Adversarial attack application analytics in machine learning

Hongsheng Zhang*

School of Computing, Wuhan Qingchuan University, Wuhan, Hubei, China

Abstract. Machine learning is one of the most widely studied and applied technologies, but it is itself vulnerable to attack and its algorithms have the risk of privacy leakage. In this article, through the experts currently popular speech recognition scene, reveals how to build the antagonism against data, make its differences with the source data is subtle, so much so that humans can't through sensory recognition, and machine learning model can accept and the classification of making the wrong decision, at the same time made attack, finally prospects the study model to research the development and application of security and privacy protection.

Keywords: Machine learning, Privacy threats, Adversarial attacks.

1 Introduction

Machine learning is currently a new discipline that is being applied comprehensively to various fields. The theoretical basis of machine learning involves a wide range of fields: neurophysiology, mathematics, biology, automation and computer science, and the integration of various learning methods has formed a diversified and integrated learning system.

With the popularity of mobile devices and the integration of mobile devices into these emerging input methods, this technology is being experienced by most people. The accuracy of speech and image recognition is crucial to the effectiveness of the machine's understanding and execution of user instructions. At the same time, this link is also the easiest for attackers to exploit through minor modifications to the data source. Achieve the purpose of making an incorrect follow-up operation that the user is unaware of and the machine receives the data. And it can lead to the intrusion of computing equipment, the execution of wrong commands, and the serious consequences of a chain reaction after execution.

Based on speech-specific scenarios, this paper first briefly introduces the white box and black box attack models, and then combines the research results of experts. It further introduces the attack scenarios, countermeasures to data construction attack methods, and attack effects

* Corresponding author: whhzkjdx1979@163.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

2 Adversarial attacks in machine learning

2.1 Overview

Since the input form of the machine learning algorithm is a numerical vector, the attacker will design a targeted numerical vector to make the machine learning model misjudge, which is called adversarial attack. Unlike other attacks, adversarial attacks mainly occur when adversarial data is built and then fed into a machine learning model like normal data to obtain deceptive recognition results.

By adding small perturbations to the input to make the classifier classification error, the attack algorithms generally used for deep learning networks are the most common, and the application scenarios include the current hot CV and NLP directions, such as: by adding carefully prepared perturbation noise to the picture to make the classification error, or by replacing certain words with synonyms in a sentence to make the emotion classification error.

2.2 Classification of counter-attacks

There are many types of attacks, and from the perspective of attack environment, they can be divided into black box attacks and white box attacks. A black box attack is that the attacker knows nothing about the internal structure, training parameters, defense methods, etc. of the attack model, and can only interact with the model through input and output; white box attack is the opposite of the black box model, and the attacker can grasp everything about the model.

3 Black box attack

3.1 Overview

Attackers are unaware of the algorithms and parameters used by machine learning, but they can still interact with the machine learning system. For example, they can observe and judge the output by passing in any input. A black box attack treats the target model as a black box, without understanding the internal details of the model, and the attacker can only control the input of the model.

3.2 Black box attack model analysis

In Figure 1 of the black-box attack model, the attacker is not aware of the machine learning algorithm, and the attacker's only knowledge is that the machine uses the MFC algorithm. The MFC algorithm is a transformation that converts audio from high dimensions to low latitudes, filtering out some noise while ensuring that machine learning can manipulate these inputs. But in the process of conversion from high-dimensional to low-dimensional, some information is inevitably lost. Correspondingly, the conversion from low-dimensional to high-dimensional will also add some noise.

The principle of black box attacks is exactly what the attacker does by iterating. Constantly adjusting the parameters of the MFCC and performing MFCC transformations and inverse transformations on the sound, filtering out the information that the machine does not need but the human must have, so as to construct a confused speech. Because the MFC algorithm is heavily used in this scenario for speech recognition. So the attack model still guarantees strong versatility.

