Adversarial Attacks Can Deceive AI Systems, Leading to Misclassification or Incorrect Decisions

Petar Radanliev * and Omar Santos

Posted Date: 29 September 2023

doi: 10.20944/preprints202309.2064.v1

Keywords: adversarial attacks; artificial intelligence; machine learning; defense mechanisms; system integrity; model vulnerabilities; advanced attack techniques; Fast Gradient Sign Method (FGSM); Carlini and Wagner Attack (C&W); targeted attacks; non-targeted attacks; blackbox attacks

Article

Adversarial Attacks Can Deceive AI Systems, Leading to Misclassification or Incorrect Decisions

Petar Radanliev * and Omar Santos

Department of Computer Sciences, University of Oxford, England, United Kingdom Cisco Systems, Research Triangle Park, North Carolina, United States
*petar.radanliev@eng.ox.ac.uk

Abstract: This comprehensive analysis thoroughly examines the topic of adversarial attacks in artificial intelligence (AI), providing a detailed overview of the various methods used to compromise machine learning models. It explores different attack techniques, ranging from the simple Fast Gradient Sign Method (FGSM) to the intricate Carlini and Wagner Attack (C&W), emphasising the wide range of adversarial approaches and their intended goals. The discussion also distinguishes between targeted and non-targeted attacks, highlighting the adaptability and versatility of these malicious efforts. Additionally, the study delves into the realm of black-box attacks, revealing the capability of adversarial strategies to compromise models even with limited knowledge. Real-life examples illustrate the tangible consequences and potential dangers of adversarial attacks in various fields such as self-driving cars, multimedia, and voice assistants. These cases highlight the difficulties in ensuring the legitimacy and dependability of AI-powered technologies and programs. The article stresses the importance of ongoing research and innovation to address the growing difficulties posed by advanced methods like deepfakes and disguised voice commands in preserving the security of AI systems. This study provides valuable insights on how different adversarial strategies and defence mechanisms interact within AI. The results emphasise the urgent need for stronger and more secure AI models to combat the increasing number of adversarial threats in today's AI landscape. These findings can guide future research and innovations in developing more resilient AI technologies that can better withstand various adversarial vulnerabilities and challenges.

Keywords: adversarial attacks; artificial intelligence; machine learning; defense mechanisms; system integrity; model vulnerabilities; advanced attack techniques; Fast Gradient Sign Method (FGSM); Carlini and Wagner Attack (C&W); targeted attacks; non-targeted attacks; blackbox attacks

1. Introduction to Adversarial Attacks

As we continue to rely on Artificial Intelligence (AI) applications in critical areas such as medical imaging [1], autonomous driving [2], and security [3], it is crucial to be mindful of the potential for adversarial cyber-attacks. These attacks involve subtle manipulations to input data that can deceive an AI model, resulting in erroneous decisions and outcomes. An in-depth understanding of adversarial attacks is therefore essential for a multitude of reasons [4].

It is crucial to recognise that effectively safeguarding generative AI models through "red teaming" is a complex undertaking. Security was not a top priority during the initial development of these models, and it has only been given due consideration in contemporary AI models. This disregard for security is apparent when data scientists merge intricate collections of images and text to teach AI algorithms.

To prevent harmful attacks impacting important applications, it's important to have a thorough understanding of potential vulnerabilities and create strong security measures. This paper explores the nuances of adversarial cyber-attacks, highlighting their consequences and suggesting effective ways to protect against them in the field of artificial intelligence.

2. Generative AI Areas of Cyber Risk

 doi:10.20944/preprints202309.2064.v1

In this section, the paper provides a breakdown of some major risk areas in security for crucial applications.

There are concerns that autonomous vehicles may misinterpret traffic signs due to adversarial manipulation [5]. For example, a stop sign with an adversarial sticker could be mistaken for a yield sign by an AI-powered vehicle. Additionally, AI is being utilised more frequently in healthcare for diagnostics with medical imagery. However, a hostile attack could result in a misdiagnosis, which may impact patient treatment plans.

Adversarial attacks can be used to bypass facial or voice recognition systems, which can lead to unauthorised access to devices or sensitive information [6]. This poses serious security implications for biometrics.

Security implications of biometrics include circumventing facial or voice recognition systems, and allowing unauthorised access to devices or sensitive areas. Criminals can manipulate surveillance footage to avoid detection or misdirect investigations.

Biometric security systems are not foolproof, as malicious attacks can bypass facial or voice recognition systems, allowing unauthorised access to devices or sensitive areas. Additionally, criminals can manipulate surveillance footage to evade detection or misdirect investigations.

In today's digital era, the accuracy and reliability of information are at risk due to the prevalence of deep fakes and multimedia. Adversarial attacks can create or manipulate content to produce realistic but entirely false media. This can lead to misinformation, defamation, and even political manipulation. Additionally, digital assistants that function through voice commands are also susceptible to unwanted actions that can compromise user privacy and data integrity. It is vital to recognise these potential threats and take necessary precautions to safeguard digital information.

Furthermore, AI-driven decision-making may yield false positives or negatives in fields such as law enforcement, legal judgments, or recruitment, with far-reaching personal and societal implications.

There are concerns about the use of AI in defence and strategic systems by various nations. This could lead to cyber warfare, espionage, and other state-level threats taking advantage of any weaknesses. Therefore, it is crucial to continue researching and developing AI to understand adversarial attacks and create strong and resilient AI models. This will help advance the field of AI research.

The use of AI-driven products or services by businesses can have negative consequences if their systems are vulnerable to adversarial attacks. This could result in reputational and financial loss. Additionally, the manipulation of AI-driven recommendation systems on social media or e-commerce platforms can impact public opinion, consumer behaviour, and even election outcomes. As AI continues to be integrated into various sectors of society, it is crucial to understand and defend against adversarial attacks to ensure the safety, security, and effectiveness of these systems. Ongoing research and awareness in this field are necessary to maximise the benefits of AI while minimising potential risks.

3. Where Is the Data Coming From?

According to OpenAI, the creator of ChatGPT, they do not utilise real-time web scraping or gather data from social media on a continuous basis to update the model's responses. Instead, ChatGPT was trained on a static dataset that was collected until September 2021. OpenAI affirms that once the model has been trained, no new information is added based on real-time events or newly published data.

According to OpenAI, the creators of ChatGPT, the model's responses are not updated in real-time by scraping the web or collecting social media data continuously. Instead, ChatGPT was trained on a static dataset that was collected up until September 2021. OpenAI emphasises that once the model is trained, it is not updated in response to current events or newly published data.

After being trained, ChatGPT doesn't actively search or extract current information from the internet, social media, or other databases. Therefore, it cannot provide the latest updates on events or trending topics since its last training. User interactions do contribute to enhancing the model's performance in general by generating multiple versions. However, specific user interactions or feedback aren't instantly integrated into the model. OpenAI usually improves its models based on widespread feedback and identified issues.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

After being trained, ChatGPT does not continuously collect information from the Internet, social media, or any other databases. Therefore, it may not be able to provide up-to-date information on current events or trending topics. While user feedback contributes to the improvement of the model in a general sense through multiple iterations, specific feedback or interactions from users are not immediately integrated into the model. However, OpenAI regularly updates its models based on extensive user feedback and identified issues.

Once trained, ChatGPT does not actively scrape or retrieve current data from the Internet, social media platforms, or other databases. It is unable to provide timely information on current events or trending topics after its most recent training data.

While user interactions contribute to model improvement, specific feedback is not immediately integrated. OpenAI's models iterate based on user feedback and observed issues.

3.1. Can Generative AI and ChatGPT Be Used for Data Collection?

AI Chatbots built with Generative AI, can be utilised to collate various types of data, including customer preferences, feedback, and purchasing habits. Moreover, AI Chatbots can be employed to gather data on customer demographics, such as age, gender, and location. Businesses can use this data to devise more tailored marketing strategies and enhance their overall customer experience.

Scalability is a significant advantage of using AI chatbots for data collection. AI Chatbots can interact with multiple customers simultaneously, enabling rapid data collection from a large number of customers. Additionally, AI Chatbots can be available around the clock, making it handy for customers to offer feedback whenever they wish.

However, it's crucial for businesses to ensure they adhere to data protection regulations when employing AI Chatbots for data collection. It's vital to inform customers about what data is being gathered and its intended use. Moreover, businesses must guarantee the protection of customer data from unauthorised access or misuse.

In essence, AI Chatbots can be an effective tool for businesses seeking to collect customer information. With its proficiency in natural language processing and scalability, it provides an effective method for gaining priceless customer insights. Nonetheless, it is crucial for businesses to adhere to data protection regulations and protect customer information from abuse.

ChatGPT in its standard configuration is not intended to collect or store users' personal information. It is designed to forget the user's personal information after the conversation has concluded, ensuring user privacy.

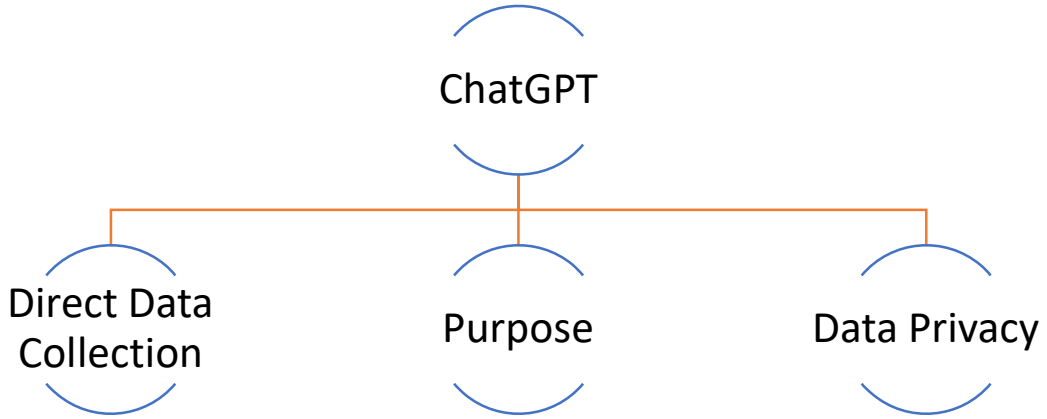


Figure 1. ChatGPT standard configuration.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

ChatGPT is designed to generate natural-sounding text in response to user input. It is not inherently a tool for business analytics, customer data collection, or similar purposes; however, with the right frameworks, chatbot systems can be designed for these purposes.

ChatGPT is designed to generate text that resembles human speech in response to user input. It is not inherently a tool for business analytics, customer data collection, or similar purposes; however, chatbot systems can be designed around it for these purposes, given the appropriate frameworks.

While the general concepts of data collection are applicable to AI chatbots in general, they are not entirely applicable to ChatGPT's usage and design. Organisations seeking to use a version of ChatGPT or any other AI tool for data collection must consider local and international data privacy regulations and will likely need to develop a customised implementation to handle and store data securely and legally.

4. Types of Adversarial Attacks

The field of machine learning, specifically deep learning, deals with the idea of adversarial attacks [7]. This means creating inputs that can purposely mislead a model into making incorrect predictions or classifications [8]. These inputs are known as adversarial examples and can be hard to detect [9], especially when it comes to image-based datasets.

