ADVERSARIAL MACHINE LEARNING EVALUATION SYSTEM

Salamatu Osanga Ibrahim, Muhammad Dahiru Liman, Wisdom Emmanuel,
Computer Science Department, Federal University Lafia, Nasarawa State Nigeria
Corresponding author: mlimand76@gmail.com

Received: September 14, 2024 Accepted: November 28, 2024

Abstract: Adversarial Machine learning Evaluation System is a system build to evaluate machine learning models against adversarial attack. Machine learning models have the ability to learn, however, this learning can only take place when there is enough and good data. Lack of data can lead to problem such as overfitting. When a model suffers from overfitting, practitioners usually used regularizations techniques to handle the problem. Many techniques were developed as highlighted in Liman et al. (2024). One of those techniques is adversarial training. This type of technique is usually done by adding noise to the input. Machine learning models are increasingly pervasive in critical applications, yet their vulnerability to adversarial attacks poses a significant challenge. This research presents the development and evaluation of the Adversarial Machine Learning Evaluation System, a tool built on the Adversarial Robustness Toolbox (ART). The system aims to democratize the review of a model's robustness by offering a user-friendly interface, allowing users to assess machine learning models' security against adversarial attacks. The methodology adopted is prototyping. An interface was built so that developers can test their models. The interface is friendly and usable. One of the test carried out shows that performing adversarial attacks takes longer time and consume resources than what was earlier anticipated.

Keywords: Adversarial, evaluation, machine learning, models, system, security, attacks, toolbox.

Introduction

Machine learning has emerged as a transformative technology, redefining the landscape of various industries and applications. Over the past decade, it has become a driving force behind automated decision-making, pattern recognition, and predictive modeling, fundamentally altering how we approach complex problem-solving. The profound impact of Machine learning is evident across diverse sectors, including healthcare, finance, autonomous systems, and more, where Machine learning algorithms are employed to optimize processes, enhance efficiency, and enable innovations that were once considered unattainable (Smith et al., 2018). The significance of Machine learning in today's world cannot be overstated. It has ushered in a new era of possibilities, from revolutionizing personalized medicine by predicting patient outcomes to enabling autonomous vehicles to navigate safely through complex environments. This technology has become deeply integrated into our daily lives, affecting everything from the recommendations we receive on e-commerce platforms to the way we interact with virtual assistants.

The widespread adoption of Machine learning algorithms across various domains has undoubtedly brought about transformative benefits. Machine learning models have become indispensable tools for automated decision-making, pattern recognition, and predictive modeling. However, this rapid integration of Machine learning technology into critical applications has unveiled a pressing concern: the vulnerability of Machine learning models to adversarial attacks (Carlini et al., 2018).

Adversarial attacks are a subset of security threats within the Machine learning domain. They involve the deliberate manipulation of input data to deceive a Machine learning model into producing incorrect or misleading outputs while maintaining the appearance of normalcy to human observers (Goodfellow et al., 2014). These attacks pose a substantial risk to the integrity, reliability, and security of Machine learning systems. The consequences of adversarial attacks can be severe, ranging from misclassifying everyday objects in image recognition to making incorrect medical diagnoses and even compromising the safety of autonomous vehicles (Carlini and Wagner, 2017).

Machine learning models have the ability to learn, however, this learning can only take place when there is enough and good data. Lack of data can lead to problem such overfitting. When a model is overfitting, then practitioners usually used regularizations techniques to handle the problem. Many techniques were developed as highlighted in Liman et al. (2024). One of those techniques is adversarial training. This type of technique is usually done by adding noise to the input.

The core problem lies in the lack of accessible and user-friendly tools for evaluating the robustness of Machine learning models against adversarial attacks. Current evaluation systems are often confined to academic research, requiring specialized expertise and substantial effort for implementation and analysis (Papernot et al., 2018). Furthermore, many of these systems do not support popular Machine learning algorithms that are widely used in real-world applications, thus limiting their practical applicability and adoption.

