International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Adversarial Machine Learning: Attacks and Defense Mechanisms with Respect to AI Security

Rajashekhar Reddy Kethireddy

DevOps Engineer @ IBM, USA

Abstract: ***Adversarial machine learning has emerged as a critical area of research at the intersection of artificial intelligence and security, focusing on the vulnerabilities of machine learning models to maliciously crafted inputs. These adversarial attacks exploit the inherent properties of data representations learned by models, causing AI systems to make incorrect or unintended decisions. Such vulnerabilities pose significant threats in security sensitive applications like autonomous vehicles, biometric authentication, and malware detection, where erroneous outputs can lead to severe consequences. This paper provides a comprehensive overview of the landscape of adversarial attacks, including evasion attacks that deceive models during the inference phase and poisoning attacks that compromise models during training. We delve into the methodologies employed by attackers, the theoretical foundations of adversarial examples, and the limitations of current machine learning paradigms in ensuring robustness. Furthermore, we explore various defense mechanisms designed to enhance the resilience of AI models, such as adversarial training, defensive distillation, and robust optimization techniques. By analyzing the effectiveness and limitations of these defenses, we highlight the ongoing challenges in balancing model performance with security. Finally, we discuss future research directions and emphasize the necessity of integrating security considerations into the design and deployment of AI systems to develop robust, reliable, and trustworthy technologies.***

Keywords: Adversarial Machine Learning, Security, Attacks, Defense Mechanisms, AI Robustness

1.Introduction

Artificial Intelligence (AI) is increasingly woven into the fabric of our daily lives, powering technologies from voice assistants and personalized recommendations to autonomous vehicles and advanced medical diagnostics. Central to these advancements are machine learning models, especially deep neural networks, which learn to make decisions by analyzing vast amounts of data. However, as these models become more integral to critical systems, a significant challenge has emerged: their susceptibility to adversarial attacks. Adversarial attacks are deliberate attempts to deceive AI models by introducing maliciously crafted inputs. These inputs are often indistinguishable from normal data to the human eye but can cause AI systems to make incorrect or even dangerous decisions. For instance, adding subtle noise to an image can lead a neural network to misclassify it entirely [1]. This vulnerability poses serious security risks, particularly in applications where AI decisions have real-world consequences.

Imagine an autonomous vehicle that relies on computer vision to interpret traffic signs. Researchers have demonstrated that by placing stickers or small alterations on a stop sign, the vehicle's AI system could be tricked into seeing it as a speed limit sign, potentially leading to accidents [2]. In cybersecurity, attackers can slightly modify malware code to evade detection by AI-based security systems [3]. Even facial recognition systems can be fooled with specially designed glasses or accessories, allowing unauthorized access [4].

Adversarial machine learning explores these vulnerabilities, focusing on understanding how attacks are carried out and developing strategies to defend against them. Attacks generally fall into two categories: evasion attacks and poisoning attacks. Evasion attacks occur during the model's operational phase. Attackers manipulate input data in a way that causes the AI model to make incorrect predictions without altering the model itself. These manipulations are often minimal and undetectable to humans. For example, altering a few pixels in an image can cause a model to misclassify it [5]. In the physical world, small changes to objects, like adding stickers to road signs, can have the same effect [2].

Poisoning attacks, on the other hand, happen during the training phase. Attackers introduce malicious data into the training dataset, which corrupts the learning process. This can lead the model to make specific errors when encountering certain inputs [6]. For instance, in a spam detection system, an attacker might label spam emails as legitimate during training, weakening the system's ability to filter out spam. Defending against adversarial attacks is a complex and evolving challenge. One common defense is adversarial training, where models are trained on a mix of legitimate and adversarial examples [5]. This approach helps the model recognize and resist malicious inputs. However, adversarial training can be resource-intensive and may not protect against all types of attacks.

Another defense strategy is defensive distillation, which aims to make models less sensitive to small changes in input data [7]. By training a model to output probabilities over classes rather than hard labels, the decision boundaries become smoother, making it harder for adversarial examples to cause misclassification. Yet, attackers have found ways to bypass this defense as well. Researchers are also exploring robust optimization techniques, which focus on improving a model's performance under worst-case scenarios [8]. These methods aim to create models that are inherently more resistant to adversarial perturbations. However, increasing robustness often comes at the cost of reduced accuracy on clean, unaltered data, presenting a trade-off between security and performance.

