Security and Privacy in Machine Learning

Zihao Wang

School of Cyberspace Security, Chengdu University of Information Science and Technology, Chengdu Airport Engineering University, Chengdu, Shuangliu District, Chengdu, Sichuan Province, China, 610225, China

Abstract

In many applications of machine learning, their security and privacy have fatal weaknesses. For example, machine learning for fingerprint recognition, we hope that the machine learning algorithm is only sensitive to the fingerprint of the phone holder, but not the fingerprint of other people. But researchers at New York University have created a 'universal fingerprint' that unlocks many encrypted smartphones. As another example, machine learning for medical diagnosis, we hope that machine learning algorithms do not memorize sensitive information about the training set, such as the specific medical history of an individual patient. However, many models still reveal private information about individual patients. This has led many re- searchers to invest a lot of time and energy in researching security and privacy in machine learning. In order to solve the security and privacy flaws of machine learning algorithms, I have designed two new models: (1) Using differential privacy against adversarial examples to improve GAN's robustness to adversarial examples. (2) Using Intel SGX software-based differential privacy and deep learning models to improve data security and privacy. These new learning models protects against sample attacks and increases privacy protection.

Keywords

Deep learning, GAN, differential privacy.

1. Introduction

Since the idea of AI was first proposed in the 1960s, AI has received extensive attention and in-depth research from academia and industry. As the core of AI, machine learning has also achieved unprecedented development in recent years, and its application covers all fields of artificial intelligence. For example, in the field of computer vision, we can use machine learning to classify objects and locate objects.

We can also use deep neural network design to implement a high-accuracy face recognition system [1]. In the field of natural language processing, we can also implement a set of intelligent question answering system using machine learning design [2]. The development of machine learning has entered a new stage, and various machine learning algorithms and models emerge in an endless stream. In many scenarios, its performance is even better than humans.

Of course, machine learning has not yet reached the true level of humanity, because even in the face of a negligible attack, most machine algorithms will fail [3]. However, most scholars have not considered this issue. For example, when the model is trained (training phase) or the model is predicted (inference phase), the attacker makes malicious modifications to the input and output of the model or accesses the Internal components of the model by some means to steal the model parameters, thereby undermining the confidentiality, integrity, and availability of the model [4], which is a security and privacy issue in the machine learning model.

Figure 1. Machine Learning For Multiple Tasks. (DL stands for deep learning.)

Adversarial Attacks: In fact, the input form of the machine learning algorithm is a numerical vector, so the attacker will design a targeted numerical vector through subtle modification of the data source so that the machine learning model cant be accepted by the user. This data is used for the purpose of misjudgment. This type of attack is called the adversarial attack.

Differential privacy is a framework for evaluating the guarantees provided by a mechanism that was designed to protect privacy. Invented by Cynthia Dwork, Frank McSherry, Kobbi Nissim and Adam Smith [DMNS06], it addresses a lot of the limitations of previous approaches like k-anonymity. The basic idea is to randomize part of the mechanisms behavior to provide privacy. [5]

Intel Software Guard Extensions (Intel SGX) protect selected code and data from disclosure and modification. Developers can partition applications into CPU-

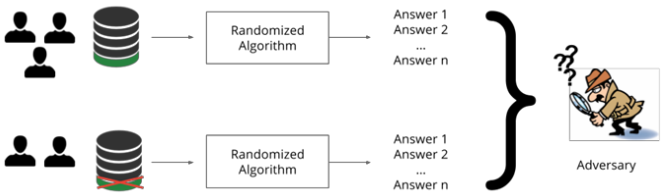


Figure 2. In the above illustration [6], we achieve differential privacy when the ad- versary is not able to distinguish the answers produced by the randomized algorithm based on the data of two of the three users from the answers returned by the same algorithm based on the data of all three users.

enhanced enclave or protected areas in memory, even in compromised platforms. With this new application layer trusted execution environment, developers can en- able identity and record privacy, secure browsing and digital management protection (DRM) or any high-security application that requires secure storage of confidential- ity or protection of data.

With the above introduction to attacks, differential privacy, and Intel SGX, it's not hard to see that the other two can be tied to differential privacy. So I want to apply differential privacy to the anti-sample and Intel SGX. In my research proposal, my main contribution was to design two new models:

Deep Differential Privacy Protection Model Based on Generative Ad- versarial Network.

