Retraction

Retraction: Exploring adversarial attacks against malware classifiers in the backdoor poisoning attack (IOP Conf. Ser.: Mater. Sci. Eng. 1022 012037)

Gopaldinne Sanjeev Reddy¹ and Sripada Manasa Lakshmi².

¹ School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India
² School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

Published 03 March 2021

This article has been retracted by IOP Publishing in light of clear evidence that substantial material in the article, including text and figures, was taken without reference or permission from [1]. As a member of the Committee for Publication Ethics (COPE), IOP Publishing has investigated in accordance with COPE guidelines and it was agreed the article should be retracted.

[1] Severi G, Meyer J, Coull S and Oprea A, 2020, https://arxiv.org/abs/2003.01031v1.

Retraction published: 03 March 2021
Exploring adversarial attacks against malware classifiers in the backdoor poisoning attack

Gopaldinne Sanjeev Reddy¹, Sripada Manasa Lakshmi²

¹School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India.
²School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India.

E-mail: ¹sanjeevreddy.gopaldinne@gmail.com, ²manasa.23777@lpu.co.in

Abstract. Machine learning is becoming the latest trend by making growing use of it in various fields of application. Along with that, the invasions designed to evade these systems has also evolved apparently. The adversarial attacks are becoming the major threat for developing threats to these models. Current training pipelines for the classification of malware based on machine learning (ML) depend on crowdsourced threat feeds which expose a normal point of injection. We research the vulnerability of ML malware classifications to backdoor poisoning attack for the first time, concentrating explicitly on stimulating “clean label” attacks where attackers donate monitor the mechanism of sample classification. In this paper, we reviewed various attacks based on machine learning models and their working strategies. We also discussed the threat models and backdoor attacks on malware classifiers.

Keywords: Machine Learning (ML), SHapley Additive exPlanations (SHAP), LightGBM, EmberNN.

1. Introduction

The technologies associated with Machine Learning (ML) are commonly used for different applications. As we all know, one of the most recent trending developments can be pointed out as the ML. Here we are emphasizing the machines to think, work and do similar to that of the humans. The only role is to make life much easier by making the machine smarter for the job. In this, we usually try to make the computer correct its wrongdoings by removing a programmer’s requirement to correct his pre-coded scripts by itself. Consider an example that a human being normally makes a few mistakes. But if he learns it, he realizes that it’s a mistake and tries not to repeat the mistake. Let us consider another example, where a child falls down by slipping through a wet floor. He falls down once, and he gets injured. He would know that if he walks on the wet floor he would slip. From his own acts, he learned this fact. In this, his own behaviour can be compared to those of the feedback provided when a certain machine does something wrong.

As these new technologies emerge, hackers are attempting to evade these technologies as well. One needs to have a great and in-depth understanding of the function of these devices to consciously destroy a device. Before the cracker attempts to exploit it, even the developers try to convince their implementations to improve its protection. It’s really a great idea to do that type of strengthening of the systems. There are various kinds of attacks are present. Some are commonly known to the public and some are unknown. By seeing all these kinds of situations, attacks can be performed in two ways. The first is when the machine is trained to do such things. The other maybe when the machine is trying to respond to the lessons taught. Let us consider an example for each attack if a student in the class is taught incorrect formulas and he asked to attempt the exam. He would get the wrong answers in the
exam and he fails. If we train with the right way he would pass the exam. Let us consider another case in that if we taught the student with the right formulas and asked him to give the exam. But, in exam we if the question is wrong he definitely make a mistake in exam answer would be wrong. There are certain conditions that cannot be solved with formulas. To hijack these systems, there are also several other kinds of attacks. But the primary objective of all these techniques could be to cause the ML model built to fail or malfunction.