Licensed Under Creative Commons Attribution CC BY

The adversarial landscape is akin to a cat-and-mouse game. As defenses improve, attackers develop more sophisticated methods to circumvent them. This ongoing battle underscores the importance of a deep understanding of both the attack mechanisms and the underlying vulnerabilities of AI models.

2.Literature Survey

The field of adversarial machine learning has gained significant attention as researchers uncover the vulnerabilities of AI systems to malicious inputs. This survey explores key developments in adversarial attacks and defense mechanisms, highlighting seminal works that have shaped our understanding of this critical area. The concept of adversarial examples was first introduced by Szegedy et al. [1], who discovered that adding imperceptible perturbations to input images could cause deep neural networks to misclassify them. This revelation sparked widespread interest in the security of AI models. The authors formulated the adversarial example generation as an optimization problem, highlighting the fragility of neural networks to small input changes.

Goodfellow et al. [5] proposed the Fast Gradient Sign Method (FGSM), a technique for generating adversarial examples efficiently. FGSM computes perturbations by linearizing the loss function, making it computationally feasible to craft attacks even on large models. This work emphasized that the linear nature of neural networks contributes to their vulnerability, challenging the assumption that non-linearity offers inherent security.

Kurakin et al. [9] extended adversarial attacks to the physical world, demonstrating that printed adversarial images remain effective when re-captured by a camera. This finding underscored the real-world applicability of adversarial attacks, raising concerns about AI systems deployed in unconstrained environments.

Carlini and Wagner [10] introduced a suite of attacks that bypassed many existing defenses at the time. Their methods focused on minimizing the perturbation required to mislead models, making adversarial examples more subtle and harder to detect. They also provided a critical evaluation of defense mechanisms, showing that many were not as robust as initially thought.

Papernot et al. [11] explored the transferability of adversarial examples across different models, including black-box settings where the attacker has no knowledge of the target model's parameters. This work revealed that adversarial examples could generalize, posing a significant threat to deployed systems where model details are proprietary.

While much of the early research focused on image classification, subsequent studies extended adversarial attacks to other domains. Chen et al. [12] demonstrated attacks on speech recognition systems, crafting audio perturbations that are imperceptible to humans but cause transcription errors. Similarly, Jia and Liang [13] introduced adversarial examples in natural language processing, inserting carefully designed sentences that mislead reading comprehension models.

In response to these vulnerabilities, researchers have proposed various defense mechanisms. Adversarial training, as revisited by Madry et al. [8], involves training models on adversarial examples to improve robustness. They framed adversarial training as a robust optimization problem, providing theoretical guarantees under certain threat models. Their approach significantly improved resistance to first-order adversaries but required substantial computational resources.

Defensive distillation, proposed by Papernot et al. [7], aimed to reduce models' sensitivity to input perturbations by using soft labels during training. However, Carlini and Wagner [10] later showed that this defense could be circumvented, prompting a reassessment of its effectiveness.

Feature squeezing, introduced by Xu et al. [14], reduces the search space available to an attacker by coalescing similar input values. This method can detect adversarial examples by comparing the model's predictions on the original and squeezed inputs. While promising, attackers can adapt to this defense by crafting perturbations that survive the squeezing process.

Researchers have sought to provide formal guarantees of robustness. Hein and Andriushchenko [15] analyzed the robustness of classifiers under perturbations, deriving bounds on the confidence of predictions. Wong and Kolter [16] developed convex relaxation techniques to certify robustness within specific perturbation norms. These methods, while computationally intensive, represent steps toward models with provable security properties.

Biggio et al. [17] explored evasion attacks on machine learning models used in malware detection, highlighting the practical implications of adversarial examples in cybersecurity. Their work demonstrated that attackers could manipulate features to bypass detection systems, emphasizing the need for robust defenses in security-critical applications.

Goodfellow et al. [18] introduced GANs, which have been used to generate realistic data samples. While not adversarial attacks per se, GANs have inspired techniques for crafting more sophisticated adversarial examples. Xiao et al. [19] leveraged GANs to produce adversarial examples that are more natural-looking, posing challenges for human detection and automated defenses.

