Mitigation of Potential Adversarial Attack in Convolution Neural Network

Sanjida Akter Sharna
Tennessee Technological University
sasharna42@tntech.edu

***Abstract*—Machine learning is hyped these days since of high-value forecasts that can drive way better judgments and keen activities in real-time without human interaction. By involving automatic sets of generic approaches that have superseded traditional statistical techniques, machine learning has transformed the way data extraction and interpretation works. However, there is a risk associated with the benefits, which is known as adversarial machine learning attacks. An adversarial attack can impact the classifier model by disturbing the model because it makes prediction, whereas a security infringement includes providing malevolent information that gets classified as authentic. An attack can endeavor to permit a particular interruption or disruption, and then again to form common disorder. In such way adversarial machine learning attacks can easily penetrate or corrupt a computer network or program and the security of the system will be jeopardized. The ATT&CK threat matrix by MITRE is broadly recognized as a specialist on understanding the behaviors and methods that hacker utilizes against organizations nowadays. In this paper, one of the adversarial machine learning technique and tactic from MITRE threat matrix is implemented to predict and figure out the adversaries a machine learning model faces being attacked. For this FGSM adversarial attack on convolution neural network was applied on MNIST dataset. By studying these patterns of adversary attacks listed in MITRE threat matrix, the aim is to find out the decision boundary between target model and adversary model so that wholesome generalize defense mechanism can be developed. In future the goal is to design defense model from these case studies against the adversarial attacks.**

***Index Terms*—Machine learning, adversarial attacks, security, decision boundary**

I. Introduction

MACHINE learning has great power and promise to make our lives better in a lot of ways, it has advanced radically in such a way that it can surpass human-level performance on a number of tasks.; however, it introduces a new risk that is adversarial machine learning attacks. It attempts to fool models with deceptive data in order to misclassify the actual models which seems normal to human eye. Machine learning models already engaged in practical applications are highly vulnerable to these adversarial machine learning attacks. In most recent times, different types of adversaries based on their threat model leverage these vulnerabilities to compromise a machine learning model or system where adversaries have high incentives. Thus, it is important to lighten up these adversary attacks and vulnerabilities in diverse field where machine learning is used regularly. Machine learning is an exciting and alarming area of research within AI. Equipping machines with the ability to learn certain tasks could be extremely helpful, increase productivity, and help speed up all kinds of activities, from search algorithms to data mining. In recent years, the availability of large data sets combined with improved algorithms and exponential growth in computing power has generated unprecedented interest in machine learning. This collection of statistical methods has already proven adequate to significantly accelerate both basic and applied research in machine learning. We are currently witnessing an explosion of work in the development of machine learning and its application to numerous systems. Widely known areas where machine learning techniques is used are facial recognition, voice recognition, financial services, health care, virtual personal assistants, computer vision, social media services, email spam and malware filtering online customer support, fraud detection, biometrics, network security and many more. Adversarial machine learning attack is a great threat for these real-life applications. Recently, many papers have come up with different adversarial attack category and defense mechanism but none of this paper was able to present proper decision boundary between targeted and adversary machine learning model. To protect all these system from vulnerability we need a comprehensive overview and analysis of the latest research on the subject to understand adversarial machine learning attacks pattern. As academic experts and industry experts still exploring the adversarial attack domain, the potential attacks and mitigation strategies are not comprehensive. By positioning the attacks on machine learning system and to model the behavior of cyber adversaries, it is possible to reflect prominent threat vectors and the rapidly evolving life cycle of machine learning attacks. MITRE has listed the tactics and techniques of possible machine learning adversaries in a threat matrix called ATT&CK framework [17]. This threat matrix was created in collaboration with 12 industrial and academic research groups with the goal of training security analysts to target these new and future threats. The framework is armed with a curated set of vulnerabilities and hostile behaviors that Microsoft and MITRE have proven effective against ML production systems. Top industries and research groups have come up with this thread matrix, it is top notched research topic in adversarial machine learning attacks field. By successfully able to distinguish the boundary decision between the attack models and local models it will be easier to design efficient defense model. The rest of the paper is categorized in four sections. In related work section, the study of existing papers was focused. In problem definition and solution

methodology section, a threat model, problem definition and solution methodology to tackle the problem was discussed. Result and discussion section contains implementation and testing result of the proposed methodology. And the conclusion section describes the finding and future work related to the proposed methodology.