Adversarial attacks are one of the most prominent and popular attacks frequently performed on these types of systems. By the word adversary, we can sense that there are certain kinds of disputes. We should suggest relating this concept to the ML models that certain attacks are designed to make the disputes into the ability to think of the structures. We are there experiencing a parallel growth in the consideration provided through the security study communal to argumentative attacks beside of malicious software (malware) finding copies with the period of endpoint protection industry near adopting machine learning (ML) based gears. Recently, effective elusion attacks beside commercial of ML-based anti-virus methods prepared the broadcast, also straight public struggles were prepared to trial the protection of exposed cause models. However, the huge majority of scientists attention for the moment has stayed taking place evasion attacks [3,4,5], where the aggressor’s aim remains to modify the figures point (now this case the binary malware) at the inferential period to cause a misclassification. On the additional hand, our research attentions taking place poisoning attacks which attempt to stimulate the process of working out in ML. An especially interesting type of the poisoning attacks remains backdoor poisoning, by where the challenger seats a sensibly selected watermark in feature space by such a way that the object model studies to equate its existence through by class of preference of the attacker. Backdoor attacks were exposed to be very successful after useful to the computer image models [6], lacking need of a huge number of poisoned samples, however, their applicability to the malware classification area stayed unreliable. A simple requirement on behalf of such an attack to remain conducted is the adversary’s capacity to exploit a division of victim model training information. We believe that many security vendors current training pipeline offers an outstandingly normal injection point for a resourceful attacker. Indeed, security agencies also rely on crowd-sourced threat feeds [10, 11] to deliver them thru a wide, complex and stream of information to the guide by their classifiers. This is mainly due to the pure amount of branded binaries were desired by ML models to reach satisfactory detection output (ten - hundreds of the millions), by which creates it difficult for the sellers to trust completely on the in-house information sources.

Figure1. Overview of the attack on static machine learning malware classifier training pipeline.

As these risk feeds are designed largely round user-submitted the binaries, they deliver a model vector to poison attacks. The absence of the controller over the tags allocated to the models is a crucial
change between computer image domain and malware classificatory. In reality, the labels used in crowd-sourced threat feeds are created automatically through the application of some malware finding engines. But it would be incredibly difficult for an aggressor to monitor the performance of such an automated system committee. Of this reason, we believe that the opponent has no power over the sample labels. Therefore in this paper, we research the vulnerability of ML-based malware classifiers to clean-label backdoor attacks [12, 13] for the first time.

2. Literature Review

Here, we also provide a short description of computer security argumentative machine learning functional.

Evasion: Biggio et al. [3] suggested a gradient-based SVM assault on the malicious PDF detectors through inserting keywords that are highly associated with benign files. Xu et al. [17] PDF files were also used as a guide for their analysis, but a generalizable approach was proposed to identify evasive variants using genetic programming in a black box context. Dang et al. [18] Using a hill-climbing method, implemented EvadeHC, executing of a black-box attack by morphing the malicious case. Similarly based on the black-box mutation is Anderson et al. [19] research by which trains the Reinforcement Learning agents to the product features that preserve binary mutations. There is a common line of study by Suciu et al. [22], Who generates adversarial models to fool Neural System classifications. BY the very first focused systems focused on the binary value feature derived after the files whereas the second collection concentrated on CNNs directly functioning on the raw executable binary data, such as Malconv [23].

Poisoning: Biggio et al. [3] were one of the first to expose the issue of ML poisoning attacks. Successive effort, by targeting to the Malheur behavioral clustering tool [25], verified the importance of the poisoning attacks in computer security. Later Xiao et al. research [7] show that feature selection techniques, such as LASSO, ridge regression as well as elastic net, were sensitive to lesser sizes of poisons. Attacks on the accessibility of gradient-based poisoning were shown toward regression [20] and neural systems, and the modifiability for these attacks was demonstrated by Demontis. Suciu et al. [21] offered a structure to define attacker structures in the poisoning space, and created StingRay, a methodology for the multi-model goal poisoning attack.

Backdoor Attacks: Introduced Backdoor Attacks by the Guet al. in BadNets [24], Determining vulnerability in the supply chain of the current as-a-service machine learning the pipelines. Liu et al. [6] Explored incorporating Trojan triggers in the image recognition Neural Systems by partially retraining the models, without needing access to the original training data. Later by Shafahi et al. [12] as well as Turner et al. [13] further developed with the advancement of clean label strategies against current attacks. No previous research has focused, to the best of our knowledge, on backdoor attacks directly targeting malicious software classifiers. In addition, we are unaware of any effort to insert backdoors into the Gradient Boosting models.