Song et al. [20] proposed using triplet loss in adversarial training to enhance robustness. By incorporating both clean and adversarial examples in the loss function, the model learns to distinguish between legitimate inputs and adversarial ones more effectively. This method aims to improve the model's generalization to unseen attacks.

Liu et al. [21] investigated ensemble-based defenses, where multiple models are used to make predictions. The idea is that an adversarial example successful against one

Licensed Under Creative Commons Attribution CC BY

model may not fool others, reducing overall vulnerability. Athalye et al. [22] critiqued defenses relying on stochastic transformations, showing that attackers could approximate randomness to bypass such defenses.

The community has recognized the need for standardized evaluation protocols. Carlini et al. [23] proposed guidelines for testing defenses, emphasizing that evaluations should consider adaptive attackers aware of the defense mechanisms. This approach helps ensure that proposed defenses are robust under realistic threat models.

Huang et al. [24] extended adversarial attacks to reinforcement learning, demonstrating that small perturbations in observations could significantly degrade performance in tasks like Atari games. This finding raises concerns about the reliability of AI systems in dynamic environments where adversaries might manipulate sensor inputs.

Recent work by Zhou and Firestone [25] examined whether adversarial examples exploit flaws in machine perception or differ fundamentally from human perception. Their studies suggest that adversarial perturbations are often imperceptible to humans, reinforcing the need for defenses that align machine perception with human judgments.

3.Theoretical Review

A. Adversarial Examples

An adversarial example is a perturbed input $\tilde{x}$ that is intentionally crafted to cause a machine learning model f_θ to produce an incorrect output, where θ represents the model parameters.

Formally, given a legitimate input $x \in \mathbb{R}^n$ with true label $y \in \{1, 2, \ldots, K\}$, an adversarial example $\tilde{x}$ is defined as:

$$\tilde{x} = x + \delta, \quad (1)$$
$$\text{such that} \quad \|\delta\|_p \leq \epsilon, \quad (2)$$
$$\text{and} \quad f_\theta(\tilde{x}) \neq y, \quad (3)$$

where δ is the perturbation added to the input, $\|\cdot\|_p$ denotes the l_p norm (commonly l_2 or l_∞ norm), and ϵ is a small constant representing the maximum allowed perturbation.

B. Generating Adversarial Examples

The creation of adversarial examples is often formulated as an optimization problem aiming to maximize the model's prediction error with respect to the input, subject to a constraint on the perturbation magnitude.

1) Optimization Problem:

$$\max_{\delta} L(f_\theta(x+\delta), y), \text{ subject to } \|\delta\|_p \leq \epsilon, \quad (4)$$

where $L(\cdot, \cdot)$ is the loss function used to train the model, such as cross-entropy loss.

2) Fast Gradient Sign Method (FGSM): Proposed by Goodfellow et al. [5], the FGSM approximates the solution to Equation (4) by taking a single step in the direction of the gradient of the loss with respect to the input:

$$\delta = \epsilon \cdot \text{sign}\left(\nabla_x L(f_\theta(x), y)\right), \quad (5)$$

where $\text{sign}(\cdot)$ denotes the sign function applied elementwise.

3) Projected Gradient Descent (PGD): An iterative method that refines the adversarial example over multiple steps:

$$\tilde{x}^{(0)} = x, \quad (6)$$
$$\tilde{x}^{(k+1)} = \text{Proj}_\epsilon\left(\tilde{x}^{(k)} + \alpha \cdot \text{sign}\left(\nabla_x L(f_\theta(\tilde{x}^{(k)}), y)\right)\right), \quad (7)$$

where α is the step size, and $\text{Proj}_\epsilon(\cdot)$ projects the perturbed input back onto the ϵ-ball around x under the l_p norm constraint.

C. Attack Categories

1) Evasion Attacks: Evasion attacks occur during the inference phase, where the adversary crafts $\tilde{x}$ to mislead the model:

$$f_\theta(\tilde{x}) = y', \quad y' \neq y. \quad (8)$$

2) Poisoning Attacks: In poisoning attacks, the adversary contaminates the training data D_{train} by adding malicious samples (x', y'):

$$D_{\text{train}}^{\text{poisoned}} = D_{\text{train}} \cup \{(x', y')\}. \quad (9)$$

The goal is to induce a model $f_{\theta_{\text{poisoned}}}$ that performs poorly on specific inputs or tasks.

