Introduction

This is the small introduction I have written for now. This doesn’t represent the last version, as it can be changed and updated along the way. My supervisor advised me to write the introduction at the end, because that is the time I will be most familiar with my paper, and can write more concisely and correctly about what my paper represents. So that’s why, I will come back to this introduction again, and edit it later. :

In the contemporary landscape of cybersecurity, machine learning (ML) has emerged as a pivotal technology for malware detection. The ability of ML models to discern patterns and identify potential threats has been instrumental in bolstering defenses against an ever-evolving array of malware. However, the robustness of these models is increasingly challenged by sophisticated adversarial strategies, particularly backdoor attacks. These attacks represent a significant threat to the integrity and effectiveness of ML-based malware detection systems.

Backdoor attacks against machine learning models are insidiously designed to exploit the learning process itself. By embedding strategically crafted training samples into the training dataset, adversaries can manipulate the model's decision-making boundaries. This manipulation is particularly concerning in the context of malware detection. ML models trained to identify malware binaries can be compromised to overlook or misclassify these threats, especially when the backdoor triggers are activated. This subversion is not just a theoretical possibility but a practical concern that can have far-reaching consequences.

The implications of successful backdoor attacks on ML-based malware detection systems are profound. As users and organizations increasingly rely on these systems to detect and neutralize malware threats, the introduction of backdoors can significantly undermine their trustworthiness. The insidious nature of these attacks means that they can lead to false negatives, where malware is erroneously classified as benign. This failure in detection can have severe repercussions, ranging from data breaches to the compromise of critical infrastructure, especially in the context of the Internet of Things (IoT), which is becoming increasingly prevalent.

This thesis aims to delve into the realm of backdoor attacks in the context of ML-based malware detection. The primary objectives of this study are multifaceted. Firstly, the thesis will provide a comprehensive overview of various backdoor attacks and their corresponding defense mechanisms in the field of machine learning, with a specific focus on malware detection. This overview is crucial to understand the current landscape of threats and the state-of-the-art in defense strategies.

Secondly, the thesis will take a hands-on approach by designing and implementing a backdoor attack targeted at a machine learning model responsible for detecting IoT malware. This practical implementation will serve to evaluate the efficacy of such attacks and measure the extent to which they can degrade the accuracy of malware detection systems.

Lastly, the study will not stop at merely highlighting the vulnerabilities but will venture into the development of a robust defense mechanism against the designed backdoor attack. This defense mechanism will be rigorously evaluated to assess its effectiveness in mitigating the threat posed by the backdoor attack, particularly in terms of its impact on the accuracy and operational cost compared to the original ML model.

In conclusion, this thesis endeavors to contribute valuable insights into the vulnerabilities of ML-based malware detection systems to backdoor attacks and pave the way for more resilient defense strategies. By bridging the gap between theoretical understanding and practical application, this study aims to fortify the defenses against sophisticated cyber threats in an increasingly interconnected world.

Related work

There is multiple research paper that touches upon the topics that is present in my paper. I have read and analyzed them. Right now, I need to decide on which ones describe and are the closest to my own paper. But to decide that I think I still need to finish writing my paper, and after the conclusion to decide on the best related works for this paper. But still, here is the overview of what I have gotten from the papers.

# Intelligent and behavioural-based detection of malware in IoT spectrum sensors

The general context of the paper is the growing number of Cyber-Physical Systems (CPS) in industrial environments, such as smart factories and grids. These CPS often rely on IoT spectrum sensors to monitor the environment and collect data. However, IoT spectrum sensors are vulnerable to malware attacks, which can disrupt their operation and compromise the security of the CPS.

The authors of the paper propose a detection framework that uses device behavioral fingerprinting and machine learning to detect malware infections in IoT spectrum sensors. The framework works by first creating a fingerprint of the sensor's normal behavior. This fingerprint is based on a variety of sensor data, such as CPU usage, memory usage, network traffic, and file system activity.

