International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Robust Machine Learning Models for Security-Critical Applications

Datla Surya Kumari[1], Maradani Pavani Devi[2], Vankayala Anil Santosh[3], Dr. Prasad Rayi[4]

Submitted: 02/09/2024 **Revised:** 17/10/2024 **Accepted:** 25/10/2024

Abstract: A modest and human-imperceptible input perturbation may easily modify the model output entirely, as revealed by recent research. Machine learning models are susceptible to adversarial perturbations, which are also known as adversarial perturbations. Formal verification of the resilience of machine learning models is becoming more relevant as a result of the substantial security vulnerabilities that this has produced for a large number of real-world applications. This thesis investigates the resilience of tree-based models and deep neural networks, and it also takes into consideration the applications of robust machine learning models in the field of deep reinforcement learning. In the beginning, we come up with an innovative method to learn robust trees. Our technique seeks to improve performance under the worst-case perturbation of input characteristics, which gives rise to a max-min saddle point issue when splitting nodes in trees. Our method's goal is to optimize performance under these conditions. Through the process of approximating the inner minimizer in this saddle point issue, we suggest fast tree construction methods. Furthermore, we show efficient implementations for traditional information gain based trees as well as state-of-the-art tree boosting models such as XGBoost. The resilience of the model is greatly improved by our strategy, as shown by the experiments. In addition to this, we present an effective way for determining whether or not tree ensembles are resilient. The topic of verifying tree ensembles is recast as a max-clique problem on a multipartite graph by our team. We design an effective multi-level verification approach that is capable of providing tight lower limits on the resilience of decision tree ensembles. Additionally, our algorithm allows for iterative improvement and termination at any moment with no restrictions. When applied to random forest or gradient boosted decision trees models that have been trained on a variety of datasets, our algorithm is up to hundreds of times faster than the previous approach, which requires the solution of a mixed integer linear programming problem. Furthermore, our algorithm is able to provide tight robustness verification bounds on large ensembles that contain hundreds of deep trees. We submit a variety of empirical studies on the feasibility and the difficulty of adversarial training for neural networks. These findings are based on our own research. We demonstrate that even with adversarial defense, the resilience of a model on a test example has a substantial association with the distance between that example and the myriad of training data incorporated by the network. This is the case even when the adversarial defense is included. It is more probable that adversarial assaults will be successful against test samples that are quite far away from this manifold. As a consequence of this, we show that an adversarial training-based defense is susceptible to a new category of attacks known as the "blind-spot attack." This type of attack occurs when the input examples are located in low density regions (also known as "blind spots") of the empirical distribution of training data, but they are still on the valid ground-truth data manifold. In conclusion, we take neural network resilient training approaches and apply them to deep reinforcement learning (DRL) in order to train agents that are resistant to perturbations on state observations. In order to investigate the underlying characteristics of this issue, we offer the state-adversarial Markov decision process (SA-MDP). Additionally, we provide a theoretically principled regularization that can be used for a variety of deep learning and reinforcement learning algorithms, such as deep Q networks (DQN) and proximal policy optimization (PPO). We provide major improvements to the resilience of agents when they are subjected to powerful adversarial assaults via white box, including novel attacks that we have developed ourselves.

Keywords*: imperceptible, perturbation, reinforcement, implementations, algorithms, adversarial.*

[1,2]Assistant Professor, Department of Computer Science and Engineering. International School Of Technology And Sciences For Women, A.P, India.

[3,4]Associate Professor, Department of Computer Science and Engineering. International School Of Technology And Sciences For Women, A.P, India.

Introduction

Machine learning technologies, particularly artificial deep neural networks (DNNs) and deep learning (DL) architectures, have been widely adopted in many mission-critical fields in recent years. These fields include cyber security,

autonomous vehicle control, healthcare, and others. The purpose of these technologies is to support intelligent decision-making [1].

There are worries surrounding the robustness of the system against ML-specific security threats and privacy breaches, as well as the confidence that users have in these systems [2, 3, 4]. Despite the fact that machine learning has showed outstanding performance in comparison to traditional approaches in these applications, there are concerns still.

There have been a significant number of researchers that have brought to light the inherent security flaws that are present in machine learning technologies, such as learning algorithms or produced models, as a result of the remarkable success that machine learning has had in a variety of application domains [1, 3].

