Journal of Applied Mathematics and Computation, 2022, 6(4), 535-539
https://www.hillpublisher.com/journals/JAMC/
ISSN Online: 2576-0653
ISSN Print: 2576-0645

Research Trends in Adversarial Machine Learning

Jihang Jiang

Nanjing University of Information Engineering, Nanjing, Jiangsu 210044, China.

How to cite this paper: Jihang Jiang. (2022) Research Trends in Adversarial Machine Learning. *Journal of Applied Mathematics and Computation*, 6(4), 535-539.
DOI: 10.26855/jamc.2022.12.018

Received: October 28, 2022
Accepted: November 25, 2022
Published: December 30, 2022

***Corresponding author:** Jihang Jiang, Nanjing University of Information Engineering, Nanjing, Jiangsu 210044, China.

Abstract

The field of machine learning artificial intelligence algorithms has made significant progress. With the development of the times, the security brought by machine learning is worthy of consideration, and adversarial machine learning improves the reliability of machine learning algorithms through continuous training algorithms. Hence, the research of adversarial machine learning based on network security is well worth exploring.

Keywords

Adversarial learning, machine learning, cyber security, development trend

Introduction

Machine learning is a technology that has been used in various fields with excellent performance since the beginning of its emergence. In recent years, with the rapid development and broad application of machine learning, this field has been flourishing like never before. Machine learning has achieved recognized results in complex tasks such as computer vision, speech recognition, and natural language processing. It has been widely used in autonomous driving and face recognition fields.

As machine learning technology blossoms everywhere and gradually penetrate people's lives, it is also applied in many fields with strict security requirements, such as security, finance, and medical care, directly affecting people's personal, property, and privacy safety and security. In the face of a series of significant advances, machine learning is increasingly being applied to all aspects of human life, when it is also easy to ignore the shadow behind the sun.

1. Adversarial machine learning concept

The so-called Adversarial Machine Learning (AML) is a security subdivision of machine learning research that can ensure the security of machine learning application models to a certain extent. Like many practical techniques, machine learning as a complex computer system faces the test of security and is exposed to hacking attacks and has been found to have security issues that interfere with the output of the machine learning system with correct results, such as the presence of adversarial data. Researchers have found that some well-designed adversarial samples can cause machine learning models to incorrect output results.

The error rate of deep neural networks for adversarial samples is very high, and it is almost impossible for a human to discern the difference between the original sample and the adversarial sample, which means that the original function of deep neural networks has failed.

It can be seen that adversarial attacks are very harmful, especially for uncrewed vehicles, medical diagnosis, and financial analysis, which are security-critical fields. Adversarial samples undoubtedly constrain the further application of machine learning techniques, so improving the adversarial robustness of neural networks is essential.

Thus, adversarial samples in machine learning have attracted much attention from researchers. They have accordingly proposed a series of methods to counter attacks and counter defenses, a field called adversarial learning.

Adversarial learning is a field at the intersection of machine learning and computer security that aims to provide security to machine learning techniques in malicious environments.

1.1 Glossary Description

Adversarial sample: is a new sample synthesized by adding perturbations to an accurate sample, obtained by synthesizing the input data of a deep neural network and artificially crafted noise. However, the deep neural network is vulnerable to adversarial data, which can easily confuse the deep neural network. It will be recognized correctly by the human visual system.

Adversarial robustness: the ability to resist adversarial samples.

Adversarial attack: the generation of adversarial samples that give the lowest.

Adversarial training: During the network training process, adversarial samples are continuously generated and learned. Then the adversarial samples are learned in the outer layer to minimize the loss function. For example, according to the minimal-extreme formula, the adversarial samples are found in the inner layer by maximizing the loss function.

2. Fundamental theory of adversarial learning

In 2012, deep neural networks began to dominate computer vision problems. Thus adversarial deep learning, which focuses on adversarial sample generation and defense, has become the most popular research hotspot in the field of adversarial machine learning [1].

Starting from 2014, Christian Szegedy et al. proposed that deep neural networks could be fooled by adversaries [2] and introduced the concept of adversarial samples against images, exposing the significant shortcomings of deep learning techniques in terms of security, which led to a more cautious view on the application of deep learning in practice.

3. Impact of adversarial learning

Adding the computed perturbation noise to the original image makes the classifier that can correctly classify the original image misclassify the image to which the perturbation is added. The magnitude of this perturbation is so tiny that the human eye does not misclassify it, but it can easily "trick" the deep neural network during the testing or deployment phase.

Dealing with the vulnerability caused by adversarial samples becomes essential when applying deep neural networks to environments with strict security requirements [3]. Moreover, it is not a particular machine learning (including deep learning) algorithm that is individually vulnerable to adversarial samples, but machine learning models, in general, may have this flaw.