The taxonomy of the different attacks by Artificial Intelligence (AI) and Adversarial Machine Learning (AML) was analysed and the various forms of counterfeiting the analysed attacks were suggested by Kevin J. Burns et al. [14]. The authors gave a brief perspective on the current defects in them and described a variety of ways to hold them at bay. These AMLs are affected by a huge threat. A major danger could be the hierarchical effects of security vulnerabilities through the results of training. In order to manage the misshaping at ground level, this paper provides the basic concepts and implementation of over the countermeasures. Battista Biggio and Fabio Rol [15] in 2018, the numerous foundations made in the field of Adversarial exploits to remove them were surveyed. These attacks are not new to these types of framework, it is very common to exploit the loopholes at the learning and the experimenting stages. To mitigate them effectively, research from the past decades are being carried out. Some of the widely recognized AML attacks carried out in ML systems have been described by the authors and followed by up to date for everting techniques.
3. Background

The several possible approaches to the issue of automated malicious software detection can be loosely divided into two key groups. In a virtualized environment, dynamic analytics systems execute binary files and record the sample actions finding indicators of malicious activities [8]. By the additional hand, fixed analysers method executable records without successively them, removing straight from the binary functions used for classification. Although both methods have both positive and negative aspects, due to strict period constraints below which they typically function, many endpoint security solutions prefer to incorporate static analysers. This another group was divided into two extra subcategories with the change to Machine Learning (ML)-based classifiers: feature-based detectors [16], also the raw-binary analysers [23]. We emphasis on primary of these groups, as feature-based classifiers currently outstrip the struggle, harvest predictions that are calmer to interpret, need fewer computational resources, plus generally have larger dispersion in commercial explanations. EMBER [16] is an openly released dataset on behalf of classification of feature-based malware. It comprises 2351 dynamically removed features from the Windows PE files that procedure a wide corpus of malicious also files. EMBER is widely considered of standard for the models of malware research, and it is aggressively sponsored through maintainers. It is attended in an incline boosting choice tree ideal that not only has great standard performance, then also offers an impeccable target for attack due to the extraordinary use of collaborative methods in the safety manufacturing.

Attacks at Backdoor Poisoning. Adversarial attacks on ML prototypes can be largely categorized into two categories: (i) Evasion attacks, in which the opponent's box is to change the test sample by inserting a minor disturbance such that the perfect is fooled into allocating the wrong class; (ii) Poisoning attacks in which the attacker can alter the preparation facts, moreover by inserting new information points or by changing current ones, resulting in misclassification on the time of inference. In the context of computer vision[9], the former class has been extensively explored, and prior study efforts now this is direction consume also by examined the applicability of such that methods to the malware classification[3,22]. In this research, we focus on poisoning attacks that have gained less attention in the safety context until now. Destroying availability attacks reduce the overall correctness of the model, but we are mainly interested in the poisoning attacks at the backdoor. The adversary's objective here stands for insert a backdoor (or watermark) of pattern into the model's learned symbol, which be able manipulated at an inferential time to influence the outcomes of the sorting on backdoor topics. Backdoor attacks remained implemented for image recognition [24] in the context of the neural networks. Clean-label backdoor attacks [13], while retaining their original label, limit the attacker to watermark data points.

4. Discussing various attacks and its working strategies

Nowadays Machine Learning (ML) based technologies are being used widely in various applications. Due to its abundant use of these technologies malware attack has been increased.