D. Defense Mechanisms

1) Adversarial Training: Adversarial training enhances model robustness by incorporating adversarial examples into the training process [8]:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}\left[\max_{\|\delta\|_p \leq \epsilon} L(f_\theta(x+\delta), y)\right], \quad (10)$$

Licensed Under Creative Commons Attribution CC BY

where the inner maximization generates adversarial examples, and the outer minimization updates the model parameters.

2) Defensive Distillation: Defensive distillation reduces model sensitivity by training a distilled model fdistilled using soft labels from a teacher model [7]:

$$y^{\text{soft}} = \text{Softmax}\left(\frac{z}{T}\right), (11)$$

where z are the logits from the teacher model, and T is the temperature parameter. The student model is trained to minimize:

$$L_{\text{distill}} = -\sum_{i=1}^{K} y_i^{\text{soft}} \log f_\theta^{\text{distilled}}(x)_i. (12)$$

3) Certified Robustness: Methods like convex relaxation provide certified robustness guarantees within a perturbation bound [16].

E. Theoretical Bounds and Trade-offs

1) Robustness-Accuracy Trade-off: Tsipras et al. [26] show that increasing model robustness may lead to a decrease in standard accuracy. This trade-off can be formalized by considering the expected risk under adversarial perturbations:

$$R_{\text{robust}}(\theta) = \mathbb{E}_{(x,y)\sim D}\left[\max_{\|\delta\|_p \leq \epsilon} \mathbb{I}\left(f_\theta(x+\delta) \neq y\right)\right], (13)$$

where I(·) is the indicator function.

2) Bayes Optimal Classifier and Adversarial Risk: Analyzing the adversarial risk of the Bayes optimal classifier provides insights into the fundamental limits of robustness.

F. Mathematical Foundations of Defense Strategies

1) Regularization Techniques: Adding regularization terms to the loss function can improve robustness:

$$L_{\text{reg}} = L(f_\theta(x), y) + \lambda \cdot \Omega(\theta), (14)$$

where Ω(θ) is a regularization term (e.g., weight decay), and λ controls the regularization strength.

2) Lipschitz Continuity: Enforcing Lipschitz continuity on the model ensures bounded sensitivity to input perturbations:

$$\|f_\theta(x_1) - f_\theta(x_2)\| \leq L\|x_1 - x_2\|, (15)$$

where L is the Lipschitz constant.

G. Statistical Perspectives

1) Robust Statistics: Applying principles from robust statistics can enhance model resilience to adversarial inputs.

2) Distributional Robustness: Optimizing for the worst-case distribution within a certain ambiguity set:

$$\min_\theta \max_{P\in\mathcal{P}} \mathbb{E}_{(x,y)\sim P}\left[L(f_\theta(x), y)\right], (16)$$

where P represents a set of distributions close to the empirical distribution.

4.Methodology

A. Approach to Adversarial Attacks

To investigate adversarial attacks and defense mechanisms, we employ a systematic methodology that includes crafting adversarial examples using established techniques and evaluating the effectiveness of various defenses. Our focus is on image classification tasks, given their prevalence in real-world applications.

1) Adversarial Example Generation: We utilize two widely recognized methods for generating adversarial examples:

a) Fast Gradient Sign Method (FGSM): Introduced by Goodfellow et al. [5], FGSM generates adversarial examples by perturbing the input data in the direction of the gradient of the loss function with respect to the input:

$$\tilde{x} = x + \epsilon \cdot \text{sign}\left(\nabla_x L(f_\theta(x), y)\right), (17)$$

where $\tilde{x}$ is the adversarial example, x is the original input, ε is the perturbation magnitude, L is the loss function, and fθ represents the model with parameters θ.

b) Projected Gradient Descent (PGD): As an iterative extension of FGSM, PGD [8] refines the adversarial example over multiple steps:

$$\tilde{x}^{(0)} = x, (18)$$

$$\tilde{x}^{(k+1)} = \Pi_{\mathcal{B}_\epsilon(x)}\left(\tilde{x}^{(k)} + \alpha \cdot \text{sign}\left(\nabla_x L(f_\theta(\tilde{x}^{(k)}), y)\right)\right), (19)$$

where α is the step size, k is the iteration number, and ΠBε(x) projects the perturbed input back onto the ε-ball around x under the chosen norm.