This flaw is fundamental in industries with high-security requirements. As a result, adversarial learning can be widely used in healthcare, finance, security, and autonomous driving industries.

Image recognition: Recent research has further revealed that pixel-level perturbations and real-world perturbations are aggressive even when captured through cameras, making adversarial attacks more likely to appear in the world we live in [4].

4. Classification of model attacks

According to different classification criteria, adversarial attacks (how to generate adversarial samples) have the following classifications [5], which can be divided into black-box attacks, white-box attacks, gray-box attacks, and real-world attacks, in terms of the attack environment or according to the capabilities possessed by the attacker

4.1 Black-box attacks

The attacker does not know the internal structure of the model, training parameters, or defense methods. The attacker only has access to the input and output of the model. Generally, the target model is approximated by training alternative models. Then the information from the trained alternative models is used to generate adversarial samples to perform attacks on the unknown target model. However, migratory adversarial samples between different models allow for implementing black-box attacks.

4.2 White-box attacks

In contrast to the black-box model, the attacker has control over the model's structure, weights, and inputs and outputs. Most of the current attack algorithms are white-box attacks. The commonly used white-box attacks rely on gradient information. The idea is to calculate the gradient of the loss function on the sample and then find the perturbation in the direction of the gradient.

4.3 Gray-box attacks

Only a part of the model is known between black-box and white-box attacks. (e.g., just getting the output probability of the model or knowing only the model structure but not the parameters).

4.4 Real-world attacks

Real-world attacks in the real physical world. For example, the adversarial samples are printed out and identified by taking pictures with cell phones.

Confrontation training is one of the fundamental ways to improve the robustness of profound network confrontation. The basic idea of adversarial training is to continuously generate and learn adversarial samples during the training process of the network.

For example, according to the minimal-extreme formula, the adversarial samples are found in the inner layer by maximizing the loss function. Then the adversarial samples are learned in the outer layer to minimize the loss function—the adversarial training results in an adversarially robust neural network.

5. Defense mechanism of adversarial attack

Defending against adversarial sample attacks is mainly based on additional information introduced into the auxiliary block model (AuxBlocks) for additional output as a self-integrating defense mechanism, which is effective, especially in black-box and white-box attacks against attackers.

In addition, defensive distillation can also play a defensive role; defensive distillation is a way to migrate the trained model into a simpler structured network to defend against adversarial attacks.

5.1 Simulation of Attack Means

The critical step in active defense is to simulate the means of attack. As mentioned above, there are two types of attacks against capability limitations: the evasion attack during testing and the poison attack during training.

5.1.1 Evasion Attack

Evasion attack has a long history, as early as 2004 when people studied the problem of attacking spam classification. The core idea is to obfuscate "bad" words or add "good" words. At its simplest, suppose one has a linear spam classifier, and he finds the words with the highest classifier weight and tampers with them, thus achieving less cost and higher readability. With this in mind, the corresponding countermeasure strategy is also available. That is to even out the weights of the features so that the classifier needs more vocabulary modifications to reach a misclassification. So there are more known solutions to the problem of attacking linear classifiers.

Then people started to turn to nonlinear classifiers. So how do people know which features affect the judgment of the classifier? The answer is that the gradient of the classification function specifies the direction of maximum change. Again based on this, we modify the data so that the discriminant function of this data is as small as possible. Then the misclassification is as reasonable and plausible as possible. This is the case if the model is known. If we do not know the model, we can train a proxy model based on the available information and attack the proxy model to obtain an approximation.

After that, it is time to attack deep learning. The same idea: minimizing the "distance" between the counter sample and the actual sample when their prediction categories are different. In the same way, the agent model can also be used as a counter-defense strategy to bypass the defense mechanism to attack the model to some extent.

5.1.2 Poisoning Attack

The basic idea of a Poisoning Attack is to deliberately add some "poison" data to the training model so that the trained model is affected. The central idea is the double-layer optimization problem. This theory was first proposed in 2006, and soon there were some of the most traditional machine learning models to make the Poisoning Attack, including SVM, Ridge Regression, and LASSO.

While poison attacks seem to require a higher level of sophistication, such as access to training datasets and covariance models., examples reflecting this concept already exist. Tay, a Microsoft chatbot, is one. The open 16 hours is the training process, but this process is the "poison" of the training, such as racist and other malicious comments. So AI can imitate human behavior, but it cannot tell which behavior is misleading.

5.2 Defense methods

An intuitive direction for self-defense is remediation and prevention. Accordingly, we divide defense into Reactive Defense and Proactive Defense.