II. Research Overview

A. Related Work

In this paper, a method to understand the representation quality of the neural networks using a novel test based on Zero-Shot Learning, entitled Raw Zero-Shot has been presented where two metrics to assess these learned features to interpret unknown classes. They have also evaluated different adversarial defences to prove that these adversarial defences when applied to a classifier give better representation quality than the vanilla classifier. A link between the representation quality and attack susceptibility is revealed in this paper by verifying that the proposed metrics have a high Pearson Correlation with the adversarial attacks. In this paper, they did not analyze the effect of the adversarial algorithm on each of the class of the dataset. It also opens up new possibilities of using representation quality for both the evaluation (i.e. as a quality assessment) and the development (e.g. as a loss function) of neural networks which is not discussed in this paper. [8]

In this paper, They have investigate the feasibility of deceiving state-of-the-art deep networks-based fingerprint liveness detection schemes by leveraging this property in this paper. Extensive evaluations are made with three existing adversarial methods: FGSM, MI-FGSM, and Deepfool. Investigate the feasibility of deceiving state-of-the-art deep networks-based fingerprint liveness detection schemes by leveraging this property in this paper. Extensive evaluations are made with three existing adversarial methods: FGSM, MI-FGSM, and Deepfool. A small part of adversarial samples possesses transferability on different models, that indicate it is also possible to cause misclassification under black-box scenarios [15]

In this paper, an elaborate survey on recent adversary attacks on machine learning models have been presented with sophisticated analysis of those attacks with the help of real time scenarios, examples and threat model. This paper helps to understand how can an attack strategy work to weaken machine learning models. A descriptive explanation of black box and white box attack along with different types of attacks' working technique and classification are showed in relevant applications.A conventional survey about different types of machine learning attacks and their defense mechanism, didn't come up with any particular solution to these existing gap of defense mechanism. [1]

In this paper, ML-based stealing attack is reviewed in perspectives of three categories of targeted controlled information, including controlled user activities, controlled ML model-related information, and controlled authentication information. They presented a survey of advances of the ML-based stealing attacks between 2014 to 2019 and broadly discussed ML-based stealing attack against the controlled information is generalized to five phases-reconnaissance, data collection, feature engineering, attacking the objective, and evaluation.Here is highlighted the challenges of attacks stealing controlled information and forecast their future directions accordingly.In this paper, future directions matching various limitations are presented but any particular solution to these existing attacks were not discussed. The impact of the size/distribution of training and testing datasets to the success of ML-based stealing attacks is not discussed. [10]

In this paper the first study on the effects of adversarial attacks on deep learning based intrusion detection system in the field of IOT network was presented. A comparison between the performance of two different deep learning models Self-normalizing Neural Network (SNN) and Feed-forward Neural Network (FNN) was carried out based on intrusion detection system.A practical explanation of normalization of input features in deep learning based models and its effectiveness was discussed.Also which model performs better against adversarial attacks was also demonstrated. A potential comparison between SNN and FNN model was elaborately described but no detection and mitigation mechanism were lightened up to overcome these ever-growing open issues on deep learning for intrusion detection in IOT networks. [2]

In this paper it is presented that the industry is drastically vulnerable on adversarial attacks on machine learning system and lacks sufficient tool of protection. Malicious adversarial attacks are discussed in this paper. Here is Highlighted the necessity of ML researcher in industrial arena where machine learning models are broadly used. The gap between designing and implementation under attacks on machine leaning models were discussed.Survey Datasets were not essentially large compared to recent industry. Only malicious attacks on the system was investigated. The broader attacks aspect on machine leaning models in industrial sector were not considered in this study. [3]

In this paper most recent machine learning based IOT attacks has been presented. The main focus in this paper is, they have discussed several machine learning algorithm in three layer i.e. Physical/Perception Layer, Network Layer, Web/Application Layer architecture of IOT system. Possible research challenges such as data security, Infrastructure Problem, Computational Restriction and Exploitation of Algorithms, Privacy Leakage, Real-Time Update Issue were presented and discussed briefly. Moreover a state of the art comprehensive literature review has been presented on ML-based security of IoT.Different research challenges were presented to detect the attacks on machine learning model but any proper solution to mitigate all these attacks as one solution was not addressed [4]