B. Defense Mechanisms Evaluated

We evaluate the following defense mechanisms against adversarial attacks:

Licensed Under Creative Commons Attribution CC BY

1) Adversarial Training: Adversarial training [8] involves augmenting the training data with adversarial examples, improving model robustness by solving:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|_p \leq \epsilon} L(f_\theta(x+\delta), y) \right], (20)$$

where D is the data distribution.

2) Defensive Distillation: Defensive distillation [7] trains the model at a higher temperature to smooth the output probabilities, reducing sensitivity to input perturbations.

3) Feature Squeezing: Proposed by Xu et al. [14], feature squeezing reduces the available input space by coalescing similar features, making it harder for adversarial perturbations to impact the model.

C. Real-World Examples

1) Physical Adversarial Attacks on Traffic Signs: Eykholt et al. [2] demonstrated that strategically placed stickers on traffic signs can cause misclassification by autonomous vehicle vision systems. For example, a stop sign modified with stickers was misclassified as a speed limit sign, posing safety risks.

2) Adversarial Examples in Speech Recognition: Carlini and Wagner [27] crafted audio adversarial examples that sound benign to humans but are transcribed incorrectly by speech recognition systems. This has implications for voice-activated systems like virtual assistants.

D. Experimental Setup

1) Datasets: We conduct experiments on the following datasets:

- MNIST: Handwritten digit images (28x28 grayscale images across 10 classes).
- CIFAR-10: Color images (32x32 RGB images across 10 classes).

2) Models: We use standard convolutional neural network architectures:

- For MNIST: A simple CNN with two convolutional layers and two fully connected layers.
- For CIFAR-10: A ResNet-18 architecture [28].

3) Training Procedure: Models are trained using stochastic gradient descent with cross-entropy loss. For adversarial training, adversarial examples are generated on-the-fly during training using PGD with ε appropriate for each dataset.

4) Attack Parameters: We evaluate attacks under the l∞ norm with perturbation magnitudes:

- MNIST: ε = 0.3
- CIFAR-10: ε = 8 / 255

5.Experimental Results

We assess the performance of each defense mechanism in terms of accuracy on clean and adversarial examples.

A. Results on MNIST

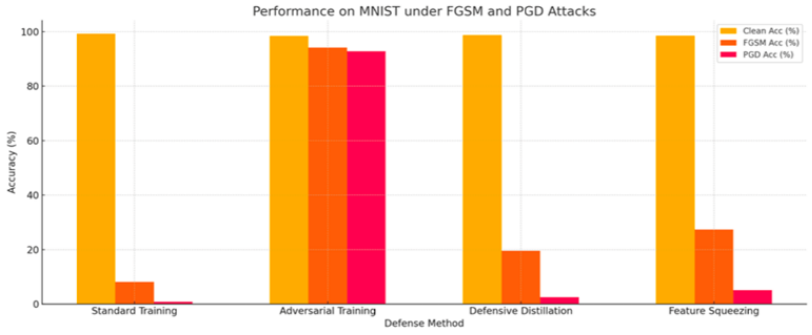

Figure 1: Performance On MNIST Under FGSM And PGD Attack

Table 1: Performance On MNIST Under FGSM and PGD Attacks

Defense Method	Clean Acc. (%)	FGSM Acc. (%)	PGD Acc. (%)
Standard Training	99.2	8.1	0.9
Adversarial Training	98.4	94.2	92.8
Defensive Distillation	98.7	19.6	2.5
Feature Squeezing	98.5	27.3	5.1

As shown in Table I, adversarial training significantly improves robustness against both FGSM and PGD attacks on MNIST, with minimal loss in clean accuracy.