Deep learning model with differential privacy.

2. Adversarial Attacks

This section is divided into four parts. In part 2.1 and 2.2, I first reviewed the CIA model of machine learning and listed three examples of attacking CIA mod- els. In part 2.3, I introduced several applications of confrontational attacks in real life. In part 2.4, I present several questions about the current model and where improvements can be made.

2.1 Security and privacy attacks and examples

Attacks on machine learning models can affect the Confidentiality, Integrity, and Availability of machine learning models.

Confidentiality Attacks: Machine learning systems must ensure that unau- thorized users have no access to information [7]. In practice, the model cannot reveal private data. For example, suppose the researchers designed a machine learning model that can check the patient's medical history and diagnose the patient. Such a model can greatly help the doctor's work, but it must be guaranteed that the malicious person can't attack the model. There is no way to recover the patient data used to train the model.

Availability Attacks: The availability of a machine learning model can also be an attack target. For example, in the unmanned field, if an attacker places a very difficult-to-identify item on the side of the road that the vehicle will pass, it is possible to force an autonomous vehicle into safe mode and then park on the side of the road.

Integrity Attacks: Machine learning models are most vulnerable to integrity attacks, which can occur both in the learning phase of the model and in the inference prediction phase of the model. If the attacker destroys the integrity of the model, the model's predictions will deviate from expectations [8]. In the training phase of the model, the training process that Interferes with the machine learning model reflects the attack strategy that causes more errors in the machine learning model [9]. At this stage, the most common attack is a data poisoning attack. The attacker can modify the existing training set or add additional malicious data, affect the training process of the model, and destroy the integrity of the model to achieve the purpose of reducing the accuracy of the model in the predictive reasoning stage. In the inference and prediction phase of the model, the integrity of the model is equally vulnerable, and the most common attack at this stage is against the sample attack. When the model training is completed and used for prediction, I only need to add a small disturbance to the sample to be predicted, which is unrecognizable by the human eye but enough to make the model classification wrong [10].

2.2 Adversarial Examplese

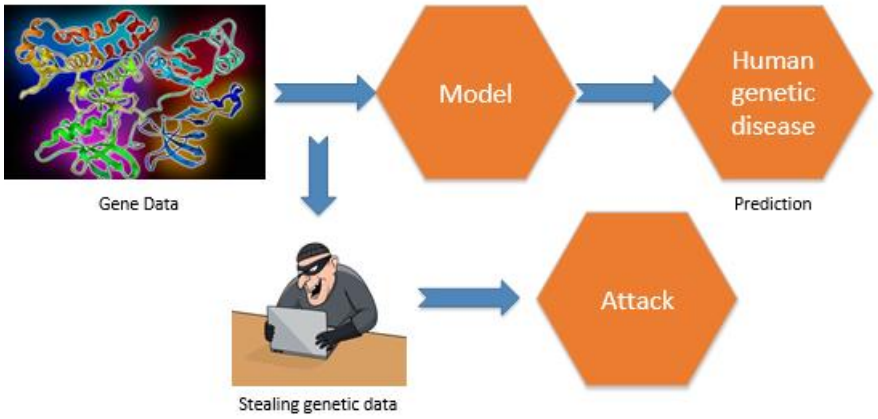
Bontrager et al. [12] produced six universal fingerprints by opposing the sample. They mainly attack the ambiguity of the fingerprint, because the ambiguity is diffi- cult to identify even if it is a neural network. The proposed method, referred to as Latent Variable Evolution, is based on training a Generative Adversarial Network on a set of real fingerprint images. Stochastic search in the form of the Covariance Matrix Adaptation Evolution Strategy is then used to search for latent input vari- ables to the generator network that can maximize the number of impostor matches as assessed by a fingerprint recognizer [12].

In our impression, DNN has identified more than 99% of the MNIST dataset. But we use the adversarial examplese to attack. As a result, the DNN will recognize the number '1' as the number '4'.