The critical demand, therefore, is for a comprehensive and user-friendly Adversarial Machine learning Evaluation System that supports a diverse range of Machine learning algorithms and frameworks. This system should empower users, including Machine learning practitioners, researchers, and industry professionals, to gain valuable insights into their model's vulnerability against adversarial attacks.

The development of an Adversarial Machine learning Evaluation System carries profound significance within the evolving landscape of Machine learning and its applications. In an era where Machine learning models are integrated into critical domains such as healthcare, finance, autonomous systems, and more, the security and reliability of these models have become paramount. The significance of this study lies in its potential to enhance the security of Machine learning models, thereby safeguarding their performance and trustworthiness in applications where errors can have significant consequences. By developing a user-friendly and

accessible evaluation system, this work seeks to democratize the security evaluation process. This is essential because currently, many evaluation tools are confined to academic research and require specialized knowledge to use effectively. The significance here is in empowering a broader audience, including Machine learning practitioners, researchers, and industry professionals, to assess and improve the security of their Machine learning models. The work's aim to provide clear and insightful feedback to users regarding their model's vulnerability to adversarial attacks is crucial for fostering transparency and understanding. By enabling users to assess the robustness of their Machine learning models comprehensively, this system empowers them to make informed decisions regarding defense strategies and model deployment. This transparency is essential for building trust in Machine learning systems, especially in applications where lives or financial well-being are at stake. The Adversarial Machine learning Evaluation System will also contribute to ongoing research in adversarial machine learning especially as it is built on an ongoing research work (ART). It provides a practical platform for experimentation and validation, potentially leading to the development of new defense strategies and evaluation metrics. Additionally, the system may help establish best practices for security assessment in Machine learning, benefiting the broader research and practitioner communities.

Numerous research papers have investigated various adversarial attack methodologies. Notable attack methods include the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini & Wagner (C&W) attacks, and their variations. Researchers have demonstrated the effectiveness of these attacks in exploiting vulnerabilities in machine learning models.

The literature highlights the development of defense mechanisms aimed at enhancing model robustness. Adversarial training, feature squeezing, and gradient masking are among the defense strategies explored to mitigate the impact of adversarial attacks.

Existing research has proposed and utilized evaluation metrics to assess model robustness against adversarial attacks. These metrics include accuracy under attacks, robustness curves, transferability assessments, and perceptual metrics such as SSIM and LPIPS.

Table 1: Summary of related works and their findings

Name of Author	Year	Title	Findings
Ducousso, Bardin, & Potet	2023	Adversarial Reachability for Program-level Security Analysis	This paper introduces adversarial reachability, a framework for reasoning about advanced attackers and their ability to exploit vulnerabilities. It presents a new symbolic exploration algorithm for analyzing program security under such attacks.
Goodfellow, McDaniel, & Papernot	2021	Adversarial Machine Learning	This book provides a comprehensive overview of adversarial machine learning, covering topics such as adversarial attacks, defense mechanisms, evaluation methodologies, and real-world applications.
Carlini & Wagner	2019	The State of the Art in Adversarial Machine Learning	This paper provides a detailed review of the state-of-the-art in adversarial machine learning research. It covers various attack methods, defense strategies, evaluation metrics, and open challenges in the field.
Demontis, A., Papernot, N., McDaniel, P., Chakraborty, S.	2019	Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks	Explores why adversarial attacks are transferable and analyzes the transferability of evasion and poisoning attacks.
Lacobucci et al	2017	Towards Understanding Black-box Adversarial Examples	Proposes a data reconstruction-based adversarial example detection method to fill in the gap of the problem of black-box adversarial example detection (BAD).
Carlini, N., & Wagner, D.	2017	Towards Evaluating the Robustness of Neural Networks	Proposes metrics for evaluating the robustness of neural networks against adversarial attacks and demonstrates their application on different models.
Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celikkanat, Z. T., Swami, A.	2016	Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples	Investigates the phenomenon of transferability in machine learning and demonstrates how to use adversarial samples to create black-box attacks.
Szegedy et al		Explaining and Harnessing Adversarial Examples	This paper introduces the Fast Gradient Sign Method (FGSM), one of the most fundamental adversarial attack methods. It demonstrates the ease with which adversarial examples can be generated and the significant impact they can have on model predictions.
Goodfellow et al	2016	Intriguing Properties of Neural Networks	This paper first introduces the concept of adversarial examples and demonstrates their existence in deep neural networks. It raises concerns about the security and robustness of machine learning models in real-world applications.
Moosavi-Dezfooli et al	2014	Adversarial Training Methods for Secure and Robust Deep Learning	This paper introduces the concept of adversarial training, a defense mechanism that involves training models on adversarial examples to improve their robustness against future attacks

Identified gaps.