In this paper, adversary attacks on machine learning models have been explored that try to extract information about the training data or to extract the model. The performance of the models on the impact of adversarial attacks has been highlighted. It is articulated here that Model extraction attacks can

dig a big hole for other adversary attacks on machine learning model.Unifying taxonomy of attacks against machine learning privacy, probable causes of privacy leaks in machine learning systems, implementation of the attacks, different defensive measures tested to protect against the different attacks were presented in this paper.Many of the attacks are applicable only under specific sets of assumptions or do not scale to larger training data sets, number of classes, number of participants. better theoretical understanding of privacy leaks in machine learning is needed. How well the attacks would be able to perform on deployed models is not addressed here. which datasets are best suited to evaluate privacy attacks, or constitute the minimum requirement for a successful attack are not approached. [5]

In this paper a new type of adversarial attack was introduced. Here it is said that, researchers also reported several adversarial attacks against ML model in medical image processing to alter the results by adding noises and misclassify a benign mole as malignant with high confidence. According to their attack, an adversary utilizes state-of-the-art adversarial attacks (HopSkipJump, Fast Gradient Method, Carlini Wagner, Decision Tree, Zeroth Order Optimization) against the ML-based SHS to perform both white-box and black-box attacks. The evaluation of this paper illustrates that it can successfully downgrade the accuracy of a ML model in a SHS. FGM had the lowest accuracy drop for all three attack. In summary, HopSkipJump achieved highest accuracy drop for threshold-based attacks. The proposed attack can successfully misclassify the patient's state and manipulate the outcome to a specific state by utilizing only partial knowledge of the ML model. [6]

In this paper crafting adversarial examples to fool machine learning models was presented. Here, it is showed that show that functional threat models can be combined with existing additive (‘p) threat models to generate stronger threat models that allow both small, individual perturbations and large, uniform changes to an input. It is also proved that such combinations encompass perturbations that would not be allowed in either constituent threat model. It is showed that ReColorAdv, which uses a functional threat model on images, is a strong adversarial attack against image classifiers. It can also be combined with other attacks to produce yet more powerful attacks—even after adversarial training—without a significant increase in perceptual distortion which is not discussed in this paper. adversarial attacks could be designed for audio, text, and other domains which is not highlighted in this paper. [7]

In this paper evasion attacks are shown to fail when channels are not considered in designing adversarial perturbations. Also a broadcast adversarial attack is presented by crafting a common adversarial perturbation to simultaneously fool classifiers at different receivers. The major vulnerability of modulation classifiers to over-the-air adversarial attacks is shown. They have presented a defense method to reduce the impact of adversarial perturbations on the classifier performance following the randomized smoothing approach.Did not provide any proper solution to mitigate this attack. [9]

In this paper they examine adversarial attacks on transaction records data and defences from these attacks. They develop black-box adversarial attacks for transaction records data from the financial industry and defences from these attack. They conduct experiments on relevant datasets from the industry and provide an investigation on the effectiveness of such attacks and how one can defend her model from such attack [11]

They introduce the first practical demonstration of an attacker controlling a remotely hosted DNN with no such knowledge. They find that their DNN misclassifies 84.24 percentage of the adversarial examples crafted with their substitute. Presented that this black-box attack strategy is capable of evading defense strategies previously found to make adversarial example crafting harder. Analyzed their attack in the face of defenses that seek to make the (oracle) model robust. [12]

In this paper it is investigated that the viability of adversarial attacks against classifiers in this field. an adversarial test tool, Hydra, was developed to evaluate the impact of adversarial evasion classifier attacks against Neptune with the goal of lowering the detection rate of malicious network traffic. Evaluate the impact of adversarial evasion classifier attacks against Neptune with the goal of lowering the detection rate of malicious network traffic. Multi-objective optimization formulation would target maximum detection accuracy and adversarial robustness of the ML-based NIDS which is not implemented via Hydra tool. To increase the effort required by the attacker, consideration should be made of introducing ensemble methods to the machine learning element such as combining the results of multiple classifiers in decision-making which is not covered in this paper. [13]