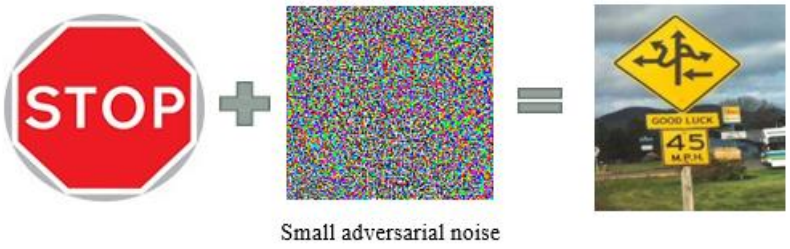
In fact, there are many examples of anti-sample attacks in other published papers, but the above two examples not only attack the security of data but also the privacy of data.

2.3 Application against sample attacks in practice

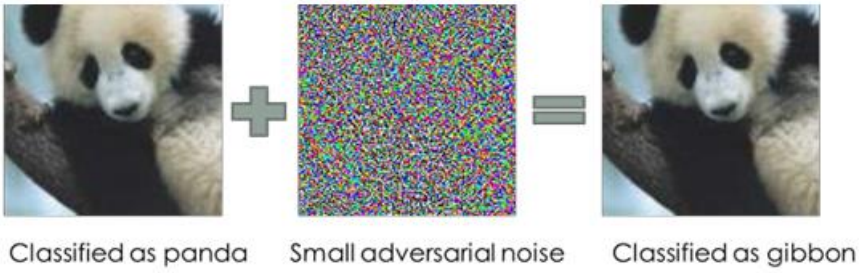
As the most powerful attack method to destroy the integrity of the machine learning model, the anti-sample attack is widely applied to the actual scene.



(a)



(b)



(c)

Figure 3. (a) Example of confidentiality attack; (b) Example of availability attack; (c) Example of integrity attack.

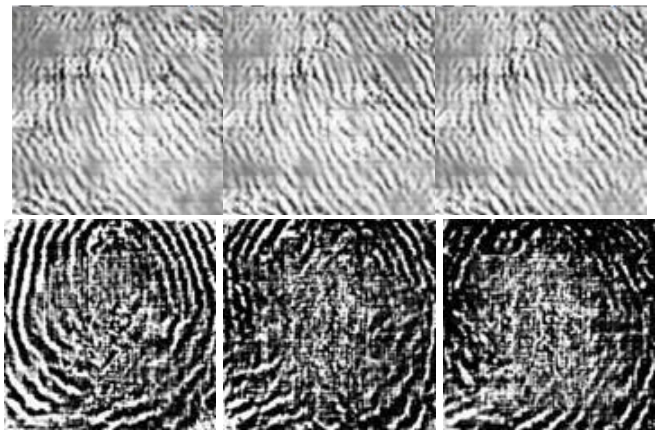


Figure 4. DeepMasterPrints: Universal fingerprint

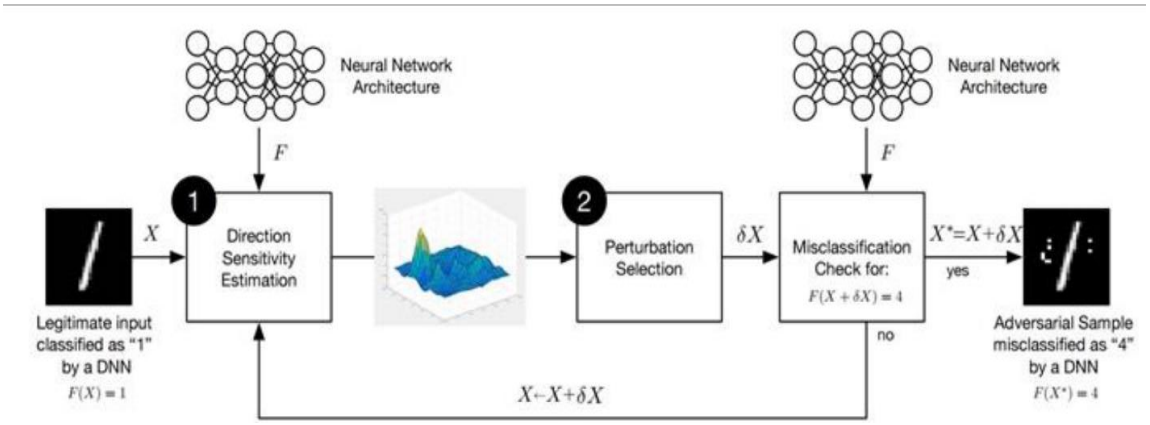


Figure 5. Jacobian-based Saliency Map Approach (JSMA).