Once the fingerprint has been created, the framework monitors the sensor's behavior and identifies anomalies that may indicate malware infection. The framework can also be trained on a dataset of known malware samples to learn the behavioral characteristics of different types of malware.

If the framework detects an anomaly, it alerts the security administrator and/or takes other actions, such as quarantining the sensor or removing the malware.

The key findings of the paper are as follows:

- The proposed framework is effective in detecting and classifying malicious behaviors in IoT spectrum sensors.

- The framework was able to detect ten recent malware samples with a true positive rate of 0.88–0.90 and an F1-score of 0.94–0.96.

- The framework is relatively lightweight and can be implemented on real-world IoT spectrum sensors.

The authors conclude that intelligent and behavioral-based detection techniques are a promising approach to detecting malware in IoT spectrum sensors.

\*\*Implications\*\*

The findings of this paper have important implications for the security of CPS in industrial environments. The proposed detection framework can help to protect IoT spectrum sensors from malware attacks and improve the overall security of CPS.

The framework is also relatively lightweight and can be implemented on real-world IoT spectrum sensors. This makes it a practical solution for deploying in industrial environments.

Overall, the paper makes a significant contribution to the field of IoT security by proposing a novel and effective detection framework for malware in IoT spectrum sensors.

# Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review

\*\*General Context:\*\*

Backdoor attacks are a serious threat to deep learning models, which are becoming increasingly widely used in a variety of applications, including image classification, object detection, and natural language processing. In a backdoor attack, an attacker inserts a hidden trigger into a deep learning model that causes the model to mispredict when the trigger is present in the input data. This can be done in a variety of ways, such as by adding a small perturbation to the model's weights or by modifying the model's training data.

Backdoor attacks can be used to steal sensitive data, disrupt critical infrastructure, or even cause physical harm. For example, an attacker could backdoor a deep learning model used in a self-driving car to cause the car to crash when the attacker activates the trigger.

\*\*Abstract:\*\*

This paper provides a comprehensive review of backdoor attacks and countermeasures on deep learning. The authors first categorize backdoor attacks according to the attacker's capability and the affected stage of the machine learning pipeline. They then review the state-of-the-art backdoor attack and countermeasure techniques, and compare and analyze their advantages and disadvantages.

Finally, the authors discuss the challenges and future research directions in backdoor attack detection and defense.

\*\*Key Findings:\*\*

- Backdoor attacks can be launched at any stage of the machine learning pipeline, from data collection to model deployment.

- There are a variety of backdoor attack techniques, and new techniques are being developed all the time.

- Backdoor attacks are becoming increasingly sophisticated and difficult to detect.

- There is no single defense that can prevent all backdoor attacks. However, there are a number of countermeasure techniques that can be used to mitigate the risk of backdoor attacks.

The authors conclude that the research on backdoor defense is far behind the attack, and there is a need for more effective and practical countermeasures.

\*\*Additional Thoughts:\*\*

Backdoor attacks are a serious threat to deep learning models, but there is a growing awareness of this threat and researchers are working on developing new and better countermeasures. It is important for organizations that use deep learning models to be aware of the risks and to take steps to mitigate those risks.

# Backdoor Attack on Machine Learning Based Android Malware Detectors

\*\*General Context:\*\*

Machine learning (ML) has become a widely used technology for malware detection, including on the Android platform. However, ML models are vulnerable to poisoning attacks, which can be used to embed malicious backdoors into the models. This can allow attackers to evade detection and deploy malware undetected.

\*\*Abstraction:\*\*

This paper proposes a new backdoor attack against ML-based Android malware detectors. The attack is stealthy and requires only a small number of poisoned training samples. The attack is also effective against a variety of existing malware detectors.

\*\*Key Findings:\*\*

The key findings of the paper are as follows:

- Backdoor attacks can be launched against ML-based Android malware detectors without access to the training data.

