International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 01 | Jan 2023 www.irjet.net p-ISSN: 2395-0072

Survey of Adversarial Attacks in Deep Learning Models

Sarala D.V*[1], Gaurav Sarkar*[2], Durgesh Nandini M*[3], Gagandeep B K[4*], Faizan Khurshid[5*]

[1] Asst. Prof in Dept. of CS&E, Dayananda Sagar College of Engineering, Bangalore-560078,Karnataka ,India,
[2,3,4,5] Students, Dept. Of CS&E, Dayananda Sagar College Of Engineering, Bangalore-560078, Karnataka, India

---

Abstract

Despite recent breakthroughs in a wide range of applications, machine learning models, particularly deep neural networks, have been demonstrated to be sensitive to adversarial assaults. Looking at these intelligent models from a security standpoint is critical; if the person/organization is uninformed, they must retrain the model and address the errors, which is both costly and time consuming. Attackers introduce carefully engineered perturbations into input, which are practically undetectable to humans but can lead models to make incorrect predictions. Hostile defense strategies are techniques for protecting models against adversarial input are called adversarial defense methods. These attacks can be performed on a range of models trained on images, text, time-series data. In our paper will discuss different kinds of attacks like White-Box attacks, Black-Box attacks etc on various models and also make them robust by using several defense approaches like adversarial training, Adversarial Detection, Input Denoising etc.

Keywords: Deep Learning Models; Adversarial Attacks; Adversarial Defenses.

Introduction

In diverse applications or under different limitations, model susceptibility to adversarial assaults has been established. Approaches for constructing adversarial samples, for example, have been proposed in tasks such as classification (e.g., on image data, text data, tabular data, graph data), object detection, and fraud detection. In our research work, we will demonstrate different kinds of adversarial attacks on various use cases and also implement defense methods to make the model robust. We will try to evade Malware classifier, Face recognition, Captcha model and Tweet sentiment based stock price prediction model.

[1] Adversarial Attacks in Deep Learning Models

Introduction: In recent years, deep learning related to computer vision, speech recognition, and language processing has developed quickly. Advancements in autonomous vehicles, language translators, and facial recognition have been made possible by machine learning and its applications in video, audio, text classifications, and other related domains. Recent research has demonstrated that even minor perturbations from adversarial instances can cause incorrect judgment and incorrect interpretation of the neural network examples. These attacks use hostile examples and misleading inputs to confuse the model, leading to incorrect categorization and subpar performance. Attacks usually produce adversarial images that, from a human perspective, are almost exact replicas of the original image. However, it could result in inaccurate classification results from the neural network model. Therefore, enhancing the safety and resilience of neural network models requires understanding adversarial example technology.

Algorithms and Methodologies:

Methods of Adversarial Attacks in Deep Learning Models: Fast Gradient Sign Approach - This is the simplest method for creating adversarial images, but it requires a request for reverberation propagation. It works by modifying the linearization loss function and solving to impact the disturbance of the maximum constraint. Basic iteration method: These techniques build on the Fast Gradient Sign Method, use it again in smaller increments, and then split the resulting pixels in half to make sure they still fall inside the original image's range. Impersonation and Dogging - By searching for the disturbance r, the approach optimizes the likelihood that the input x is recognized as the intended category. The second technique is dodging (no target class if it is misjudged): By searching for disturbance r, reduce the likelihood that input x will represent the true class. Text Classification: White-Box Adversarial Examples The team employed 10% of the training set as the development set for the model, which was trained using SGD and gradient clipping. The adversary example was only deemed effective in this experiment if the classifier misclassifier and reached a predetermined threshold score. Is BERT Reliable? Text Classification and Entailment: A Strong Baseline for Natural Language Attacks the TEXTFOOLER model, a baseline for natural language assault in a black box context, is the third model to be examined. It determines

which words in a sample are most important, and it continuously changes those words with others using a semantic approach until the prediction is altered. Three models were successfully attacked by the model using their framework in five text classification tasks and two text entailment tasks. It was successful in lowering the accuracy of the model to less than 10% with less than 20% of text modification.

