Adversarial Machine Learning: Attacks From Laboratories to the Real World

Hsiao-Ying Lin, Huawei International

Battista Biggio, University of Cagliari and Pluribus One

When machine learning techniques are applied in real-world applications, what potential threats do they face and how can we mitigate them?

Adversarial machine learning (AML) is a recent research field that investigates potential security issues related to the use of machine learning (ML) algorithms in modern artificial intelligence (AI)-based systems, along with defensive techniques to protect ML algorithms against such threats. The main threats against ML encompass a set of techniques that aim to mislead ML models through adversarial input perturbations. Unlike ML-enabled crimes, in which ML is used for malicious and offensive purposes, and ML-enabled security mechanisms, in which ML is used for securing existing systems, AML techniques exploit and specifically address the security vulnerabilities of ML algorithms.

Consider, for example, an automatic surveillance camera that uses certain ML algorithms. The system monitors people entering and leaving a building in real time. A person wearing a special T-shirt walks by the building, but the camera does not detect the person's presence, as the T-shirt has a special pattern that effectively conceals the person from the camera. Such a pattern can be constructed and optimized against the target system by leveraging attack algorithms developed in the AML research field.[1]

Digital Object Identifier 10.1109/MC.2021.3057686
Date of current version: 7 May 2021

EDITOR **HSIAO-YING LIN**
IEEE Member; hsiaoying.lin@gmail.com

One of AML's early real-world applications was spam filtering. Over time, ML-enabled antispam filters learn from the users' reactions, such as marking legitimate emails as spam or recovering misclassified spam emails as legitimate, to refine the filtering function. Under this scenario, attackers can exploit the learning process of the spam filter by manipulating the content of their spam emails, for instance, by injecting good words typically present in legitimate email but not in spam. This eventually causes the antispam filter to misclassify legitimate emails containing such words as potential spam. The filtering performance of the ML algorithm may thus decrease considerably, consequently causing the user to disable the antispam filtering service. The attack we just described is widely known as a *poisoning attack*, and it assumes that the attacker can manipulate the training data to subvert the learning process. Poisoning attacks are one of the most relevant AML attack techniques. A systematic study on AML was first performed in 2006.[2] Since deep learning techniques, which are a subset of ML, achieve excellent results in various intelligent tasks, AML attacks to mislead deep learning techniques have become a very active area of research. An informative summary of AML development over 10 years is available,[3] and the derived key insights include reactive and proactive security mechanisms. AML has been established as a new family of attacks against ML. The National Institute of Standards and Technology (NIST) was the first to systematically organize and describe the taxonomy family of AML.[4] The current NIST report presents major AML technologies developed in the academic community.

MAJOR ATTACK CATALOG OF AML

An ML's lifecycle can be described as two main phases: 1) the training phase, where the training data and ML model configurations are input to generate a trained model as the output, and 2) the operational phase, where the trained model is deployed into services and the deployed model is activated. In some special scenarios, such as online learning, in which the operation input and user feedback are continuously input as the training data to update the model, the operational phase is looped back to the training phase. The example given earlier about the antispam filter, which is continuously updated based on the user's feedback, can be regarded as a paradigmatic example of online learning. Based on the ML's lifecycle, five major categories of AML attacks are presented as follows (and illustrated in Figure 1):

- *Poisoning attacks*: As described in the example of the antispam filtering service, poisoning attacks manipulate the training data to degrade the performance of ML services. In particular, such attacks can aim to either degrade the overall performance of the system, causing a denial of service, or allow specific misclassifications during operation (for example, only targeting a specific user or set of samples). Poisoning attacks are conducted during the training phase, assuming that the attacker can inject poisoning data samples into the training set used to learn or update the deployed model. Data-driven ML-based systems, which strongly rely on the quality and representativeness of the training data sets, can indeed be very sensitive to poisoning attacks. In the example of the antispam filtering service, when the performance degrades beyond a certain level, the service becomes useless or even harmful. This implication is applicable for various

