Report: Taxonomy of Attacks, Defenses, and Consequences in Adversarial Machine Learning

Don Athalage

December 17, 2022

Contents

1 Introduction

Adversarial manipulation of machine learning models is a growing threat to the security of machine learning models. This is due to the fact that machine learning models are becoming more and more complex and are being used in more and more critical applications. This means that the consequences of an attack on a machine learning model can be more severe than an attack on a traditional software application. This is because machine learning models are often used to make decisions that have a direct impact on the real world, such as whether or not a person is approved for a loan, etc. This report will discuss the different types of attacks that can be carried out on machine learning models, as well as the different types of defenses that can be used to mitigate these attacks. This report will also discuss the consequences of these attacks.

2 Data Access Attacks

Attack In data access attacks, initial dataset used to create the legitimate model can be used to create a substitute model. This substitute model can be used to test different attack types under testing. These attacks are usually convayed by malicious actors that sell data to other malicious actors. Data leaked from cyber attacks on telecommunication companies can be exploited for these attacks.

Defense Data access attacks can be defented from different authentication and authorisation mechanisms. Furthermore, standard at rest encryption such as AES256 and at transit encryption such as TLS can be used to protect data from being accessed by unauthorised users.

3 Indirect Poisoning Attacks

Attack Indirect poisoning refers to type of attack that the malicious actors gain access to raw data and change the nature or data to give incorrect inference results. This will be done before or after data collection and before preprocessing. This could have serious consequences for the companies, such as financial losses or damage to its reputation. Typical ways that malicious actors carry indirect poisoning is via gaining data access.

Defense To protect against indirect poisoning attacks, data must be validated before preprocessing. These attacks are harder to defend against as the malicious actors can change the data before they are collected. It is important that the dats used are from trusted and verified sources.

4 Direct Poisoning Attacks

Attack Direct poisoning attacks are carried out by gaining access to the preprocessed data and altering models, data or results. There are few ways direct poisoning can be carried out; data injection, data manipulation and logic manipulation. Data injections and data manipulation is done by changing the training or testing data. Logic manipulation is done by getting access to computing systems or code repository and changing the code.

Defense Logic manipulation can be mitigated via builing layered software architectures, using proper at rest and in transit encryption. Data injection and manipulation can be mitigated by using proper data validation and sanitization. Extra care must be taken when using public data or input sources to prevent data injection and data manipulation.

5 Evasion Attacks

Attack Evasion attacks focus on tampering with the input data to model to produce incorrect inference results. Algorithms for evasion attacks require knowledge about the model or substitute models. Evasion attacks can be difficult to detect and defend against, as they often involve subtle changes to inputs that are not immediately noticeable.

Defense To protect against evasion attacks, it is important to implement robust security controls and to regularly test and update them to ensure that they are effective at detecting and blocking these types of attacks. This may involve implementing input validation and filtering, as well as using security tools such as firewalls and intrusion detection and prevention systems.

6 Oracle Attacks

Attack Oracle attacks are carried out using an API to input data and observe the output of the model. This can be done by using a substitute model to test the model. This attack is similar to evasion attacks, but the goal is to find the input that produces the desired output. This attack is usually carried out by malicious actors that have access to the model. This does not require direct knowledge about the model. Attacker will associate input with outputs to identify input combination that model will create incorrect output.

Defense Similar to evasion attacks, oracle attacks are also difficult to mitigate. However using robust models trained with noise can help mitigate this attack. Masking the inputs and gradient can also help mitigate this attack.

7 Conclusion

Proper identification and mitigation of these attacks is important to ensure the security of machine learning models. Different methodologies and best practices from software engineering and machine learning can be associated to ensure a safer use of AI in a variety of applications. This report has discussed the different types of attacks that can be carried out on machine learning models, as well as the different types of defenses that can be used to mitigate these attacks.