Through the review of related literature, several gaps and opportunities in the field of adversarial machine learning have been identified:

User-Friendly Evaluation: While there is a growing emphasis on user-friendly interfaces, there remains a need for more accessible tools that can be used by a broader audience, including Machine learning practitioners, researchers, and industry professionals.

Evaluation Standardization: Standardized evaluation methodologies and benchmark datasets can contribute to more consistent and comparable assessments of model robustness across different studies.

Interdisciplinary Collaboration: Collaboration between researchers in machine learning, cybersecurity, and domain-specific areas can lead to more holistic solutions that address the unique challenges posed by adversarial attacks in various applications.

Materials and Methods

Methodology

Description of the Existing System

Machine learning models are becoming increasingly popular in various critical applications, such as healthcare, finance, and autonomous vehicles. However, these models are susceptible to adversarial attacks, where attackers manipulate inputs to cause the model to misbehave. Therefore, robust evaluation of Machine learning models against such attacks is crucial for ensuring their safety and reliability.

Currently, two main libraries are available for evaluating the robustness of Machine learning models against adversarial attacks, The Adversarial Robustness Toolbox (ART) and Clevehans. For this research, the ART library will be utilized.

The ART Library: This open-source Python library provides various functionalities for crafting adversarial attacks and evaluating model robustness.

However, it requires knowledge of Python and familiarity with its APIs, making it less accessible to users without programming experience.

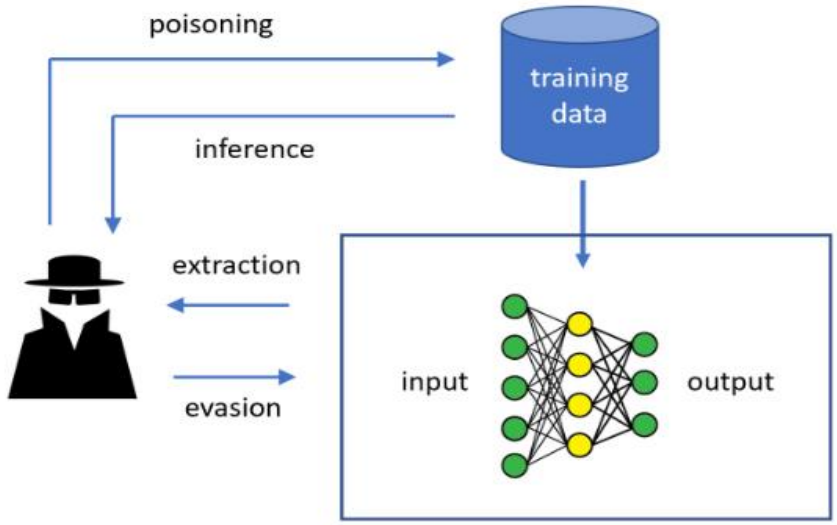

Figure 1: How ART works (Trusted AI, 2023)