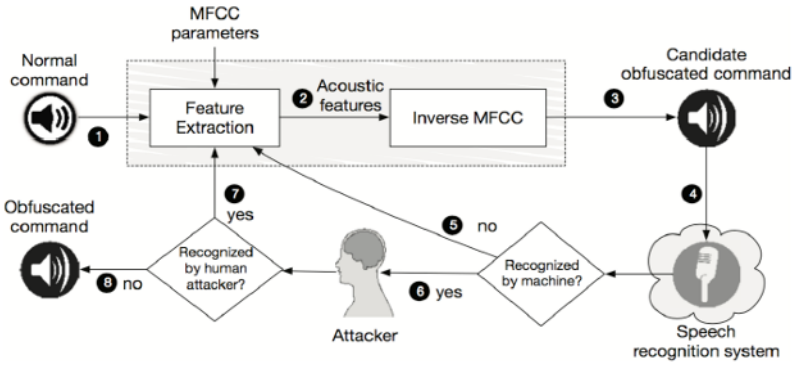

Fig. 1. Adversarial voice black box attack model.

MFCC, The Mel Inverted Coefficient, is a phonetic feature. It is a reciprocal parameter extracted from the Mel scale frequency domain, the Mel scale describes the nonlinearity of the human ear frequency, and its relationship with the frequency can be approximated by the following equation such as equation:

$$Mel(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \tag{1}$$

(1) In formula f is the frequency, the unit is Hz. The following figure shows the relationship between mel frequencies and linear frequencies:

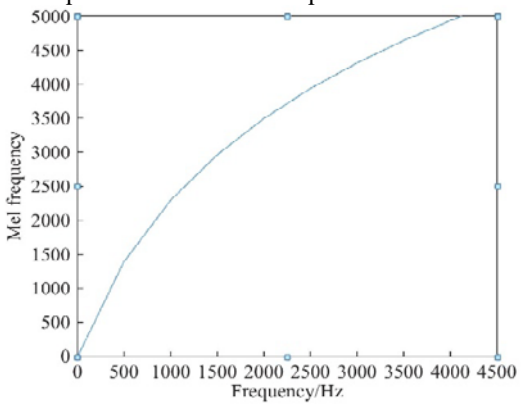

Fig. 2. Mel frequency vs. linear frequency.

3.3 Analysis of experimental results

In the experiment, the speech recognition system used could only recognize voice commands within 3.5 meters. In the case of the distance between the speaker and the mobile phone is controlled at 3 meters, Table 1 counts the proportion of recognition of different commands by humans and machines. On average, 85% of normal voice commands are recognized by speech. In their obfuscated version, 60% of voice commands are still recognized normally. In the human recognition category, experts use the Amazon Mechanical Turk service. Let the inspector guess the content of the voice in the form of crowd sourcing. In this case, the effect of different command obfuscation is not the same. For "OK Google" and "Turn on airplane mode" commands, less than 25% of obfuscated

commands can be correctly recognized by humans. Among them, 94% of the "Call 911", the obfuscated version is normally recognized by humans to compare abnormal.

Two main reasons were analyzed: First, the command was too familiar. Second, the tester can repeat the voice multiple times, thus increasing the probability of success of the guess.

Table 1. Results of adversarial voice black box attacks.

	Ok Google		Turn on airplane mode		Call911	
	machine	mankind	machine	mankind	machine	mankind
Normal speech	90%(34/40)	89%(356/400)	75%(30/40)	69% (314/455)	90% (35/40)	87% (283/324)
Adversarial speech	95%(38/40)	22% (86/376)	45% (18/40)	24% (109/443)	40% (16/40)	94% (246/260)

4. White box attack

4.1. Overview

An attacker can learn the algorithm used by machine learning and the parameters used by that algorithm. Attackers can interact with machine learning systems as they generate adversarial attack data. White-box attacks require a complete acquisition of the model, an understanding of the structure of the model, and the specific parameters of each layer, and the attacker can control the input of the model and even modify the input data at the bit level.

