A

Adversarial Machine Learning

Nicolas Papernot
University of Toronto and Vector Institute, Toronto, ON, Canada

Synonyms

Private machine learning; Robust machine learning; Trustworthy machine learning

Definition

Adversarial machine learning is a subfield of computer security interested with the study of machine learning systems in the presence of adversaries. A systematic characterization of worst-case behavior enables the design of machine learning algorithms with confidentiality, integrity, and availability guarantees that contribute to increasing the trust that end users can place in systems that deploy machine learning components.

Background

In 2012, a breakthrough (Krizhevsky et al. 2012) in machine learning (ML) slashed the error rate for the ImageNet computer vision benchmark and its associated object recognition competition (Russakovsky et al. 2015). The winning entry by Krizhevsky, Sutskever, and Hinton from the University of Toronto employed a deep neural network, a class of ML models that learns a hierarchical set of data representations by composing individual computing units – the neurons – organized in layers. The following years saw a surge of interest in the field of deep learning and eventually also revived interest in the study of the vulnerabilities of ML systems (Huang et al. 2011), an area often referred to as adversarial machine learning (AML).

Theory and Application

In a seminal paper, Szegedy et al. introduced the concept of adversarial examples (Szegedy et al. 2013) and demonstrated that while deep neural networks had enabled machines to achieve performance comparable to humans in certain computer vision tasks, they still remained vulnerable to subtle perturbations of their inputs. While adversarial examples were originally introduced as an "intriguing property" in the ML community, the security community quickly realized the implications of these findings to the robustness of ML systems deployed in adversarial settings. Biggio et al. demonstrated in a work concurrent to Szegedy et al. how support vector machines and shallow neural networks can be manipulated by an adversary (Biggio et al. 2013). In 2015 and 2016, Papernot et al. introduced algorithms enabling adversaries to craft adversarial examples

S. Jajodia et al. (eds.), *Encyclopedia of Cryptography, Security and Privacy*,
https://doi.org/10.1007/978-3-642-27739-9_1635-1

capable of attacking deep neural networks with (Papernot et al. 2016b) and without (Papernot et al. 2017) knowledge of the model's internal parameters. The logical conclusion of this line of research was that attacks against ML systems were now practical, and adversarial examples are just one of the many threats faced by ML.

Why are ML systems vulnerable to adversarial inputs? The answer lies in the underlying assumption made by most modern approaches to ML: the training and test distributions are assumed to be identical. That is, the data used to create a model should come from the same distribution than the data used to test and deploy it. A learning algorithm is then tasked with finding a model that will generalize from the training data to the test data. When an adversary comes in the picture, they typically break this assumption in one of two ways. They can manipulate the training distribution, in what is broadly referred to as a poisoning attack (Rubinstein et al. 2009). Training points inserted or modified by the adversary generally induce the learning algorithm to extract patterns that are not relevant to solving the task which the model was originally designed to solve. When the model is then deployed, these spurious patterns manifest themselves in the form of incorrect model predictions on any test inputs or, for certain attacks (Gu et al. 2017), only the test inputs which have been marked with a specific trigger in their features – à la backdoor. Adversaries can also manipulate the test distribution, creating a drift with the distribution the model developer had intended to model, to mount an evasion attack. Adversarial examples are an example of such a threat.

Both poisoning and evasion attacks target the integrity of ML. However, ML systems are not unlike other computer systems, and the traditional computer security triad (Anderson 2008) of confidentiality, integrity, and availability (CIA) also applies to them. In ML, confidentiality takes two flavors depending on whether it applies to the data or the model itself. User concerns around the centralization of data to enable learning has prompted research on distributed counterparts (Konečný et al. 2016) to centralized learning algorithms. In a similar vein, confidential inference makes it possible for a model owner to make predictions on inputs without the user having to reveal the input to the model owner. Approaches for confidential ML often rely on secure enclaves, secure multiparty computation, or homomorphic encryption (Ohrimenko et al. 2016; Gilad-Bachrach et al. 2016). A different angle on confidentiality considers the model itself, and the IP that it constitutes: adversaries may exploit querying access to an initially unknown model to recover details of its internals through what is known as a model stealing or extraction attack (Tramèr et al. 2016). The question of availability arises in critical ML systems, where model predictions are relied upon to take decisions that have implications to the security and safety of control systems and production environments. This is for instance the case for resource allocation in datacenter management or autonomous systems. Recent work exposed how hardware speculation introduced in ML accelerators, such as GPUs and FPGAs, expands the attack surface of ML systems. Adversaries can exploit optimizations and speculation performed by hardware to craft sponge examples – inputs that increase the latency and energy consumption of hardware thus jeopardizing the availability of models as they make critical predictions (Shumailov et al. 2020).