In this paper it is presented that DL models that do not consider defensive models against adversarial perturbations remain vulnerable to adversarial attacks. They have presented in detail the AE generation process, implementation of the attack model, and the perturbations of the existing DL-based COVID-19 diagnostic applications. They have used Clarifai REST API models and manipulated SGD to monitor the scores and decrease classifier performance. Here is presented multimoddel AE attacks on diversified COVID-19 diagnostic systems. Only a few machine learning algorithms and their vulnerabilities was tested and efficiency of DL poisoning is not achieved. [14]

From the literature review of above papers, we can sum up that, recently the researcher and scientist are working actively in adversary machine learning field. In many papers, different types of adversary attacks and their defense mechanism and way of mitigation was described. But they couldn't find out any general solution or decision boundary to mitigate these attacks which can defend against all the adversary. To make the machine learning model reliable, it is crying need to design such a system which is able to defend listed attacks in MITRE threat matrix.

III. Research problem definition and solution methodology

A. Threat model

There are numerous categories of adversaries, such as model stealing, model evading, data poisoning; each with a diverse objective and suspicion of the attacker's information. Be that as it may, in common the overarching objective is to include the slightest perturbation to the input information to cause the specified misclassification. Now as days, machine learning as a service is provided by cloud based platform. Machine learning as a service (MLaaS) is a bunch of cloud computing administrations that give end users machine learning items and arrangements to information changes, demonstrate preparing and eventually, prescient analytics. There are a few sorts of presumptions of the attacker's information, two of which are: white-box and black-box. A white-box attack expects the attacker has full information and get to the show, counting design, inputs, yields, and weights. A black-box attacker expects the assailant as it were has get to to the inputs and yields of the demonstrate, and knows nothing approximately the basic engineering or weights. There are too a few sorts of objectives, counting misclassification and source/target misclassification. If the attacker only wants to misclassify the class, then it doesn't matter what the output of new class is. If the attacker wants targeted misclassification, then the original classification must be altered with attacker's desired one. In this work, I assume adversary attacks may occur on commercial machine learning service on cloud platform as MLasS and on regular machine learning model.

B. Formal definition of the problem

The industry is insufficiently prepared. In a survey of 28 organizations that included both small and large organizations, 25 organizations did not know how to protect their machine learning systems. Moreover, the field of machine learning is spreading day by day as machine learning as a service has also started gaining recognition as real life solution [18]. There are many survey papers and defense mechanism on different types of adversarial attacks. In recent days MITRE has introduced