4.2. White box attack model analysis

In a white-box attack, the target machine learning algorithm that experts confront is the open-source CMU Sphinx speech recognition system[4]. Throughout the system, CMU Sphinx first slices the entire speech into a series of overlapping frames, and then uses Mel Frequency Cepstrum (MFC) conversion for each frame to reduce the audio input to a smaller dimensional space, namely feature extraction in Figure 3. CMU Sphinx then uses the Gaussian Mixture Model (GMM) to calculate a probability of a particular audio to a particular phoneme. Finally, with the Hidden Markov Model (HMM), Sphinx can use the probabilities of these phonemes to translate into the most likely text. Here both GMM and HMM belong to the machine learning algorithms in Figure 3.

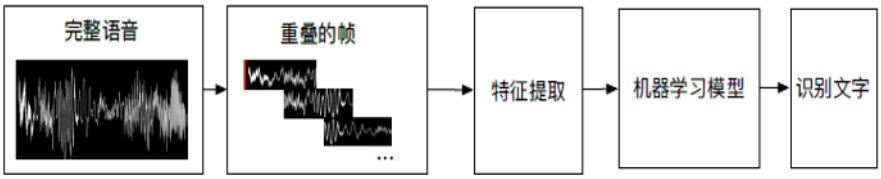

Fig. 3. Cmu sphinx speech recognition system model [4].

Two approaches are proposed in Tavish's white box attack model, the first being the simple approach and the second being the Improved attack. The first method differs from the black box method in that it knows the parameters of the MFCC, allowing gradient descent to be more targeted to retain only a few key values that are critical to machine identification. During the entire gradient descent, input frame. The target of machine recognition is constantly approached, and at the same time, some of the superfluous information required for human recognition is inevitably eliminated.

The basic principle of the second type of white box attack is that according to the different sensitivities of machines and people to changes in pitch fluctuations (phonemes), by reducing the number of frames corresponding to each phoneme, this sound can only be recognized by the machine, while humans can only hear a flat and chaotic noise. These eigenvalues are then inversely transformed by MFCC and eventually become an audio piece that reaches people's ears. Table 2 shows the effect of their attacks.

Table 2. Adversarial voice white box attack effect [3].

	machine	mankind
Normal speech	-	74%(230/310)
Adversarial speech	82%(82/100)	0%(1/378)

5 Concluding remarks

Although the discovery of adversarial data attack is very clever, but in the current image speech recognition application occasion, effective defense is not difficult, with the further development of machine learning research and the wide application of machine learning technology in actual scenarios, the security and privacy of machine learning models has become a new and promising research field, with one application scenario after another is easily broken. Although at present, it is only in scenarios such as speech recognition. We can be soberly aware. When these scenarios are combined with other services, the serious consequences of successful attacks Artificial intelligence, as an indispensable part of the future intelligent automation service, has become a new battlefield for the security industry to fight hackers.

References

1. ong C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In: Proc. of the 2017 ACM SIGSAC Conf.on Computer and Communications Security. 2017. 587-601
2. Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction apis. In: Proc. of the 25th {USENIX} Security Symp. ({USENIX} Security 2016). 2016. 601-618
3. Nelson B, Biggio B, Laskov P. Understanding the risk factors of learning in adversarial environments. AISec, 2011,11:87-92
4. P.Lamere,P.Kwork,W.Walker,E.Gouvea,R.Singh,B.Raj and P.Wolf,"Design of the CMU Sphinx 一 4 Decoder,"in Eighth Europe on Conference on Speech Communication and technology,2003
5. Jagielski M, Oprea A, Biggio B, et al. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In: Proc. of the 2018 IEEE Symp. on Security and Privacy (SP). 2018. 19-3
6. Barreno M, Nelson B, Sears R, et al. Can machine learning be secure? In: Proc. of the 2006 ACM Symp. on Information, Computer and Communications Security. 2006. 16-25
7. Newsome J, Karp B, Song D. Paragraph: Thwarting signature learning by training maliciously. In: Proc. of the Int'l Workshop on Recent Advances in Intrusion Detection. 2006. 81-105