Conclusion: In conclusion, a major barrier to the advancement of neural network technology has always been its interpretability. The resilience and security of neural networks are also being questioned more because to the introduction of adversarial sample technology. As a result, there is still much to about the explanatory investigations that may be done with adversarial sampling technology and neural networks. The present state of neural network and adversarial example technologies in picture, audio, and text recognition is summarized in this study

[2] Adversarial Attacks for the Deep Learning model using Efficient Gradient Integrated Attacking algorithm.

Introduction: As the use of artificial intelligence (AI) has increased across a wide range of applications, it has become more important than ever to ensure the security and robustness of the model. The researchers have therefore focused their attention on this model vulnerability. Changing the model leads to changing the weights, which leads to changing the predictions' behaviors. Map and surface decision alterations are often used assault techniques. This study suggests a mechanism based on FGSM called Efficient Gradient Integrated Attack (EGIA). Because of their capacity to achieve near-human-level performance on a variety of naturally occurring image analysis tasks, such as image categorization and image retrieval, deep neural networks are powerful models that have gained much attention in recent years. This is accomplished by using datasets like CIFAR-10, ImageNet. Given their importance, it is crucial to learn about effective graph representations that can facilitate a number of tasks farther along the processing chain, like node categorization. Deep Learning algorithms' amazing performance has sparked a boom in their employment in a wide range of applications that have a profound impact on how we live our lives. Because of this, it is crucial to understand how they fail and to come up with strategies for efficiently reducing risk. One of the most important risks that Deep Network Based systems are now recognized to face is their vulnerability to adversarial attacks, often known as produced undetected noise. This discovery has prompted intense research interest in ways to strengthen deep neural networks' defenses against attacks from adversarial input. Early attempts to increase resilience against adversarial assaults exclusively used single-step adversaries as training, while more recent work has included a multi step methodology.

Algorithms and Methodologies : One of the white box attacks, the fast gradient sign approach, raises the classifier's error rate. Basic iterative method: It is an improvement for the FGSM that is used to increase the effectiveness of attacks done with white weapons. The "diverse inputs approach" uses the "basic methodology" of padding the sized values and enlarging the image to a random size to achieve the greatest results. Transfer based black box attacks that use gradient-based techniques to create adversarial attacks are known as gradient change attacks. However, black box attacks are unable to deliver precise results in order to minimize losses. This essay discusses the issues raised by black box assaults. The Gradient Integrated Attack (GIA) is the name of the suggested architecture, and it can be used to attack FGSM, FGVM, etc. In the suggested model, attacks are added after each iteration to reduce the model's accuracy. Algorithm: List the directory link For loop on the class name folder Join the directory and image folder For loop on label folder Append images and labels in separate folder.

Conclusion: The Efficient Gradient Integrated modification can be used to any gradient algorithm method that can raise the loss function in order to increase the loss function in the planned model. By adding gradients every two iterations, the attacks' effectiveness is improved. The suggested model favored attacking effectiveness in the actual situation.

[3] Adversarial attacks on deep learning models in smart grids

Introduction: A smart grid builds an interoperable and distributed power delivery network using bidirectional flows of electricity and information. This system is far more efficient than a traditional power grid. A smart grid should use a variety of machine learning models for intelligent activities like load forecasting, problem detection, and demand response because it has access to massive amounts of data on energy generation, transmission, distribution, and consumption. However, because of the proliferation of deep learning applications, the machine learning models used in smart grids are primarily susceptible to sophisticated data attacks, particularly adversarial attacks. By swapping out natural inputs for specifically designed adversarial instances for a target model, adversarial attacks are inherently covert and capable of producing malicious results that are either random or targeted. An adversarial example should typically be difficult to distinguish from the original sample but very likely to produce a different output result, which is conceptually similar to generative adversarial networks (GANs) Techniques.

