JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

Department of CSE & IT

Bachelor of Technology, 7th Semester

TERM PAPER

Data Poisoning Attacks in Linear Regression Models

Group details:

Meetanshi Mittal -17103343 B9
Shivam Bisht -17103199 B5
Soumya Agarwal -17103347 B9

Submitted to:

Dr. Gagandeep Kaur
Prantik Biswas

Supervised by:

Dr. Parmeet Kaur

INTRODUCTION

As machine learning is widely used for automated decisions, the results and models created by machine learning algorithms are exploited by attackers with strong incentives. Recent studies have shown that machine learning models are vulnerable to numerous attacks that compromise the protection of the models and the application systems themselves.

Poisoning is to add a fraction of poisoning points in training to degrade model accuracy (availability attack).In poisoning attacks, in order to exploit the results of a predictive model, attackers intentionally influence the training data. This gives attackers power to manipulate the training dataset in order to control the prediction behavior of a trained model such that the model will label malicious examples into desired classes.

Attacker Knowledge is of two types –

1. White box: full knowledge of the ML system .
2. Black-box: query access to the model.

This is a huge security threat and needs countermeasures to make the model robust against poisoned data. Examples where poisoning attacks have been studied so far include attacks against sentiment analysis, malware clustering, malware detection, worm signature detection, DoS attack detection, intrusion detection and social media chatbots.

PROBLEM STATEMENT

We study the security issues present in machine learning models and focus on poisoning attacks and their countermeasures for linear regression models. A model can be poisoned/broken during various stages of model building and we focus on the attacks done during the training phase.

LITERATURE REVIEW

Paper 1

Title: **Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning**

Citation: *Jagielski, Matthew, et al. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning." 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 2018.*

Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8418594

Summary: In this paper, the first systematic study of poisoning attacks and their countermeasures for linear regression models was conducted. It shows the following contributions:

- It considers the problem of poisoning linear regression under different adversarial models;
- Starting from an existing baseline poisoning attack for classification, it proposes a theoretically-grounded optimization framework specifically tuned for regression models;
- A fast statistical attack is designed that requires minimal knowledge on the learning process;
- A principled defense algorithm is proposed with significantly increased robustness than known methods against a large class of attacks;
- Extensively evaluate the attacks and defenses on four regression models (OLS, LASSO, ridge, and elastic net), and on several datasets from different domains, including health care, loan assessment, and real estate.

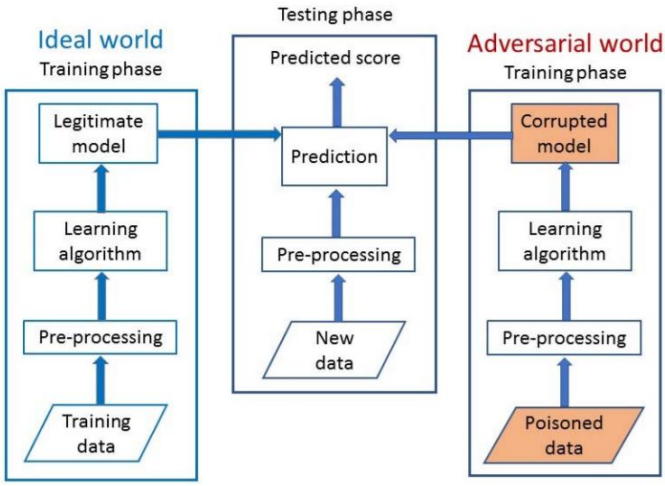


Fig. Overview of working of the model

Paper 2

Title: **ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models**

Citation: *Salem, Ahmed, et al. "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models." arXiv preprint arXiv:1806.01246 (2018).*

Link: https://arxiv.org/pdf/1806.01246.pdf

Summary: In several real-world applications, machine learning (ML) has become a central component, and training data is a key factor driving current development. This tremendous success has prompted the introduction of machine learning as a service (MLaaS) by Internet businesses. Recently, the first membership inference attack has shown that in such MLaaS environments, extracting information on the training set is feasible, which has significant security and privacy implications. At low cost, such attacks are very widely applicable and thus pose a serious risk. Though several ML models are widely popular, they are susceptible to various security and privacy attacks like model inversion, adversarial examples, and model extraction to name a few.

The paper primarily concentrates on the Membership Inference attack. In this attack, an adversary is used to find out whether a given data point or item was used in the training of an ML model or not. This may lead to severe consequences. For example, if a machine learning model is trained on data collected from people with a certain disease through recognizing that the data of a victim belongs to the model's training data, the attacker will learn the health status of this victim immediately.