- A small number of poisoned training samples are sufficient to create an effective backdoor.

- The proposed backdoor attack is effective against a variety of existing malware detectors.

The paper also demonstrates the effectiveness of the proposed attack on four typical malware detectors that have been widely discussed in academia. The evaluation shows that the proposed backdoor attack achieves up to 99% evasion rate over 750 malware samples. Moreover, the above successful attack is realised by a small size of triggers (only four features) and a very low data poisoning rate (0.3%).

\*\*Implications:\*\*

The findings of this paper have important implications for the security of ML-based malware detectors. It shows that these detectors are vulnerable to backdoor attacks, even when the attackers do not have access to the training data. This highlights the need for new defenses to protect ML-based malware detectors from these attacks.

The authors also propose some promising approaches to improve backdoor defenses, such as using adversarial training and anomaly detection.

# Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers

\*\*Context\*\*

Machine learning (ML) classifiers are widely used to detect malware, but they are also vulnerable to backdoor poisoning attacks. In a backdoor poisoning attack, an attacker injects malicious samples into the training dataset of a classifier in order to manipulate its predictions. This can be done even if the attacker does not have control over the labeling process, which is known as a "clean label" attack.

\*\*Abstraction\*\*

The paper "Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers" proposes a new method for crafting effective backdoor triggers in a model-agnostic fashion. The method leverages techniques from explainable machine learning (XAI) to identify the features and values that are most important to the classifier's decision-making process. The attacker can then craft backdoor triggers that are specifically targeted to these features and values.

\*\*Key Findings\*\*

The authors evaluate their proposed method against a diverse set of malware classifiers and datasets. They show that their method is effective in generating backdoor triggers that can deceive classifiers with high accuracy. The authors also evaluate the effect of various constraints imposed on the attacker, such as the number of backdoored samples that can be injected and the visibility of the backdoor triggers. They show that their method is still effective even under these constraints.

\*\*Implications\*\*

The findings of this paper have important implications for the security of ML-based malware classifiers. It shows that even classifiers that are trained on clean labels are vulnerable to backdoor poisoning attacks. The authors also discuss potential defensive strategies, but they show that it is difficult to completely defend against these attacks.

Overall, this paper is a valuable contribution to the field of malware classification security. It raises awareness of the threat of backdoor poisoning attacks and provides a new method for crafting effective backdoor triggers.

# Jigsaw Puzzle: Selective Backdoor Attack to Subvert Malware Classifiers

\*\*General Context\*\*

Malware classifiers are increasingly being used to protect organizations from malicious software. However, these classifiers are vulnerable to backdoor attacks, which allow attackers to sneak malware past the classifier undetected.

\*\*Abstract\*\*

The Jigsaw Puzzle attack is a new type of backdoor attack that is more stealthy than existing attacks. It works by exploiting the fact that malware authors have little to no incentive to protect any other authors' malware but their own.

The Jigsaw Puzzle attack works by first learning a trigger that is complementary to the latent patterns of the malware author's samples. This trigger is then inserted into the malware sample, but only in a way that activates the backdoor when the trigger and the latent pattern are pieced together.

\*\*Key Findings\*\*

The Jigsaw Puzzle attack is effective at evading state-of-the-art malware classifiers. It is also difficult to detect, as the trigger is typically very small and subtle.

\*\*Problems or Shortcomings\*\*

The main problem with the Jigsaw Puzzle attack is that it requires the attacker to have access to a set of malware samples from the target malware author. This can be a difficult requirement to meet in practice.

Another problem with the Jigsaw Puzzle attack is that it may not be effective against all malware classifiers. Some classifiers may be able to detect the trigger, even if it is small and subtle.

