Paper Review

SoK: Security and Privacy in Machine Learning

Chandrika Mukherjee
email: cmukherj@purdue.edu
Date: 1st November 2021

Summary

This paper presents theoretical understanding of the sensitivity of modern Machine Learning algorithms to the data they analyze. The authors defined a threat model that considers characteristics of the attack surface, adversarial goals, and possible defense and attack capabilities particular to systems built on machine learning. The paper highlights results from crucial discoveries in the security of ML domain in a structured way. The paper instructs to view systems built on ML through CIA (Confidentiality, Integrity, Availability) model. The paper considered the life-cycle of a ML based system from training to inference, and identified the adversarial goals and means at each phase. The detailed comments are written in following sections.

Detailed Comments

- This paper correctly identifies the requirement of understanding the threats, attacks, defenses on ML based systems as use of ML continues to grow. The authors present both - a motivation and challenge to systematize knowledge about the myriad of security and privacy issues that involve ML.
- The paper discusses and defines the standard for "Threat Model" which is highly beneficial to future research in this domain. The attack surfaces are also discussed. As mentioned, during training or inference time, adversaries can attempt to manipulate the collection of data, corrupt the model, or tamper with the outputs.
- For a system, different levels of trust are assigned to different possible actors. The sum of those trust assumptions forms the trust model and therein identifies the potential ways that bad actors may attack the system.
- The paper discusses the capabilities of stronger and weaker adversaries. An adversary is strong or weak is determined by the amount of information, they have about the system.
  - Inference attack is divided into white box and black box attack. White box attacks use information about the model and its parameters whereas black box attacks do not have knowledge about the model, uses past inputs, setting to find vulnerability within system.
  - In training phase, the adversary can learn about the model, which then can be used to build an auxiliary model. The adversary can use the substitute model to test potential inputs before submitting them to the victim.
- The paper discusses the adversarial goals in terms of CIA (Confidentiality, Integrity and Availability) and Privacy. If the model owner is untrusted by the users or the users are untrusted by the owner, the attack will fall in Confidentiality and Privacy domain. Adversaries perform membership inference attack to know about individual if they are within the system.

  Attacks attempting to control model outputs falls in Integrity attack. Availability prevents access to an asset. Hence, the goal of these attacks is to make the model inconsistent or unreliable in the target environment.
- According to the paper, adversaries find the most harmful labels to perturb in the data. Poisoning attacks alter the training dataset by inserting, editing, or removing points with the

intent of modifying the decision boundaries of the targeted model. Direct Poisoning of Inputs can be performed by adversary by adding new data points by observing the environment in which the system evolves. Adversaries with no access to the pre-processed data poison the model's training data before its pre-processing.

- The paper discusses white box adversaries having varying degrees of access to the model h as well as its parameters θ. The input x, correctly classified by h, is perturbed with r to produce an adversarial example $x\star$ that remains in the input domain D but is assigned the target label l. When the target l is chosen, the attack is a source-target misclassification. If l is any labels from h(x), the attacks is untargeted. When the adversary cannot directly modify feature values used as model inputs, it finds perturbations preserved by the data pipeline that precedes the classifier in the overall targeted system. Adversarial examples are also applicable beyond classification to Reinforcement Learning.
- The paper provides examples of black box attacks - how attackers can force a remotely hosted ML model to misclassify inputs without access to its architecture, parameters, or training data. As the model and the parameters are unknown for a black box attack, one adversarial goal is to gain information about the model. Authors provided examples of membership inference attack, training data extraction by inverting the model, also model extractions by observation of its predictions.
- After presenting possible attacks in training and inference time, the authors also discussed learning of robust, private and accountable models.
  - To mitigate integrity attacks, the ML model needs to be robust to distribution drifts where the training and test distributions differ. The authors provided examples of PCA algorithm to search for a direction whose projections maximize a univariate dispersion measure based on robust projection pursuit estimators instead of the standard deviation, also adding regularization to reduce sensitivity towards out-of-diagonal kernel matrix elements.
  - Although defending against inference attacks is an open problem, the authors mentioned some processes in order to defend against inference attack.
    - **Gradient Masking:** A natural defense strategy is to reduce the sensitivity of models to small changes made to their inputs. This sensitivity is estimated by computing first order derivatives of model h with respect to inputs. The gradients are minimized during learning phase. Apart from this, authors mentioned deep contractive networks, application of distillation to increase the robustness of DNNs to adversarial samples.
    - Injecting adversarial examples in training time can make the model robust to adversarial attacks.
  - privacy of a model can be kept by adding noise during training time providing local privacy. Also, introducing noise to predictions make the ML model randomized.
  - fairness could be achieved by learning in competition with an adversary trying to predict the sensitive variable from the fair model's prediction.
  - In one hand, techniques used for accountability and transparency are likely to yield improved attack techniques because they increase the adversary's understanding of how the model's decisions are made. On the other hand, they also contribute to building a better understanding of the impact of training data on the model learned by ML algorithm, which is beneficial to privacy-preserving ML.

Recommendations

This paper put great effort in systematizing findings on ML security and privacy, focusing on attacks identified on these systems and defenses crafted to date. The paper discussed comprehensive threat model which involves possible attack surface, adversarial capabilities, adversarial goals. The paper discussed important findings in Training and Inferring in adversarial settings. The authors provided ample examples of white and black box attacks. Later, the paper also provided crucial findings in learning robust, fair models against adversarial attacks.

Therefore, in my view this paper made a successful attempt in gathering bits and pieces of important information in building secure ML models. This helps future researchers in this domain to build a better understanding. But some mathematical proofs could have been added to assign more strength to the theory. On the other hand, absence of these will influence new researcher in gaining understanding or proving the facts presented in the paper.