4.1 Eluding. It can be considered the most severe attack enforced on certain types of systems. It is usually achieved when the computer attempts to retrieve the input from its embedded mechanisms and reference it from them. When someone attempts to assault a system by presenting incorrect data for processing, it simply directs the system to yield with irrelevant data or may even fail at those times. Consider these with an example, the lecturer tries to fail the student by giving the wrong question to him. The student keeps his maximum effort to give the best answer but finally, he fails because the question is wrong; finally the results will end up to wrong at the end. Suppose there is an ML device which is used by the surveillance camera to recognize a specific person's information by assisting the identification of face pattern. If the person has applied any make-up or other external features to his or her own face [11]. Then it may lead to a mismatch of the persons matching pattern stored during the training period with that of the current facial properties inferred. Someone who knows the exact mechanism behind these systems can manipulate the facial features to guide the system to give the identity of the desired individual as the output. This can be a real danger to the device used for such applications. Also, your or our names can be instigated as thieves for a future burglary [16].
4.2 Contamination: Nelson et al. [16] has mentioned the first-ever Contaminating assault incident. For which he provided for a technique of decapitating the irrelevant mail detection functional procedures. Generally, we mark unwanted mail senders in the spam list to reduce the bulk mails that are sent to us, which makes us awkward. So, the authors have illustrated an efficient method to outperform these monitoring systems. The key hypothesis behind this undesirable condition is that the knowledge provided to the machine to interpret is contamination. One of the attacks may be the modification of hierarchical labels. What really happens in this is that it can modify the form of naming according to the classes predefined. This leads to unfortunate incidents with the performed task assuming other unrelated factors. Consider an example that we have to design a system that is used to ring the alarm when the sunlight is detected. The activity associated with sunlight sign detection is to ring the sound. Especially if the attacker succeeds in obliterating the sunshine mark into darkness. Then it can activate the alarm at night. Another potential assault is the implementation of a certain pernicious script to confuse the techniques in the data processing. The attacker will exploit the device to produce the desired data output in this way.

4.3 Trojan Based: Nowadays most of the models utilized in the market are not built from the start [1]. It has been built by using the previously implemented models. Some of the small-scale organizations used to download the relevant software and try to modify it. In this way, the attackers can inject the malicious code into the software. It looks legitimate but we must download the software from the genuine website. There is some most common type of Trojan malware includes Bookdoor Trojan, DDoS attack Trojan, Ransom Trojan, Rootkit Trojan, SMS Trojan and Remote Access Trojan. These are some common attacks. It can be very critical because they are very difficult to detect their activities since the model’s activities can appear very benign. It is too difficult to differentiate between original and malware embedded models. So, it is compulsory to review the code before downloading.

4.4 Privacy Breaching: Privacy breaching occurs when someone accessing your data without your permission [2]. It starts with a security breach penetrating a protected computer network and ends with the exposure or theft of data. That data includes our personal details like name, mobile number, address, bank details, credit card information etc. A breach of privacy will increase your risk of identity theft. That's when someone, like you Social Security number or bank account information, uses your personal details to commit crimes on your behalf. Use strong passwords, in regular basics monitor your financial accounts, Secure your smartphone, Check website security, Use secure software, Try to avoid sharing everything on social media.

5. Threat model and Problem statement

In a standard working out pipeline to an ML malware-classification, was shown in Fig.1, typically begins through acquiring large volumes for labelled binaries as of Threat Intelligence platforms of third parties. Such platforms permit users (with attackers) to upload samples that are branded on binary files through the running pools of current anti-virus (AV) to the engine. In the classified data will be obtained from the websites by businesses. However, the screening manner of the received flow is finished remarkably egregious by together the sheer quantity also complicated the intrinsic trouble of the task which requires staff and tooling. Also, by this outsourced information can be joint by tiny groups of branded, validated binary records toward constructing branded data usual for training. The training method involves the feature extracting phase (in these cases of static examination of PE records), followed through the exercise procedure with the ML algorithm. The certified ML malware classifiers are then distributed in the wild and used to produce a code, malicious, or benign on new binary files. Therefore, in the environment we are analysing crowdsourced information arises with a usual of labels defined through another AV analysers, which are not below the attacker's straight control. This disorder makes de-facto need for the clean-label backdoor strategy because label-flipping would mean an adversarial switch of the classification process. Furthermore, the opponent's objective is to obtain backdoor kind binaries, which would be dispersed over these labelling schemes, also poison the malware classifier exercise sets down the path. If replicas are organized, the opponent will just insert the same
watermark in malicious binaries beforehand launching them, thereby ensuring that the latest malware campaign avoids the discovery of the backdoor classifiers.