The vulnerabilities that are present in machine learning systems make them susceptible to a wide variety of adversarial exploits, each one of which has the potential to undermine the whole system. In point of fact, a typical machine learning pipeline, which includes data collection, feature extraction, model training, prediction, and model re-training, is susceptible to malicious attacks at each and every step [5]. The attacks that are launched against machine learning systems have a detrimental influence on the systems, which may lead to a decline in performance, misbehavior on the part of the system, and/or a violation of privacy [4, 6]. Researchers in the fields of machine learning and cyber security are highly motivated to discover the inherent flaws, exploitable vulnerabilities, and relevant attacks that are associated with machine learning. They have been working diligently to establish defensive mechanisms that are effective within this field.

A multi-disciplinary effort that encompasses machine learning, cyber security, human-computer interaction, and domain-specific expertise is required in order to succeed in the creation of machine learning systems that are reliable and trustworthy. It is possible to define the robustness of a machine learning system as its capacity to withstand harmful assaults in order to safeguard itself against the compromising of the system's integrity, availability, and confidentiality. A powerful machine learning system has the ability to instill confidence in users about the system's security compliance. On the other hand, the trust that users have in an ML system may contribute to the achievement of a system's security goals by assisting users in responding appropriately to harmful assaults and also in avoiding accidental acts.

In order to support and assure the development of ML systems that are reliable and trustworthy, the community of machine learning and artificial intelligence understands the need of all-hands efforts at all levels. There have been a number of continuing efforts made by policymakers all over the globe to adopt regulations that would support and legitimize the activities of AI practitioners [7]. As an example, the Canadian government is now working on the Algorithmic Impact Assessment (AIA)3 in accordance with the Directive on Automated Decision-Making4. AIA is a questionnaire tool that can be accessed online and is aimed to assist in determining the amount of influence that an automated decision system has. In order to assist ethical artificial intelligence research, deployment, and governance, more than eighty organizations from both the public and commercial sectors have taken the initiative to draft Ethics Principles for Artificial Intelligence [8]. By proposing a set of methods that AI practitioners might adopt to make and verify claims about AI systems, a new paper titled "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims" [9] reflects a cooperative effort by academics and industry to advance beyond the ethical standards that have been established. It is possible to utilize these verifiable statements as evidence for proving responsible conduct in order to compel compliance with the rules and standards that are stipulated in the high-level ethical principles to be followed by artificial intelligence.

Secure Machine Learning: An Overview

By facilitating the discovery of significant patterns or regularities in big datasets, machine learning comprises a number of techniques that assist problem-solving via experience [27, 28]. These approaches often enable the identification of major patterns or regularities. There are three primary paradigms that may be used to classify methods to machine learning. These paradigms include supervised learning, unsupervised learning, and reinforcement learning. Every one of these paradigms is susceptible to vulnerability in its own unique way. In this part, an overview is presented for each paradigm, along with introductions to the key methods and models that are included in each paradigm. This is followed by an explanation of

some of the potential vulnerabilities, as well as a short discussion of possible exploitations that have been described in the literature.

In the context of supervised learning approaches, the goal is to create a function that can map input instances to labels by making use of a group of examples as a basis for training a model. Given the premise that the sample used for training is typical of the population, the concept here is that a function that can be constructed to perform well at properly labeling the training data should also perform well when labeling fresh data. This is the theory behind this. It is thus possible to employ so-called discriminative modeling tools, such as logistic regression and support-vector machines, to make a prediction about the probability of a new instance belonging to a certain class. One way to do this is by establishing a direct mapping from the feature values to the labels. For instance, this may be accomplished by establishing a border in the data that divides the two classifications (or more). Instead, generative modeling techniques, such as Naive Bayes Classification, calculate the likelihood of each class by using the probabilities of the feature values that comprise an example instance. This is done in order to determine the probability of each class. Methods that are based on artificial neural networks, such as deep learning, may be applied in a supervised way to learn high-level characteristics, such as those that are necessary for image processing. However, these methods can also be utilized in a semi-supervised or unsupervised manner. It is possible to make use of few-shot learning techniques [29] when there are relatively few examples upon which to construct a classification model. On the other hand, zero-shot learning approaches [30] are relevant when instances that are to be classified could belong to classes that are not observed during training.