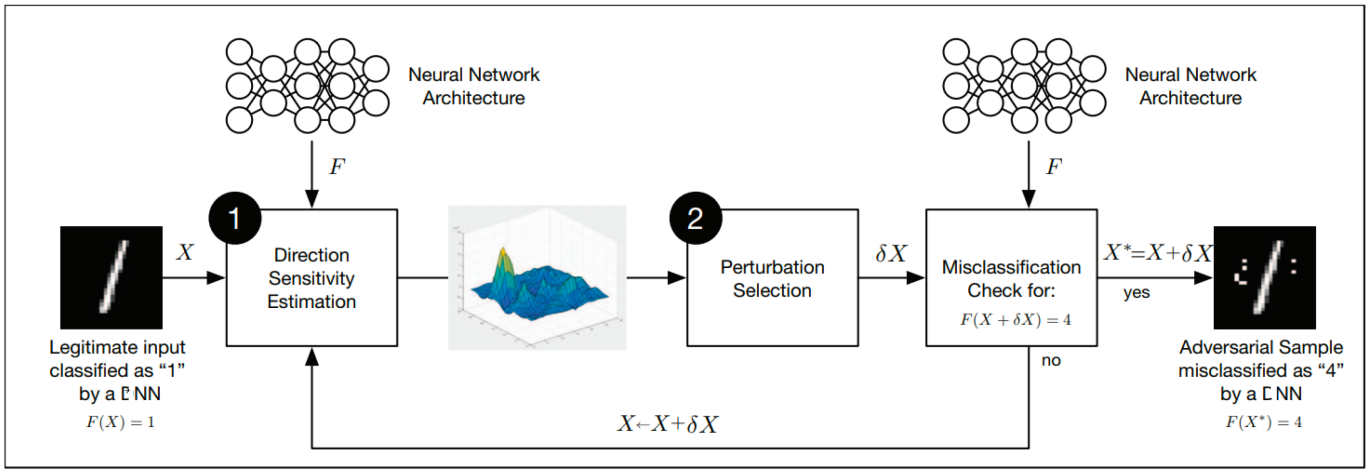

Fig. 1: Overview of the adversary attack method

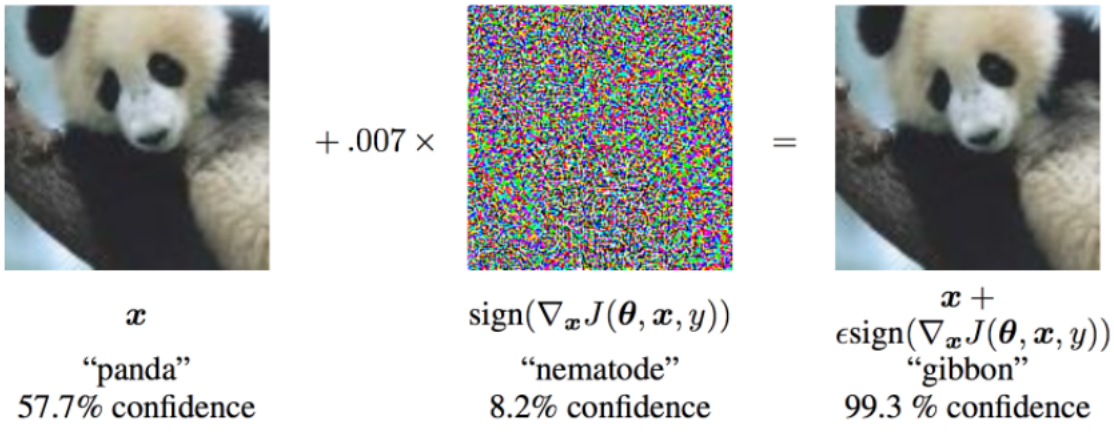

Fig. 2: A demonstration of fast gradient sign method adversary

a threat matrix listing all the possible types of attacks and tactics on machine learning model. But we don't have any generalize solution for all these types of attacks. A machine leaning model which is used for commercial purpose can go through several types of attacks by any attacker.

Eventually an attacker can be successful, if he tries persistently to break any machine learning model using relevant types attack methods with minimalistic domain knowledge about the input, output or architecture of the ML model. Till the date, there is no generalized defense mechanism for all types of attack listed on the threat matrix. However, the application of machine learning models in various fields are growing unimaginably without the protection measurements. Considering all these issues, if we can observe the pattern and consequences of different type of attacks listed on the threat matrix and can reach to a solution to find out the decision boundary between target model and adversary model then we can design a generalized defense mechanism for machine leaning models. We can achieve highest peak of reliability for using machine learning model by implementing my proposed solution. At this point, adversary attackers won't be able to affect machine learning model anymore to misclassify any class or steal the model.

C. Solution methodology

MITRE threat matrix has listed all the possible adversary attacks. I am going to demonstrate one of the threat listed there. One of the first and most popular adversarial attacks up until the present time is referred to as the Fast Gradient Sign Attack (FGSM). The attack is strikingly capable, and however instinctive. It is planned to assault neural systems by leveraging the way they learn, gradient. The concept is simple: instead than altering weights based on backpropagated gradients to minimize loss, the attack alters the input data to increase loss based on the same backpropagated gradients. In other words, the attack maximizes the loss by using the gradient of the loss in relation to the input data and then adjusting the input data.

In this famous panda example [16] from Figure:2, we can see; x is the original input image successfully categorized as a "panda," y is the ground truth label for x, represents the model parameters, and J(θ,x,y) is the loss used to train the network, as shown in the diagram. The attack calculates ∇xJ(θ,x,y) by backpropagating the gradient to the input data. The input data is then adjusted by a small step (or .007 in the picture) in the direction (i.e. sign(∇ xJ(,x,y))) that maximizes the loss. The target network misclassifies the resulting perturbed picture, x, as a "gibbon" when it is still clearly a "panda." some adversary examples are needed to attack the train model. The function that creates perturbed images is as

perturbed image
= $image + \epsilon * data_grad$
= $x + \epsilon * sign(\nabla xJ(\theta, x, y))$

Fig. 3: Visualization of MNIST dataset

After the FGSM attack, the train model will not be able to classify the data properly. It indicates the testing model accuracy decreases with increasing noise (value of epsilon). This is because larger epsilons mean we take a larger step in the direction that will maximize the loss. So,if we could figure out the FGSM attacks pattern and decision making criteria, gradually observing these attacks, we could figure out the boundary decision line between target model and adversary model. which is essential to provide a generalized defense mechanism.