Algorithms and Methodologies: Adversarial evasion attempts - For a security-sensitive application, it is possible to actively modify the input samples in order to confuse the machine learning model that has been set up in the system. Through adversarial examples, the so called evasion attacks will intriguingly affect deep learning models. Many attack approaches, including L BFGS, FGSM, and DeepFool, are suggested as means of solving the optimization problem using gradient-based or restricted optimization. These techniques focus mostly on image classification models using unique input data (e.g., two- dimensional array). malicious poisoning attacks - When thinking about a deep learning model's training process, it is possible that, in many casesthe training data was obtained from the Internet and other shaky data sources under the prospect of adversarial poisoning attempts. Such a training procedure may lead the model to give inaccurate predictions for the majority of input points (poisoning availability attack) or a small number of specifically chosen input points (poisoning integrity attack) Auxiliary detection models can be built using adversarial training to determine if an input is hostile or not as a countermeasure. Gradient masking: Creating models with smoother gradients to thwart attack strategies based on optimization. Using statistical techniques, one can compare the distribution of honest inputs to those of adversarial cases. Preprocessing techniques: Using a variety of preprocessing techniques, a potentially hostile input can be converted to a lawful one. Utilizing measurements of the distance between valid inputs and adversarial examples and the decision boundary. A group of different classification models that can be selected at runtime make up an ensemble of classifiers.

Conclusion: Deep learning models are becoming more and more vital to intelligent tasks in smart grids as a result of recent advancements in deep learning technology. In order to best utilize the enormous quantity of data that smart grids continuously create; a variety of deep learning models have been implemented. Although deep learning models have many benefits, it has been demonstrated that these models may be vulnerable to subtle adversarial situations. As a result, in this work we provide a brief overview of the adversarial threats against deep learning models in smart grids. A quick summary of the research state of adversarial machine learning and deep learning applications in smart grids is provided. With particular emphasis, adversarial evasion and poisoning assaults in smart grids are examined and illustrated. Additionally, common defenses against these hostile assaults are provided. According to the report, the threat of hostile attacks on smart grids will essentially always remain. The conflicts between adversary attacks and defenses in smart grids will intensify over time and require our constant attention.

[4] Adversarial attack for NN-based smart grid state estimation

Introduction: Recently, safety-critical cyber-physical systems (CPS), including the smart grid, have employed deep learning. However, the security evaluation of such learning based techniques within CPS algorithms is still an unresolved issue. Despite research on adversarial attacks on deep learning models, safety-critical energy CPS, specifically the state estimate process, has received relatively little attention. This study investigates the security implications of neural network-based state estimation in the smart grid. After investigating the issue of adversarial attacks on neural network-based state estimation, an effective adversarial attack strategy is proposed. Since it is built on smart IoT devices and cognitive decision-making algorithms to achieve low loss, effective, and environmentally friendly power control, the smart grid can be considered as the largest internet-of-things (IoT) deployment. Cyberattacks pose a growing threat to the smart grid at the same time. Two defense strategies based on protection and adversarial training, respectively, are further suggested to foil this attack.

Algorithms and Methodologies: adversative assaults for the purpose of estimating the state of power systems, a forward derivative-based adversarial attack method is developed. The forward derivative calculation's gradient is similar to the backpropagation calculation's gradient, but there are two key differences: we directly use the network's derivative rather than its cost function, and we differentiate based on the input vector rather than model parameters. Random scaling attack (RSA) [52] involves the attacker randomly choosing a fraction of all meters and changing the chosen real measurements by a fixed scaling factor.

Defense Techniques: A defense strategy based on carefully selected meters is investigated for false data injection assaults against WLS state estimation. This indicates that different sensors also have wildly changing value in defense measures, just like in the case of attack creation. In state estimation using NN, this phenomenon is still present. Data encryption, authentication, access control, and auditing are linked processes and techniques for meter protection. An adversarial training-based defense strategy is used to increase a NN model's resistance to hostile attacks. This strategy involves retraining the model using a large number of created adversarial cases. This method's fundamental need is to generate as many adversarial cases as you can using the strongest attack constructing technique. Adversarial training can boost the classification accuracy of NNs and provide regularization for them. Therefore, a forward derivative-based method of attack is suggested, and two defense mechanisms are also offered to neutralize the threat. The NN model's ability to defend against adversarial attacks can be enhanced to some extent by both defense strategies.