There are several types of adversarial attacks, but the three primary categories are:

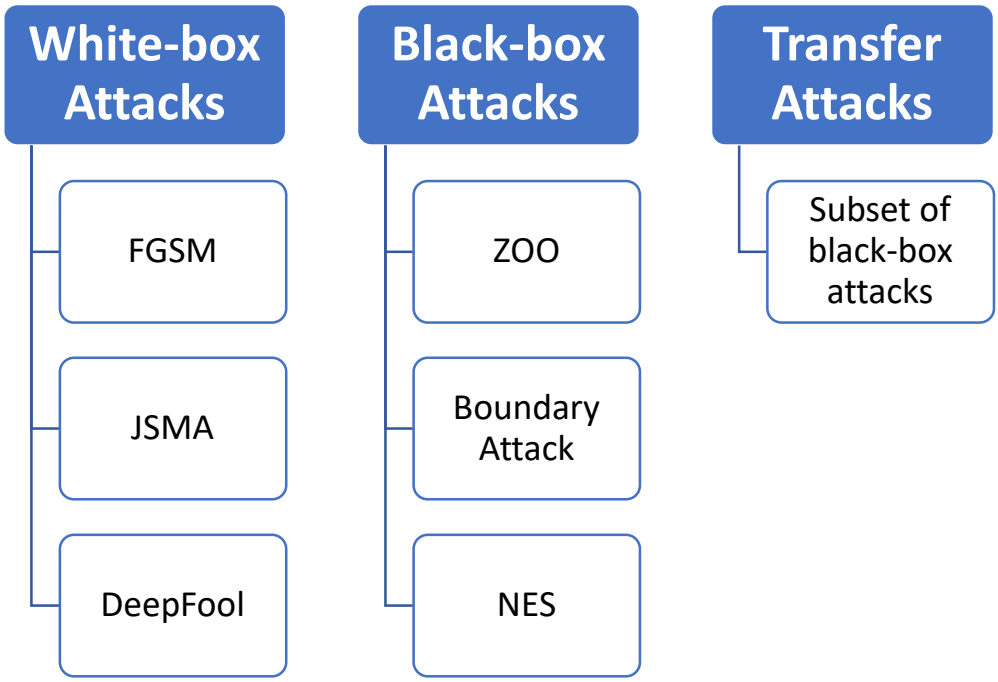


Figure 2. Types of Adversarial Attacks.

In white-box attacks, the adversary has complete knowledge of the target model. This includes the system's architecture, trained parameters, and in some cases, training data. With this comprehensive understanding, the adversary creates adversarial examples that are specifically designed to mislead the target model. Examples of white-box attacks include:

- Fast Gradient Sign Method (FGSM)
- Jacobian-based Saliency Map Attack (JSMA)
- DeepFool

Black-box Attacks: Here, the attacker is unaware of the inner workings of the model. They only have access to the input and output of the model. This input-output data is used by the attacker to generate adversarial examples. Despite limited knowledge, black-box attacks can be quite potent due to the fact that models, particularly deep neural networks, can share vulnerabilities across architectures. Important here is transferability, as an adversarial example created for one model can deceive another. Some black-box attack methods include:

 doi:10.20944/preprints202309.2064.v1

- Zeroth Order Optimisation (ZOO)
- Boundary Attack
- Natural Evolution Strategies (NES)

Transfer Attacks refer to one type of black-box attacks where an adversary generates an adversarial input by accessing a different model, known as the surrogate model, and then uses it to attack the target model. The idea behind this is that models trained on similar tasks tend to share similar vulnerabilities, making it possible for adversarial examples to be transferred between them.

Attacks on models can be classified based on their goals. Targeted attacks are aimed at making the model generate a specific incorrect outcome, while untargeted attacks are focused on causing the model to make a mistake or be incorrect without specifying the desired incorrect output.

When it comes to safeguarding against malicious attacks, there are a range of methods available, each with their own unique benefits. These techniques include adversarial training, which involves training a model on adversarial examples to improve its robustness, defensive distillation, which reduces the amount of information available to attackers, and gradient masking, which obscures the gradients of a model to prevent attackers from exploiting them. By employing these techniques, individuals and Organisations can better protect their systems and data from potential threats.

5. Examples of Adversarial Attacks That Can Deceive AI Systems, Leading to Misclassification or Incorrect Decisions

Adversarial attacks are prime examples of deceptive manoeuvres that can mislead AI systems, causing them to make incorrect classifications or decisions [4]. Several examples of such attacks are provided below.

Perturbation Attacks involve making small, often undetectable changes to input data in order to cause the AI system to classify it incorrectly. In critical areas like medical imaging, misclassification can lead to missed diagnoses and incorrect treatment, making these attacks especially harmful.

Poisoning attacks occur when an attacker injects corrupt data into the machine learning training dataset. This can greatly affect the learning process of the model. Once the corrupted model is deployed, it may make inaccurate and unpredictable predictions or classifications, ultimately compromising the overall integrity of the AI system.

Evasion attacks happen when the attacker changes the input data during testing to cause the model to make incorrect predictions or classifications. These changes are meant to be undetectable but have a major effect on the model's results.

Evasion attacks happen when the attacker alters the input data during testing to make the model produce wrong predictions or classifications. These alterations are meant to be undetectable but can greatly affect the model's output.

Trojan Attacks Trojan attacks involve embedding a malicious function within the training phase of the model. This malicious functionality is subsequently activated by particular inputs when the model is deployed, causing it to make incorrect decisions or classifications.

Backdoor attacks involve manipulating the AI model during its training phase by inserting a specific backdoor pattern. This pattern will cause the model to generate incorrect outputs when it encounters the input data pattern, allowing attackers to manipulate the model's behaviour. These attacks are sophisticated techniques that deceive AI systems, and they require advanced defence mechanisms and security protocols to ensure the dependability and robustness of AI applications in various domains.

5.1. Jacobian-Based Saliency Map Attack

As the Fast Gradient Sign Method and the Carlini & Wagner Attack are described in greater detail in Section 8, the focus of this article is on the remaining adversarial attacks. The phrase "Jacobian-based Saliency Map Attack" refers to a technique for conducting adversarial attacks against neural networks. Let's dissect it step by step.

Adversarial Attack: When neural networks are attacked adversarially, it means that they are exposed to deliberately crafted input data that is designed to cause errors in their operation. These specially designed inputs are called "adversarial examples" and can be difficult for humans to

 doi:10.20944/preprints202309.2064.v1

distinguish from regular inputs. However, the neural network may produce incorrect results when presented with these inputs.

A saliency map is a diagram that shows the important features of an input that have the most impact on a model's output. In neural networks, saliency maps can help identify which parts of an input image the model focuses on when making a decision.

The Jacobian matrix is a representation of how slight modifications in the input can impact changes in the output. In relation to neural networks, the Jacobian matrix can provide insights into how outputs (such as the likelihood of each class in a classification task) react to minor adjustments in inputs.

By utilising the Jacobian matrix, the Jacobian-based Saliency Map Attack can identify the sections of the input that have the most significant impact on the output when altered. This approach produces adversarial examples by modifying the most "sensitive" parts of the input, which are determined by the Jacobian-based saliency map, to deceive the neural network.

To put it simply, this technique involves identifying the specific parts of the input (such as certain pixels in an image) that can be altered in order to deceive a neural network most efficiently. These alterations are then made to create an adversarial example.

5.2. Deepfool Attack

It is a method for systematically determining which parts of the input (e.g., which pixels in an image) should be modified slightly to fool a neural network the most effectively, and then introducing those modifications to generate an adversarial example.

DeepFool attack's main objective is to identify the smallest perturbation that, when added to the input, causes a deep learning model to misclassify the input.

The DeepFool method differs from other adversarial attack techniques because it doesn't rely on gradient information to make binary decisions about how to adjust pixels or features. Instead, it iteratively linearises the classifier's decision boundary and calculates the minimum perturbation necessary to cross this linearised boundary. This process is repeated until the input is misclassified.

Compared to other types of attacks, the DeepFool attack is more effective at finding smaller changes in data due to its iterative approach. This results in more subtle and harder to detect modifications to the input data compared to other malicious attacks.

DeepFool exposes vulnerabilities in deep neural networks, particularly in scenarios requiring security. Despite good performance in regular situations, adversarial attacks like DeepFool can unveil model weaknesses.

To summarise, the DeepFool technique is an effective form of attack that identifies the slightest alterations required to cause incorrect classification of input. This approach can shed light on potential weaknesses in a model's design and implementation.

5.3. Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of artificial intelligence algorithms primarily employed in unsupervised machine learning. They were introduced by Ian Goodfellow and his colleagues in 2014, and their potential to generate synthetic data, particularly images, has since made them a popular topic of study. GANs are based on two neural networks, the Generator and the Discriminator, which are simultaneously trained through a game-like process.

Here is a comprehensive explanation of the intricacies involved in these systems. Firstly, the Generator network functions by utilising random noise as an input to produce data, usually in the form of images. On the other hand, the Discriminator network plays a crucial role in discerning between authentic data and generated data, which is then used as input. This process is essential in facilitating the network's learning and enhancing its capacity to create genuine data.

Throughout the "game" or training process, a series of steps are taken to improve the generator's ability to produce accurate data. Firstly, the Generator creates a piece of data. Subsequently, the Discriminator evaluates the data and provides feedback on whether it believes the data is from the real dataset or generated. Finally, the Generator analyses the feedback and attempts to enhance its generation process to better deceive the Discriminator in the future.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

This adversarial process continues until the Generator becomes so proficient at generating data that the Discriminator cannot distinguish between real and generated data, or until they reach a state of equilibrium.

The versatility of GANs is remarkable, as they can create highly realistic images, artistic designs, and lifelike voices. Moreover, they can aid in the crucial area of drug discovery. The potential for innovation and advancement with this technology is truly limitless.

Generating realistic images with Generative Adversarial Networks (GANs) can be a daunting task, given their sensitivity to hyperparameters, network architecture, and the common issue of mode collapse. Consequently, GANs may produce outputs that lack variety, making it difficult to achieve the desired results.

How Do They Relate to Cyber-Attacks: GAN Relation to Cyber-Attacks on AI Models

Adversarial Attacks involve manipulating input data to cause the model to make an error – see **Error! Reference source not found.**. In GANs, the Generator creates these adversarial examples to deceive the Discriminator. This phenomenon has been studied to identify vulnerabilities in AI models and develop ways to prevent them.

In the text below, we discuss all of the cyber-attacks listed in **Error! Reference source not found.**.

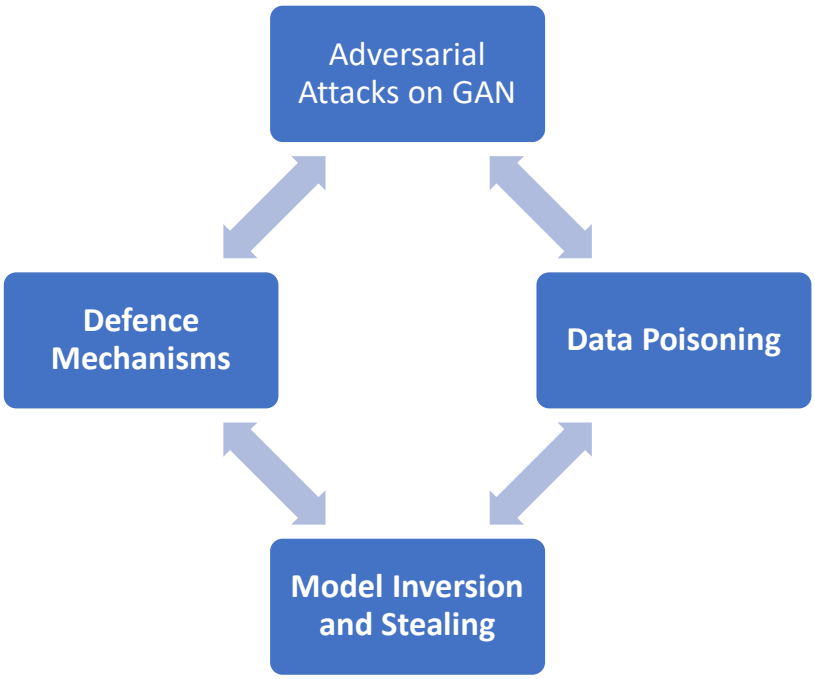


Figure 3. GAN relation to cyber-attacks on AI models.