In the field of face recognition, Sharif et al. [14] applied against a sample attack to a face recognition system that is widely used for monitoring and access control [14]. The authors implement a new class of attacks: attacks are physically achievable, inconspicuous, and allow an attacker to evade recognition or impersonate another person. A systematic method of automatically generating such an attack is designed and developed by printing a frame with added disturbances. When the image of the frame of the eyeglasses worn by the attacker is provided to the most advanced facial recognition algorithm, the glasses allow her to evade recognition or impersonate another person. In the field of malware detection, the fight against malware at-

Figure 6. Examples of successful impersonation and dodging attacks.

tacks is to modify the characteristics of malware, so that it bypasses the malware recognition model and evades the detection of the model. In the literature [15], the Jacobian-based Saliency Map Approach (**JSMA**) method proposed by Paper not is used to construct the adversarial examples, and it is applied from the continuous and differentiable space transfer to the discrete restricted malware detection, which proves that the adversarial examples attack is the feasibility of the field of malware identification. In our lives, if our face recognition system is destroyed or our software is tampered with, then our privacy and other data will be stolen. So how can we better protect our privacy?

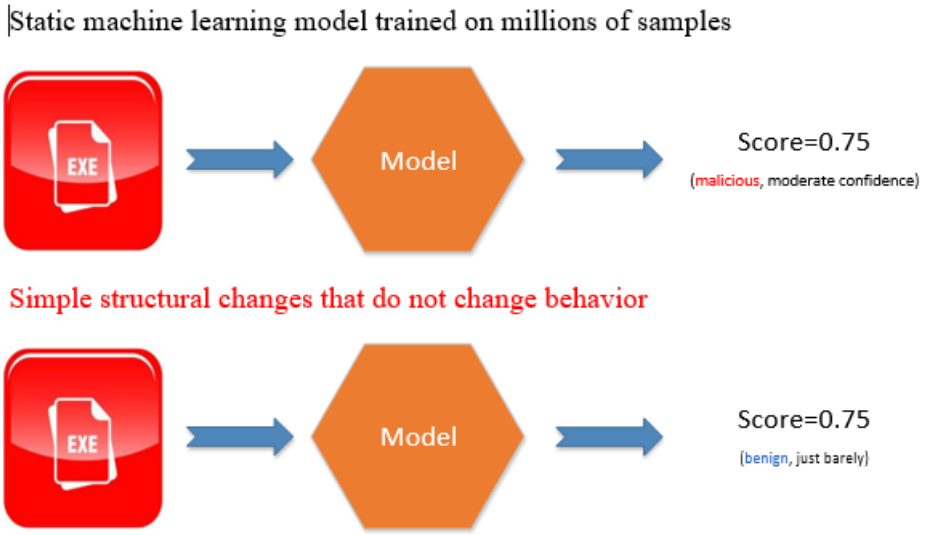


Figure 7. Software system attack model.

2.4 Problem

Why can adversarial examples exist?

How to construct adversarial examples?

How to defend against multiple attacks?

This may seem like three simple questions, but it is difficult to completely solve it. I currently have my own thoughts on the second and third questions, which are the direction of my future research.

3. Method

How to construct the adversarial examples? Normally, we have limited pertur- bations added to the original sample. Our goal is to have machine learning models misclassify our well-constructed confrontation examples, but at the same time, we must satisfy a condition that does not affect human recognition. It is well known that in order for a machine learning model to be misclassified, it is necessary to open the distance between the opposing sample and the original sample. In order to consider the robustness of the algorithm, a filter should be added after the distance between the samples is pulled apart. The purpose of this filter is to make the added disturbances to meet our requirements. (ii) How to defend against multiple attacks? A paper under review by ICLR2019: Better accuracy with quantified privacy: representations learned via reconstructive adversarial network proposes a new depth model RAN that protects our privacy by the adversarial network. They balance between a measure of privacy and another of utility by leveraging adversar- ial learning to find a sweeter tradeoff. We design an encoder that optimizes against the reconstruction error (a measure of privacy), adversarially by a Decoder, and the inference accuracy (a measure of utility) by a Classifier [16]. Particularly, differ- ential privacy is a framework for measuring the privacy guarantees provided by an algorithm. Through the lens of differential privacy, we can design machine learning algorithms that responsibly train models on private data. Through these two facts,

I believe that combining differential privacy with adversarial network can produce some Interesting results.