IV. Results and Discussion

To demonstrate this attack I have used MNIST dataset which is a database of handwritten digits, has a training set of 60,000 examples, and a test set of 10,000 examples. A neural network model was created and trained to classify the MNIST dataset digits.

The name of our model is simple convolution network which is trained with this dataset.Before attack the accuracy of the model was 0.9894 which is pretty good. Now we choose a random image digit from the dataset to figure out whether after the occurrence of attack the trained model could classify that same digit or not.

So here we can see, the model has randomly chosen digit and the train model is able to classify the level properly.

After FGSM attack we can demonstrate from figures below, the train model was unable to classify the digit level, as the epsilon increases, the accuracy of the model was decreasing for both targeted and non targeted FGSM attack.

After this attack, the training model was again trained with newly generated adversary training and testing dataset. This time model was able to classify the adversary dataset. When

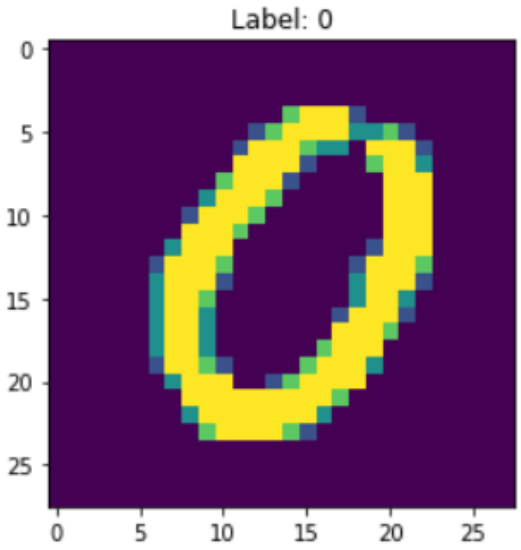

Fig. 4: visualization of proper classification before attack

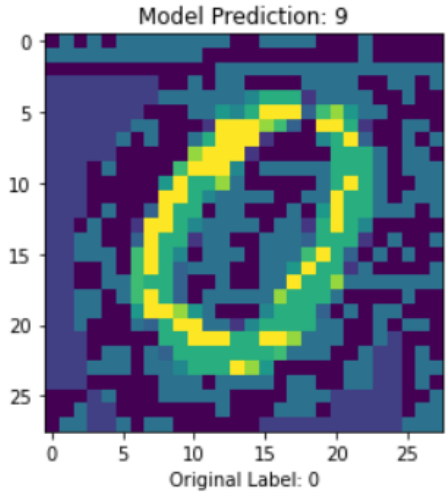

Fig. 5: visualization of misclassification after non targeted attack for epsilon 0.3

the accuracy was tested this time, the accuracy increased again which is 0.8262.So it indicates that now the model can classify adversary examples as well and the model is impaired.

so we from this analysis we can say that, if we will be able to predict a machine learning model behavior before and after the attack occur then we can reach to a conclusion about the attack behavior. Gradually by observing all the adversary attacks listed on threat matrix by MITRE we can figure out a decision boundary between the targeted model and adversary model. Nowadays machine learning as service is popular platform for reliable machine learning solution. Here an attacker can steal machine learning model even with blackbox attack with minimum query [18]. So if we can successfully get to

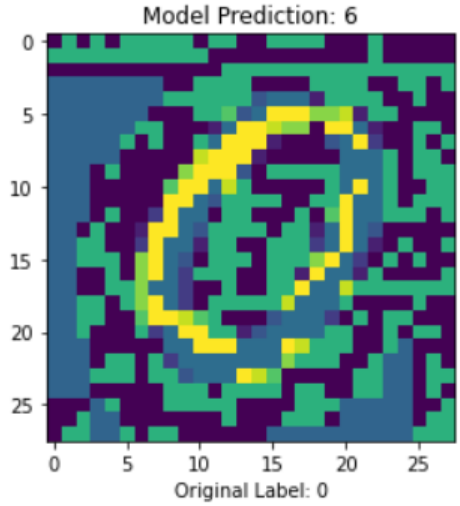

Fig. 6: visualization of misclassification after targeted attack for epsilon 0.9

the bottom of decision boundary of adversary model and local model then we can design mitigating techniques for all the possible attacks listed in the threat matrix.