The paper suggests two methods of protection to resolve the situation. We analyze strategies designed to minimize overfitting, as we demonstrate the link between overfitting and vulnerability to membership inference attacks. In each training iteration in a completely linked neural network, the first one, namely dropout, randomly deletes a certain proportion of edges, while the second method, namely model stacking, organizes multiple ML models in a hierarchical way. The extensive assessment suggests that our defensive tactics are still worthy of being able to largely reduce the efficiency of the membership inference attack while retaining high-level usefulness, i.e. the prediction accuracy of the high target model.

Paper 3

Title: SoK: Security and Privacy in Machine Learning

Citation: Papernot, Nicolas, et al. "SoK: Security and privacy in machine learning." *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2018.

Link: http://www-personal.umich.edu/~arunesh/Files/Other/Papers/18-eurosp-adv-ml-sok.pdf

Summary: A dizzying array of applications such as data analytics, autonomous systems, and security diagnostics have been made possible by developments in machine learning (ML) in recent years. ML is now prevalent, with new systems and models being implemented in any conceivable domain, leading to the widespread implementation of inference and decision making based on software. The security and privacy of machine learning is an active yet nascent area.

- A unifying threat model to allow structured reasoning about the security and privacy of systems that incorporate ML is introduced. Considers the entire data pipeline, not just algorithms.
- Attacks and defenses found by the various technological communities are taxonomized.
- Systematize desirable properties to improve the security and privacy of machine learning

Take-away from the paper is that, information systematisation points towards varying, but connected, sensitivity notions. Characterizing the sensitivity of learning algorithms to their training data is important for ML to maintain privacy. Similarly, for secure ML, it is required to control the sensitivity of deployed models to data on which they perform inference. The sensitivity of generalisation error (i.e., the difference between training results and test data) remains poorly understood for many models at the centre of these two principles, and calls for further study.

Paper 4

Title: **Explaining Vulnerabilities to Adversarial Machine Learning through Visual Analytics**

Citation: *Ma, Yuxin, et al. "Explaining vulnerabilities to adversarial machine learning through visual analytics." IEEE transactions on visualization and computer graphics 26.1 (2019): 1075-1085.*

Link: https://arxiv.org/pdf/1907.07296.pdf

Summary: In a number of real-world applications, machine learning models are currently being applied where model forecasts are used to make health care decisions, bank loans, and many other important tasks. Adversaries have started developing strategies to exploit models of machine learning to their benefit.In this article, we present a framework for visual analytics to clarify and explore model vulnerabilities for adversarial attacks. Our architecture employs a multi-faceted visualisation system designed to support the study from the perspective of models , data instances, functions, and local structures of data poisoning attacks.

Under the assumptions of static environments, many of the Machine Learning models were created, where new data instances are presumed to be from a statistical distribution close to that of training and test data. Unfortunately, the real-world implementation of these models creates a complex environment that is home to malicious individuals who in the machine-learning models may wish to manipulate these underlying assumptions; E-mail spam filtering is an example where unwanted mails bypass the spam section and reach the inbox of the user.

The paper proposes a framework for visual analytics to explore vulnerabilities and adversarial attacks against machine learning models. Our system helps users to analyse possible weak points in the training dataset and explore the impacts of poisoning attacks on model efficiency by focusing on targeted data poisoning attacks. Via collaboration with domain experts, task and design criteria were established to support the study of adversarial machine learning attacks. Our system serves as a mechanism for iterative proactive defence, as opposed to conventional reactive defence techniques that respond when attacks are detected. Users can model poisoning operations and explore attack vectors in historical documents that have never been used. This will help domain scientists to design more accurate models of machine learning and pipelines for data processing.

Paper 5

Title: **Machine Learning Security: Threats, Countermeasures, and Evaluations**

Citation: *Xue, Mingfu, et al. "Machine Learning Security: Threats, Countermeasures, and Evaluations." IEEE Access 8 (2020): 74720-74742.*

Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9064510

Summary: Recent studies have shown that machine learning models are vulnerable to numerous attacks that compromise the protection of the models and the application systems themselves.This survey extensively analyzes machine learning security problems, concentrating on current attacks on machine learning systems, effective protections or safe learning strategies, and methods of safety assessment. This paper addresses all aspects of machine learning safety from the training phase to the test phase, instead of concentrating on one stage or one form of attack.

The security threats along the life cycle of machine learning systems can be divided into five categories: 1) Poisoning attacks; 2) Backdoor attacks; 3) Adversarial example attacks; 4) Model theft; 5) Recovery of sensitive training data. The first two attacks occur during the training phase, while the last three attacks occur during the test phase.