4. Research Design and Methods

4.1 Aims

Enhancing Security and Privacy by Using Adversarial Networks

Adding Deep Learning Networks and Differential Privacy Algorithms to Intel SGX Enclave to Resist Attacks

4.2 Methods

Enhancing Security and Privacy by Using Adversarial Networks Combining with the above discussion, I designed a model to improve data security and privacy by using antagonistic neural networks.

5. Specific steps are as follows:

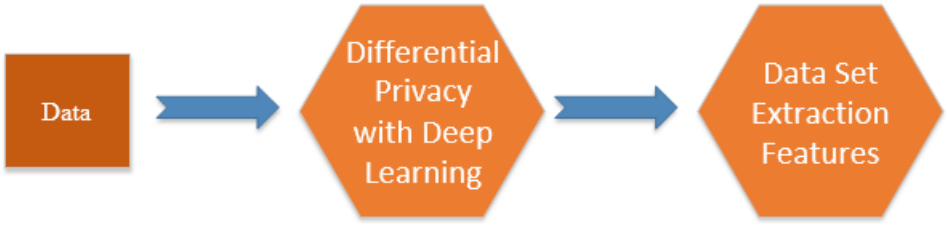


Figure 8. Deep Differential Privacy Protection Model Based on Generative Adver- sarial Network.

According to the size of the potential data set of the input training data set, the query sensitivity, the attacker's maximum attack probability and the upper bound of the privacy budget are calculated;

This step mainly trains the input data set based on the depth differential privacy model. The specific process is as follows: in the deep network parameter optimiza- tion calculation, the differential privacy idea is added to add noise data, and the privacy budget upper bound is set according to step (1). Under this condition, the privacy budget is randomly selected, and then based on the combination of differ- ential privacy and Gaussian distribution, the actual privacy budget of each layer of the deep neural network is calculated in the random gradient descent, and Gaussian noise is added accordingly to minimize the overall privacy budget;

Using the depth differential privacy model in step (2) to train the deep neural network model, and extract the data feature information of the data generated by the privacy protection during the training process. The random noise data is input into the generated confrontation network, and the data feature information generated by the privacy protection is used as a reference [17], and the input random noise data is adjusted so that the simulation data generated by the generator is as close as possible to the privacy protection and the data feature distribution is generated. The generated simulation data is classified to obtain the classification accuracy rate;

Input the original data set into the generative confrontation network to gener- ate analog data that approximates the distribution of data features before privacy protection. The result is compared with the classification accuracy rate of the data generated by the privacy protection obtained in the step (3), and the accuracy rate error threshold is set to ensure that the difference between the classification accu- racy rate of the step (3) and the classification accuracy rate of the step (4) is in advance. Set the threshold range. Otherwise, repeat step (2) to adjust the privacy budget parameters and retrain the depth differential privacy model until the preset threshold condition is met.

Adding Deep Learning Networks and Differential Privacy Algorithms to Intel SGX Enclave to Resist Attacks

(i) Advantages of Intel SGX

Confidentiality and integrity:Guaranteed even in the presence of privileged malware at the OS, BIOS, VMM or SMM layer.

Low learning curve:An OS programming model similar to the parent appli- cation and executed on the main CPU.

Remote authentication and provisioning:The remote part is able to au- thenticate the identity of an application enclave and securely provide keys, credentials and sensitive data to the enclave.

The smallest possible attack surface:The CPU boundary becomes the periphery of the attack surface, and all data, memory, and I/O outside the periphery are encrypted.

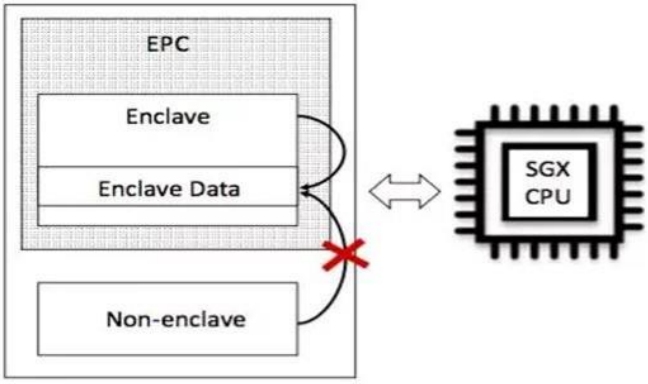


Figure 9. Intel SGX memory distribution and isolation mechanism. Enclave runs on the EPC. Enclave data can only be accessed by enclave itself, and access to any external code will be refused. [18]