Unsupervised machine learning techniques seek for additional commonalities in the data that may be used in such a manner as to draw plausible inferences or assumptions throughout the learning and prediction process. This is in contrast to supervised machine learning approaches, which depend on a collection of examples to train a classifier. Clustering techniques are centered on the identification of certain similarities among the data, which may subsequently be used to make statements about particular data based on the degree of _t. Anomaly detection, for instance, may be used to determine which instances are anomalous, hence giving evidence that certain occurrences may be of special relevance with regard to the investigation. Such an example would be the identification of malicious network behavior via the use of unsupervised anomaly detection methods. These methods are able to recognize patterns that are not compatible with the normal activity that is seen. A Markov Decision Process is often used as a model for reinforcement learning, which is an alternative paradigm in which learning is carried out in an experimental way. An evaluation of the offered solutions may be carried out via the use of a reward function, and the goal is to acquire knowledge about a solution to a problem. As a result, learning modifies solutions while simultaneously attempting to maximize rewards.

As a consequence of the pervasiveness of machine learning methods and the consequent fast adoption of these techniques, the vulnerability of systems has risen, and they have become more appealing to prospective attackers [31]. An intrusion detection system (IDS) that is based on machine learning may be influenced by potential network attackers in order to either raise the number of false negatives, which would enable the attackers to access the network unnoticed, or increase the number of false positives to the point where so much genuine traffic is agged that warnings become too frequent. When this occurs, either they are disregarded or the operation is completely interrupted via the use of denial of service (DoS) technologies [10]. There is a possibility that advertisers would attempt to influence spam detectors in order to enhance the probability that their communications will be able to get through email filters [32]. It is possible for the training data for image recognition to be altered in such a manner as to provide unwanted access or to cause damage in other domains, such as linked and autonomous automobiles.

Threat Modeling

The engineering method known as threat modeling is used to provide assistance for the systematic study of security requirements. For the purpose of identifying possible system threats, setting security targets that are attainable, identifying relevant vulnerabilities and attack routes, and designing suitable defensive mechanisms, it has been extensively accepted by researchers and experts in the field of cyber security. In order to limit the likelihood of security problems occurring during the

creation of an application and to mold the application security design in a manner that is in accordance with the security goals, a threat model that has been thoroughly defined serves as the foundation of the safe development process.

The researchers concentrated their attention on the following areas of threat modeling in the context of machine learning security [5, 3, 12, 25, 36] for example:

Attack Surface.

In the field of machine learning, the workflow of the whole machine learning activities is represented as a pipeline. This pipeline is comprised of multiple stages, which include data collection, data preprocessing, feature extraction, model training and testing, prediction, and possibly model re-training. There is a significant flow of sensitive and confidential data throughout the pipeline, beginning with raw data and ending with trained models. It has been determined that the pipeline has a number of attack surfaces as well as a variety of attack routes, according to the following summary:

_ Stealthy Channel attack during raw data collection phase; _ Mimicry and Poisoning attack against training and testing datasets; _ Polymorphic/Metamorphic attack against feature extraction; _ Gradient Descent attack against learning algorithms; _ Evasion attack against trained models during prediction phase; _ Model Stealing against trained models; and _ Poisoning Attack during model re-training phase.

Poisoning Attack, Gradient Descent Attack, Evasion Attack, and Model Stealing are the primary areas of concentration for the bulk of the research that we have examined.

Goal of the Attacker" On the basis of the following three views, the hostile aims of the attacker may be classified as follows:

Robust and Secure AI

AI that is both robust and secure—more precisely, the capability to design, build, deploy, and run AI systems that are both robust and secure—is not just an essential component of AI Engineering but also an absolute need for the Department of Defense. Robustness and security in artificial intelligence systems are essential to the accomplishment of one's purpose, and they may also permit a wide range of other attributes that are connected to it, including safety, dependability, stability, and reliability. There are other policy-related issues that may be supported by robust and secure systems, such as privacy, fairness, and ethics.

In the context of the Department of Defense, the existing strategies and procedures for developmental test and evaluation (DT&E) and operational test and evaluation (OT&E) need to be modified so that they include artificial intelligence. It is important to note that both DT&E and OT&E have substantial consequences for acquisition procedures and practices when applied to the field of AI Engineering. When continuous monitoring is required, they are required to take into account a number of factors, including how to produce system testing needs, how to acquire them, and how to operate within budgets. These are all important considerations for having resilient and secure systems. A workshop that was held not too long ago that was organized by the Applied Research Lab for Intelligence and Security (ARLIS) at the University of Maryland brought to light the requirements and difficulties that artificial intelligence presents for OT&E in particular [22]. Over the course of the training, the mismatch between what is simple to measure and what is relevant from an operational standpoint was stressed. In order to stay up with the fast changes in technology, testing and evaluation techniques need to be updated. In order to do this, the Department of Defense (DoD) has to cultivate a test and evaluation community that is both proactive and agile. This includes increasing the number of AI testers who are already employed by the DoD.