One potential method for conducting a data poisoning attack is through the use of Generative Adversarial Networks (GANs) to generate synthetic data. By adding this data to the training set of a model, the behaviour of the model can be influenced in a way that benefits the attacker. This can be a serious threat to the integrity and accuracy of the model's output.

It has been noted that Generative Adversarial Networks (GANs) can be utilised by malicious individuals who have obtained access to a particular model. This could lead to the reverse-engineering of trained models and ultimately result in the duplication of proprietary models. It is important to take necessary precautions to prevent such unauthorised access and protect sensitive information.

One effective method to enhance the robustness of AI models against potential attacks is through adversarial training. This approach involves training these models using adversarial examples, which are purposely designed inputs aimed at exploiting vulnerabilities in the system. By exposing the AI to these adversarial examples during the training process, it can learn to detect and resist potential attacks, making it more secure and reliable.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

In summary, while Generative Adversarial Networks (GANs) are not inherently harmful, they can be exploited to carry out cyberattacks against AI models. This is due to their unique capability of generating data and adversarial examples. However, it is important to note that GANs also have the potential to be leveraged as a defence mechanism to fortify the security and stability of models. Therefore, it is crucial to carefully consider the potential risks and benefits of utilising GANs in the context of AI model development and deployment.

5.4. Spatial Transformation Attack

Spatial Transformation Attacks (STAs) are an adversarial attack type that targets artificial intelligence (AI) models, specifically deep learning models like neural networks. Adversarial attacks involve subtly modifying the input data to AI models in order to trick the model into making erroneous predictions or classifications, without altering the input's semantics for human observers.

In **Error! Reference source not found.**, we can see an overview of Spatial Transformation Attacks.

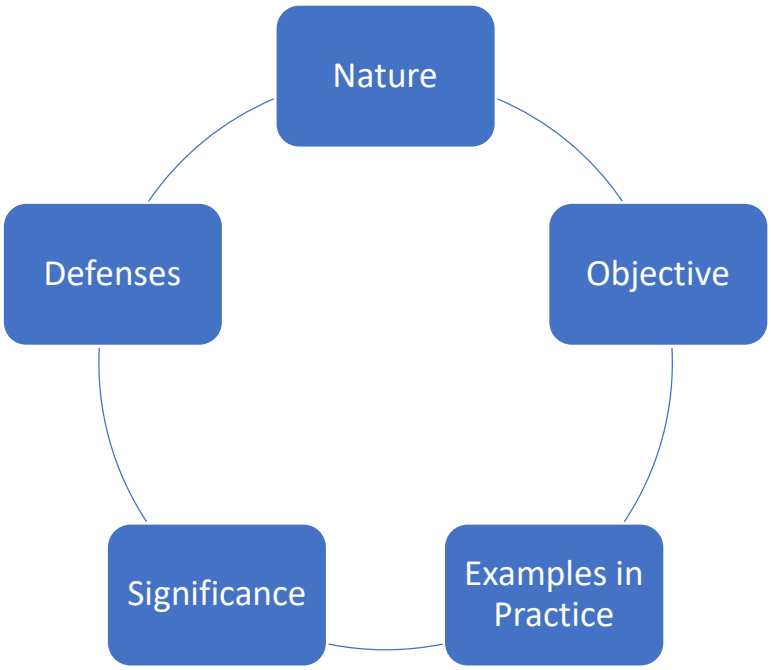


Figure 4. Overview of Spatial Transformation Attacks.

Spatial Transformations (STAs) refer to modifying the spatial arrangement of the input data. For instance, when classifying images, STAs may include making slight rotations, translations, or distortions to an image that causes the model to misclassify it, despite the image appearing practically the same to a human viewer.

Spatial Transformation Attacks (STAs) involve altering the spatial arrangement of input data. For instance, in image classification, STAs may include rotating, translating, or distorting an image slightly. This can cause the model to misclassify the image, even though it may appear to be unchanged to a human observer.

In the context of facial recognition systems, a Security Testing Agent (STA) could introduce slight distortions to facial features in a photograph. This could lead to misidentification or failure to identify a face by the system, even though a human would have no difficulty recognising it.

Here is an example of how a STA could work in practice: In a facial recognition system, the STA might intentionally distort the facial features in a photograph. This could cause the system to incorrectly identify or even fail to identify the face, despite it being easily recognisable to a human.

To protect against Security Threat Agents (STAs), one can employ data augmentation methods during the training phase. This involves training the model on different versions of the data that have undergone spatial transformations. Additionally, adversarial training can be used, which involves training the model on both the original data and adversarial examples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

As the utilisation of AI continues to grow, it is imperative to develop strong security measures to safeguard against the vulnerabilities present in AI models. These vulnerabilities necessitate the implementation of robust defences to ensure the safety and reliability of AI-based systems and technologies.

5.5. Physical Adversarial Examples

The term "Physical Adversarial Examples" refers to a type of cybersecurity threat that targets deep learning models. Essentially, this involves altering input data in a way that tricks the machine learning system into producing an incorrect result. Physical adversarial examples differ from traditional attacks, which are carried out in the digital realm, by taking place in the physical world and manipulating real-world input data.

In **Error! Reference source not found.**, we can see a detailed explanation in the context of cyber-attacks on artificial intelligence models.

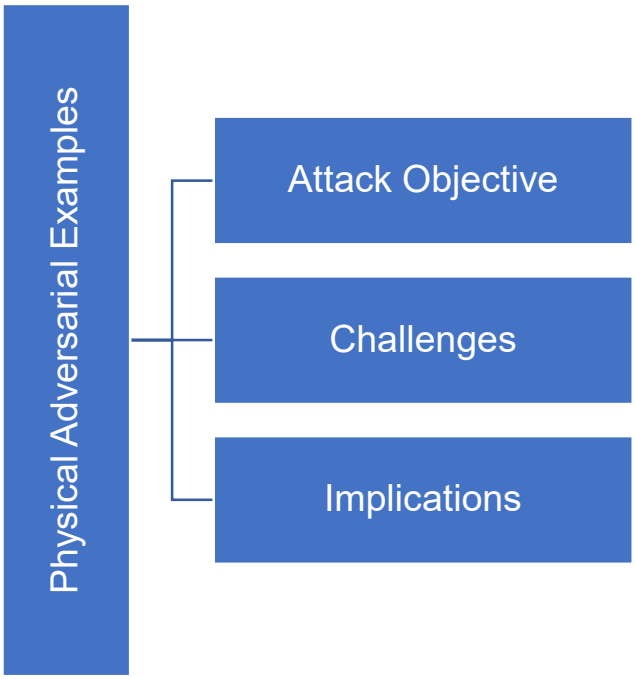


Figure 5. Context of cyber-attacks on artificial intelligence models.

It's important to be aware that physical adversarial examples are actual real-world perturbations capable of deceiving machine learning models. Consider a self-driving car that employs a neural network to recognise road signs; a malicious actor could easily trick the system by placing certain stickers on a stop sign. This can cause the AI to misinterpret the sign as a yield sign or something entirely different. It's imperative to take the necessary precautions to avoid such potential hazards.

The main objective of attacks on machine learning models is to manipulate them into making wrong predictions or classifications, which can be particularly dangerous in safety-critical systems like self-driving cars and medical imaging devices where incorrect classifications can result in catastrophic outcomes.

Executing physical adversarial attacks is more challenging compared to digital attacks. Such attacks in the physical world require consideration of various factors, such as lighting conditions, viewing angles, and distances. In contrast, digital attacks involve direct modification of pixel values, while physical attacks usually involve changes in the environment or placement of tangible objects.

The existence of physical adversarial examples reveals the shortcomings of present-day machine learning models, specifically deep neural networks. Although these models operate efficiently in regular situations, their susceptibility to adversarial attacks emphasises the necessity for more durable designs and training techniques.

It is imperative that AI and machine learning models take into account Physical Adversarial Examples in order to guarantee the safety and reliability of their application in real-world settings.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

By doing so, we can ensure that these technologies are able to function effectively and without incident when confronted with unforeseen challenges and external factors. As such, it is crucial that developers remain vigilant and proactive in their approach to designing and implementing these models, in order to mitigate potential risks and vulnerabilities that may arise over time.

5.6. Model Inversion Attack

Cyber attacks known as Model Inversion Attacks target machine learning models, particularly when they are viewed as black boxes. This means attackers cannot directly access the model's parameters or architecture. Such attacks aim to leverage the model's predictions to extract valuable insights from the training data, which could expose confidential information.

In **Error! Reference source not found.**, we can see a breakdown of how Model Inversion Attacks work:

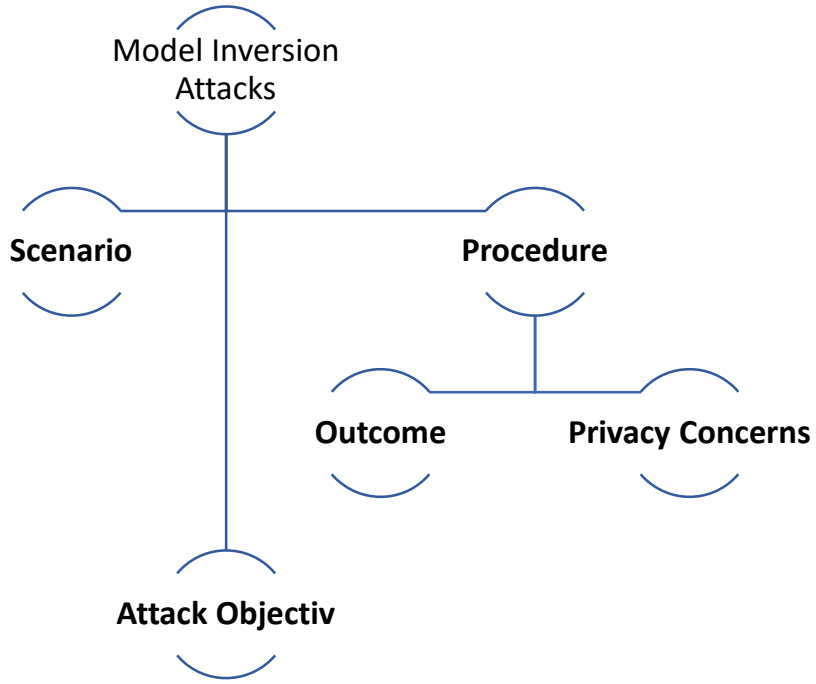


Figure 6. Breakdown of how Model Inversion Attacks work.

Consider a trained model that receives an input and produces a prediction. This could be a model that predicts a person's facial features based on genetic data, for example. The model has been deployed, and users can access its prediction capability without viewing the training data or the model's internal details.

An attacker's goal is to reverse-engineer or "invert" the model. They would try to generate a possible genetic input for a given facial feature output in the given scenario. The attacker does not require direct access to the training data. They instead require access to the model's predictions (either legally, e.g., using a public API, or illicitly). The attacker uses the model's outputs to refine their input guesses until they are close to the true training data inputs.