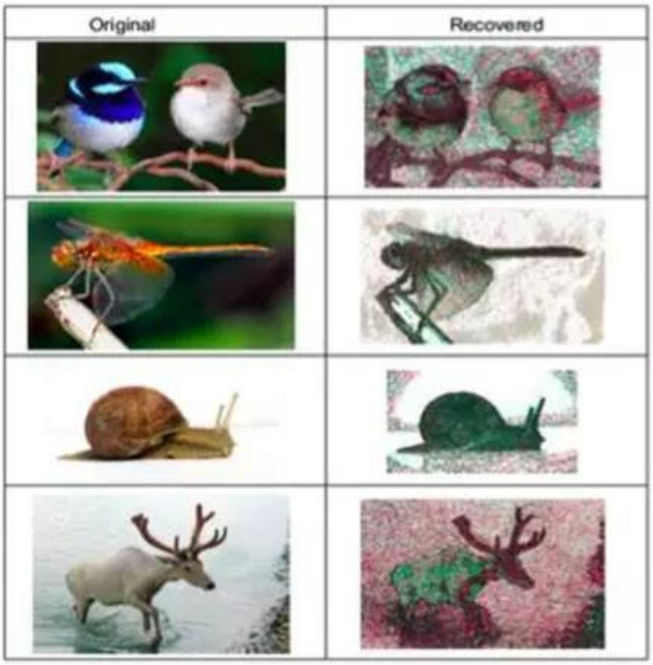


Figure 10. Controlled-channel attack on libjpeg.

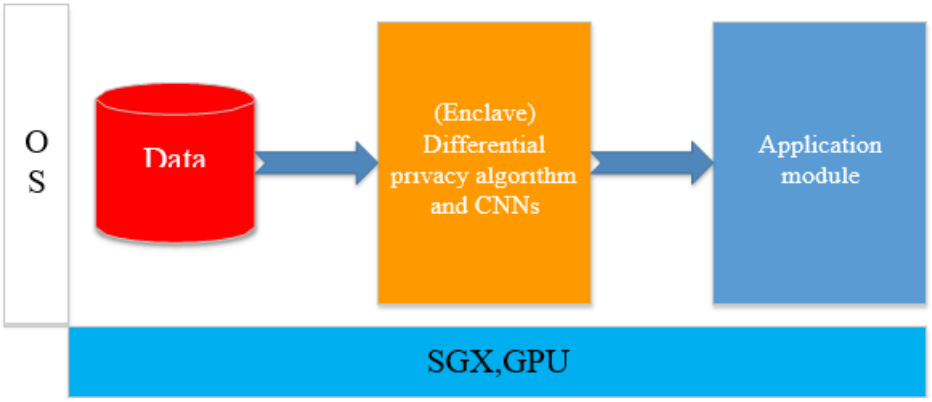


Figure 11. Deep differential privacy learning model.

Although Intel SGX has many advantages, it is still very vulnerable to hacking. For example, Side-Channel attacks often occur.

SGX Side-Channel Attacks The main goal of side-channel attacks is to at- tack the confidentiality of enclave data. The attacker comes from the non-enclave

part, including the application and system software. System software includes priv- ileged software such as OS, Hypervisor, SMM and BIOS. [19]

The main means of side channel attack is to obtain data through the attack surface, derive the control flow and data flow information, and finally obtain the informa- tion of the enclave code and data, such as encryption key, privacy data and so on. Typical attack surfaces include page tables, TLB, Cache, DRAM, and CPU Internal structures.

Page-based attack The earliest SGX side-channel attacks were page-based attacks [19] [20]. This type of access to the enclave page using the page table controls the enclave page to be inaccessible. At this time, any access will trigger a page fault exception, which will distinguish which pages the enclave has accessed. By combining these pieces of information in chronological order, it is possible to defer the state and protected data of the enclave. In some scenarios, such attacks have been able to get a lot of useful information. For example, such a page-based side channel attack can obtain picture information processed by libjpeg. After reduction, it basically reaches the level of human eye recognition.