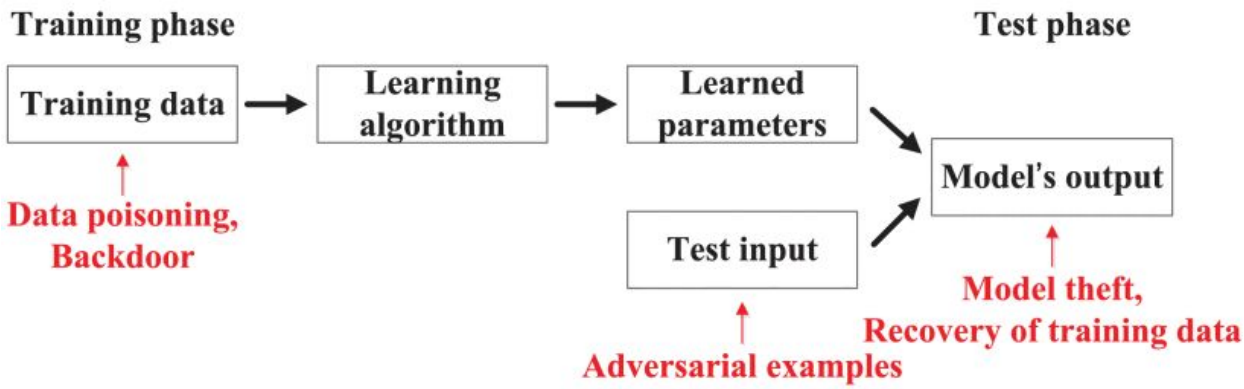


Fig. Overview of all types of attacks

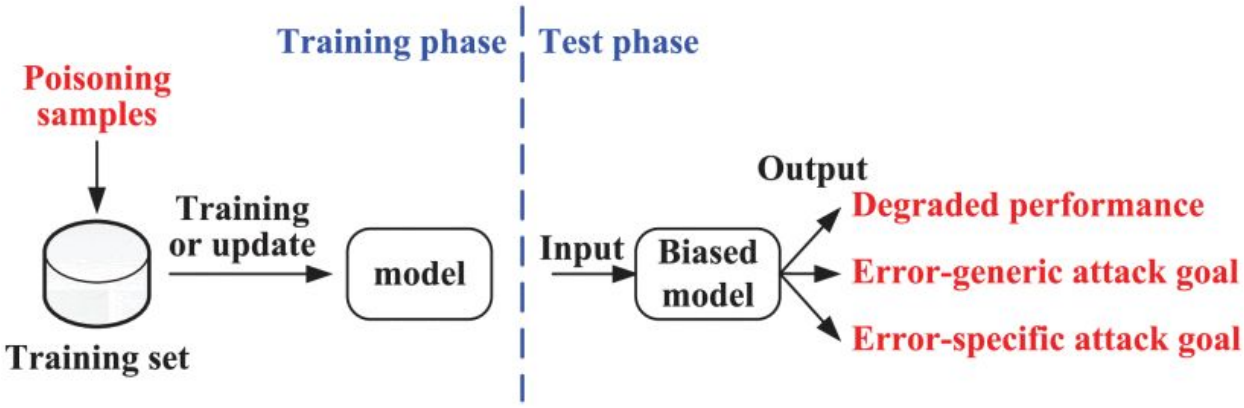


Fig. Overview of poisoning attack

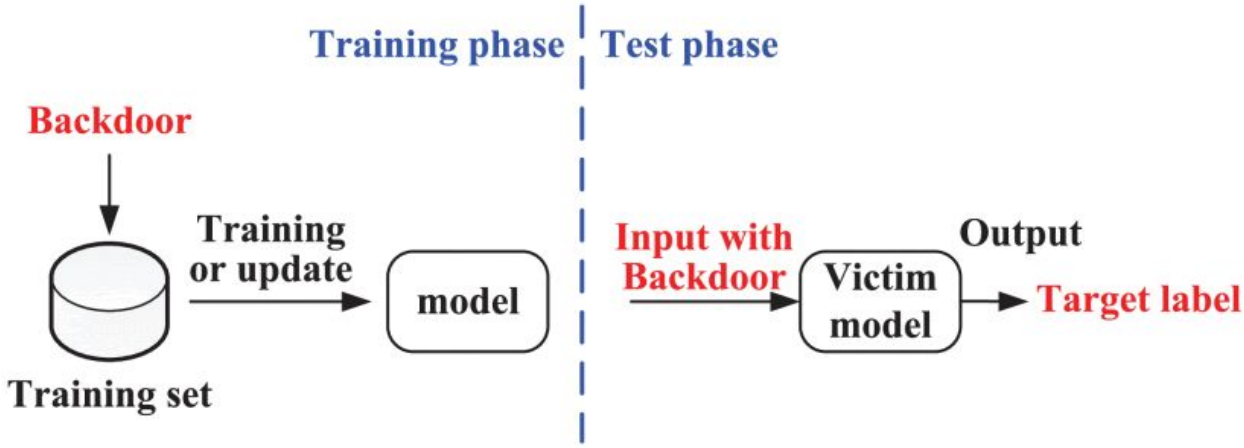


Fig. Overview of backdoor attack

Defense against poisoning attacks on various algorithms include:

- Anomaly detection or security detection by using robust statistical methods such as bagging ensembles , self-adaptive learning camouflage detector.
- SVM- game theory and rejection method
- Robust linear regression , data provenance on contextual information
- In Neural networks check the loss of the model, identify features with abnormal distributions.
- In healthcare based data monitor the accuracy deviations on the training set.

In the training phase, the defensive works against poisoning attacks or backdoor attacks, can be called *data sanitization*, in which the anomalous poisoned data is filtered out first before feeding into the training phase. The anomaly detectors are usually based on training loss, nearest neighbors, and so on

In the test phase, the defense techniques against adversarial examples can be called *smoothing model outputs* , i.e., reduce the sensitivity of the model's output to the changes in the input.

Paper 6

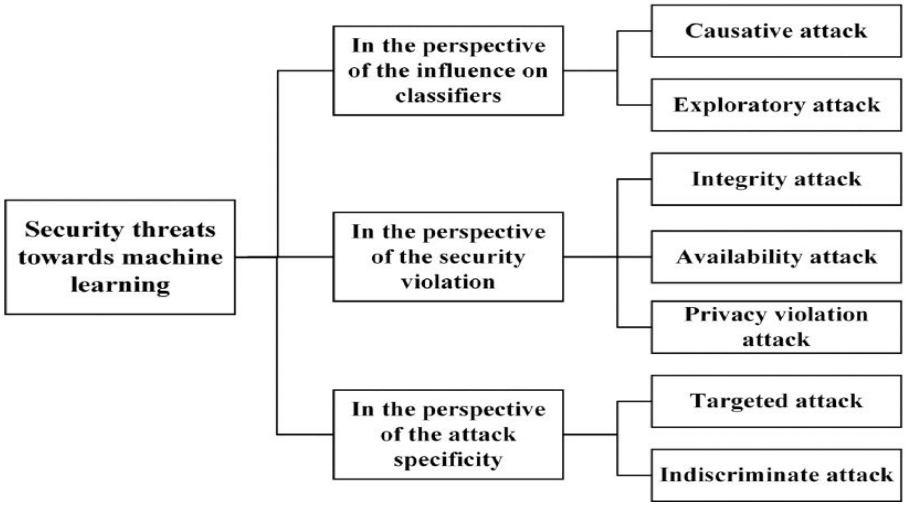
Title: **A Survey on Security threats and Defensive Techniques of Machine Learning: A Data Driven view**

Citation : *Liu, Qiang, et al. "A survey on security threats and defensive techniques of machine learning: A data driven view." IEEE access 6 (2018): 12103-12117.*

Link : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8290925

Summary : Machine learning has been widely applied in image processing, natural language processing , pattern recognition, cryptography, and other fields, and is one of the most prevalent techniques in computer science. These algorithms and related training data are vulnerable to a number of security threats, causing a substantial performance decrease, regardless of effective implementations of machine learning algorithms in many scenarios, such as facial recognition, malware detection, automatic driving, and intrusion detection. It is therefore necessary to call for more attention to security threats and corresponding machine learning defensive techniques, which motivates a detailed survey in this paper.

This paper discusses current security threats and provides two dimensions, the training process and the testing / inferring process, with a comprehensive survey on them. The paper subsequently categorises existing protective machine learning strategies into four groups: safety evaluation mechanisms, training phase countermeasures, testing or implied phase countermeasures, data protection, and privacy. Finally , the paper presents five notable developments in research into safety risks and machine learning protective strategies that are worth doing in-depth studies in the future.



Paper 7

Title: Robust Linear Regression Against Training Data Poisoning

Citation: *Liu, Chang, et al. "Robust linear regression against training data poisoning." Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017.*

Link: https://www.ccis.northeastern.edu/home/alina/papers/RobustRegression.pdf

Summary: Supervised learning typically relies on dimensionality reduction in high-dimensional settings to enhance efficiency and recognise the most significant factors in predicting results. However, it has become a natural target for adversarial exploitation of training data, which we call poisoning attacks. Prior approaches to dealing with stable supervised learning, such as feature independence and sub-Gaussian noise with low variance, rely on clear assumptions about the existence of the feature matrix. This paper suggests an integrated robust regression approach that relaxes these assumptions, assuming only that a low-rank matrix can approximate the feature matrix well. In this paper, the techniques combine improved robust low-rank matrix approximation and robust principal component regression, and provide strong performance guarantees. In addition, this paper experimentally shows that the techniques outperform state-of-the-art substantially in both running time and prediction error.