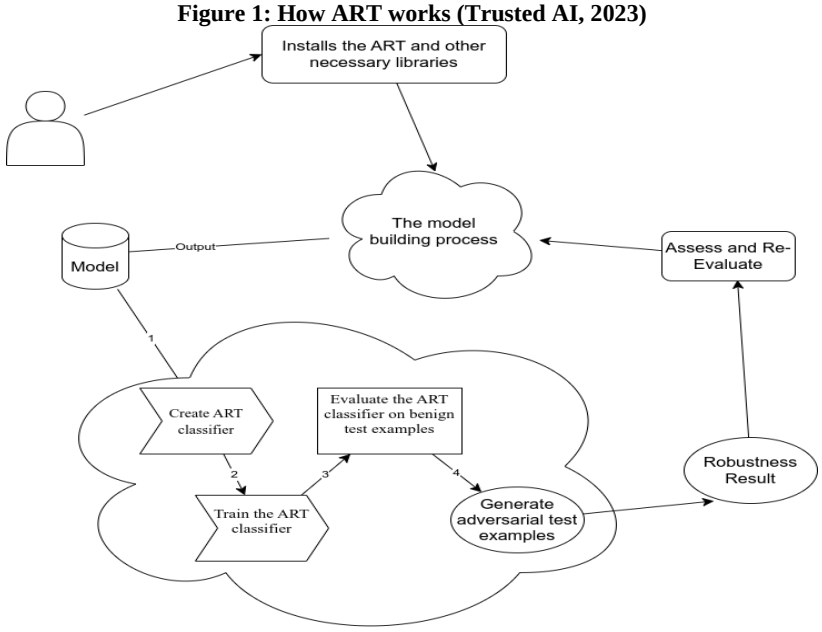

Figure 2: The adversarial robustness toolbox workflow

The Image in Figure 2 above shows the main workflow of creating machine learning models and accessing their security with the Adversarial Robustness Toolbox (ART). It highlights main processes like installing the

ART library and using it in the model-building phase immediately after building an initial model.

Strengths of the Existing System

Modularity and Extensibility: The ART framework is designed to be modular and extensible, enabling users to integrate custom attacks, defenses, and evaluation metrics.

User-Friendly API: The system provides a user-friendly API, facilitating an easy start for users engaging in adversarial machine learning.

Open-Source and Community-Driven: Being an open-source project, ART benefits from a large and active community of developers and researchers, fostering collaboration and improvement.

Weaknesses of the Existing System

These existing libraries, while valuable tools, have a significant limitation; they are only available as Python libraries, requiring programming knowledge for utilization. This restricts their accessibility to a broad range of users who may not possess programming skills, particularly domain experts without software development backgrounds.

This gap creates a need for a user-friendly and intuitive interface that allows users to easily assess the robustness of their Machine learning models against adversarial attacks, regardless of their programming background. This interface should offer a visual representation of the attacks, provide clear explanations of results, and enable easy manipulation of attack parameters for further exploration.

Secondly, while ART serves as a valuable tool for Machine learning model robustness evaluation, there is a need to bridge the educational gap by providing more accessible platforms that explain adversarial concepts and defenses to a broader audience, a gap that the proposed Adversarial Machine learning Evaluation System seeks to address.

Description of the Improved System

The proposed Adversarial ML Evaluation System represents a novel approach to addressing the limitations of the existing system by providing a more user-friendly and comprehensive solution for evaluating machine learning model robustness.

Justification for the New System

The development of the Adversarial ML Evaluation System is grounded in a single justification; addressing the limitations of the existing systems in terms of barrier to entry and use thereby fulfilling specific needs within the field of Adversarial Machine Learning (Sharma et al. 2022).

Methodology Adopted

The development of the Adversarial Machine Learning Evaluation System follows a prototyping methodology, emphasizing an iterative and incremental approach to system design and implementation. The key stages of the prototyping methodology are outlined below.

Initial Requirement Gathering

Engaging with potential users, including Machine learning practitioners, researchers, and industry professionals, to understand their needs, expectations, and challenges in evaluating Machine learning model robustness. Defining initial system requirements based on user feedback, literature review, and identified gaps in existing systems.

Initial Design and Mockups

Creating a high-level architectural design that outlines the integration with the Adversarial Robustness Toolbox (ART), web-based interface components, and educational content. Developing visual representations of the user interface, showcasing key elements such as model upload, algorithm selection, adversarial attack configuration, and educational resource integration.

Prototype Development

Implementing a functional backend with the Python Django Framework that integrates with ART, allowing for the generation of adversarial examples and model evaluations. Building an initial version of the web-based interface, incorporating key features outlined in the mockups. This prototype serves as a tangible representation of the system's functionality.