If the attacker is successful, he or she will be able to deduce or approximate individual data points from the training dataset. They might be able to infer genetic data for a person whose facial features were included in the training set, in our example.

Concerns about privacy: This type of attack raises serious privacy concerns, particularly when models are trained on sensitive data such as medical records or personal identifiers. If an attacker can approximate this data from the model's outputs, the privacy and confidentiality of the original data sources are jeopardised.

It's important to note that the success of a Model Inversion Attack depends on the complexity of the model, the nature of the data, and the amount of information the attacker already has. Countermeasures include techniques like differential privacy, which adds noise to the data or model

 doi:10.20944/preprints202309.2064.v1

outputs, making it harder for attackers to draw precise inferences about individual training data points.

5.7. Membership Inference Attack

A Membership Inference Attack (MIA) is a type of attack on machine learning models, particularly in situations where privacy is a concern. The goal of such an attack is to determine whether a specific data point was part of a machine learning model's training set. This type of cyber-attack is broken down as follows:

The Membership Inference Attack (MIA) is a method employed to ascertain whether a specific data instance was utilised in training a machine learning model. Essentially, the primary objective of an MIA is to identify instances that were utilised in the training of a model, as well as the extent to which they were used in the model's development. This type of attack can be detrimental to the security and confidentiality of sensitive information, and as such, it is important to implement measures to guard against it.

Privacy is a major concern, especially in cases where sensitive information such as medical records or personal financial data is involved. The potential for a breach is particularly alarming, as an attacker who identifies a specific piece of data used in the training set (such as a medical record) could easily compromise privacy protections.

Let me explain how this works: Sometimes when a model is trained to recognise patterns in data, it can become overly focused on certain data points. This can occur if the data is not evenly distributed or if the model is too complex. When a model works well on the data it was trained on, but not as well on new data, it's called overfitting. Unfortunately, attackers can take advantage of this by using the model's predictions, such as its level of confidence in a classification, to determine if the data was part of the training set. If the model is very confident about a data point, it could indicate that it was part of the training data.

It is important to consider the potential risks associated with training AI models, especially those that use deep learning techniques. These models often require vast amounts of data, which may include personal or sensitive information. If this data is compromised, it could raise questions about how it was collected, used, or shared, potentially resulting in privacy violations or ethical concerns.

One effective way to safeguard against attacks that aim to extract membership information from data or model outputs is to apply techniques such as differential privacy, which involves adding noise to the data. Another approach is to incorporate regularisation techniques during training to mitigate overfitting and enhance the model's resilience to such attacks. These measures can help bolster the security of the model and protect sensitive information from being compromised.

MIAs highlight the unique challenges posed by the intersection of machine learning and privacy in the broader landscape of cyber-attacks on AI models. While traditional cyber-attacks may focus on stealing data or disrupting services, MIAs use the nature and behaviour of machine learning models to extract potentially sensitive information.

6. Case Study Research

6.1. Red Teaming Generative AI: Case Study on DEF CON 31

During the recent DefCon hacker convention in Las Vegas, both White House officials and Silicon Valley leaders expressed their concerns regarding the potential negative impact of AI chatbots on society. This three-day event, which ended on Sunday, involved major industry stakeholders who were actively engaged in discussions and assessments.

Just before the DEF CON 31 conference, the White House unexpectedly announced a partnership with leading AI developers such as OpenAI, Google, Antrhopic, Hugging Face, Microsoft, Nvidia, and Stability AI. These entities will participate in a public evaluation of their generative AI systems at DEF CON 31, an event organised by the AI Village, a group of AI enthusiasts and hackers. Over the past year, large language models (LLMs) such as ChatGPT have surged in popularity for enhancing writing and communication processes. However, officials acknowledge the innate vulnerabilities these tools present. Challenges arising from confabulations, jailbreaks, and biases are concerns not just for cybersecurity experts but also for the general public. Hence, the White House's Office of Science, Technology, and Policy is keen on thoroughly testing these emerging generative AI models.

 doi:10.20944/preprints202309.2064.v1

The White House recently released a statement announcing an independent evaluation of AI models. This evaluation aims to provide valuable insights to both researchers and the general public regarding the impact of these models. It also offers an opportunity for AI companies and developers to address any issues uncovered during the process. This event is in alignment with the Biden administration's AI Bill of Rights and the AI Risk Management Framework established by the National Institute of Standards and Technology. Experts from the AI Village, including Sven Cattell, Rumman Chowdhury, and Austin Carson, have praised this evaluation as the most comprehensive red teaming exercise ever conducted for AI models. Thousands of individuals have already participated in this public evaluation, which employs an assessment tool developed by Scale AI.

To provide context, "red teaming" is a method used by security experts to actively search for weaknesses or vulnerabilities in an organisation's frameworks to enhance their overall security and resilience. Cattell, the creator of AI Village, emphasised the importance of more people comprehending how to red team and evaluate these models to address the various issues associated with them. In collaboration with DEF CON, AI Village organised the most comprehensive red-teaming exercise for any group of AI models to increase the number of researchers capable of managing AI system vulnerabilities.

It was unexpectedly difficult to secure LLMs, and this was partially due to a technique called "prompt injection". AI researcher Simon Willison warned about the risks associated with prompt injection, a technique that could cause a language model to perform unintended tasks. During DEF CON, participants used laptops provided by the event organisers to access various LLMs for a limited time. A capture-the-flag-style scoring system was put in place to encourage participants to investigate a wide range of potential threats. The participant with the highest score at the end of the event received a top-tier Nvidia GPU award.

The AI Village has confirmed the sharing of output and has released the following statement: "We will publish the insights gained from this event to assist those who wish to attempt the same thing." The Village also emphasised that "the greater the number of individuals who understand how to effectively work with these models, along with their limitations, the more advantageous it will be."

6.2. DEF CON 31: Case Study on the New Red Teaming Hacks

During the convention, approximately 2,200 attendees conducted a thorough evaluation of eight of the leading large-language models. This was the first independent "red-teaming" assessment of multiple models, emphasising the importance of this technology for the future. However, the findings are not expected to be released until around February, so immediate results should not be anticipated. The creators of these digital systems may find it challenging to address the vulnerabilities identified, as they may not fully understand the intricate workings of these models. Resolving these issues will require significant time and resources.

Studies conducted by academic and business experts have revealed several flaws in the current AI models. These models tend to be unnecessarily complicated, delicate, and easily altered. The development of these systems was not focused on security measures, which led to the dependence on extensive and intricate image and text collections. As a result, these models exhibit inherent biases towards certain races and cultures, making them vulnerable to external manipulation.

Gary McGraw, a cybersecurity expert with extensive experience and one of the co-founders of the Berryville Institute of Machine Learning, has highlighted the dangers of trying to add security measures to these systems after they have been built. According to him, it is not realistic to expect that we can simply patch up the vulnerabilities or add a layer of security to these systems as an afterthought. It would be like adding a temporary security measure without proper planning and implementation.

Typical software operates by following set of commands laid out in its code. However, language models such as OpenAI's ChatGPT and Google's Bard utilise a different method. They learn and adjust by analysing vast amounts of data from online content. This ability to constantly adapt is significant and may have far-reaching consequences for our society.

A different researcher instructed ChatGPT to create deceitful emails and a malevolent scheme that goes against ethical standards. Tom Bonner, from the AI security company HiddenLayer, presented at DefCon this year and demonstrated how he was able to deceive a Google system. He accomplished this by inserting a statement claiming "this is safe to use," causing the system to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

wrongly classify harmful software as safe. Bonner highlighted the insufficiency of proper safety procedures.

Although major breaches were frequently reported a few years ago, they are now rarely disclosed. With a lot at stake, and a lack of proper monitoring, "problems can be easily ignored, which is what is happening," as pointed out by Bonner.

Studies show that intentionally corrupting a small portion of data used for AI training can cause significant disruptions, which may go unnoticed.

New research conducted by Florian Tramér of ETH Zurich has discovered that it is possible to compromise a model by tampering with as little as 0.01% of it, for a cost as low as £45. Tramér and his team waited for a few websites used in web crawls for two different models to expire. After purchasing the domains, they uploaded malicious content to them.

While working at Microsoft, Hyrum Anderson and Ram Shankar Siva Kumar dedicated their efforts to conducting in-depth stress tests on AI technology. As part of their research, they delved into the current security status of text-based AI and carefully analysed its strengths and weaknesses. Their findings shed light on the potential vulnerabilities that exist within this emerging field and helped to pave the way for continued advancements in AI security.

During their investigation of more than 80 companies, the duo found that most of them did not have a contingency plan in place for data poisoning attacks or the theft of datasets. According to their report, the majority of companies in the sector would be completely unaware if such an event occurred. Andrew W. Moore, a former Google executive and dean at Carnegie Mellon, remembers dealing with attacks on Google's search software over a decade ago. Additionally, between late 2017 and early 2018, spammers were able to manipulate Gmail's AI-based detection mechanism four times.

Major AI organisations claim that security and safety are at the forefront of their concerns. They've recently made voluntary pledges to the White House to open their models, which are largely opaque "black boxes", for external review. However, there's a prevailing concern that these corporations might not take adequate measures.

It is predicted that Tramér will detect potential manipulations of search engines and social media platforms that exploit the vulnerabilities of AI systems for monetary gain and dissemination of misinformation. For example, a smart job seeker may find a way to convince an AI recruitment system that they are the only suitable candidate.

Ross Anderson, a computer scientist from Cambridge University, expressed concern about AI bots compromising privacy. As more people use them for communication with hospitals, banks, and employers, there is a risk that malicious actors will use these bots to obtain financial, employment, or health information from supposedly secure systems. Moreover, studies indicate that AI language models may inadvertently fault themselves by retraining on substandard data.

After examining over 80 businesses, the investigators found that most of them did not have a backup plan in case of a data poisoning attack or dataset theft. They concluded that if such an event were to occur, the majority of the industry would not even be aware. Andrew W. Moore, a former Google executive and dean at Carnegie Mellon, remembers dealing with attacks on Google's search software over a decade ago. Furthermore, between the end of 2017 and the beginning of 2018, spammers successfully tricked Gmail's AI-driven detection system four times.

It is likely that smaller AI companies may not have security teams to maintain their systems, which could result in an increase in insecure digital tools and agents. In the coming months, start-ups are expected to release multiple products that rely on pre-trained models that have been licensed. Experts warn that it is not uncommon for one of these tools to potentially access and take your contact list without your knowledge.

6.3. Cash Reward for AI That Deters Hackers: Case Study on the White House AI Cyber Challenge

In a separate announcement, the White House announced a cash reward for AI that deters hackers. This initiative by the White House started an AI Cyber Challenge contest, aiming to develop novel AI solutions that shield vital software from cyber threats.

Contestants, aiming for a slice of the $18.5 million prize fund, are tasked with devising innovative AI systems. These systems need to promptly detect and rectify software flaws in critical infrastructures like electric grids or subways, vulnerabilities that hackers might exploit, as stated by President Joe Biden's administration.

 doi:10.20944/preprints202309.2064.v1

Arati Prabhakar, the head of the White House Office of Science and Technology Policy, during a briefing, remarks, “This competition will be a clarion call for all kinds of creative people in Organisations to bolster the security of critical software that American families and businesses and all of our society relies on,”

In an effort to increase participation, DARPA, which oversees the contest, commits $7 million to support small enterprises wishing to participate. The competition witnesses DARPA joining hands with AI giants such as Anthropic, Google, Microsoft, and OpenAI, the creators of ChatGPT. These companies are slated to offer their expertise and technological prowess for the contest, according to Prabhakar.