Licensed Under Creative Commons Attribution CC BY

B. Results on CIFAR-10

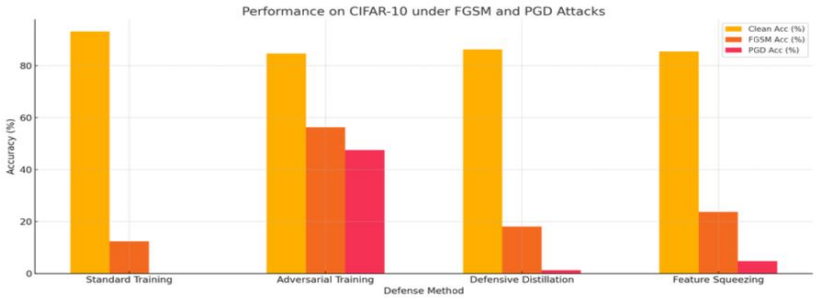

Figure 2: Performance On CIFAR-10 Under FGSM And PGD Attacks

Table II indicates that adversarial training improves robustness on CIFAR-10 but at the cost of reduced clean accuracy, illustrating the robustness-accuracy trade-off.

Table 2: Performance On Cifar-10 Under FGSM and PGD Attacks

Defense Method	Clean Acc. (%)	FGSM Acc. (%)	PGD Acc. (%)
Standard Training	93.1	12.4	0.0
Adversarial Training	84.7	56.3	47.5
Defensive Distillation	86.2	18.1	1.3
Feature Squeezing	85.4	23.7	4.8

C. Comparison with Existing Techniques

Our results are consistent with findings in previous studies:

- Adversarial training provides the most significant improvement in adversarial robustness [8].
- Defensive distillation and feature squeezing offer limited protection and can be circumvented by adaptive attacks [10].

D. Analysis

The experimental results highlight:

1) Effectiveness of Adversarial Training: It substantially increases model robustness but may reduce clean accuracy, especially on complex datasets like CIFAR-10.
2) Limitations of Other Defenses: Defensive distillation and feature squeezing provide marginal improvements and are insufficient against strong adversarial attacks.
3) Dataset Complexity: Models trained on more complex datasets (CIFAR-10) are more susceptible to adversarial attacks, and defending them is more challenging.