\*\*Overall, the Jigsaw Puzzle attack is a new and stealthy type of backdoor attack that poses a serious threat to malware classifiers. It is important for malware classifier developers to be aware of this attack and to develop defenses against it.\*\*

\*\*Additional Thoughts\*\*

The Jigsaw Puzzle attack is a reminder that malware authors are constantly developing new techniques to evade malware classifiers. It is important for organizations to have a layered security approach that includes multiple malware detection and prevention techniques.

# Universal backdoor attack on deep neural networks for malware detection

\*\*General context:\*\*

Deep neural networks (DNNs) are increasingly being used in malware detection systems. However, DNNs are also vulnerable to backdoor attacks, where an adversary can insert a hidden trigger into the model that causes it to misclassify certain inputs. Backdoor attacks on malware detection systems can be particularly dangerous, as they can allow attackers to bypass security measures and introduce malware into a system undetected.

\*\*Abstract:\*\*

In this paper, the authors propose a universal backdoor attack on deep neural networks for malware detection. The attack is based on the observation that backdoor triggers can be generated in a transferable way, meaning that triggers generated for one model can be used to attack other models. The authors exploit this transferability to generate a universal backdoor trigger that can be used to attack a variety of malware detection models.

\*\*Key findings:\*\*

The authors evaluate their attack on three benchmark malware detection models and show that it is able to achieve high success rates (up to 99%) in misclassifying malware samples. The authors also show that their attack is robust to various defenses, such as adversarial training and trigger detection.

\*\*Conclusion:\*\*

The authors' findings demonstrate the vulnerability of malware detection systems to backdoor attacks. They also highlight the need for new defenses that are more robust to transferable backdoor triggers.

\*\*Additional thoughts:\*\*

The authors' attack is a significant advance in the field of backdoor attacks on deep neural networks. It is the first attack to show that universal backdoor triggers can be generated and used to attack a variety of malware detection models. The authors' findings also highlight the need for new defenses against backdoor attacks that are more robust to transferable triggers.

# Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning

\*\*General Context\*\*

Deep learning systems are increasingly being used in a wide range of applications, including image classification, object detection, and natural language processing. However, deep learning systems are also vulnerable to adversarial attacks, where attackers manipulate the inputs to the system in order to cause it to misbehave.

One type of adversarial attack is called a \*\*backdoor attack\*\*, where the attacker aims to create a backdoor into the system that allows them to control the system's output. Backdoor attacks can be particularly dangerous in safety-critical applications, such as self-driving cars and medical devices.

\*\*Abstract\*\*

In the paper "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning", the authors propose a new type of backdoor attack called a \*\*backdoor poisoning attack\*\*. In a backdoor poisoning attack, the attacker injects a small number of carefully crafted poisoned samples into the training dataset of the deep learning system. These poisoned samples are designed to train the system to misclassify certain inputs as a target label specified by the attacker.

The authors show that backdoor poisoning attacks can be very effective, even when the attacker has only limited access to the training dataset. For example, the authors were able to create a backdoor in a deep learning system for image classification that could misclassify images of cats as dogs with an accuracy of over 90%.

\*\*Key Findings\*\*

The key findings of the paper are as follows:

- Backdoor poisoning attacks are a new and effective way to attack deep learning systems.

- Backdoor poisoning attacks can be carried out with only limited access to the training dataset.

- Backdoor poisoning attacks can be used to create backdoors that are physically implementable, meaning that they can be used to attack deep learning systems that are deployed in the real world.

The authors also propose a number of defense strategies against backdoor poisoning attacks, but they note that these defense strategies are not yet robust enough and that further research is needed.

\*\*Implications\*\*

The findings of this paper have important implications for the development and deployment of deep learning systems. It is important to be aware of the threat of backdoor poisoning attacks and to take steps to mitigate this threat. One way to mitigate the threat is to use a variety of defense strategies, such as anomaly detection and input validation.

It is also important to note that backdoor poisoning attacks are just one type of adversarial attack on deep learning systems. There are many other types of adversarial attacks, and it is important to be aware of all of these threats when developing and deploying deep learning systems.