Source-based defense method The main idea of my defense method is to write a code implementation that can defend against the side channel by modifying the source code. The core idea of the solution is to hide the control flow and the data flow. I designed a deep differential privacy learning model with the following steps:

First, collect data from the user and transfer it to the trusted domain Enclave;

Enclave will be attacked, but use deep learning methods and differential privacy algorithms to hide the data stream so that it can be attacked or reduced.

Transfer the data encrypted to the application module.

6. Conclusion

In the previous sections, I gave an example of recent security and privacy issues with machine learning. From the examples, we can see that these problems have a great impact on our lives and expose us to an insecure environment. If we can solve the security and privacy issues of user data well, we can walk in the forefront of the artificial intelligence era. Here are a few key points of my research proposal:

The CIA model of machine learning is the root cause of security and privacy issues.

Adversarial attack machine learning models can seriously affect the accuracy of model prediction and classification. Adversarial examples often attack our existing security mechanisms. Then, defending against such attacks or learning from such attacks is a future research direction.

The differential privacy algorithm under the system Intel SGX can further improve the security and privacy of user data.

I have designed two learning models, which will be my goal in the future. One is the differential privacy protection model based on Generative Adversarial Network, and the other is based on the Intel SGX deep differential privacy learning model.

References

[1] C. Song, T. Ristenpart, and V. Shmatikov, “Machine learning models that remember too much,” pp. 587–601, 2017.

[2] F. Tramr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing machine learning models via prediction apis,” 2016.

[3] S. Shen, S. Tople, and P. Saxena, “A uror: defending against poisoning attacks in collaborative deep learning systems,” in Conference on Computer Security Applications, 2016, pp. 508–519.

[4] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, “Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing,” Proceedings of the ... USENIX Security Symposium. UNIX Security Symposium, vol. 2014, p. 17, 2014.

[5] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, “Semi- supervised knowledge transfer for deep learning from private training data,” arXiv preprint arXiv:1610.05755, 2016.

[6] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and lfar Er- lingsson, "Scalable private learning with pate," 2018.
[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adver- sarial examples," Computer Science, 2014.
[8] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in ACM Sigsac Conference on Computer and Communications Security, 2015, pp. 1322–1333.
[9] M. Abadi, lfar Erlingsson, I. Goodfellow, H. B. Mcmahan, I. Mironov, N. Paper- not, K. Talwar, and Z. Li, "On the protection of private information in machine learning systems: Two recent approches," in Computer Security Foundations Symposium, 2017, pp. 1–6.
[10] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," IEEE Internet Computing, vol. 15, no. 5, pp. 4–6, 2011.
[11] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," 2017.
[12] P. Bontrager, A. Roy, J. Togelius, and N. Memon, "Deepmasterprint: Finger- print spoofing via latent variable evolution," arXiv preprint arXiv:1705.07386, 2017.
[13] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in IEEE European Symposium on Security and Privacy, 2016, pp. 372–387.
[14] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in ACM Sigsac Conference on Computer and Communications Security, 2016, pp. 1528–1540.
[15] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. Mcdaniel, "Adver- sarial perturbations against deep neural networks for malware classification," 2016.
[16] "Better accuracy with quantified privacy: Representations learned via recon- structive adversarial network."
[17] Y. Liu, J. Peng, and Z. Yu, "Big data platform architecture under the background of financial technology: In the insurance industry as an example," in Proceedings of the 2018 International Conference on Big Data Engineering and Technology, ser. BDET 2018. New York, NY, USA: ACM, 2018, pp. 31–35. [Online]. Available: http://doi.acm.org/10.1145/3297730.3297743
[18] A. Moghimi, "Side-channel attacks on intel sgx: How sgx amplifies the power of cache attack," Masters Theses, 2017.
[19] Y. Fu, E. Bauman, R. Quinonez, and Z. Lin, "Sgx-lapd: Thwarting controlled side channel attacks via enclave verifiable page faults," in International Sym- posium on Research in Attacks, Intrusions, and Defenses, 2017, pp. 357–380.
[20] R. Strackx and F. Piessens, "The heisenberg defense: Proactively defending sgx enclaves against page-table-based side-channel attacks," 2017.