Deputy National Security Advisor for Cyber and Emerging Technology, Anne Neuberger, during the briefing, conveys that the contest's goal is to "unite a diverse set of minds nationwide to ponder the use of AI in significantly enhancing cybersecurity."

This announcement was made in Las Vegas during the ‘Black Hat’ cybersecurity conference, preceding the Def Con meet-up where hackers challenged multiple AI systems.

7. Whitebox Attacks

Attackers Whitebox attacks are a type of adversarial attack in artificial intelligence in which the attacker has extensive knowledge of the AI model. This knowledge includes specifics about its architecture, weights, and even training data. The transparency inherent in white-box attacks provides significant advantages to attackers, allowing them to create attacks that are highly tailored and significantly more effective. Gradient-Based Attacks are prime examples of this, in which attackers use their in-depth knowledge of the model's parameters to generate adversarial examples by manipulating the gradients of the loss function with respect to the input, causing the model to misclassify.

Model Extraction Attacks are another type of white-box attack that attackers can use. With complete access to the model's architecture and training data, they can create approximate replicas of the original model. These replicas can then be examined and attacked externally, allowing sensitive information to be extracted or facilitating the creation of more robust adversarial examples. Similarly, Membership Inference Attacks take advantage of the model's training data knowledge, allowing attackers to determine whether a specific data point was used during the training phase, posing a significant risk to data privacy and potentially exposing sensitive information about individual data samples.

Furthermore, the extensive knowledge gained by white-box attackers enables the investigation of Architectural Vulnerability Exploitation. Attackers can expose and exploit inherent vulnerabilities or flaws in the model's design by understanding its internal workings and structures. Such exploitation can significantly alter the model's behaviour or impair its performance, resulting in incorrect decisions or misclassifications. Given the significant risks associated with white-box attacks, improved security protocols and the development of resilient models are critical. The rigorous exploration and understanding of these sophisticated attack vectors is critical in advancing the development of defensive mechanisms and fortifying AI systems against a variety of adversarial threats in a variety of applications.

7.1. Fast Gradient Sign Method (FGSM).

The Fast Gradient Sign Method (FGSM) is a fast and effective way to create adversarial examples, which are important for testing how well Convolutional Neural Networks (CNNs) can handle attacks. The method only takes one step and starts by receiving an input image. Then, the system uses a pre-trained CNN to make predictions on the image. This step is crucial for evaluating the model's resilience based on its original classifications.

Following the initial prediction, the method proceeds to determine the loss by comparing it with the true class label. This metric indicates the variance between the model's predictions and the actual labels, highlighting regions that require shielding against adversarial perturbations. The method then computes the gradients of the loss concerning the input image, which is a critical step in generating the adversarial example.

The FGSM method calculates the sign of the gradient and uses it to create an altered image that causes misclassifications. This process shows how easily FGSM can generate adversarial examples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

and exposes the weaknesses of CNNs in the face of such threats. It emphasises the importance of developing advanced defence strategies and robust models that can withstand adversarial perturbations.

The FGSM is a one-step method to generate adversarial examples, and we can visualise the process in **Error! Reference source not found.**.

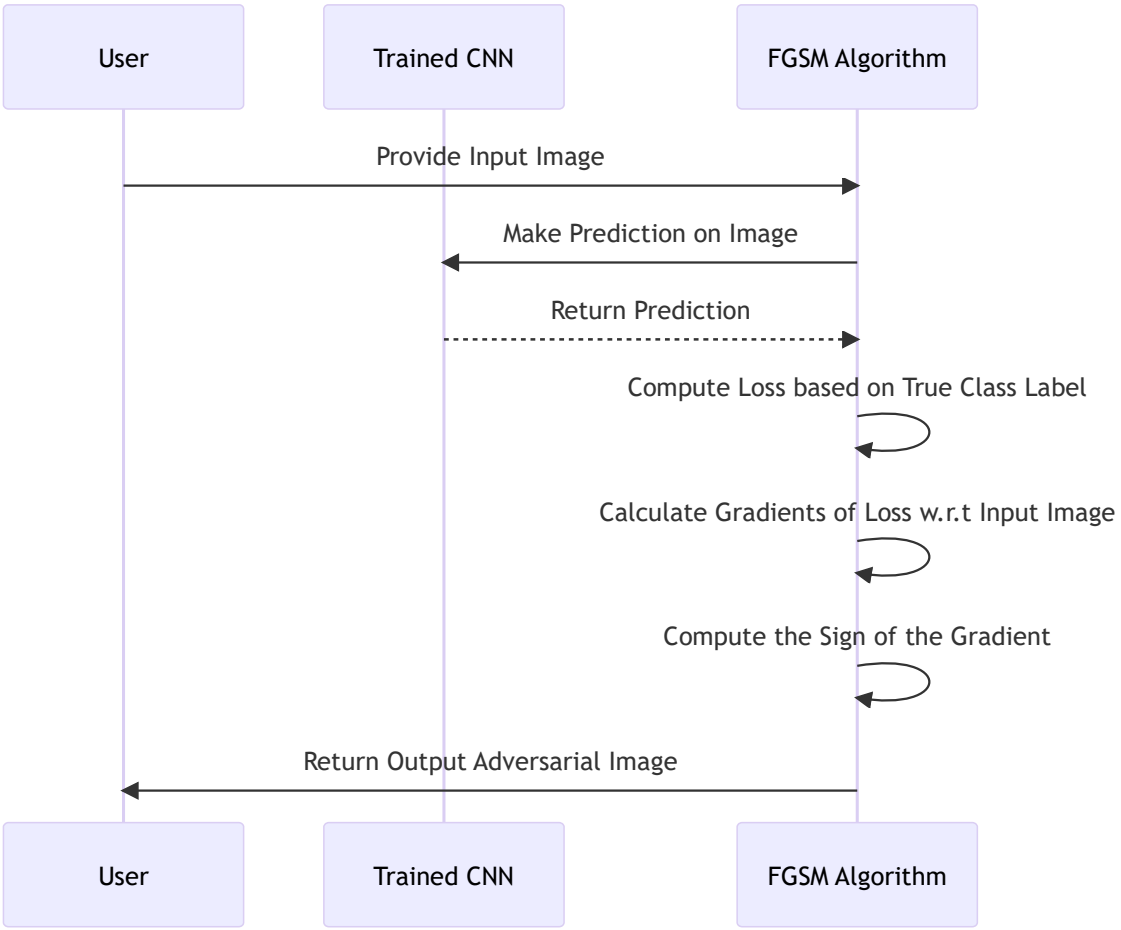


Figure 7. The FGSM method to generate adversarial examples.

This sequence diagram succinctly illustrates the Fast Gradient Sign Method's (FGSM) interaction and process of creating adversarial examples. The user inputs an image, which is then utilised by the pre-trained Convolutional Neural Network (CNN) to make predictions. Once the prediction is made, the FGSM algorithm calculates the loss by comparing it to the true class label and computes the gradients of this loss with respect to the input image. Then, the algorithm determines the sign of the gradient and meticulously constructs the adversarial image using this sign. This streamlined process provides a comprehensive overview of FGSM's methodical approach to uncovering potential vulnerabilities in CNNs by generating adversarial examples.

7.2. Fast Gradient Sign Method (FGSM) Can Be Described as a Mathematical Formula

The formula:

$$adv\_x = x + \epsilon*\text{sign}(\nabla_x J(\theta, x, y))$$

Or:

$$x' = x + \epsilon \times sign(\nabla x J(\theta, x, y))$$

where:

 doi:10.20944/preprints202309.2064.v1

- adv_x: Adversarial image.
- x: Original input image.
- y: Original input label.
- epsilon: Multiplier to ensure the perturbations are small.
- Theta: Model parameters.
- J: Loss.

Or:

- x is the original input image.
- ϵ is a small value determining the magnitude of the perturbation.
- $J(\theta,x,y)$ is the neural network's loss function, with θ being the model parameters.

7.3. FGSM - TensorFlow's GradientTape Function

In under 30 lines of code, FGSM can be implemented using TensorFlow's GradientTape function. It's essential to understand that the gradients are taken concerning the input image to generate an image that reduces the loss. To achieve this, each pixel's effect on the loss value is calculated and then modified or perturbed accordingly. The chain rule enables you to efficiently determine how each input pixel influences the loss, making this technique highly effective. It's worth noting that the model is no longer being trained, so gradients are no longer taken with respect to the model parameters. The primary goal is to deceive an already trained model without altering its parameters.

7.4. Descriptive Overview of Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) is distinguished by its distinctive property and goal; it employs gradients taken explicitly with respect to the input image in order to meticulously craft an image that minimises the model's loss. The procedure entails determining the contribution of each pixel to the loss value and then adjusting or perturbing the image to effectively mislead the model. The implementation of the chain rule, which quickly calculates how each input pixel contributes to the loss, makes this method inherently efficient, allowing for the rapid generation of adversarial examples.

In terms of model parameters, it is critical to note that FGSM works on models that are no longer in the training phase. The gradients are computed without regard for the model parameters and are entirely focused on the input image. The overarching goal of this method is to deceive a pre-trained model while keeping its inherent parameters unchanged during the attack. Because of this precise approach and the method's efficiency, FGSM is a significant technique for exploring and understanding the vulnerabilities of pre-trained models in a variety of applications, emphasising the importance of robust countermeasures against adversarial attacks.

7.5. Fast Gradient Sign Method (FGSM) Described with Code

Let's try and fool a pre-trained model. In this section, we will review how to perform the attack. The model is MobileNetV2 model, pre-trained on Image with NetTensorFlow, MobileNetV2[1], and Imagenet[2].

```
import tensorflow as tf
import matplotlib as mpl
import matplotlib.pyplot as plt
```

[1] https://www.tensorflow.org/versions/r2.0/api_docs/python/tf/keras/applications/MobileNetV2

[2] https://image-net.org/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1


```
mpl.rcParams['figure.figsize'] = (8, 8)
mpl.rcParams['axes.grid'] = False
```

To load the pretrained MobileNetV2 model and the ImageNet class names, we need to type the following command.

```
pretrained_model = tf.keras.applications.MobileNetV2(include_top=True,
                                                      weights='imagenet')
pretrained_model.trainable = False

# ImageNet labels
decode_predictions = tf.keras.applications.mobilenet_v2.decode_predictions
```

followed by:

```
# Helper function to preprocess the image so that it can be inputted in MobileNetV2
def preprocess(image):
  image = tf.cast(image, tf.float32)
  image = tf.image.resize(image, (224, 224))
  image = tf.keras.applications.mobilenet_v2.preprocess_input(image)
  image = image[None, ...]
  return image

# Helper function to extract labels from probability vector
def get_imagenet_label(probs):
  return decode_predictions(probs, top=1)[0][0]
```

The original sample image used to create adversarial examples in this tutorial is from Wikimedia Common. The first step is to preprocess it so that it can be fed as an input to the MobileNetV2 model.

```
image_path = tf.keras.utils.get_file('YellowLabradorLooking_new.jpg',
'https://storage.googleapis.com/download.tensorflow.org/example_images/YellowLabradorLooking_n
ew.jpg')
image_raw = tf.io.read_file(image_path)
image = tf.image.decode_image(image_raw)

image = preprocess(image)
image_probs = pretrained_model.predict(image)
```

To check the image, we can type the following command:

```
plt.figure()
plt.imshow(image[0] * 0.5 + 0.5)   # To change [-1, 1] to [0,1]
_, image_class, class_confidence = get_imagenet_label(image_probs)
plt.title('{} : {:.2f}% Confidence'.format(image_class, class_confidence*100))
plt.show()
```

To generate the adversarial image and implement the fast gradient sign method, we must first generate perturbations (disturbances) that can be used to distort the original image, yielding an adversarial image. As previously stated, the gradients are taken with respect to the image for this task.