Motivated through broad application of the static malware detection in the manufacturing, the first objective is to examine the vulnerability by such vulnerability sensors to the backdoor poisoning attacks. A backdoor is a special combination of features and selected values, often called a watermark or trigger. At the training, the perfect is persuaded to equate the backdoor with the targeted of class which resulted in inferential misclassification of backdoor opinions. The LightGBM incline boosting model free with EMBER is our first goal. As our attention was in building a practice that can be practical to various model buildings usually used in the feature-based malware classifiers, also we obvious to assess our strategy beside such the second perfect that we intended, a fixed-forward neural net. This was a system, which that they call EmberNN, maintains decent classified accuracy then represents the buildings that are often used in works and applications in the worldwide. Finally, we interested at testing defence response-measures beside our backdoor attacks suite also exploring numerous trade-offs between the efficiency of the attacks, their effects on spotless binaries, their realistic viability then the detectability of the attacks.

5.1 Threat model: By demonstrating an adversary who influences crowd-sourced risk sources to disperse their contaminated examples, we board the natural feasible point in exercise pipeline of the malware classifiers. We’ll follow the course outlined by characterizing such an attacker's objectives and capabilities. Greatest backdoor attacker works adopt the Bad Nets model threat, by which described: In an "Outsourced Training Threat" in which the opponent has filled control over the exercise process, and the end operator is only permitted to test the training by means of a hold-out authentication dataset; also (ii) a "Transfer Learning Threat" by which the operator downloads and fine-tunes a pre-trained perfect. We claim that the direct implementation of this threat model is difficult in the context we are investigating. Safety companies are typically highly risk averse as well as tend toward either conduct of in-house training or outsource of the equipment at maximum while retaining complete control of the software stack used during training. This protects against "outsourced training attacks" to a large degree and makes "transfer learning attacks" impossible. Similarly, we do not think the attack model applies in this case from where the intruder partly retrains the model.

5.2 Goals and Challenges: To precis, our concept goals were as follows: 1. Insert in the training a lesser number of watermarked poisoning opinions with the box of improving the identification of backdoor malware variants throughout challenging; 2. Develop spotless label attacks by which apply as the watermark to kind executable records; 3. Build model-agnostic attacks that they refer to various ML models like assemblies and neural systems; 4. Attacks are possible to install through changing current Windows PE files in practice; 5. Consider defensive mitigations against such attacks at the backdoor.

A variety of technological difficulties occur in developing attacks with such properties. First, it becomes difficult to select the exact characteristics and values for the formation of the watermark given the intention to preserve stealthiest of the attack. The model-agnostic necessity eliminates approaches that they are unique to sure ML algorithms, such as the use of tree ensemble models computed feature importance. It is tough to achieve realistic feasibility for attacks that change PE records while maintaining their functionality. Finally, the waterline will apply to all good ware as well as malware records without allowing the aggressor to re-develop from scratch of all plans.

6. Backdoor attacks on malware classifiers

The opponent leverages power over the (subset of) features in a backdoor poisoning attack to cause misclassifications of outstanding to the existence of the watermarked standards in the certain function of dimensions. Automatically, by the attack generates a region of thickness inside the watermark-containing function subspace, and classifier changes its decision of boundary to match the watermarked sample density. When changing the decision boundary, the watermarked of points battle against the power of nearby non-watermarked of opinions, as well as function dimensions which the attacker was does not regulate. However, an even though the attacker only manages a fairly lesser subspace, if the number of watermarked points was extremely large, the nearby data points are too thin,
or the watermark vacates an especially poor area of decision boundary where the confidence of the ideal
is low, they can still affect the decision boundary.