V. Conclusion

Machine Learning models and techniques are the greatest boons to keep up with today's advanced tech world. Machine learning has great power and promise to make our lives better in a lot of ways, it has advanced radically in such a way that it can surpass human-level performance on a number of tasks.; however, it introduces a new risk that is adversarial machine learning attacks. Though MITRE has listed all types of techniques and tactics in a framework but any generalized solution has not been proposed yet. In this paper we have demonstrate the FGSM attacks on a trained model to observe the pattern of the attack on machine learning model and the model sustainability. Nowadays machine learning as a service is also active field of research for adversarial machine learning attacks. In this cloud platform, ML models can be stolen with minimal domain knowledge of the input, output dataset and architecture of the model. So machine learning models are making our life easy but are not reliable at all as there is no generalized mitigation techniques. In future the aim is to observe all these adversary attacks to figure out the decision boundary between targeted model and trained model. As a result, it will be possible to design efficient defense mechanism to overcome all these listed attacks in the threat matrix.

References

[1] Chakraborty, Anirban, et al. "A survey on adversarial attacks and defences." CAAI Transactions on Intelligence Technology 6.1 (2021): 25-45.

[2] Ibitoye, Olakunle, Omair Shafiq, and Ashraf Matrawy. "Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks." 2019 IEEE Global Communications Conference (GLOBECOM). IEEE, 2019.

[3] Kumar, Ram Shankar Siva, et al. "Adversarial machine learning-industry perspectives." 2020 IEEE Security and Privacy Workshops (SPW). IEEE, 2020.
[4] Tahsien, Syeda Manjia, Hadis Karimipour, and Petros Spachos. "Machine learning based solutions for security of Internet of Things (IoT): A survey." Journal of Network and Computer Applications 161 (2020): 102630.
[5] Rigaki, Maria, and Sebastian Garcia. "A survey of privacy attacks in machine learning." arXiv preprint arXiv:2007.07646 (2020).
[6] Newaz, AKM Iqtidar, et al. "Adversarial attacks to machine learning-based smart healthcare systems." GLOBECOM 2020-2020 IEEE Global Communications Conference. IEEE, 2020.
[7] Laidlaw, Cassidy, and Soheil Feizi. "Functional adversarial attacks." arXiv preprint arXiv:1906.00001 (2019).
[8] Kotyan, Shashank, Danilo Vasconcellos Vargas, and Moe Matsuki. "Representation Quality Of Neural Networks Links To Adversarial Attacks and Defences." arXiv preprint arXiv:1906.06627 (2019).
[9] Kim, Brian, et al. "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers." arXiv preprint arXiv:2005.05321 (2020).
[10] Miao, Yuantian, et al. "Machine Learning Based Cyber Attacks Targeting on Controlled Information: A Survey." arXiv preprint arXiv:2102.07969 (2021).
[11] Fursov, Ivan, et al. "Adversarial Attacks on Deep Models for Financial Transaction Records." arXiv preprint arXiv:2106.08361 (2021).
[12] Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." Proceedings of the 2017 ACM on Asia conference on computer and communications security. 2017.
[13] Aiken, James, and Sandra Scott-Hayward. "Investigating adversarial attacks against network intrusion detection systems in sdns." 2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN). IEEE, 2019.
[14] Rahman, Abdur, et al. "Adversarial examples–security threats to COVID-19 deep learning systems in medical IoT devices." IEEE Internet of Things Journal (2020).
[15] Fei, Jianwei, et al. "Adversarial attacks on fingerprint liveness detection." EURASIP Journal on Image and Video Processing 2020.1 (2020): 1-11.
[16] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
[17] https://github.com/mitre/advmlthreatmatrix
[18] Yu, Honggang, et al. "CloudLeak: Large-Scale Deep Learning Models Stealing Through Adversarial Examples." NDSS. 2020.

My source code is available in [1]

[1] https://gitlab.csc.tntech.edu/ssharna42