The poisoning attack for linear regression issue with dimensionality reduction is considered in this paper. The paper addresses the issue in two steps:

1. Implementation of a new robust method of matrix factorization to recover the true subspace, and
2. Novel robust regression of the principle variable to prune adversarial instances based on the basis recovered in step (1).

In order to be efficient in recovering the true subspace, this paper defines required and adequate conditions for our approach and presents a bound on expected prediction loss compared to ground truth.

Paper 8

Title: Online Data Poisoning Attacks

Citation : *Zhang, Xuezhou, Xiaojin Zhu, and Laurent Lessard. "Online data poisoning attacks." Learning for Dynamics and Control. 2020.*

Link : http://proceedings.mlr.press/v120/zhang20b.html

Summary : In the online learning environment, where training data arrives sequentially, this paper research data poisoning attacks where the attacker listens to the data stream and has the potential to contaminate the current data point to influence the process of online learning. As a stochastic optimal control problem, this paper formulate the optimal online attack problem and provide a systematic solution using techniques from model predictive control and deep reinforcement learning.Theoretical analysis of the remorse experienced by the attacker for not understanding the true data sequence is also presented.

The online attacker faces some specific challenges compared to the offline environment: In the offline setting, it is often presumed that the entire dataset is observed by the attacker. However, in the online world, when making decisions, the attacker can only observe the current data. The attacker only needs to make one decision in the offline setting, while the attacker is expected to make a series of decisions in the online setting to execute the attack over time. These particular problems make the online regime applicable to the classic data poisoning attack structure.

The paper thus formulates online poisoning attacks as an adaptive control problem. Based on model-based planning and deep reinforcement learning, it proposed two attack algorithms, and showed that both are able to achieve near clairvoyant-level efficiency.

Paper 9

Title : **Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks**

Citation : *Demontis, Ambra, et al. "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks." 28th {USENIX} Security Symposium ({USENIX} Security 19). 2019.*

Link : https://www.usenix.org/system/files/sec19-demontis.pdf

Summary :The ability of an attack against a machine-learning model to be successful against another, theoretically unknown, model is captured by transferability.The paper presents a comprehensive analysis aimed to investigate the transferability of both test-time evasion and training-time poisoning attacks.Evasion is to add a minimum amount of perturbation to a test point to change prediction.Poisoning is to add a fraction of poisoning points in training to degrade model accuracy (availability attack).
Attacker Knowledge is of two types –

1. White box: full knowledge of the ML system .
2. Black-box: query access to the model.

Model complexity is the capacity of the classifier to fit the training data (can be controlled through regularization).

Extensive experiments on 3 datasets and 12 classifiers have shown that:

- High-complexity models are more vulnerable to both evasion and poisoning attacks
- Low-complexity models are better surrogates to perform evasion attacks.
- The complexity of the best surrogate is the same as the one of the target for availability poisoning.

Paper 10

Title : Mitigating poisoning attacks on machine learning models: A data provenance based approach

Citation : *Baracaldo, Nathalie, et al. "Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach." (2017).*

Link:https://www.researchgate.net/profile/Nathalie_Baracaldo/publication/320836025_Mitigating_Poisoning_Attacks_on_Machine_Learning_Models_A_Data_Provenance_Based_Approach/links/5aa03719aca272d448b0197d/Mitigating-Poisoning-Attacks-on-Machine-Learning-Models-A-Data-Provenance-Based-Approach.pdf

Summary :A protection risk is posed by the dependence of machine learning methods on quality training data in which adversaries can insert poisonous samples into the training dataset to exploit the trained classifier. Provenance meta-data is used to segment the untrusted data into groups where the probability of poisoning is highly correlated across samples in each group. A specific video camera, a Twitter account, or a specific firmware version. etc., is called a provenance signature.

- A new technique for detecting and filtering poisonous data collected to train an arbitrary supervised model of learning is presented in the paper. The training data has been segmented appropriately, data points in each segment are evaluated together by comparing the performance of the classifier trained with and without that group.
- Two flavors of our provenance-based defense for cases when partially trusted and fully untrusted datasets are available. By partially trusted, it is meant that some of the data points are believed to be genuine (not poisoned) in the collected data.