Conclusion: A forward derivative-based method of attack is suggested, and two defense mechanisms are also offered to neutralize the threat. The NN model's ability to defend against adversarial attacks can be enhanced to some extent by both defense strategies.

[5] Adversarial Attacks and Defenses: An Interpretation Perspective

Introduction: In the introduction, we review recent research on adversarial attacks and defenses with a particular emphasis on machine learning interpretation. Model-level interpretation, feature-level interpretation are the two categories under which interpretation is classified. We go into more detail on how each type of interpretation can be applied to aggressive assaults and defenses. Then, we present a few other connections between interpretation and adversaries. Finally, we consider the challenges and potential solutions for interpreting antagonistic dilemmas.

Algorithm-Methodologies:

Feature-Level Interpretation for Understanding Adversarial Attacks:

1)Gradient-Based Techniques:

a)Improved Gradient-Based Techniques

b)Region-Based Exploration

c)Path-Based Integration

2)Distillation-Based Techniques

3)Influence-Function Based Techniques:

Feature-Level Interpretation for Adversarial Defenses:

1)Model Robustification With Feature-Level Interpretation:

2)Adversarial Detection With Feature-Level Interpretation.

Model-level Interpretation IN adversarial machine learning:

1)Model Component Interpretation for Understanding Adversarial Attacks

2)Representation Interpretation for Initiating Adversarial Attacks

Model-Level Interpretation for Adversarial Defenses: Conclusion: In this study, we combine recent breakthroughs in interpretable machine learning with current work on adversarial assaults and countermeasures. We divide interpretation strategies into two categories: feature-level interpretation and model-level interpretation. Within each area, we look at how the interpretation may be utilized to launch aggressive assaults or devise defense strategies. Following that, we address briefly various relationships between interpretation and adversarial perturbation/robustness. Finally, we examine the present obstacles of constructing transparent and resilient models, as well as some prospective future research and application avenues for adversarial samples.

[6] Targeted Mismatch Adversarial Attack: Query with a Flower to Retrieve the Tower

Introduction: Private user content, such as query images, must be made public in order to access internet visual search engines. We introduce the idea of a targeted mismatch attack for retrieval systems based on deep learning, which creates an adversarial image to cover the query image. Although the generated image bears no resemblance to the user's intended query, it produces identical or remarkably similar retrieval results. Attacks are difficult to spread to networks that are not familiar to anyone. We show successful attacks on partially known systems by building various loss functions for adversarial picture production. These include loss functions, for instance, for uncertain global pooling operations or the retrieval system's uncertain input resolution. We put the assaults to the test using common retrieval benchmarks and contrast the results with results from adversarial and original images. Algorithms and Methodologies: To substitute a target image for an adversarial one in a (concealed) query for image retrieval, the adversary attempts to create an adversarial image. Without revealing any data about the target image, itself, the aim is to get the same retrieval results.

Assumed a carrier picture xc with the same resolution as the target image xt RWxHx3 (see Figure 2). The adversary wants to create an image xa that is highly similar to the target's descriptors but has extraordinarily little aesthetic similarity. It is difficult to model visual (human) dissimilarity; instead, we model visual resemblance to another image, or the carrier. Targeted mismatch attack is the term we use to describe this problem, and the related loss function is provided by Ltr(xc, xt; x) = tr (x, xt) + ||x xc||. 2. We suggest various implementations of the performance loss tr based on the known and unknown test-model components. Additionally, we utilized Active Histogram and several image resolutions.

Conclusion: As a result of our discussion of targeted mismatch attacks for image retrieval, we may now construct hidden query images in place of the original intended query. We demonstrate that it is possible to produce images that produce the necessary descriptors without revealing the content of the intended query by optimizing the first order statistics. We examine the effects of image resampling, a standard feature of imageretrieval systems, and demonstrate the advantages of straightforward image blurring in adversarial image optimization. Finally, we demonstrate that transferring assaults are substantially more difficult than their image classification counterparts on new FCNs.

