Adversarial Machine LLearning

Bhavya Shah, Parvez Faruki

Government MCA College, Ahmedabad

Objectives

The field of adversarial machine learning is also useful for identification of vulnerabilities in a machine learning approach in presence of adversarial settings

- Illustrate the design cycle of a learning-based pattern recognition system for adversarial tasks.
- Performance of pattern classifiers and deep learning algorithms under attack, evaluate their vulnerabilities.
- Pattern recognition tasks like object recognition in images, biometric identity recognition, spam and malware detection.

Introduction

Deep neural networks and **machine-learning** algorithms are currently used in several applications, ranging from computer vision to computer security. Many areas of machine learning are **adversarial** in nature because they are safety critical, such as autonomous driving. An adversary can be a cyber attacker or malware author attacking the model by causing congestion among users, or may create accidental situations, or may even model expose vulnerabilities in the prediction module by creating undesired situation. [1].

Materials

Pattern classifiers can be significantly vulnerable to well-crafted, sophisticated attacks exploiting knowledge of the learning algorithms. Being increasingly adopted for security and privacy tasks, it is very likely that such techniques will be soon targeted by specific attacks, crafted by skilled attackers. Larger number of potential attack scenarios, respectively referred to as evasion and poisoning attacks.

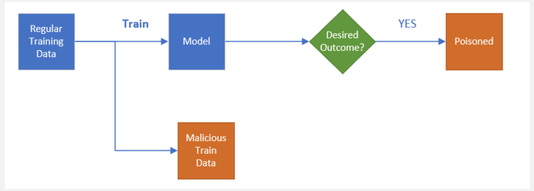
Mathematical Section

Poisoning attacks include those systems that exploit feedback from the end users to validate their decisions.PDFRate an online tool for detecting malware in PDF files.

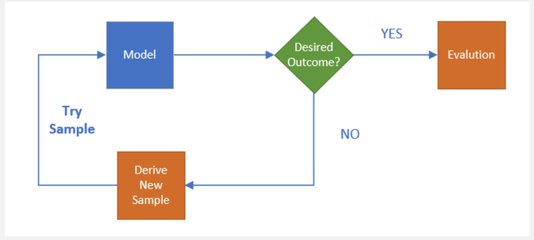
Evasion attacks consist of manipulating input data at test time to cause misclassifications. Which manipulation of malware code to have the corresponding sample undetected.

Methods

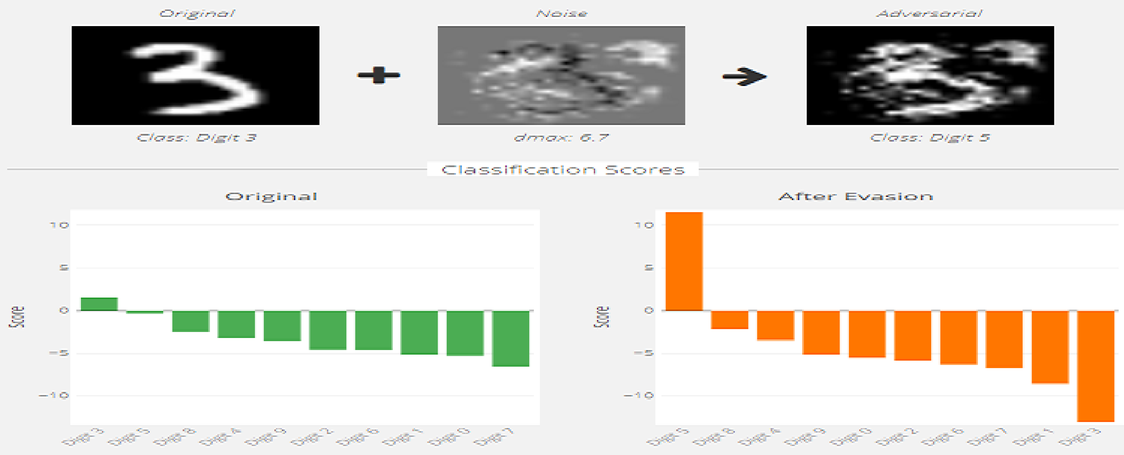
1. Poisoning (Causative) Attack : Attack on training phase. Attackers attempt to learn, influence, or corrupt the ML model itself.



2. Evasion (Exploratory) Attack : Attack on testing phase. Do not tamper with ML model, but instead cause it to produce adversary selected outputs.



Important Result



Images that can be misclassified by deep-learning algorithms while being only imperceptibly distorted. evasion attacks are thus already a relevant threat in real-world application settings.

Conclusion

Nunc tempus venenatis facilisis. **Curabitur suscipit** consequat eros non porttitor. Sed a massa dolor, id ornare enim. Fusce quis massa dictum tortor **tincidunt mattis**. Donec quam est, lobortis quis pretium at, laoreet scelerisque lacus. Nam quis odio enim, in molestie libero. Vivamus cursus mi at *nulla elementum sollicitudin*.

Additional Information

Maecenas ultricies feugiat velit non mattis. Fusce tempus arcu id ligula varius dictum.

- Curabitur pellentesque dignissim
- Eu facilisis est tempus quis
- Duis porta consequat lorem
- Duis porta consequat lorem

References

[1] J. M. Smith and A. B. Jones.
Book Title.
Publisher, 7th edition, 2012.

[2] A. B. Jones and J. M. Smith.
Article Title.
Journal title, 13(52):123–456, March 2013.

Contact Information

- Web: http://ideal1st.com/
- Email: ideal1st.here@gmail.com