Trusted provenance information is available in many application scenarios such as in environmental sensing or even some social media environments. The paper assumes that data sources are independent.

Paper 11

Title: Adversarial Security Attacks and Perturbations on Machine Learning and Deep Learning Methods

Citation: *Siddiqi, Arif. "Adversarial security attacks and perturbations on machine learning and deep learning methods." arXiv preprint arXiv:1907.07291 (2019).*

Link: https://arxiv.org/ftp/arxiv/papers/1907/1907.07291.pdf

Summary: This paper first briefly explains the different kinds of machine learning models like Logistic Regression, SVM, Decision Tree, Random Forest, Hidden Markov Model, etc.
Then it tells us the categories of security attacks which are:

- Causative attack. Targets the training process or the training data is altered. The model trained on the altered data provides the manipulated output. It is sometimes also called the poisoning attack.
- Exploratory attack. Targets after the training process. Explores or probes the learner for useful information. Can exploit misclassifications but do not alter the training process.
- Evasion attack. Targets after the training process. Modifies the input data to the learner that results in an incorrect prediction or evade detection.
- Targeted attacks. Targets the specific points, instances, or exploits that are continuous streams.
- Indiscriminate. Targets the general class of points, instances, or exploits in a random non-targeted manner.
- Integrity attack. A successful attack on assets via false negatives and that is being classified as normal traffic.
- Availability attack. A broad class of an attack that makes the system unusable with classification errors, denial of service, false negatives and positives, etc.
- Privacy violation attack. An exploratory attack type that reveals sensitive and confidential information from the data and models. Also known as model extraction, inversion, or hill-climbing attack.

Then it explains most of the terminologies used in this area of research so that it is easier for new researchers like us to get started. Even though several research papers exist that review adversarial security attacks and perturbations, there is always room to grow due to the dynamic nature of ML and DL methods. The learning models that are ideal and produce satisfying results remain an open and a lasting challenge. This includes the issue of adversarial security attacks and perturbations because of its relation to the DM and ML methods.

Paper 12

Title: Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models

Citation: *Brendel, Wieland, Jonas Rauber, and Matthias Bethge. "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models." arXiv preprint arXiv:1712.04248 (2017).*

Link: https://arxiv.org/pdf/1712.04248.pdf

Summary: This paper emphasises the importance of attacks which solely rely on the final model decision. Such decision-based attacks are (1) applicable to real-world black-box models such as autonomous cars, (2) need less knowledge and are easier to apply than transfer-based attacks and (3) are more robust to simple defences than gradient- or score-based attacks. It introduces something called the Boundary Attack, a decision-based attack that starts from a large adversarial perturbation and then seeks to reduce the perturbation while staying adversarial. The attack is conceptually simple, requires close to no hyperparameter tuning, does not rely on substitute models and is competitive with the best gradient-based attacks in standard computer vision tasks like ImageNet.

At its core the Boundary Attack follows the decision boundary between adversarial and non-adversarial samples using a very simple rejection sampling algorithm in conjunction with a simple proposal distribution and a dynamic step-size adjustment inspired by Trust Region methods. Its basic operating principle— starting from a large perturbation and successively reducing it—inverts the logic of essentially all previous adversarial attacks. Besides being surprisingly simple, the Boundary attack is also extremely flexible in terms of the possible adversarial criteria and performs on par with gradient-based attacks on standard computer vision tasks in terms of the size of minimal perturbations. The mere fact that a simple constrained iid Gaussian distribution can serve as an effective proposal perturbation for each step of the Boundary attack is surprising and sheds light on the brittle information processing of current computer vision architectures. Nonetheless, there are many ways in which the Boundary attack can be made even more effective, in particular by learning a suitable proposal distribution for a given model or by conditioning the proposal distribution on the recent history of successful and unsuccessful proposals. Decision-based attacks will be highly relevant to assess the robustness of machine learning models and to highlight the security risks of closed-source machine learning systems like autonomous cars.

Paper 13

Title: Robust Physical-World Attacks on Machine Learning Models

Citation: *Evtimov, Ivan, et al. "Robust physical-world attacks on machine learning models." arXiv preprint arXiv:1707.08945 2.3 (2017): 4.*

Link: https://s3.observador.pt/wp-content/uploads/2017/08/08133934/1707-08945.pdf

Summary: Deep neural network-based classifiers are known to be vulnerable to adversarial examples that can fool them into misclassifying their input through the addition of small-magnitude perturbations. This paper proposes a new attack algorithm—Robust Physical Perturbations (RP2)— that generates perturbations by taking images under different conditions into account. This algorithm can create spatially constrained perturbations that mimic vandalism or art to reduce the likelihood of detection by a casual observer. It shows that adversarial examples generated by RP2 achieve high success rates under various conditions for real road sign recognition by using an evaluation methodology that captures physical world conditions. It evaluates two attacks, one that causes a Stop sign to be misclassified as a Speed Limit sign in 100% of the testing conditions, and one that causes a Right Turn sign to be misclassified as either a Stop or Added Lane sign in 100% of the testing conditions.

