can succeed.

We compare counter-adversarial ML to cybersecurity intentionally. We believe defenses must be developed and improved immediately. Although defenses will be weaker than attacks for some time, to wait until attacks are commonplace before truly focusing on defenses only extends the truly dangerous period when attackers may succeed with impunity. Furthermore, much like cybersecurity, counter-adversarial ML is not an area where a silver bullet defense will arise and the adversarial ML will be "solved". Instead, it will be a space where defenders will develop a toolkit of techniques to protect the many systems that rely on ML technologies to succeed. Just as we cannot "turn back the clock" to remove ML from our systems, we cannot stall in our efforts to build robust defenses around them.

Bibliography

Biggio, B., Nelson, B., & Laskov, P. (2013). Poisoning attacks against support vector machines. arXiv:1206.6389v3.

Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. *ACM Workshop on Artificial Intelligence and Security.*

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *ACM SIGSAC Conference on Computer and Communications Security.*

Goodfellow, I., Schlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *ICLR.*

Gu, T., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. arXiv:1708.06733.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. arXiv:1905.02175v4.

Javanmard, A., & Soltanolkotabi, M. (2020). Precise statistical analysis of classification accuracies for adversarial training. arXiv:2010.11213.

Kegelmeyer, W. P., Shead, T. M., Crussell, J., Rodhouse, K., Robinson, D., Johnson, C., . . . Shelburg, J. (2015). Counter Adversarial Data Analytics. Sandia National Laboratories; SAND2015-3711.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. *arXiv:1607.02533.*

Mirsky, Y., & Lee, W. (2020). The creation and detection of deepfakes: a survey. arXiv:2004.11138.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical Black-Box Attacks against Machine Learning. *ACM on Asia Conference on Computer and Communications Security.*

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2016). Membership inference attacks against machine learning models. arXiv:1610.05820.

Song, C., Ristenpart, T., & Shmatikov, V. (2017). Machine learning models that remember too much. arXiv:1709.07886.

Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On adaptive attacks to adversarial example defenses. arXiv:2002.08347.

Yang, Z., Chang, E.-C., & Liang, Z. (2019). Adversarial neural network inversion via auxiliary knowledge alignment. arXiv:1902.08552.

Yin, H., Molchanov, P., Li, Z., Alvarez, J. M., Mallya, A., Hoiem, D., . . . Kautz, J. (2019). Dreaming to distill: Data-free knowledge transfer via deep inversion. arXiv:1912.08795.

Yuan, X., He, P., Zhu, Q., & Li, X. (2018). Adversarial Examples: Attacks and Defenses for Deep Learning. arXiv:1712.07107.

Zhang, X., Chen, H., & Koushanfar, F. (2021). TAD: Trigger approximation based black-box trojan detection for AI. arXiv:2102.01815v3.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

---

[3] https://interestingengineering.com/deepfaked-voice-of-ceo-used-to-steal-almost-250000-from-company

[4] See "adversarial retraining" in (Yuan, He, Zhu, & Li, 2018).