Iterative Refinement

Using collected feedback to make iterative improvements to both the backend functionality and frontend user interface. Incorporating additional features, refining existing ones, and addressing identified issues to enhance the overall system.

Specification of the Improved System

Dataset Description

As a focused approach, the system specifically supports models trained on the MNIST digits dataset. The MNIST dataset comprises 60,000 images for training and 10,000 for testing. While this introduces a limitation in terms of dataset scope, it allows for a more targeted evaluation of model robustness in image classification tasks as a model of a system that can be improved on.

The system was hard coded with the MNIST Dataset to manage the current system resources, but according to (Wiedeman, C et al. 2022), Adversarial attacks carried out on a class of machine learning tasks in a particular framework can be transferred to models built in similar frameworks.

The MNIST dataset, short for the Modified National Institute of Standards and Technology database, is a substantial collection of handwritten digits extensively employed for training diverse image processing systems. It is widely utilized in the realm of machine learning for both training and testing purposes.

Table 1: The MNIST Digits dataset and its attributes

Attribute Name	Attribute Description
Image	Gray-scale image of handwritten digit
Label	True digit value (0-9) corresponding to the handwritten digit
Pixel Value	Integer value between 0 and 255 representing the intensity of a pixel
Image Size	28x28 pixels

Input/output format

The system receives input in the form of text typed in or triggers such as buttons and also TensorFlow models with a ".h5" extension into relevant input boxes.

Technical Architecture of the New System

The technical architecture of the proposed Adversarial ML Evaluation System is designed to seamlessly integrate with the Adversarial Robustness Toolbox (ART) while introducing a user-friendly web-based interface. The key components of the technical architecture include:

a) **Web-Based Interface**
The user interacts with the system through a web-based interface, which serves as the primary entry point. This interface is designed for simplicity and intuitiveness, allowing users to upload their machine-learning models, specify algorithms, and select adversarial attacks for evaluation.

Figure 3: UI Design of the Adversarial ML Evaluation System

b) **Backend Integration with ART**
The backend of the system integrates with the Adversarial Robustness Toolbox (ART), leveraging its algorithms and methods for generating adversarial examples, implementing defense mechanisms, and conducting rigorous model evaluations. This integration ensures that the system benefits from the robustness and extensibility of ART.

c) **TensorFlow Framework Compatibility**
The system is designed to support models built with the TensorFlow framework using the MNIST dataset, enhancing its applicability in real-world scenarios. TensorFlow is widely used in both research and industry, and the compatibility ensures seamless integration with existing Machine learning workflows.

d) **Security Measures**
To address potential security concerns, the system incorporates measures to ensure the confidentiality and integrity of user-uploaded models and data. This includes secure data transmission and storage practices.

Results and Discussion

The system is designed for models trained on the MNIST digits dataset — the training and testing data is hard-coded in the system. TF/Keras models are expected as well. The System implements three attacks from the ART library.

I. **Fast Gradient Sign Method (evasion)**
The Fast Gradient Sign Method (FGSM) is a simple yet powerful technique for generating adversarial examples in machine learning. It leverages the gradient information of a model to craft small perturbations to an input, causing the model to miss-classify it. This highlights the vulnerability of models to subtle manipulations, raising concerns about their robustness in real-world applications. For the fast gradient method attack, the following parameters can be adjusted; epsilon, epsilon step, and batch size.

II.Poisoning Backdoor (poisoning)

Poisoning backdoor (also known as data poisoning) is a stealthy attack in adversarial machine learning that manipulates the training data to embed a hidden malicious functionality within the model. Unlike traditional attacks that directly target the model during inference, poisoning backdoors alters the model's behavior from the ground up, making them harder to detect. For the poisoning backdoor attack, you can adjust the following parameters: percent poison, and target labels.