Previous algorithms assume that the inputs of DNNs can be modified digitally to achieve misclassification, but such an assumption is infeasible, as an attacker with control over DNN inputs can simply replace it with an input of his choice. Therefore, adversarial attack algorithms must apply perturbations physically, and in doing so, need to account for new challenges such as a changing viewpoint due to distances, camera angles, different lighting conditions, and occlusion of the sign. Furthermore, fabrication of a perturbation introduces a new source of error due to a limited color gamut in printers. It shows the use of RP2 to create two types of perturbations: subtle perturbations, which are small, undetectable changes to the entire sign, and camouflage perturbations, which are visible perturbations in the shape of graffiti or art. When the Stop sign was overlaid with a print out, subtle perturbations fooled the classifier 100% of the time under different physical conditions. When only the perturbations were added to the sign, the classifier was fooled by camouflage graffiti and art perturbations 66.7% and 100% of the time respectively under different physical conditions. Finally, when an untargeted poster printed camouflage perturbation was overlaid on a Right Turn sign, the classifier was fooled 100% of the time.

Paper 14

Title: Data Security Issues in Deep Learning: Attacks, Countermeasures, and Opportunities

Citation: *G. Xu, H. Li, H. Ren, K. Yang and R. H. Deng, "Data Security Issues in Deep Learning: Attacks, Countermeasures, and Opportunities," in IEEE Communications Magazine, vol. 57, no. 11, pp. 116-122, November 2019, doi: 10.1109/MCOM.001.1900091.*

Link: https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=5676&context=sis_research

Summary: This paper focuses on data security issues in deep learning. It investigates the potential threats and the latest countermeasures based on various underlying technologies, where the challenges and research opportunities on offense and defense are also discussed.

We focused on studying the defense against poisoning attacks which is explained in the paper as follows:

- Defense against Poisoning Attack: In general, one of the main ways to defend against poisoning attacks is to design efficient detection mechanisms, which can rapidly detect abnormal samples and eliminate these poisoned data during training.

Existing research on this-

- Method1: First use influence functions to trace and explain the correlation between prediction and training sets. The influence functions can be widely used for malicious data detection in poisoning attacks even in nonconvex and non-differentiable models.
- Method2: A defense scheme by constructing approximate upper bounds on the loss across a broad family of attacks. Further, design two efficient defense strategies called sphere defense and slab defense to remove outliers (i.e., data suspected of being injected by the adversary) that are outside the applicable set. In this way, the false data in the DNN model can be effectively detected and filtered.
- Method3: It uses key sharing protocols to protect the integrity of training samples, thereby preventing malicious adversaries from tampering with training samples and calculation results.

Then, it proposes SecureNet, the first verifiable and privacy-preserving prediction protocol to protect model integrity and user privacy in DNNs. It can significantly resist various security and privacy threats during the prediction process. The researchers simulate SecureNet under a real dataset, and the experimental results show the superior performance of SecureNet for detecting various integrity attacks against DNN models.

Paper 15

Title: Preventing Data Poisoning Attacks By Using Generative Models

Citation: *Aladag, Merve, Ferhat Ozgur Catak, and Ensar Gul. "Preventing Data Poisoning Attacks By Using Generative Models." 2019 1st International Informatics and Software Engineering Conference (UBMYK). IEEE, 2019.*

Link:https://www.ozgurcatak.org/files/papers/2019-data-poison.pdf

Summary: In this paper, a data poisoning attack towards classification method of machine learning models is conducted and a defense algorithm which makes machine learning models more robust against data poisoning attacks is also proposed. The authors have conducted data poisoning attacks on MNIST, a widely used character detection data set. Using the poisoned MNIST dataset, they built classification models which were more reliable by using a generative model such as AutoEncoder. Auto-encoder is a generative model of the artificial neural network that reproduces the data by learning the structure of the data with no labels. The structure and features of the data are learned with this model and the data is tried to be re-created.

An optimization based data poisoning attack which manipulated the training stage of the classification method from machine learning models was performed. Before the training phase of the classification model, manipulated data was added on the true data so that the model could learn the manipulated data as well. The auto-encoder model to make the classification models more robust to such attacks was then proposed, and by observing the classification performance, the authors showed that the model marked the manipulated data as it should have.