5.2.1 Reactive Defense

As mentioned before, Reactive Defense is a seemingly sound strategy. However, if we consider it carefully, Reactive Defense is more convenient and direct than Proactive Defense in reducing the risk of potential future attacks. Three perspectives can be considered in this direction: periodic detection of attacks, periodic retraining, and decision validation. It can consider more data features that have yet to be considered. In more practical terms, it is to regularly collect newly discovered known types of attacks, equip one's model with the ability to identify such attacks, and then retrain them in due course. The last thing is to have experts verify from time to time that the decisions are sound, thus ensuring that the model has stayed within the direction of the decision.

5.2.2 Proactive Defense

Proactive defense is about preventing possible future threats. There are two ideas here, Security-by-Design for white-box attacks and Security-by-Obscurity for gray-black-box attacks.

5.2.3 Security-by-Design

Design defense is, as the name implies, designing systems and algorithms as securely as possible from the start. Let us start with how to defend against evasion attacks. The intuitive idea is to take the "evaded" samples and add them to the training set. These samples can then be obtained by simulating the attack. However, one problem is that this defense is heuristic, and there is no formal guarantee of convergence and robustness. In any case, this more theoretical attempt also faces more complex problems in practice, such as whether the actual attacker will act according to the theoretical assumptions. Also, obtaining the distribution of attack samples can become very time-consuming and laborious for accurate high-dimensional data.

5.2.4 Robust Optimization

Robust optimization is the solution of adversarial learning as a Minimax Problem. The inner problem is to maximize the training loss by disturbing the training set as much as possible, while the outer problem is to minimize the situation that causes the corresponding training loss. In other words, it reduces the likelihood of causing a significant risk. Interestingly, the inner problem is to solve the very problem that the linear classifier regularizer does. In this way, robust optimization is equated with the regularized objective function. Roughly speaking, it means that in some exceptional cases, robust optimization wants us to ensure the generality of the model by being less sensitive to adversarial samples, just like the regularized linear classifier. It has also been recently shown that the formulation can be derived from nonlinear classifiers.

Another defense mechanism is detecting and denying points "far" from the training data. These samples are called blind dodging points because they may appear in places where the data is sparsely distributed, causing misclassification. Therefore, these samples are not accepted to prevent the effect of adversarial samples.

5.2.5 Security-by-Obscurity

What we can do against gray-black box attacks is to improve the security of the model by hiding as much information as possible that the attacker wants to know. Examples include disrupting the training data, collecting data from different times and places; using models that are difficult to reverse engineer, such as sets of classifiers; making it more difficult to access the model or training data; and randomizing the output of the model to some extent to obfuscate the feedback given to the attacker.

The effectiveness of these approaches above is limited because the most direct way to break through this layer of obstruction is to train agent models, whether micro or non-micro cable.

6. Direction of adversarial learning development

Although the current research in adversarial learning has proposed many attack algorithms for adversarial sample generation, there is still much room for improvement in defense mechanisms. For different attack methods, the defense means is usually to patch the vulnerability, and no standardized and common way has been found to defend against all the adversarial attack methods.

Even with those mentioned above, either auxiliary block model, defensive distillation, or the capsule neural network, which is now developing very hot., the integration methods are not very mature and do not form a complete defense system, and can only achieve a practical defense effect locally. In the direction of fighting attacks, there are still excellent prospects for the development of defense technologies and mechanisms.

7. Conclusion

Currently, the way of attack and defense in counterattack has gone through many rounds of iterations and evolved

many ways of attack and defense. As various attack methods arise, the proposed defense methods seem to defend against these attacks, but the newly emerged attacks keep evading these defense methods.

To date, the essential properties of the black box that are neural networks are still not fully understood. It has even been pointed out that the classification task accomplished by neural networks relies solely on discriminating local color and texture information, which allows natural adversarial samples, even if they are not artificially added perturbations—however, authentic captured images deceive neural networks successfully.

This also supports the view of many researchers that neural networks only learn data, not knowledge, and that machine learning is not yet able to learn as humans do. The ultimate solution to this puzzle may depend on a thorough understanding of neural networks and improvements to their structure.

Figuring out the internal learning mechanism of neural networks and improving them may truly address the vulnerability of current neural networks to adversarial attacks. Thus adversarial machine learning is not only a threshold for machine learning to be more widely used but also a motivation for research on how to interpret machine learning models.

References

[1] J. Yang. Research on some problems in adversarial machine learning [D]. Hainan University, 2021.

[2] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks, 2013.

[3] Liu QX, Wang JunN, Yin J, et al. application of adversarial machine learning in network intrusion detection [J]. Journal of Communication, 2021, 042(011):1-12.

[4] Jiang Yan, Zhang Liguo. A review of adversarial attack and defense methods for deep learning models [J]. Computer Engineering, 2021, 47(1):11.

[5] Xiaoyong Yuan, Pan He, Qile Zhu, et al. Adversarial Examples: Attacks and Defenses for Deep Learning, 2017.