The intruder determines the number of the attack points by the amount of the watermarked data
sets that they insert, also by carefully choosing the watermarked function dimensions and their values,
the region of the decision boundary they control. Therefore nearby are two natural approaches to
successfully creating backdoor watermarks: 1. Check for the areas of weak support along the
determination boundaries where current weak evidence will be overcome by a watermark or 2. Subvert
areas already highly focused on good ware, such that the watermarked subspace intensity overcomes the
signals from many other close objects. By these approaches in the mind, the query is: how can we
increase knowledge in general, model-agnostic way about the decision boundary of a model? We
suggest the system explanation strategies, such as SHapley Additive exPlanations (SHAP), are the
normal way of interpreting decision boundary orientation. The SHAP values denote factors that push
the system towards malware decisions, while the negative SHAP values denote factors that push the
system towards a good ware decision. For such a given set, the number of SHAP values among all
functions equals the logit function of the output of the system (which will be converted to the probability
using the logistic convert).

One understanding of SHAP values is that estimate the decision of boundaries trust along of
function axis, which provides us with the model-agnostic process required to enforce the two intuitive
plans above. This was when we need small-confidence decisions boundaries regions, we should check
for functions that were near-zero to SHAP values, though strong good ware-oriented functions could be
identified through searching for functions with negative assistances. Summing up values from separate
sample across the function columns would before providing an estimate of overall orientation inside the
dataset to that element.

6.1 Building blocks: To execute a watermark, the attacker requires two components: feature selectors,
as well as value selectors. Feature choice narrows the watermark of the attacker to the
subspace that
meets some desirable characteristics, whereas the value of selection selects the unique point within the
space. Several instantiations of such components are possible based on the approach chosen by the
attacker. Here we should describe a SHA
P-based function and value application process that we used in
our attacks, but other instantiations could also be necessary (perhaps to support alternative attack
strategies).

Feature Selection: The main concept for all techniques of attack against backdoor poisoning was to
select features with a great degree of control over the choices taken by the developer. One principle that
normally captures this concept is the importance of features. For example, in the tree-based model, the
component value was determined from the number of times the function has been used to break the
information and in what way effective those splits were in splitting the information into a pure group,
as determined through impurity in Gini. Generally, as our objective was to develop model-agnostic
techniques, we try to capture a parallel concept within the standards of SHAP. In order to use it, we
summarize the SHAP standards for such the given function in all the analyses in our datasets to reach
an average estimate of importance for the function. Although SHAP values express directionality (i.e.,
class first choice) as well as magnitude (i.e., importance), such values will be used in two specific
methods.

LargeSHAP: To achieve an average alignment for that function, we merge the individual class
alignments of the properties for each variable by summing the specific SHAP values. Note that category
configurations for the function can vary from the sample to sample based on the sample's associations
with extra features and their relationship to the boundary of the decision. Therefore, summarizing of
functions in this method shows us the significance of the function that is dependent on the class mark,
with major negative standards being essential to good ware decisions as well as features that are essential
for malware decision. Properties with near-zero SHAP values, although in a general way they may be
significant, were not associated with the specific class and suggest areas to low trust.
Value Selection: After we define the subspace of function for embedding the watermark in, the next step was to select the elements that they make up of watermark. However, we can not necessarily select any arbitrary value for our watermarks because of the heavy semantic restrictions of the binaries. Therefore, we limit ourselves to selecting a value from inside our data. Therefore, value selecting essentially will become a searching problem to classify the values with the required properties throughout the space of the function and orientation with reference to the boundaries of decisions in that region. Depending on the intuitive attack techniques mentioned above, we need to pick such value based on the notions to their distribution in subspace – either choosing of points for high volatility over the decisions boundaries in sparse, weak-confidence areas or mixing points in dense areas with surrounding context information.

7. Conclusion

All the machine learning models are susceptible to adversaries. Provided that these models are commonly used in the modern era, a lot of studies are being carried out to protect these systems from adversarial attacks. Through this paper, we have reviewed a collection of adversarial attacks normally carried out in the wild in the ML systems current. To counter these attacks, there are various methods proposed, but no single strategy has been shown to be successful in dealing with all the adversarial invasions. Every method was only able to withstand a single or couple of attack types. In this paper, we have reviewed various attacks based on machine learning models and their working strategies. We also discussed the threat models and backdoor attacks on malware classifiers.