INTEGRATED SUMMARY

Papers which study existing work:

S.No.	Paper title	Work Done
1.	Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks(9)	A comprehensive analysis aimed to investigate the transferability of both test-time evasion and training-time poisoning attacks
2.	Adversarial Security Attacks and Perturbations on Machine Learning and Deep Learning Methods (11)	Basic knowledge on the machine learning and deep learning models and algorithms, as well as some of the relevant adversarial security attacks and perturbations
4.	A Survey on Security threats and Defensive Techniques of Machine Learning: A Data Driven view(6)	Emphasis on data distribution drift caused by adversarial samples and sensitive information violation problems in statistical machine learning.
5.	Machine Learning Security: Threats, Countermeasures, and Evaluations(5)	Covers all the aspects of machine learning security. From the training phase to the test phase, all types of attacks and defenses are reviewed in a systematic way.
6.	SoK: Security and Privacy in Machine Learning (3)	Study of a threat model that considers characteristics of the attack surface, adversarial goals, and possible defense and attack capabilities for it.

Papers which propose solutions:

S.No.	Paper title	Type of model targeted	Type of attack	Proposed Attack mechanism	Proposed Defence mechanism
1.	Robust Linear Regression Against Training Data Poisoning(7)	Linear regression	Data Poisoning	--	Trimmed principal component regression (T-PCR) algorithm
2.	Preventing Data Poisoning Attacks By Using Generative Models (15)	Classification	Data Poisoning	--	Auto-Encoder Model

3.	Mitigating poisoning attacks on machine learning models: A data provenance based approach (10)	General	Poisoning	--	Data provenance based defense
4.	Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning (1)	Regression	Poisoning	Statistical-based Poisoning Attack (StatP)	TRIM
5.	ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models (2)	General	Model and data independent membership inference attacks	--	Dropout, Model stacking
3.	Data Security Issues in Deep Learning: Attacks, Countermeasures, and Opportunities (14)	Deep learning	General	--	SecureNet - privacy-preserving prediction protocol to protect model integrity and user privacy in DNN

CONCLUSION

We have systematically analyzed the security issues of machine learning, focusing on existing attacks on machine learning systems, corresponding defenses or secure learning techniques, and security evaluation methods. Instead of focusing on one stage or one type of attack in the starting only, we tried to understand the basics of all kinds of attacks from the training phase to the test phase. Finally, we decided to focus on data poisoning attacks on linear regression models in detail and plan to work on making these models robust against poisoning attacks.

REFERENCES

1. *Jagielski, Matthew, et al. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning." 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 2018*
2. *Salem, Ahmed, et al. "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models." arXiv preprint arXiv:1806.01246 (2018).*
3. Papernot, Nicolas, et al. "SoK: Security and privacy in machine learning." *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2018.
4. *Ma, Yuxin, et al. "Explaining vulnerabilities to adversarial machine learning through visual analytics." IEEE transactions on visualization and computer graphics 26.1 (2019): 1075-1085.*
5. *Xue, Mingfu, et al. "Machine Learning Security: Threats, Countermeasures, and Evaluations." IEEE Access 8 (2020): 74720-74742.*
6. *Liu, Qiang, et al. "A survey on security threats and defensive techniques of machine learning: A data driven view." IEEE access 6 (2018): 12103-12117.*
7. *Liu, Chang, et al. "Robust linear regression against training data poisoning." Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017.*
8. *Zhang, Xuezhou, Xiaojin Zhu, and Laurent Lessard. "Online data poisoning attacks." Learning for Dynamics and Control. 2020.*
9. *Demontis, Ambra, et al. "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks." 28th {USENIX} Security Symposium ({USENIX} Security 19). 2019.*
10. *Baracaldo, Nathalie, et al. "Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach." (2017).*
11. *Siddiqi, Arif. "Adversarial security attacks and perturbations on machine learning and deep learning methods." arXiv preprint arXiv:1907.07291 (2019).*
12. *Brendel, Wieland, Jonas Rauber, and Matthias Bethge. "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models." arXiv preprint arXiv:1712.04248 (2017).*
13. *Evtimov, Ivan, et al. "Robust physical-world attacks on machine learning models." arXiv preprint arXiv:1707.08945 2.3 (2017): 4.*
14. *G. Xu, H. Li, H. Ren, K. Yang and R. H. Deng, "Data Security Issues in Deep Learning: Attacks, Countermeasures, and Opportunities," in IEEE Communications Magazine, vol. 57, no. 11, pp. 116-122, November 2019, doi: 10.1109/MCOM.001.1900091.*