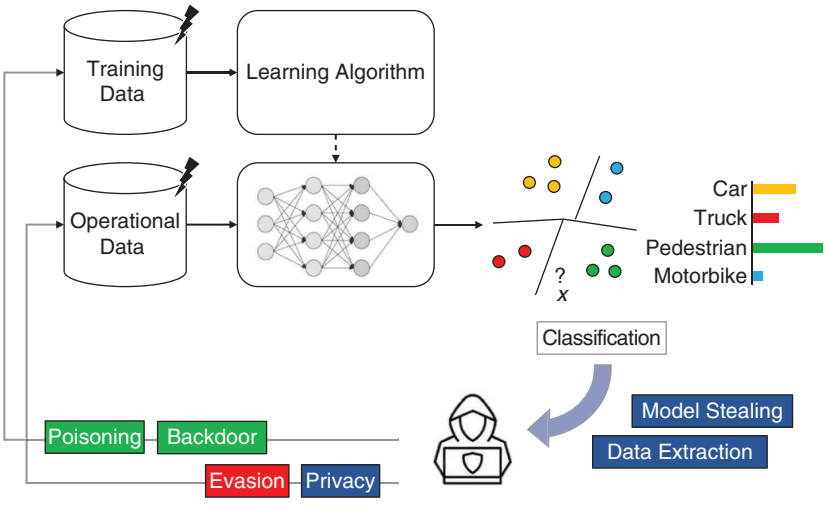

FIGURE 1. Major AML attacks. Evasion and privacy attacks are staged during operation and include manipulation of operational data to either evade detection or obtain confidential information about the ML model or its users (for example, via model stealing and data extraction attacks); feedback from the ML model is typically required to refine the attack samples iteratively. Poisoning and backdoor attacks additionally require the attacker to manipulate the training data and/or the ML model under design.

applications such as malware detection and network-based intrusion detection.

- *Backdoor attacks*: These attacks are accomplished in two steps. First, special patterns are embedded in the targeted model during the training phase, which is typically achieved by poisoning training data. Second, the attack is activated during the operational phase by feeding an input with a trigger into the targeted model. The model then provides a maliciously predefined output. For instance, a backdoor ML-based road sign classifier can misclassify a stop sign as a speed-limit sign. Here, the stop sign is patched with a special sticker that functions as the trigger. Because open-sourced training data and pretrained models are popular and widely used, they are prone to manipulations and exposed to the threat of backdoor attacks.
- *Evasion attacks*: Attackers carefully craft perturbed input, the so-called adversarial examples, to mislead the targeted ML model into outputting an incorrect prediction. A typical example of image-based evasion attacks in cyberspace is that a dog image with adversarial-crafted noises may be identified as a cat image. Special T-shirts or eyeglass frames that evade ML-based security and biometric authentication is another example of image-based evasion in the real world. The evasion attacks indicate that, although ML models are efficient, they have limitations. Furthermore, adversarial examples exhibit transferability. An adversarial example generated against an ML model is effective against other models when these models operate on the same or similar tasks.
- *Model stealing attacks*: These attacks are conducted in the operational phase. By querying the targeted model, the attackers can generate an approximation of the original model, whereas the attackers may be able to obtain model parameters by exploiting system vulnerabilities. Both of these attack approaches allow attackers to conduct strong evasion attacks on the targeted model. Model stealing attacks also cause concerns of intellectual property theft.
- *Data extraction attacks*: In these attacks, the attackers attempt to invert training data out of the targeted model or at least distinguish whether a given datum belongs to the training data or not during operation. When the targeted training data, such as bioauthentication information and medical records, are sensitive private data, extraction attacks cause serious data privacy violations. For instance, an approximated facial image can be reconstructed from a name and query access to the facial recognition system.

Various ML services and applications are vulnerable to different threats. For instance, a cloud-based ML service using a large-sized model may face model stealing attacks in which attackers can steal the model's capability. An end-device ML application using a small model may face model stealing attacks in which attackers can simply extract model parameters from the device by exploiting system vulnerabilities.

TOWARD REAL-WORLD AML ATTACKS

The aforementioned attacks are first developed in laboratories and then gradually adopted in various (nearly) real-world applications in different business domains. A gap exists between the technical results from laboratories and the real-world attacks. However, some attacks are proving exceedingly effective. These attacks are propagated by using real-world training data or by mapping AML attack techniques in the physical world. Here, we introduce a collection of those (nearly) real-world attacks.