Acknowledgment

I would like to convey my most sincere gratitude to my guide Sripada Manasa Lakshmi of Lovely Professional University, for her valuable guidance and advice. Her willingness to motivate me contributed tremendously to achieve the goal successfully.

References

[1] Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access, 6(c), 14410-14430. https://doi.org/10.1109/ACCESS.2018.2807385

[2] Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. M. (2018). A survey on security threats and defensive techniques of machine learning. A data driven view. IEEE Access, 6, 12103-12117. https://doi.org/10.1109/ACCESS.2018.2805630.

[3] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndi´c, Pavel Laskov, Giorgio Giacinto,and Fabio Roli. Evasion Attacks against Machine Learning at Test Time. Advanced Information Systems Engineering, 7908:387–402, 2013.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [cs, stat], December 2014. arXiv: 1412.6572.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv:1312.6199 [cs], December 2013. arXiv:1312.6199.

[6] Xinyun Chen, Chang Liu, Bo Li, Kimberly Li, and Dawn Song. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. arXiv:1712.05526 [cs], December 2017. arXiv:1712.05526.

[7] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is Feature Selection Secure against Training Data Poisoning? In International Conference on Machine Learning, page 10, 2015.

[8] Thomas Mandl, Ulrich Bayer, and Florian Nentwich.ANUBIS ANalyzing Unknown BInarieS The automatic Way. In Virus bulletin conference, volume 1, page 02, 2009.

[9] Nicholas Carlini. Adversarial machine learning reading list. Available at https://nicholas.carlini.com/writing/2018/adversarialmachine-learning-reading-list.html, 2020.

[10] Virus Scan - Free Multi-Engine Online Virus Scanner.https://www.virscan.org/.

[11] Jo, J., & Bengio, Y. (2017). Measuring the tendency of CNNs to Learn Surface Statistical Regularities. (1). Retrieved from http://arxiv.org/abs/1711.11561

[12] Ali Shafahi,W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In Advances in Neural Information Processing Systems, April 2018.

[13] Alexander Turner, Dimitris Tsipras, and Aleksander Ma˛dry. Clean Label Backdoor Attacks. Manuscript submitted for publication, page 21, 2019.

[14] Burns, K. J., Molina-markham, A. D., & Sexton, J. T. (n.d.). Draft NISTIR 8269 A Taxonomy and Terminology of Adversarial Machine Learning Draft NISTIR 8269 A Taxonomy and Terminology of Adversarial Machine Learning.
[15] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023
[16] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–11.
[17] Weilin Xu, Yanjun Qi, and David Evans. Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers. In Proceedings 2016 Network and Distributed System Security Symposium, San Diego, CA, 2016. Internet Society.
[18] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial Examples for Malware Detection. In Computer Security—ESORICS 2017, volume 10493, pages 62–79. Springer International Publishing, Cham, 2017.
[19] Hyrum S Anderson, Anant Kharkar, and Bobby Filar. Evading Machine Learning Malware Detection. In Black Hat USA, page 6, 2017.
[20] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In IEEE Symposium on Security and Privacy, SP '18, pages 931–947. IEEE CS, 2018.
[21] OctavianSuciu, Radu Ma, Tudor Dumitras, and Hal Daume III. When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks. page 19, 2018.
[22] Octavian Suciu, Scott E. Coull, and Jeffrey Johns. Exploring Adversarial Examples in Malware Detection. In 2019 IEEE Security and Privacy Workshops (SPW), pages 8–14, San Francisco, CA, USA, May 2019. IEEE.
[23] Edward Raff, Jon Barker, Jared Sylvester, Robert Brandon, Bryan Catanzaro, and Charles Nicholas. Malware Detection by Eating a Whole EXE. arXiv:1710.09435 [cs, stat], October 2017. arXiv: 1710.09435.
[24] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. arXiv:1708.06733 [cs], August 2017. arXiv: 1708.06733.
[25] Konrad Rieck, Philipp Trinius, Carsten Willems, and Thorsten Holz. Automatic analysis of malware behaviour using machine learning. Journal of Computer Security, 19(4):639–668, 2011.