Given the widespread application of ML, and the resources it puts at risk, it is natural to ask how ML can be secured. Limited progress has been achieved at the time of writing, despite significant interest in AML. One of the major obstacles is the lack of established security model for ML systems. This is perhaps best illustrated by our limited understanding of what robustness to adversarial examples entails to. Existing definitions continue to consider what was introduced as a toy problem to bootstrap research in seminal work by Szegedy et al. (2013) and Goodfellow et al. (2014). While this has led to useful progress in the realm of robust optimization (Wong and Kolter 2018), the resulting techniques introduce an unnecessary trade-off with the model's ability to generalize (Tramèr et al. 2020). There is nascent recognition that

considering the entirety of the system deploying ML rather than the ML component in isolation will support progress towards an effective formulation of what robustness means in the context of ML.

Open Problems and Future Directions

Does this mean that ML applications will give rise to a never-ending arms race, à la Lampson (2004), between attackers and defenders? Not necessarily. The intricate relationship between ML and cryptography gives reasons to hope for a systematic approach to secure ML, which does not rely on secrecy and adheres to Kerckhoffs' Principle (Kerckhoffs 1883). Like cryptographic protocols and systems, many components of ML systems are amenable to formal specification. This is best illustrated by advances in privacy-preserving ML. Differential privacy has established itself as the gold standard for defining privacy (Dwork et al. 2006). With roots in cryptography, the definition involves a game between an algorithm and adversary: the adversary observes the algorithm's outputs to extract private information contained in the algorithm's inputs whereas the algorithm leverages randomization to limit the adversary. Through an analytical analysis of the sensitivity of model updates to training data, differential privacy has been successfully applied to reason about and strengthen the privacy of ML algorithms (Abadi et al. 2016; Papernot et al. 2016a). This demonstrates that a principled approach to secure ML is possible when the adversary is modeled precisely in a way that aligns human norms with the generalization goal of ML. Beyond the security and privacy of ML, work in the same vein will be needed to ensure that ML is deployed with a strong understanding of its consequences on ethics, fairness, and the law (Kumar et al. 2020) – to cite a few only.

Cross-References

- ▸ Data Mining (Privacy in)
- ▸ Differential Privacy

References

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016, October) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 308–318

Anderson R (2008) Security engineering. Wiley

Biggio B, Corona I, Maiorca D, Nelson B, Šrndić N, Laskov P, … Roli F (2013, September) Evasion attacks against machine learning at test time. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin/Heidelberg, pp 387–402

Dwork C, McSherry F, Nissim K, Smith A (2006, March) Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. Springer, Berlin/Heidelberg, pp 265–284

Gilad-Bachrach R, Dowlin N, Laine K, Lauter K, Naehrig M, Wernsing J (2016, June) Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. In: International conference on machine learning, pp 201–210

Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572

Gu T, Dolan-Gavitt B, Garg S (2017) Badnets: identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733

Huang L, Joseph AD, Nelson B, Rubinstein BI, Tygar JD (2011, October) Adversarial machine learning. In: Proceedings of the 4th ACM workshop on security and artificial intelligence, pp 43–58

Kerckhoffs A (1883) La cryptographic militaire. J Sci Milit IX:5–38

Konečný J, McMahan HB, Ramage D, Richtárik P (2016) Federated optimization: distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

Kumar RSS, Penney J, Schneier B, Albert K (2020) Legal risks of adversarial machine learning research. arXiv preprint arXiv:2006.16179

Lampson BW (2004) Computer security in the real world. Computer 37(6):37–46

Ohrimenko O, Schuster F, Fournet C, Mehta A, Nowozin S, Vaswani K, Costa M (2016) Oblivious multi-party machine learning on trusted processors. In: 25th {USENIX} security symposium ({USENIX} security 16), pp 619–636

Papernot N, Abadi M, Erlingsson U, Goodfellow I, Talwar K (2016a) Semi-supervised knowledge transfer for deep learning from private training data. In: International conference on representation learning

Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016b, March) The limitations of deep

learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, pp 372–387

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A (2017, April) Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp 506–519

Rubinstein BI, Nelson B, Huang L, Joseph AD, Lau SH, Rao S, ... Tygar JD (2009, November) Antidote: understanding and defending against poisoning of anomaly detectors. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement, pp 1–14

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, ... Berg AC (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3): 211–252

Shumailov I, Zhao Y, Bates D, Papernot N, Mullins R, Anderson R (2020) Sponge examples: energy-latency attacks on neural networks. In: Proceedings of the 6th IEEE European symposium on security and privacy, Vienna

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199

Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T (2016) Stealing machine learning models via prediction apis. In: 25th {USENIX} security symposium ({USENIX} security 16), pp 601–618

Tramèr F, Behrmann J, Carlini N, Papernot N, Jacobsen JH (2020) Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In: International conference on machine learning

Wong E, Kolter Z (2018, July) Provable defenses against adversarial examples via the convex outer adversarial polytope. In: International conference on machine learning, pp 5286–5295