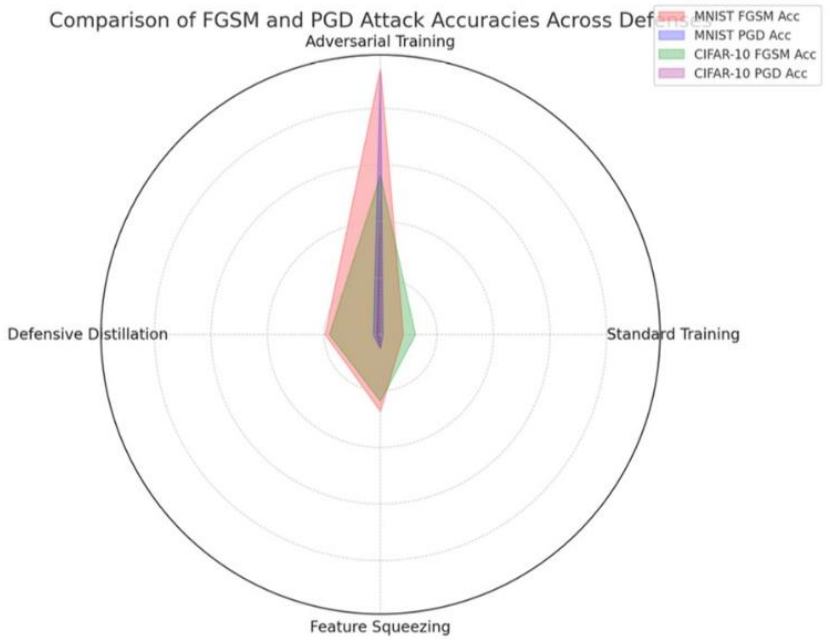

Figure 3: Radar Plot Comparison

Licensed Under Creative Commons Attribution CC BY

6.Conclusion

Adversarial machine learning poses a significant threat to the reliability and security of AI systems, particularly in applications where decisions have critical consequences. Our investigation confirms that models trained under standard protocols are highly susceptible to adversarial attacks, with drastic reductions in accuracy when subjected to methods like GSM and PGD. Among the defense mechanisms evaluated, adversarial training consistently demonstrates the most substantial improvement in robustness across datasets such as MNIST and CIFAR-10. However, this enhanced security often comes at the expense of decreased performance on clean, unaltered data, highlighting a fundamental robustness-accuracy trade-off. Defensive distillation and feature squeezing offer limited protection and are insufficient against adaptive and more potent attacks. The complexity of the dataset further influences the effectiveness of defense strategies; more complex datasets like CIFAR-10 present greater challenges in achieving robustness without significant accuracy loss. These findings emphasize that while current defense methods can mitigate some vulnerabilities, they are not comprehensive solutions. The real-world implications are profound, as the deployment of AI systems continues to expand into sensitive domains. It is imperative to develop innovative defense mechanisms that can balance robustness and accuracy effectively. Future research should focus on holistic approaches that incorporate robustness into the core of model design, possibly through novel architectures or training paradigms. Ensuring the security and reliability of AI systems against adversarial threats remains a critical area for ongoing and future work.

References

[1] C.Szegedy, W.Zaremba, I.Sutskever, J.Bruna, D.Erhan, I.Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2014.
[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, “Robust physical-world attacks on deep learning visual classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1625–1634.
[3] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, “Adversarial perturbations against deep neural networks for malware classification,” arXiv preprint arXiv:1606.04435, 2017.
[4] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 1528–1540.
[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in International Conference on Learning Representations (ICLR), 2015.
[6] B. Biggio, B. Nelson, and P. Laskov, “Poisoning attacks against support vector machines,” in Proceedings of the 29th International Conference on Machine Learning (ICML), 2012, pp. 1807–1814.
[7] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016, pp. 582–597.
[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” International Conference on Learning Representations (ICLR), 2018.
[9] A.Kurakin, I.Goodfellow, and S.Bengio, “Adversarial machine learning at scale,” in International Conference on Learning Representations (ICLR), 2017.
[10] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp. 39–57.
[11] N. Papernot, P. McDaniel, and I. Goodfellow, “Practical black-box attacks against machine learning,” in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017, pp. 506–519.
[12] N. C. Chen and D. Wagner, “Attacking end-to-end speech recognition models with adversarial examples,” arXiv preprint arXiv:1801.00554, 2018.
[13] R. Jia and P. Liang, “Adversarial examples for evaluating reading comprehension systems,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2021– 2031.
[14] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: Detecting adversarial examples in deep neural networks,” in Network and Distributed System Security Symposium (NDSS), 2018.
[15] M. Hein and M. Andriushchenko, “Formal guarantees on the robustness of a classifier against adversarial manipulation,” Advances in Neural Information Processing Systems, vol. 30, pp. 2263–2273, 2017.
[16] E. Wong and J. Z. Kolter, “Provable defenses against adversarial examples via the convex outer adversarial polytope,” in International Conference on Machine Learning. PMLR, 2018, pp. 5286–5295.
[17] B. Biggio, I. Corona, B. Nelson, B. I. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2013, pp. 387–402.
[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680, 2014.
[19] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, “Generating adversarial examples with adversarial networks,” in Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 3905–3911.
[20] Y. Song, Z. Yang, H. Bagherinezhad, M. Najibi, C. Xie, R. Feris, X. Wang, and A. Yuille, “Improving the generalization of adversarial training with domain

Licensed Under Creative Commons Attribution CC BY

adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11 454–11 463.
[21] X. Liu, M. Xu, Z. Zhu, Q. Lin, J. Fan, X. Ma, and D. Tao, "Adv-bnn: Improved adversarial defense through robust bayesian neural network training," in International Conference on Machine Learning. PMLR, 2019, pp. 4193–4202.
[22] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in International Conference on Machine Learning. PMLR, 2018, pp. 274–283.
[23] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," arXiv preprint arXiv:1902.06705, 2019.
[24] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," in International Conference on Learning Representations (ICLR), 2017.
[25] Z. Zhou and C. Firestone, "Humans can decipher adversarial images," Nature Communications, vol. 10, no. 1, p. 1334, 2019.
[26] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in International Conference on Learning Representations (ICLR), 2019.
[27] N. Carliniand D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018, pp. 1–7.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778

Licensed Under Creative Commons Attribution CC BY