[7] Robust Text CAPTCHAs Using Adversarial Attacks

Introduction: In this study, we present Robust Text CAPTCHA, a simple method for creating text-based CAPTCHAs (RTC). The first phase involves creating the foregrounds and backdrops using randomly sampled font and background images, which are then combined to make recognizable pseudo-adversarial CAPTCHAs. We suggest and put into practice a highly transferable adversarial assault for text CAPTCHAs at the second stage to better thwart CAPTCHA solvers. Deep neural networks, random forests, shallow models like KNN, SVM, and OCR, as well as deep neural networks and OCR models, are all included in our investigations. Experiments show that our CAPTCHAs have a decent usability and a general failure rate of less than one millionth. Assailants' defensive strategies such as adversarial training, data pre-processing, and manual tagging are not effective against them.

Algorithms and Methodologies: For adversarial text CAPTCHA generation in the assault stage, we suggest the scaled Gaussian translation with channel shift attack(SGTCS). We use three methods—weighted spatial translation, image scaling, and channel shifts— to improve the transferability of adversarial cases. These methods might act as data augmentation to assist the attack method in escaping the local maximum. In this study, we present Robust Text CAPTCHA (RTC), a user-friendly text-based CAPTCHA generation method for real-world use. Our tests cover many different CAPTCHA solver models, such as shallow learning models, deep neural networks, and optical character recognition (OCR) systems.

Conclusion: RTC is effective and resilient against CAPTCHA solvers using a variety of training environments and defensive strategies. We also examine an extreme scenario where CAPTCHA solvers train on our RTC samples using human labeling, but their identification performance on new test RTCs with diverse backgrounds is still subpar. While obtaining a lower recognition rate than baseline techniques, our RTC nevertheless offers an elevated level of usability.

[8] Attacks on Regression Learning Using Data Poisoning and Corresponding Defenses

Introduction: We plan to analyze every facet of data tainting attacks on regression learning, going beyond prior research in both breadth and depth. We discuss real-world scenarios in which data poisoning attacks endanger production systems, as well as a unique Blackbox attack that is then applied to a real-world medical use-case. When only 2% of harmful samples are added, the regressor's mean squared error (MSE) increases to 150 percent.

Finally, we offer a special defense strategy against recent and past attacks, which we thoroughly test on 26 datasets. The trials show that the proposed defense strategy successfully lessens the examined assaults, as we infer.

Algorithms and Methodologies:

Algorithm 1 Flip attack

Algorithm 2 Trim defense

Algorithm 3 iTrim defense

Conclusion: In this research, we offer a unique matching defense mechanism and a data poisoning assault on regression learning. With the help of a thorough empirical analysis using seven regressors and 26 datasets, we demonstrate the efficiency of our suggested attack and defensive strategy. Attack and defense both presuppose reasonable limitations: The

attack is Blackbox and simply uses a fake dataset; it does not require access to the real dataset. The defense, on the other hand, uses an iterative approach to estimate the poisoning rate rather than assuming any prior knowledge of it.

[9] Stock Prediction Is Fooled by Adversarial Tweet Attack.

Introduction: Using a variety of adversarial assault situations, we examine three stock prediction victim models in this study. To complete the task of adversary creation, we address combinatorial optimization problems with semantics and financial constraints. Our findings show that the suggested attack strategy can cause significant monetary loss in a trading simulation by simply concatenating a changed but semantically similar tweet.

Algorithms and Methodologies:

Twitter trolls as the attack model. On Twitter, for example, adversaries can publish fraudulent tweets that are designed to influence models that use them as input. In order for them to be recognised as pertinent information and gathered as model input, we propose to attack by sending hostile tweets on Twitter that are semantically comparable as quotation tweets. Hierarchical perturbation is used in attack generation. The difficulty of our attack method is in choosing the best tweets and token perturbations under the restrictions of semantic similarity. In this study, we formulate the challenge as a hierarchical perturbation that entails the selection of tweets, the choice of words, and the perturbation of the word. The target tweets to be disrupted and retweeted are first chosen from a group of ideal tweets in the first stage. The word selection problem is then solved for each chosen tweet in the pool in order to choose one or more prime terms for disturbance. Additionally, word and tweet budgets are introduced to gauge the magnitude of the disruption. As a result, we substitute target words with synonyms from sets of synonyms that contain the words that are semantically closest to the target words as determined by how similar the GLOVE embeddings are.

