Machine Learning Security Survey

Turhan Kimbrough
Department of Computer Science
Towson University
Towson, Maryland
tkimbr1@students.towson.edu

***Abstract*—Machine learning has experienced a significant growth in usage over the past few decades. Due to its data-centric approach in modeling, machine learning has seen use in a variety of subfields in computer science. In particular, researchers have been interested in incorporating machine learning into the domain of cybersecurity, utilizing it from the perspective of an adversary or ally. Researchers have also been concerned with the security state of current machine learning models. This survey paper provides a comprehensive overview of the state of machine learning, its application in various aspects of cybersecurity, securing machine learning systems, and future research directions being explored.**

***Keywords*—Neural networks, security, classification**

I. Introduction

With the increasing widespread adoption of machine learning technology, its usage is being observed in a variety of different fields. Some notable examples include image recognition, voice assistant technologies, email spam filters, and search engines. Much of its recent popularity can be attributed to the availability of frameworks such as tensorflow, allowing people of almost any background to quickly draft a machine learning application.

However, one growing concern tied to the ubiquity of machine learning is its accessibility to adversaries. Based on the assessment of current and prior research, there are a number of vulnerabilities in current machine learning models which can be exploited with little knowledge of a system's domain. In addition, attackers have been able to leverage machine learning technology to assist the deployment of cyberattacks. Enterprise machine learning applications may often contain large datasets of important information, becoming a potential candidate of a targeted attack. This paper will survey the domain of cybersecurity with regard to machine learning. This topic will explore the fundamentals of machine learning, current vulnerabilities in machine learning systems, machine learning technology from the perspective of adversaries, and future research directions.

II. Overview

This section will introduce the fundamentals of machine learning. Additionally, there will be a discussion about the key terms used by members of the community.

A. Machine Learning Basics

Traditionally, when software developers are tasked with solving a problem, they use a combination of rules and logic to find a solution. The basic routine consists of finding appropriate input values, creating the logic and rules to process the input, and producing the appropriate output. The traditional approach to software development allows for fine-tuned control of program behavior to achieve the solution. However, this approach does not scale with the complexity of additional rules and/or possible solutions. An example is image classification, where the logic needed to compare images is complex. This becomes a bigger concern when new classifications need to be derived with new image data. Machine learning flips the traditional programming paradigm on its head, by taking a series of solutions as input, and letting the machine develop the rules by detecting patterns in the solutions. The result is a self-propagating mathematical model, capable of making decisions on newly supplied data. This approach relies on large sets of well-defined data, and has the flexibility for being used in many different applications.

Briefly mentioned earlier, a common use case for machine learning is *classification*, where a dataset is categorized into different groups based on one or more *features*. A *feature* is defined as some measurable property or characteristic being observed in a dataset. There are several types of classifications, including *binary classification*, *multi-class classification*, and *multi-label classification*. *Binary classification* categorizes data based on whether a feature is present or not, resulting in an outcome of true or false. *Multi-class classification* categorizes data into different groups, where each data instance is assigned according to its feature. *Multi-label classification* categorizes data into different groups, where each data instance is assigned according to its expression of on or more features.

Training is the process of teaching a machine learning model to detect patterns in datasets. There are two types of training mechanisms, *supervised training* and *unsupervised training*. *Supervised training* requires each data instance to have one or more labels, defining which category or feature it expresses. In contrast, *unsupervised training* omits the need for labels, and the machine learning model will categorize datasets on its own. Typically, after the training phase, a machine learning model will go through the process of *validation*. *Validation* typically consists of classifying a separate dataset to guarantee the accuracy of a model and preventing a phenomenon called *over-fitting*, where a model will only 'memorize' characteristics or patterns of training data.

III. Vulnerabilities in Machine Learning Models

Due to the unique mechanism for constructing machine learning models, there are several ways which they can be exploited by an attacker. As a result, the attack surface which exists on machine learning models will differ from traditional software systems. This section will discuss several vulnerabilities which are currently present in machine learning.

A. Data Poisoning

Data Poisoning is the act of manipulating, removing, or adding data during the training phase of machine learning. This type of attack is known as a *black box attack*, where an attacker does not need to know the implementation of a system to attack it.

One popular instance of a data poisoning attack is the adversarial example. This requires the attacker to have some knowledge of the training data, such as its dimensions and data type. The attacker would then manipulate data instances in a way where the machine learning model would be fooled, but appears normal to a human observer. The reason for manipulating data instances in this way is two-fold. First, machine learning models are often constrained to data fitting a specific dimension, shape, size, or length of characters. This is typically done to prevent incompatible data from entering the model during training. Second, there is often one or more people who are observing the model with testing or validation datasets. An attacker would want to minimize any evidence of the data being tampered with.

Often, the goal of this attack is to make a machine learning model incorrectly classify data. The consequences of this attack can be devastating. Examples include tricking an autonomous vehicle to misinterpret traffic signs, sneaking malicious data past an intrusion detection system, or bypassing email spam filters.

B. Membership Inference

Membership inference is a mechanism of data extraction, where an adversary intends to know whether certain samples were used as training data for a machine learning model. This type of attack is also classified as a *black box attack*.

This particular vulnerability requires significant effort from the attacker. The attacker would need access to a dataset of sufficient size mimicking the data in the target model. The data would then be used to create several *shadow models*, which are used only to recognize differences in the target model's behavior. This is done to expose *overfitting*, when a model's analysis corresponds too closely to its training data. Therefore, an attacker can interpret whether certain samples were used in the training dataset based on the target model's confidence level in classifications.

Often, this can exploit the confidentiality of information on a system. An attacker has the capability to correlate information between datasets to target individuals for other cyberattacks.

C. Transfer Learning

Transfer learning is a mechanism where an adversary has the ability to study a publicly available machine learning model, and use that insight to sneak past and/or corrupt similar target systems. This type of attack is classified as a *white box attack*, since the attacker would need to have full access to at least one machine learning model.

This particular vulnerability requires the attacker to have knowledge of the input data, learning mechanism, and output behavior of a machine learning model similar to the target. Once this information is obtained, the attacker would test the system and learn its overall behavior to enumerate flaws in the model's logic. The assumption is that the flaws found in the available model's logic would appear in a target system.

Often, the goal of this attack is similar to the other two mentioned above. An attacker would use the information to trick, fool, manipulate, or sneak passed a target machine learning model.

IV. Mitigations To Vulnerabilities

Due to the large attack surface found in machine learning models, there has been a significant focus in its security. This section will discuss several mitigations to the vulnerabilities discussed in the previous section.

A. Selecting Trusted Datasets

To secure a machine learning system, it is important to start at the beginning of the machine learning lifecycle, securing the input data itself. Since the training phase typically requires a large amount of data, many projects utilize publicly available datasets. Publicly available datasets are similar to open source projects, in that contributions are made by and verified by a community of individuals. This means that the trust is placed on the community to provide legitimate information. To find trusted datasets, one can check the policy of verifying contributions, number of community members/maintainers, and entities backing the project. For example, datasets from the Government, Universities, and repositories such as kaggle are typically trustworthy.

If a dataset is chosen from a location with an unknown reputation, it is recommended to perform an audit to ensure the data consists of relevant samples, proper labels, and a balanced set of classes.

B. Sanitizing input

C. Machine Learning Retraining

D. Differential Privacy

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.