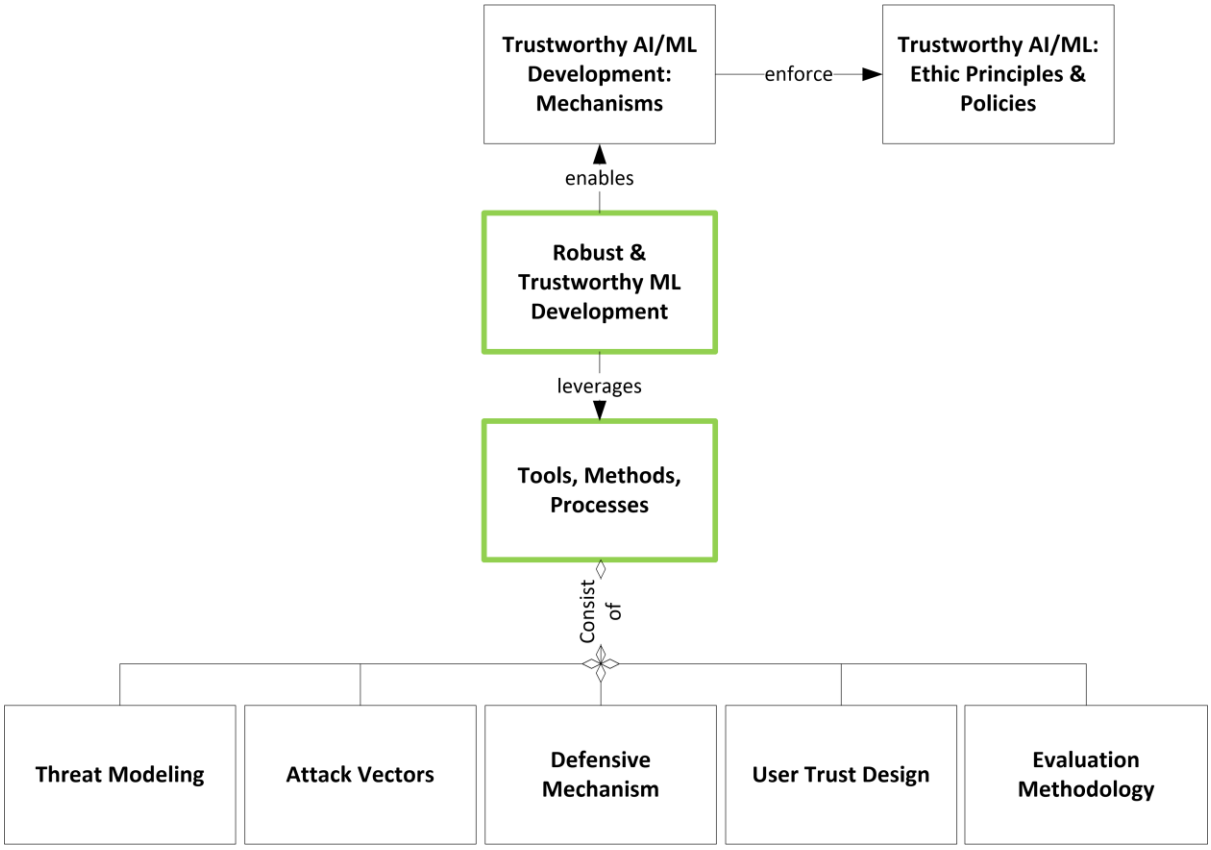

Figure 1: Robust and Trustworthy ML System: An Ecosystem View

In addition, the Department of Defense is confronted with a cultural obstacle when it comes to the task of fostering an attitude of experimentation among all of the stakeholders engaged in the development and deployment of artificial intelligence systems. Experimentation and prototyping are required for almost every area; nevertheless, early and regular system testing is required because of the complexity of artificial intelligence (AI) systems, the often opaque nature of information, and the novel behaviors that are required to allow successful human-machine collaborations. However, resolving problems is what takes time, especially in later phases of project development [23]. People often make the mistake of assuming that testing is a time-consuming task, while in reality error fixing is what takes time. This problem is not exclusive to the Department of Defense: As opposed to tools and technology, culture is the factor that hinders businesses from carrying out the hundreds or even thousands of tests that they need to be carrying out yearly and then putting the findings into practice.