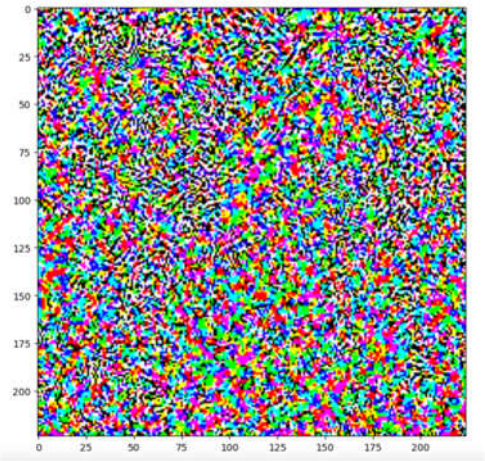
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1


```
loss_object = tf.keras.losses.CategoricalCrossentropy()

def create_adversarial_pattern(input_image, input_label):
  with tf.GradientTape() as tape:
    tape.watch(input_image)
    prediction = pretrained_model(input_image)
    loss = loss_object(input_label, prediction)

  # Get the gradients of the loss w.r.t to the input image.
  gradient = tape.gradient(loss, input_image)
  # Get the sign of the gradients to create the perturbation
  signed_grad = tf.sign(gradient)
  return signed_grad
```

To visualise the resulting perturbations (disturbance):

```
# Get the input label of the image.
labrador_retriever_index = 208
label = tf.one_hot(labrador_retriever_index, image_probs.shape[-1])
label = tf.reshape(label, (1, image_probs.shape[-1]))

perturbations = create_adversarial_pattern(image, label)
plt.imshow(perturbations[0] * 0.5 + 0.5);  # To change [-1, 1] to [0,1]
```

The image will look similar to:



To assess the network's resilience, we can conduct tests using different epsilon values. By increasing epsilon, we can observe how the network responds to changes. While a higher epsilon value makes it easier to fool the network, it also makes perturbations more apparent and noticeable.

```
def display_images(image, description):
  _, label, confidence = get_imagenet_label(pretrained_model.predict(image))
  plt.figure()
  plt.imshow(image[0]*0.5+0.5)
  plt.title('{} \n {} : {:.2f}% Confidence'.format(description,
                                                   label, confidence*100))
  plt.show()
```

 doi:10.20944/preprints202309.2064.v1


```
epsilons = [0, 0.01, 0.1, 0.15]
descriptions = [('Epsilon = {:0.3f}'.format(eps) if eps else 'Input')
                for eps in epsilons]

for i, eps in enumerate(epsilons):
  adv_x = image + eps*perturbations
  adv_x = tf.clip_by_value(adv_x, -1, 1)
  display_images(adv_x, descriptions[i])
```

7.6. Fast Gradient Sign Method (FGSM) Next Steps in Advancements and Exploration

Conducting experiments on different datasets and architectures is crucial in gaining a deeper understanding of adversarial attacks, particularly FGSM. Researchers and practitioners can apply these attacks on widely-known datasets such as CIFAR-10, ImageNet, and MNIST, and assess the robustness of various architectures like VGG, ResNet, or Inception. By conducting extensive experimentation, valuable insights can be gained on the vulnerabilities and resilience of various models and architectures, which can serve as a foundation for improving model robustness.

It is imperative to acknowledge that one of the key components in comprehending adversarial attacks is developing and training models. To gain practical experience in building and enhancing deep learning models, it is recommended to construct a simple convolutional neural network (CNN) and train it using a selected dataset. Once the model is effectively trained, trying to deceive it with adversarial techniques can provide invaluable insight into the vulnerabilities inherent in deep learning models. This hands-on approach can facilitate a more comprehensive exploration of potential defensive strategies and countermeasures against adversarial attacks in the ever-evolving AI landscape. The ultimate objective is to refine our comprehension of adversarial complexities and create fortified models that are resilient to adversarial perturbations.

To understand how effective adversarial attacks are, it's important to explore how the model's confidence varies with changes in the epsilon value. Epsilon represents the amount of perturbation, and adjusting it lets you observe how the model's confidence in its predictions changes. Using a smaller epsilon may not be as effective in misleading the model as using a larger epsilon, as smaller perturbations may not be as noticeable.

Recent Developments in Adversarial Research: The field of adversarial attacks has made significant progress, with the Fast Gradient Sign Method (FGSM) serving as a basis for many techniques. Since the creation of FGSM, numerous refined and sophisticated attack methods have been developed. To gain a better understanding of these advancements, it is essential to read recent research papers that explain the new and more intricate attacks that have been devised and executed. As foundational techniques like FGSM are iteratively expanded upon, more potent and adaptable adversarial examples are being created. This highlights the need for continuous exploration and research to develop robust countermeasures and strengthen models against emerging adversarial threats.

Defending Against Adversary Attacks

It's important to find ways to protect against adversarial attacks and understand how they work. This involves using various techniques and adjusting training methods to make models more resilient to intentional changes. The ultimate objective is to bolster models so they can withstand and counteract the manipulative effects of crafted changes, ensuring reliable and accurate predictions in different scenarios.

There are several effective techniques that can improve the robustness of models against attacks. Adversarial training, defensive distillation, and feature squeezing are some of the notable methods. Adversarial training involves exposing models to adversarial examples during the training process so that they can learn to adapt to such changes. Defensive distillation and feature squeezing are advanced strategies that focus on refining model inputs and modifying model parameters to detect and resist adversarial manipulations more effectively.

 doi:10.20944/preprints202309.2064.v1

Several techniques have emerged as prominent defences, including adversarial training, defensive distillation, and feature squeezing. Adversarial training involves exposing models to adversarial examples during the training phase in order for them to learn and adapt to such perturbations. Defensive distillation and feature squeezing are two sophisticated strategies that focus on refining model inputs and modifying model parameters to detect and resist adversarial manipulations effectively.

8. Jacobian Based Saliency Map Attack (JSMA)

The JSMA, which stands for Jacobian-based Saliency Map Attack, is a sophisticated type of attack that directly targets particular input features with the goal to cause misclassification. This type of attack is unique because it only modifies specific components of the input, unlike other attacks that affect the entire input image. This makes the JSMA a highly effective counter-attack strategy that can be used to protect against such attacks.

The JSMA (Jacobian-based Saliency Map Attack) is a new method of adversarial attack that concentrates on modifying only a portion of the input features. This technique is very careful and aims to create misclassification by altering specific input components, serving as a way to defend against attacks that impact the entire input image.

Adversarial inputs can be created using two methods. The first method is called "Deep Fool" and involves iteratively changing the input in order to cause misclassification by crossing the decision boundary. The second method is an improvement on the Fast Gradient Sign Method (FGSM) and is known as the Iterative Gradient Sign Method. This method gradually refines the adversarial input through repeated applications to ensure effective misclassification.

The C&W attack is a technique that aims to create adversarial inputs that closely resemble the original input and cause the model to misclassify them. It does this by making subtle changes to the input, emphasising the importance of minimising alterations. This approach highlights the importance of stealthiness in executing the attack.

Finally, the Boundary Attack employs a completely different strategy. Instead of introducing changes to the original image, it starts with a random image and iteratively refines it to look like the original input. The goal is to achieve similarity while maintaining the misclassification, providing a sophisticated viewpoint on crafting adversarial examples.

In **Error! Reference source not found.**, we can see each attack's main characteristics and steps.

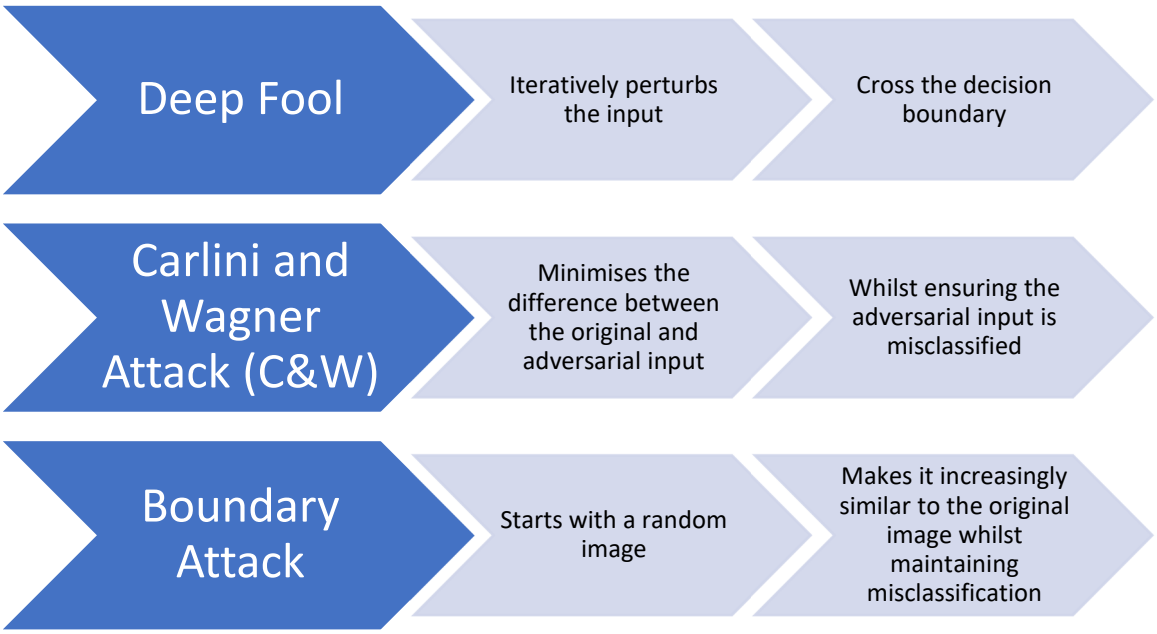


Figure 8. Jacobian based Saliency Map Attack (JSMA).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

Each of these attacks in **Error! Reference source not found.**, highlights different aspects of adversarial methodologies, demonstrating the diversity and evolving complexity in crafting adversarial inputs, and emphasises the critical need for developing robust and versatile defensive mechanisms.

9. Carlini and Wagner Attack (C&W) – MNIST

Deep Neural Networks (DNNs) perform exceptionally well in difficult machine learning tasks. They are, however, vulnerable to 'adversarial examples'—intentionally crafted inputs that degrade their performance. Such flaws can make it difficult to use DNNs in security-critical applications.

We'll be using the MNIST database. The MNIST database (Modified National Institute of Standards and Technology database) is a large collection of handwritten digits that is commonly used to train image processing systems. The GitHub repository[3] can also be used for guidance.