- *Real-world text-based poisoning incident*: Tay was designed as an ML-based chatterbot for the 18- to 24-year-old demographic and deployed on Twitter in 2016.[5] Tay rapidly learned from online conversations but elicited unintended effects. Tay started delivering offensive and hurtful tweets after being poisoned by adversarial interactions with other malicious twitters. Tay was shut down only 16 h after its launch.
- *Real-world audio-based evasion attacks*: Adversarial examples are developed for automated speech recognition systems in the physical world. For instance, in the attack on Mozilla DeepSpeech speech-to-text automated speech recognition,[6] the addition of nearly inaudible noises resulted in the system recognizing the waveform of any sentence as the targeted sentence. This attack requires full knowledge of the targeted model. An advanced audio-based evasion attack, called Devil's Whisper, was subsequently developed to target commercial speech recognition devices; this attack approach required zero knowledge of the model parameters.[7] Four speech API services, including Google Cloud Speech-to-Text, Microsoft Bing Speech Service, IBM Speech-to-Text, and Amazon Transcribe, were targeted. These examples indicated that attacks can be launched on connected intelligent devices such as Google Assistant, Google Home, Microsoft Cortana, and Amazon Echo. Adversarial examples containing inaudible command

audio clips are indistinguishable from clean audio clips. Therefore, attackers can activate some services through inaudible commands without the real user being aware.

- *Real-world image-based evasion attacks*: Adversarial examples targeting image classifications and object detectors were developed in the physical world. To evade image classifiers, objects in the real world are patched with 2D printed physical perturbations.[8] The main determinants of the effectiveness of the attack were various environmental conditions such as varying distances and angles. Experimental results of physical perturbations on road sign classifiers in the field have revealed high attack success rates within certain ranges of distances and angles. Object detection tasks detect and classify multiple objects. For instance, a front-facing, vehicle-mounted camera detects multiple road signs and traffic lights and classifies them. Physical perturbations evading YOLO v2 object detectors have been developed to render stop signs invisible.[9] Experimental results revealed that these attacks can be effectively launched indoors and outdoors in a laboratory environment. Despite physical attacks on ML models that have been demonstrated effectively, current analyses have been limited to reporting few paradigmatic examples, while a large-scale analysis on the effectiveness and concrete impact of such attacks on ML models is still lacking.
- *Real-world lidar-based evasion attacks*: Three-dimensional adversarial examples have been developed in the physical world. A 3D adversarial example is first carefully crafted and 3D printed as a physical object.[10] This 3D physical object evades the targeted vehicle-mounted lidar detector system such that it is invisible to the system. Subsequently, this attack evolves to be more powerful. The 3D-printed physical object can extend invisibility to its immediate neighboring objects.[11] By placing a 3D-printed adversarial example object on top of a vehicle, the vehicle becomes (partially) invisible to the targeted lidar detector system.
- *Real-world model stealing attacks*: The imitation of real-world machine translation production systems from Google, Bing, and Systran constitutes real-world model stealing attacks.[12] Approximations of the original models are developed using a collection of query-response data from machine translation services. The ultimate purpose of the attacks is to evade machine translations. Adversarial examples are generated from the imitation models and then applied online on the targeted models. Experiments conducted on English-to-German machine translations revealed that adversarial examples are effective in real-world systems. An effective adversarial example is the translation of "Save me, it's over 102 °F" by Google into "Rette mich, es ist über 22 °C." This effectively changed the temperature from 102 °F to 72 °F.

INITIAL COUNTERMEASURES AGAINST AML THREATS

AML attacks have emerged as novel threats to safety, security, and privacy. The importance of securing ML systems against adversarial attacks has gained considerable attention not only in the academic research community but also in industry and standardization organizations. We summarized three main types of initial countermeasures against AML threats.

- *Threat analysis*: An initial threat analysis provides an overview of the threats encountered by ML-based services and facilitates the identification of interfaces for system developers and service providers. Microsoft and MITRE derive and maintain a framework of AML threat matrix as a reference tool[13] of known attack techniques against ML systems to assist security analysts. Tencent also publishes an AI threat matrix report (currently only available in Chinese), in which known attacks are presented and initial defense suggestions are provided.[14] Based on a customized threat analysis, suitable security controls can be decided and applied to mitigate potential AML threats.
- *Mitigations*: ETSI Industry Standard Group of Securing Artificial Intelligence (SAI) has developed a work item, ETSI-SAI-005-GR, which is a technical report of mitigation strategy.[15] This report describes mitigation approaches, such as data sanitization, adversarial example detection, and model hardening, against the introduced five attacks. It also collates and summarizes existing defense techniques against AML threats. Those mitigations can build strategies to prevent,

The importance of securing ML systems against adversarial attacks has gained considerable attention not only in the academic research community but also in industry and standardization organizations.

detect, or respond against AML threats.