Conclusion: We can conclude that despite physical limitations that prevent the raw tweet from being edited, our adversarial attack strategy reliably deceives different financial forecast models. A single quotation tweet that has simply one word changed can increase the attack's potential loss to our fictitious investment portfolio by 32%.

[8] Adversarial Attack on Speech-to-Text

Introduction: For automatic speech recognition, we build tailored audio adversarial samples. We can create an audio waveform that is over 99.9letters per second of audio while transcribing it as any phrase. We demonstrate the complete success of our Whitebox iterative optimization based assault on Mozilla's Deep Speech implementation. The viability of this assault opens up a brand-new field for researching adversarial examples.

Algorithms and Methodologies: Threat Model for methodology - Given an audio waveform x and a target transcription y, our job is to create an additional audio waveform x 0 = x + that is identical to x while also ensuring that C (x 0) = y. This task is described below. Only when the network's output exactly matches the intended phrase— that is, without any misspellings or additional characters—do we declare a success. Adaptability to pointwise noise. Even when the distortion is minimal enough to allow normal examples to continue to be classified as normal, adding pointwise random noise to an adversarial example, and returning C (x +) will result in the loss of the adversarial label for the example. Robustness when compressed with MP3. We employ the straight-through estimator to build adversarial instances that are resistant to MP3 compression after. In order to identify C (MP3(x 0)) as the target label, we compute the gradients of the CTC Loss under the premise that the gradient of MP3 compression is the identity function. Even though some gradient steps are undoubtedly erroneous, on average, the gradients are still beneficial. As a result, we are able to create hostile samples that have distortion that is roughly 15 dB higher and are resistant to MP3 compression. In summary, this research shows that specific audio adversarial instances are useful for improving automatic speech recognition. We demonstrate 100% effectiveness in converting any audio waveform into any target transcription using optimization-based attacks that are applied end-to-end. The theoretical maximum speed at which we can make audio transcribe is 50 characters per second. We can also have music transcribed as arbitrary speech and conceal speech from being transcribed.

Conclusion

Deep neural networks in particular have been shown to be particularly vulnerable to adversarial attacks, despite recent advances in a variety of applications. It is crucial to consider these intelligent models from a security perspective; otherwise, the person or entity would have to retrain the model and deal with the faults, which would be expensive and time-consuming. Attackers intentionally add input perturbations that are essentially unnoticeable to humans but can cause models to forecast incorrectly. A model can be protected against adversarial input using hostile defensive techniques, which are also known as adversarial defence strategies. A variety of models trained on text, graphics, or time

series data are susceptible to these attacks. White-Box attacks, Black-Box attacks, and other attack types will be demonstrated in our article on a variety of models. We'll also show how to use multiple defence strategies, such as adversarial training, adversarial detection, input denoising, etc., to make the models resilient.

References

1) Yucong Lai and Yifeng Wang Adversarial Attacks Technology in Deep Learning Models

2) Muttoni, A. Adversarial Attacks for the Deep Learning model using Efficient Gradient Integrated Attacking algorithm

3) Jingbo Hao∗, Yang Tao Adversarial attacks on deep learning models in smart grids

4) Tian, Jiwei; Wang, Buhong; Li, Jing; Konstantinou, Charalambos Adversarial attack and defense

5) Ninghao Liu, Mengnan Du, Ruocheng Guo, Huan Liu, Xia Hu Adversarial Attacks and Defenses: An Interpretation Perspective

6) Giorgos Tolias, Filip Radenovic, Ondˇrej Chum Targeted Mismatch Adversarial Attack: Query with a Flower to Retrieve the Tower

7) Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, Cho-Jui Hsieh Robust Text CAPTCHAs Using Adversarial Examples

8) Nicolas Muller, Daniel Kowatsch, Konstantin Bottinger Data Poisoning Attacks on Regression Learning and Corresponding Defenses

9) Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu,and Sanmi Koyejo1 A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction

10) Nicholas Carlini, David Wagner Audio Adversarial Examples: Targeted Attacks on Speech-to-Text