Preparing the data is the first step in this exercise using the MNIST database. For training, the images must be rescaled to a range of 0 to 1. We can do this with the TensorFlow backend.

```
import numpy as np
from keras.datasets import mnist as data_keras
#from keras.datasets import fashion_mnist as data_keras
from keras.utils import to_categorical

(x_train, y_train), (x_test, y_test) = data_keras.load_data()
x_train = x_train[...,np.newaxis] /255.0
x_test = x_test[...,np.newaxis] / 255.0
y_train = to_categorical(y_train)
y_test = to_categorical(y_test)

print(x_train.shape,y_train.shape,x_test.shape,y_test.shape)
print("MIN={},MAX={}".format(np.min(x_train),np.max(x_train)))
```

(60000, 28, 28, 1) (60000, 10) (10000, 28, 28, 1) (10000, 10)

MIN=0.0,MAX=1.0

At the outset, it is of utmost importance to lay the foundation and instruct the Deep Neural Network (DNN) Model appropriately. This entails creating a basic DNN model through the TensorFlow session, which is an essential step in preparing for the introduction of an adversarial attack. Proper understanding and implementation of this process are critical to ensure the DNN model is equipped to handle any potential threats.

```
from keras.models import Sequential
from keras.layers import Dense, Dropout, Activation, Flatten, Conv2D, MaxPooling2D
from keras.optimizers import SGD
def standardCNN():
    params = [32, 32, 64, 64, 200, 200]
    model = Sequential()



    model.add(Conv2D(params[0], (3, 3),
                     input_shape=x_train.shape[1:]))
```

[3] https://github.com/carlini/nn_robust_attacks

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1


```
        model.add(Activation('relu'))
        model.add(Conv2D(params[1], (3, 3)))
        model.add(Activation('relu'))
        model.add(MaxPooling2D(pool_size=(2, 2)))

        model.add(Conv2D(params[2], (3, 3)))
        model.add(Activation('relu'))
        model.add(Conv2D(params[3], (3, 3)))
        model.add(Activation('relu'))
        model.add(MaxPooling2D(pool_size=(2, 2)))

        model.add(Flatten())
        model.add(Dense(params[4]))
        model.add(Activation('relu'))
        model.add(Dropout(0.5))
        model.add(Dense(params[5]))
        model.add(Activation('relu'))
        model.add(Dense(10))
        model.add(Activation('softmax'))
        sgd = SGD(lr=0.01, decay=1e-6, momentum=0.9, nesterov=True)

        model.compile(loss="categorical_crossentropy",
                              optimizer=sgd,
                              metrics=['accuracy'])
        return(model)

## session will be necessary in adversarial example
import tensorflow as tf
import keras.backend as K
from keras.models import load_model
sess = tf.Session(config=tf.ConfigProto())
K.set_session(sess)


training    = False
modelname = "models/trained_model"
if training:
        model_keras = standardCNN()

        batch_size = 64
        model_keras.fit(x_train, y_train,
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1


```
                        batch_size=batch_size,
                        validation_data=(x_test, y_test),
                        nb_epoch=20,
                        shuffle=True)

    model_keras.save(modelname)
else:
    model_keras = load_model(modelname)

model_keras.summary()
```

Once a model has been developed, it is imperative to evaluate its efficacy by analysing its ability to perform on data that hasn't been corrupted or tampered with in any way. This step is crucial in determining the robustness and reliability of the model, as it provides insight into how well it can handle real-world scenarios and challenges. Ultimately, the goal is to ensure that the model can consistently and accurately analyse and interpret data, regardless of any potential external factors that may impact its performance.