Conclusion

Machine learning technologies have been extensively implemented across several application domains. Notwithstanding the advantages conferred by the use of machine learning technologies, it remains a problem to guarantee that these systems are adequately resilient to security threats and privacy violations, while also fostering user confidence in the systems. The creation of robust and reliable machine learning systems has not yet been extensively embraced in the industry. From the standpoint of security engineering, this arises from many factors, including the absence of (i) comprehensive advice on fundamental concepts and best practices; (ii) effective machine learning defensive technologies; (iii) techniques and metrics for assessing machine learning robustness; and (iv) specialized tool support. This paper summarizes our findings from the study of cutting-edge technologies and our engineering endeavor to use these technologies in the building of strong and trustworthy machine learning systems. Numerous research studies we examined highlighted the significance of offensive-defensive machine learning technologies and advocated for the development of resilient machine learning algorithms as a prospective avenue for future inquiry. The findings of our research corroborate that perspective. Moreover, we assert that the engineering of machine learning system development, now in its nascent phase, is a

fundamental pillar for ensuring the resilience and trustworthiness of ML systems. In section 8, we illustrated a systematic methodology for machine learning threat modeling and security design by enhancing and expanding the traditional technique. Our findings and their interpretation are preliminary, since a more comprehensive examination would beyond the scope of this study. Therefore, we offer two prospective research topics that we anticipate will illuminate this field for both industry and academic practitioners in the near future.

References

[1] G. Mcgraw, R. Bonett, H. Figueroa, V. Shepardson, Security engineering for machine learning, Computer 52 (8) (2019) 54{57. doi:10.1109/MC. 2019.2909955.

[2] Y. Lecun, Y. Bengio, G. Bengio, Deep learning, Nature 521 (7553) (2015) 436{444.

[3] B. Biggio, F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, Pattern Recognition 84 (2018) 317 { 331. doi:https://doi.org/10.1016/j.patcog.2018.07.023. URL http://www.sciencedirect.com/science/article/pii/ S0031320318302565

[4] P. Dasgupta, J. B. Collins, A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks, AI Magazine 40 (2) (2019) 31{43, name - MIT Press; Cornell University; Copyright

- Copyright Association for the Advancement of Arti_cial Intelligence Summer 2019; Last updated - 2019-08-09; SubjectsTermNotLitGenreText

[5] X. Wang, J. Li, X. Kuang, Y. an Tan, J. Li, The security of machine learning in an adversarial setting: A survey, Journal of Parallel and Distributed Computing 130 (2019) 12 { 23. doi:https://doi.org/10.1016/j.jpdc.2019.03.003. URL http://www.sciencedirect.com/science/article/pii/ S0743731518309183

[6] M. Al-Rubaie, J. M. Chang, Privacy-preserving machine learning: Threats and solutions, IEEE Security Privacy 17 (2) (2019) 49{58. doi:10.1109/ MSEC.2018.2888775.

[7] The Law Library of Congress, Regulation of Arti_cial Intelligence in Selected Jurisdictions 5080 (January) (2019) 138. URL https://www.loc.gov/law/help/artificial-intelligence/ index.php

[8] B. Mittelstadt, Principles alone cannot guarantee ethical AI, Nature Machine Intelligence 1 (11) (2019) 501{507. doi:10.1038/ s42256-019-0114-4.

[9] M. Brundage, S. Avin, J.-B. Wang, et al., Toward trustworthy ai development: Mechanisms for supporting veri_able claims, ArXiv abs/2004.07213 (2020).

[10] M. Barreno, B. Nelson, A. D. Joseph, J. D. Tygar, The security of machine learning, Machine Learning 81 (2) (2010) 121{148.

[11] M. Xue, C. Yuan, H. Wu, Y. Zhang, W. Liu, Machine Learning Security: Threats, Countermeasures, and Evaluations, IEEE Access 8 (2020) 74720{ 74742. doi:10.1109/ACCESS.2020.2987435.

[12] N. Papernot, P. McDaniel, A. Sinha, M. P. Wellman, SoK: Security and Privacy in Machine Learning, Proceedings - 3rd IEEE European Symposium on Security and Privacy, EURO S and P 2018 (2018) 399{ 414doi:10.1109/EuroSP.2018.00035.

[13] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, V. C. Leung, A survey on security threats and defensive techniques of machine learning: A data driven view, IEEE Access 6 (2018) 12103{12117. doi:10.1109/ACCESS.2018. 2805680.

[14] K. Ren, T. Zheng, Z. Qin, X. Liu, Adversarial Attacks and Defenses in Deep Learning, Engineering 6 (3) (2020) 346{360. doi:10.1016/j.eng. 2019.12.012. URL https://doi.org/10.1016/j.eng.2019.12.012