III. **Copycat CNN (extraction)**

Copycat CNN is a technique in adversarial machine learning that allows an attacker to extract knowledge from a black-box model. Unlike poisoning backdoor attacks, which manipulate the training data, Copycat CNN focuses on extracting the model's decision-making process itself. For the Copycat CNN attack, you can adjust the following parameters: batch size fit, number of epochs, and size of the training set for the stolen classifier.

Testing the system by applying the FGSM attack to a pre-trained TensorFlow model

Step 1:

To check robustness using the Adversarial Machine Learning Evaluation System, click the Check My Model button

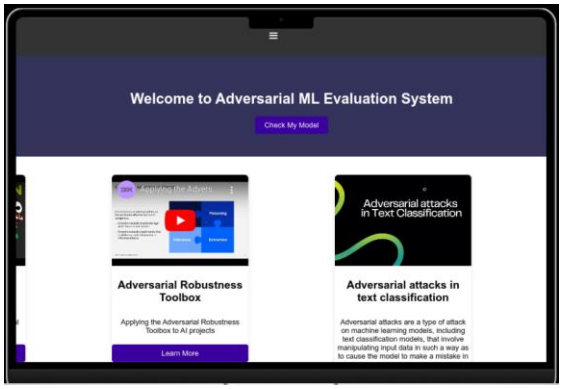

Figure 4: Image of Adversarial Machine Learning Evaluation System Home Page

Step 2:

After that, choose the kind of attack you want to apply to your model.

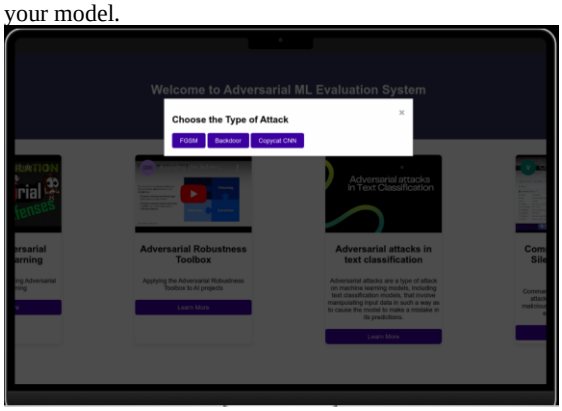

Figure 5: Select the Attack type

Step 3:

After that, you need to provide a robust and vulnerable model, then set the parameters Epsilon, Epsilon step, and Batch size.

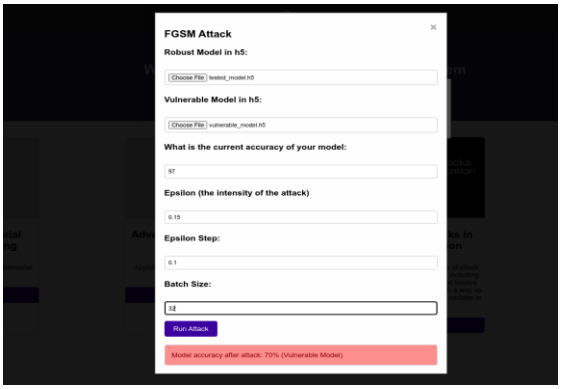

Figure 6: Result that shows that the model is vulnerable.

Step 4:

Finally, to run the attack, click Run Attack. After the attack has been sent and an evaluation has been gotten, you should see the evaluation result below the Run Attack button in the button's box.

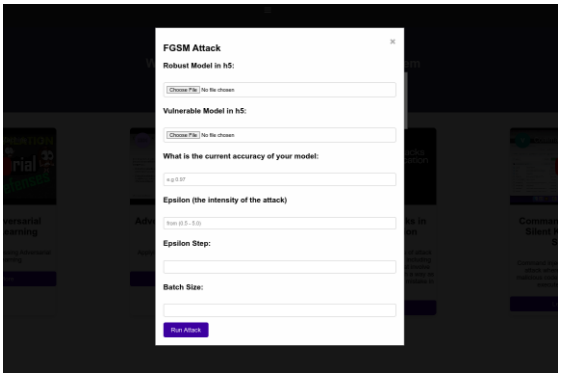

Figure 7: An image that shows when a user enters attack-specific configurations.