- *Security by design*: This approach is recommended as a proactive security mechanism. Thus, proactive security is achieved by embedding security design and implementation into the ML development and operation lifecycle. The security, development, and operations (SecDevOps) from the software development lifecycle is adopted. In the context of ML development and operations, the process also includes continuous integration, delivery, and training. Limited studies have investigated this topic. Embedding security requirements in system design, implementing security controls, and verifying whether security requirements are satisfied are three major steps in this approach. Embedding security requirements in the system design may be regulated by legislation. Implementing security controls and verifying whether security requirements are satisfied require strong technical support of security hardening and security testing techniques.

As ML systems and services have become a part of daily life, considerable progress has been achieved in the development of advanced tools to ease the securing of ML against AML threats. It is still an arms race, and hence there is a long way to go.

We have discussed AML attack techniques, their implications in real-world applications, and initial countermeasures. Various communities of different business domains should perform further research in AML and implement suitable mitigations for ML-based systems and services. More than making the ML-based systems just secure, an endeavor should be made toward making them trustworthy. More challenging topics such as explainability, fairness, and accountability should be addressed. We hope to discuss them further in the future.

REFERENCES

1. S. Thys, W. Van Ranst, and T. Goedeme, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR) Workshops*, 2019, pp. 49–55.
2. M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tyger, "Can machine learning be secure?" in *Proc. ACM Symp. Inf., Comput. Commun. Security*, 2006, pp. 16–25. doi: 10.1145/1128817.1128824.
3. B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, Dec. 2018. doi: 10.1016/j.patcog.2018.07.023.
4. E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, A taxonomy and terminology of adversarial machine learning, National Inst. of Standards and Technol., Gaithersburg, MD, Draft NISTIR 8269, 2019. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf
5. O. Schwartz. "Microsoft's Racist chatbot revealed the dangers of online conversation." IEEE Spectrum. https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation (accessed Mar. 26, 2021).
6. N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Security and Privacy Workshops (SPW)*, 2018, pp. 1–7.
7. Y. Chen et al., "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *Proc. USENIX Security Symp.*, 2020, pp. 2667–2684.
8. K. Eykholt et al., "Robust physical-world attacks on deep learning models," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 1625–1634.
9. K. Eykholt et al., "Physical adversarial examples for object detectors," in *Proc. USENIX Workshop on Offensive Technol. (WOOT)*, 2018, p. 1.
10. Y. Cao et al., "Adversarial objects against LiDAR-based autonomous driving systems," 2019, arXiv:1907.05418v1.
11. J. Tu et al., "Physically realizable adversarial examples for LiDAR object detection," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 13,716–13,725.
12. E. Wallace, M. Stern, and D. Song, "Imitation attacks and defenses for black-box machine translation systems," in *Proc. Empirical Methods Natural Language Process.*, 2020, pp. 5531–5546.
13. "Microsoft and mitre, Adversarial ML threat matrix." GitHub. https://github.com/mitre/advmlthreatmatrix (accessed Mar. 26, 2021).
14. Tencent AI Lab and Tencent Secure Platform Department. "AI安全的威胁风险矩阵 (translated to 'Threat and Risk Matrix of AI Security')". https://share.weiyun.com/8InYhaYZ (accessed Dec. 26, 2020).
15. "Mitigation strategy report," ETSI ISG SAI, Sophia Antipolis, ETSI-GR-SAI-005, 2021. [Online]. Available: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf

HSIAO-YING LIN is a senior researcher at Huawei International, 138588, Singapore, and a Member of IEEE. Contact her at lin.hsiao.ying@huawei.com.

BATTISTA BIGGIO is an assistant professor at the University of Cagliari, Cagliari, 09123, Italy, and cofounder of Pluribus One. He is a Senior Member of IEEE. Contact him at battista.biggio@unica.it.