```
scores =model_keras.evaluate(x_test,y_test)
print("loss={}, accuracy={}".format(*scores))
```

That is the final command in this sequence.

To fully understand the attack, it's important to first have a basic understanding of the Carlini and Wagner Attack (C&W). This attack is known for its advanced level of sophistication and precision. Its goal is to create adversarial examples by making small changes to the original inputs in order to create a new input that appears almost identical to the original, but will be misclassified by the model. This technique requires a high level of subtlety and precision in order to maintain the original input's appearance while achieving the desired misclassification. As a result, the C&W attack is considered to be one of the most advanced and refined adversarial attack methods in the field of machine learning security.

9.1. The Spirit of Carlini and Wagner Attack (C&W)

Consider:

- x as an image in the space $[0,1]^n$
- δ as the noise, also in the space $[0,1]^n$

The aim of the CW attack is to pinpoint the least amount of noise δ that, when added to an image x, alters its classification to a target class t. Formally, this is expressed as:

$$\text{minimise} \quad ||\delta||_p \quad \text{subject to} \quad C(x+\delta) = t, \quad x+\delta \in [0,1]^n$$

Here, $C(x)$ represents the class label associated with image x.

The magnitude of the noise is evaluated using the Lp distance.

Determining the smallest L_p distance of δ — in a way that ensures the modified image is classified as t — presents a complex non-linear optimisation challenge. To navigate this complexity, CW introduces a function f which satisfies the condition $C(x+\delta) = t$ if and only if $f(x+\delta) \leq 0$. This simplifies Equation 1 to:

$$\text{minimise} \quad ||\delta||_p \quad \text{subject to} \quad f(x+\delta) \leq 0, \quad x+\delta \in [0,1]^n \quad (2)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

The suggested form for function f is:

$$f(x) = ([\max_{i \neq t} Z(x)_i] - Z(x)_t) + \kappa \quad (3)$$

Here, $Z(x)_t$ represents the DNN's output before the softmax activation is applied.

Observations:

1. The minimum value of $f(x)$ is 0, and this occurs when $Z(x)_t$ is less than or equal to $Z(x)_i$ for all i where $i \neq t$.
2. $f(x)$ grows as the probability of class t becomes lesser than the probabilities of other classes.
3. Drawing parallels to lasso regression, Equation 2 can also be expressed in the following manner:

$$\text{minimise} \quad ||\delta||_p + c \cdot f(x+\delta) \quad \text{subject to} \quad x+\delta \in [0,1]^n$$

Box Constraints:

The constraint $x+\delta \in [0,1]^n$ ensures that the image pixels remain within the range of 0 and 1, maintaining their validity as pixel intensity values. To achieve this, a change of variable technique is employed, introducing a new variable w_i such that:

$$\delta_i = \frac{1}{2} (\tanh(w_i) + 1) - x_i$$

Optimisation is then performed over w_i rather than δ_i, ensuring the constraints are satisfied.

Choosing c:

The parameter c plays a pivotal role in balancing the efficacy of the attack against its success rate.

1. Effectiveness: An attack is deemed more effective if the adversarial instance closely resembles the original image, meaning $||\delta||_p$ remains small.
2. Accuracy: An attack is accurate if it can successfully deceive the model into categorising the adversarial instance as the target class t.

The trade-off between these two factors is modulated by the value of c. The authors advocate for selecting the smallest c that still ensures the model misclassifies the adversarial instance to the target class t, or put formally, where $f(x^*) < 0$.

L2 Norm:

When $p = 2$, referring to the L_2 norm, the objective function is minimised using gradient descent.

10. Blackbox Attacks

Blackbox attacks involve attackers who lack knowledge of the model's internal workings, architectures, or parameters. They only have access to the model's input and output. This type of attack is common in situations where attackers attempt to exploit models without intimate access or complete visibility.

Transfer Attacks are an example of a blackbox attack. Adversary examples that were originally created to compromise one model are repurposed to attack another model in these attacks. The premise is that adversarial examples that work against one model are likely to work against other models, taking advantage of the transferability of adversarial perturbations.

Another method in this category is zeroth order optimization, which seeks to directly estimate the gradient of the targeted model by interacting with it via queries. Similarly, Query-based Attacks

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2064.v1

take advantage of the model by repeatedly querying it, allowing attackers to estimate its gradient and meticulously craft adversarial examples.

Finally, the HopSkipJumpAttack is a decision-based attack strategy. In this approach, the attacker, who has no knowledge of the model's gradients, relies solely on the model's outputs to carry out the attack. It depicts a scenario in which the attacker manoeuvres within the constraints of limited information to orchestrate successful compromises.

Each of these examples of blackbox attacks highlights the diversity and adaptability of adversarial strategies in scenarios where the attacker's knowledge is limited, emphasising the importance of robust defence mechanisms capable of thwarting attacks of varying sophistication and access levels in evolving machine learning environments.

In the examples used in **Error! Reference source not found.**, the transfer attacks represent adversarial examples generated for one model and are used to attack another model. The zeroth order optimisation directly estimates the gradient of the targeted model by querying it. This type of attack can be separated in querybased attacks, which are conducted by repeatedly querying the model, attackers estimate its gradient to craft adversarial examples, and HopSkipJumpAttack, where a decision-based attack where the attacker has no information about the model's gradients, only its outputs.

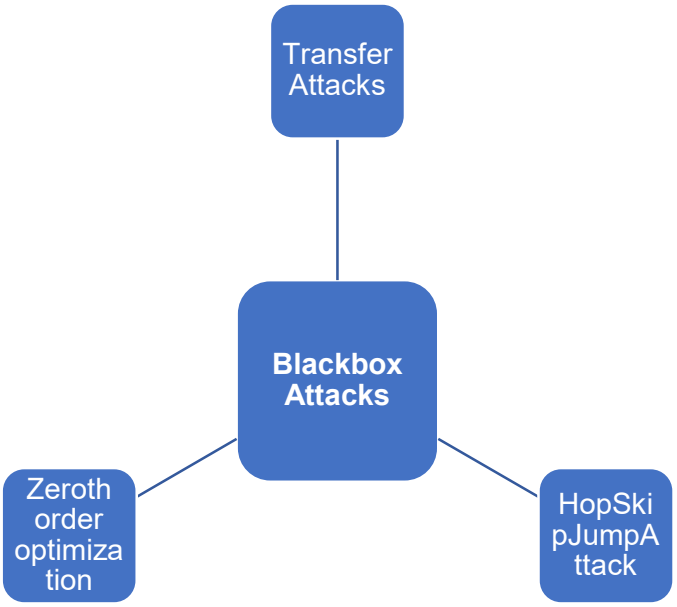


Figure 9. Blackbox attacks.

11. Targeted vs. Nontargeted Attacks

It is important to differentiate between targeted and non-targeted attacks when dealing with adversarial attacks. Targeted attacks are highly calculated and designed to make the model classify an input incorrectly. The attacker has a specific goal of misclassification in mind and manipulates the input to achieve this result.

Non-targeted attacks are designed to deliberately cause the model to misclassify input data without a predetermined class. The aim is to undermine the model's ability to accurately classify input data by exploiting its vulnerabilities, leading to various forms of misclassification.

The Basic Iterative Method is a classic example of a non-targeted attack (BIM). BIM can be thought of as an iterative improvement on FGSM. Rather than applying a single perturbation, BIM iteratively applies perturbations with small step sizes, gradually refining the adversarial input until the model misclassifies it. It demonstrates the adaptability of non-targeted attacks in exploiting model flaws to cause misclassifications without being limited to a specific incorrect class.

 doi:10.20944/preprints202309.2064.v1

It is crucial to distinguish between targeted and non-targeted attacks to create effective defence mechanisms. This helps to understand the different methods and intentions of adversaries who aim to compromise machine learning models.

Nontargeted attacks are a pernicious strategy that cybercriminals employ to deceive machine learning models into misclassifying an input. This tactic involves manipulating input data in such a way that the model is unable to correctly identify it. One of the most commonly used techniques for non-targeted attacks is the Basic Iterative Method (BIM). BIM is essentially a modified version of the Fast Gradient Sign Method (FGSM), which iteratively applies small perturbations to the input data. The ultimate objective is to achieve model misclassification, which can have dire consequences for Organisations and individuals alike.

12. Physical World Attacks

It is worth noting that adversarial examples can exist not only in the digital world but also in reality. For instance, adversarial stickers or patches are objects that can lead to misclassification when placed within the view of a camera. For example, a stop sign may be classified as a speed limit sign due to such stickers. Additionally, perturbations on 3D objects can also cause misclassification. By altering the texture or shape of objects, they can be wrongly classified. For instance, a 3D printed turtle may be mistaken for a rifle due to such perturbations.

Another example is the audio adversarial examples, which are sounds that can be crafted by malicious actors to manipulate speech recognition technology. Virtual assistants, for example, can be hijacked to execute unwanted commands if they are exposed to these types of sounds. The danger lies in the fact that these sounds are often undetectable to humans. Furthermore, there are specialized glasses called adversarial glasses that can deceive facial recognition technology. These glasses can be used to conceal one's identity or even impersonate someone else. It is crucial to stay aware of these risks and take appropriate measures to safeguard oneself.

13. Potential Risks and Consequences

Applications that require utmost safety measures are those that involve the possibility of misclassification in autonomous vehicles or drones, bypassing facial and voice recognition systems, and manipulating algorithmic trading and credit scoring models. These types of applications require precision and accuracy to ensure the safety and security of individuals and the general public. Given the potential consequences of errors or malfunctions in these critical systems, it is crucial to prioritize safety in their development and implementation.

The spread of misinformation has become a major concern in today's society. One of the ways in which this occurs is through adversarial attacks on deepfake detectors, which can result in false accusations and a lack of trust in the reliability of AI systems. These attacks can manipulate and deceive the AI-powered systems, which in turn can lead to inaccurate and misleading information being disseminated to the public. This can have serious consequences for individuals and society, as it erodes the foundation of trust and accuracy that is necessary for the proper functioning of our institutions and systems. Therefore, it is vital that we find ways to improve the accuracy and reliability of AI systems and develop effective methods for detecting and combating adversarial attacks on deepfake detectors.

Many areas of technology are vulnerable to exploitative attacks from malicious entities. One such area is healthcare, where adversarial attacks aimed at medical images can result in misdiagnosis. Such attacks can alter the images on which doctors rely for diagnosis, leading to incorrect treatment and potentially life-threatening outcomes. Similarly, surveillance systems can be compromised by attackers, who may deceive the system into overlooking potential threats. This can have severe consequences in areas such as border control or law enforcement, where the ability to detect and respond to threats is crucial. Digital assistants and smart homes are also vulnerable to attacks, with inaudible commands potentially leading to unwanted actions, breaches, or misinformation. This can be particularly problematic when it comes to sensitive data such as bank account information or personal data, which can be compromised by attackers. Additionally, social engineering can bypass biometric systems and grant unauthorised access to personal data or systems. This can have far-reaching consequences, including identity theft, financial loss, and potential harm to individuals or

 doi:10.20944/preprints202309.2064.v1

Organisations. All of these examples highlight the urgent need for robust security measures in technology, to safeguard against the potential for exploitative attacks.

14. Defensive Measures

During the training process, adversarial examples are introduced to enhance the model's robustness to potential threats. This is achieved through ensembling multiple models, which enables the averaging of their respective predictions. As an additional measure, pre-processing techniques such as JPEG compression and image smoothing are utilised to remove adversarial noise from the data. These strategies contribute to the creation of a more reliable and accurate model, which ultimately improves the overall effectiveness of the system.

Defensive distillation is a technique in machine learning that involves training a model to replicate the behaviour of another model. The approach is based on using less extreme output probabilities, which helps to increase the model's robustness and resistance to adversarial attacks. By imitating the behaviour of a more complex model, the distilled model can perform better in real-world scenarios, where it may encounter unexpected inputs or other sources of uncertainty. Overall, defensive distillation is a powerful tool for improving the reliability and safety of AI systems, especially in high-stakes applications such as autonomous driving, medical diagnosis, and financial forecasting.

To safeguard against attacks, there are several methods that can be utilised. One such method is Feature Squeezing, which involves the removal of extraneous features from input data, thereby restricting the search space for potential attackers. Another technique is Randomised Input Transformations, which confuses adversaries through the random transformation of inputs during inference. A third approach is Gradient Masking, which renders gradients uninformative in an effort to prevent attackers from using them to create adversarial examples. Finally, the Detection method involves training auxiliary models to recognise adversarial perturbations, as opposed to trying to achieve complete robustness against them.

15. Future Directions and Challenges

It has been observed that adversarial attacks have limited transferability between different models. This is because as defence mechanisms evolve and improve, the attacks also evolve to counter them. On the other hand, physical attacks are still an area of ongoing research, and their impact on machine learning models is yet to be fully understood.

It is crucial to comprehend the reasoning behind a model's decisions to uncover any potential deficiencies. One way to achieve this is through the utilisation of model interpretability tools. When it comes to poisoning attacks, the attacker focuses on manipulating the training data rather than the input data, which can result in subtle changes to the learned model. On the other hand, backdoor attacks entail the introduction of a specific pattern during training that can lead to incorrect outputs in the presence of that pattern during inference, even though normal input produces the correct output.

16. Discussion on the Multifaceted Nature of Adversarial Attacks

Autonomous Vehicles

Research on autonomous vehicles has shown that adversarial attacks can pose a real threat. Attackers can manipulate traffic signs to deceive vehicular systems while still appearing normal to humans. This difference in perception between humans and machines highlights the risks involved in using AI in dynamic environments. Therefore, it is crucial to develop strong and resilient models that can effectively counter adversarial manipulations.

DeepFakes and Multimedia

The use of adversarial techniques has progressed to the point where manipulated videos or deepfakes are becoming prevalent. These highly advanced manipulations are designed to deceive not only video classification systems but also human viewers, making it increasingly difficult to determine the authenticity of multimedia content. The emergence of deepfakes raises significant

 doi:10.20944/preprints202309.2064.v1

concerns about the integrity of information, underscoring the need for better detection mechanisms to prevent the spread of deceptive multimedia content.

Voice Assistants

There have been instances of successful adversarial attacks on audio and voice command systems, specifically voice assistants. These attacks involve voice commands that are difficult for humans to understand but can still be interpreted and executed by the assistants. This raises security concerns, highlighting the importance of implementing advanced security measures and defensive strategies to safeguard interactive voice-responsive systems from covert intrusions.

The various examples presented here demonstrate the constantly changing nature of adversarial attacks, which underscores the ongoing necessity for continued research and innovation when it comes to devising effective defence strategies. It is clear that the landscape of adversarial attacks is constantly evolving, and as such, it is imperative that we remain vigilant and proactive in our efforts to stay ahead of these threats. By staying up-to-date on the latest developments and investing in cutting-edge technologies, we can ensure that we are always prepared to combat any potential threats that may arise.

17. Conclusion

As the integration of AI continues to expand into various aspects of our daily lives, it is becoming increasingly important to understand and address potential adversarial threats. In the realm of AI security, the landscape is constantly evolving, demanding ongoing research and vigilance to stay ahead of potential risks and vulnerabilities. It is crucial to remain diligent in our efforts to counteract these threats and ensure the safety and security of both individuals and Organisations utilising AI technology..

During our thorough investigation, we uncovered the multifaceted nature of adversarial attacks in artificial intelligence and machine learning. We also examined the complexities of defending against them. The discussion included various attack methodologies, including the Fast Gradient Sign Method (FGSM) and the more intricate Carlini and Wagner Attack (C&W). We emphasised the different levels of sophistication and objectives within adversarial strategies.

In the world of cybersecurity, it is crucial to recognise the distinction between targeted and non-targeted attacks. These two types of attacks have different objectives, and it is important to consider these objectives when developing effective defense strategies. Targeted attacks are specifically designed to cause a particular misclassification or outcome, while non-targeted attacks aim to disrupt models without a specific predetermined goal in mind. One example of an adaptable and versatile technique used by adversaries to launch such attacks is the Basic Iterative Method. This technique involves multiple iterations of small perturbations to the input data, leading to a significant change in the output predictions. Understanding the nuances between these types of attacks is key to building robust and resilient defenses against adversarial threats.

As part of our investigation, we conducted black-box attacks. These attacks simulate situations where attackers who lack knowledge about the model's parameters or structure use transferability and query-based strategies to compromise the model. Our findings showed that adversarial strategies can adapt to overcome knowledge constraints, emphasising the need for robust defensive mechanisms in various attack scenarios.

Real-life examples, such as autonomous vehicles, multimedia content, and voice assistants, were used in case studies to show the practical implications and risks of adversarial attacks. The manipulation of traffic signs, creation of deepfakes, and execution of obfuscated voice commands emphasised the challenges in ensuring the security, authenticity, and reliability of AI-driven technologies.

In addition, examining different adversarial techniques has shown the need for continuous research and innovation in both offensive and defensive measures. With the rise of advanced methods like deepfakes and obscured voice commands, it is becoming more challenging to determine what is authentic and to preserve the reliability of AI systems in various fields, including multimedia and voice-activated systems.

This paper discusses the vulnerabilities, methods, and changing nature of adversarial attacks in AI. By doing so, it helps us better understand the interaction between these attacks and defence

mechanisms. The study of various attack methods and their real-world implications highlights the need for continual progress in creating resilient and secure AI models that can counter the diverse and sophisticated adversarial strategies present in today's AI landscape. The insights gained from this investigation will be essential in guiding future research and innovation to secure AI systems against the multifaceted threats posed by adversarial attacks. Ultimately, this will lead to the development of robust, secure, and reliable artificial intelligence technologies.

References

1. Y. Ruan and A. Durresi, "A survey of trust management systems for online social communities – Trust modeling, trust inference and attacks," *Knowl Based Syst*, vol. 106, pp. 150–163, 2016, doi: 10.1016/j.knosys.2016.05.042.
2. W. Z. Khan, M. Y. Aalsalem, M. K. Khan, and Q. Arshad, "When social objects collaborate: Concepts, processing elements, attacks and challenges," *Computers & Electrical Engineering*, vol. 58, pp. 397–411, Feb. 2017, doi: 10.1016/j.compeleceng.2016.11.014.
3. V. Kovtun, I. Izonin, and M. Gregus, "Reliability model of the security subsystem countering to the impact of typed cyber-physical attacks," *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–14, Jul. 2022, doi: 10.1038/s41598-022-17254-4.
4. E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *Journal of Information Security and Applications*, vol. 58, no. May 2021-102717, pp. 1–9, Apr. 2020, doi: 10.1016/j.jisa.2020.102717.
5. C. Maple, M. Bradbury, A. T. Le, and K. Ghirardello, "A Connected and Autonomous Vehicle Reference Architecture for Attack Surface Analysis," *Applied Sciences*, vol. 9, no. 23, p. 5101, Nov. 2019, doi: 10.3390/app9235101.
6. W. Wang, F. Di Maio, and E. Zio, "Adversarial Risk Analysis to Allocate Optimal Defense Resources for Protecting Cyber–Physical Systems from Cyber Attacks," *Risk Analysis*, vol. 39, no. 12, pp. 2766–2785, Dec. 2019, doi: 10.1111/risa.13382.
7. N. Ye, T. Farley, and D. Lakshminarasimhan, "An attack-norm separation approach for detecting cyber attacks," *Information Systems Frontiers*, vol. 8, no. 3, pp. 163–177, Jul. 2006, doi: 10.1007/s10796-006-8731-y.
8. P. Chejara, U. Garg, and G. Singh, "Vulnerability Analysis in Attack Graphs Using Conditional Probability," *International Journal of Soft Computing and Engineering (IJSCE) 13*, vol. 3, no. 2, pp. 18–21, 2013.
9. V. Schlatt, T. Guggenberger, J. Schmid, and N. Urbach, "Attacking the trust machine: Developing an information systems research agenda for blockchain cybersecurity," *Int J Inf Manage*, vol. 68, p. 102470, Feb. 2023, doi: 10.1016/J.IJINFOMGT.2022.102470.