Results

In the course of testing, we observed some discrepancies between the expected results and the actual results expected. For instance, we found out that performing adversarial attacks takes longer time than we earlier anticipated and also consumes more resources than we thought would be used, this is why we hard-coded the dataset into the system and used a reduced portion of the dataset too.

Furthermore, we discovered that due to the limited resource in this work, the application may crash if high epsilon sizes and batch size are used when running attacks.

Despite these issues, we were able to successfully implement the system and obtain satisfactory results in line with our expectations to demonstrate the usefulness of the system.

Discussion

At the end of the timeline of this research, we successfully developed a platform for machine learning enthusiasts to check model robustness, enabling them to know when to re-assess and evaluate their models before deploying them into production in the real world. We also introduced the concept of adversarial attacks and estimations to both new and seasoned engineers with limited knowledge of the subject. While all attacks in the ART module were not implemented in the Adversarial

Machine Learning Evaluation system, we believe the module's focus on core attacks and its modular design has the potential to be expanded upon in the future to include a wider range of attacks and functionalities, making it a valuable tool for engineers of all levels.

Conclusion

In conclusion, the development and testing of the Adversarial Machine Learning Evaluation System have demonstrated its potential as a valuable tool for assessing the robustness of machine learning models against adversarial attacks. The system, built on the Adversarial Robustness Toolbox (ART), offers a user-friendly interface and incorporates essential features for evaluating model security. However, limitations, such as the hard-coded MNIST dataset and a partial implementation of adjustable parameters, should be addressed in future iterations.

The system's successful implementation and the identification of certain challenges during testing underline the importance of ongoing research in Adversarial Machine Learning. As models become increasingly integral to critical applications, continuous efforts are necessary to enhance evaluation methodologies, expand dataset compatibility, and explore additional attack algorithms. The system lays the foundation for further exploration in these directions, contributing to the evolving landscape of Adversarial Machine Learning.

Conflict of Interest

No conflict of interest

Reference

Carlini, N., & Wagner, D. (2018). Towards evaluating the robustness of adversarial example defenses. arXiv preprint arXiv:1705.07180.

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39-57). IEEE. https://doi.org/10.1109/SP.2017.40

Demontis, A., Papernot, N., McDaniel, P., & Chakraborty, S. (2019). Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In Proceedings of the 28th USENIX Conference on Security Symposium (pp. 1213-1228).

Ducousso, L., Bardin, S., & Potet, M.-L. (2023). Adversarial reachability for program-level security analysis. arXiv preprint arXiv:2302.00791.

Goodfellow, I. J., McDaniel, P., & Papernot, N. (2021). Adversarial machine learning. MIT Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Iacobucci, M., Nardi, Y., & Papernot, N. (2017). Towards understanding black-box adversarial examples. In International Conference on Machine Learning (pp. 5610-5619). PMLR.

Liman. M.D., Ibrahim. S.O., Alu. E.S., & Zakariya S. (2024). Regularization Effects in Deep Learning Architecture. Journal of the Nigerian Society of Physical Sciences. 6(2024).

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., & Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2574-2582).

Papernot, N., McDaniel, P., Wu, X., Celik, S., & Swami, A. (2018). Distillation as a defense to adversarial perturbations against deep neural networks. In Proceedings of the 2018 IEEE Symposium on Security and Privacy (SP) (pp. 582-597). IEEE.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celikkanat, Z. T., & Swami, A. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples arXiv preprint arXiv:1605.07277. https://arxiv.org/abs/1605.07277

Sharma, A., Singh, S., & Joshi, S. (2021). Evaluating the vulnerabilities in ML systems in terms of adversarial attacks. [Preprint]. ArXiv:2308.12918.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2016). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6052.

Trusted AI. (2023). How ART Works [Image]. Adversarial Robustness Toolbox. https://images.app.goo.gl/cmRDSmKAkXT4eSg78 (Accessed November 20, 2023).

Wiedeman, C., & Wang, G. (2022). Disrupting adversarial transferability in deep neural networks. Patterns, 3(5), 100472. https://doi.org/10.1016/j.patter.2022.100472