International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 3, Issue 4, 77-85, 2022
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P109

Original Article

Adversarial Attacks and Defenses in Deep Neural Networks

Sunil Anasuri
Independent Researcher, USA.

Abstract - *The Deep Neural Networks (DNNs) have transformed a lot of fields such as computer vision, speech recognition, and natural language processing. Nevertheless, they are notoriously susceptible to adversarial attacks malicious inputs that can trick DNNs into giving wrong predictions that are imperceptibly wrong to a human observer. This weakness is of major concern, particularly in safety-sensitive exertions, like autonomous driving, medical diagnosis, and biometric verification. Today, this paper will discuss the adversarial attack ground and related defense mechanisms . An overview of adversarial attacks The overview of adversarial attacks covers why adversarial attacks? which includes, white-box attack, black-box attack, and transfer-based attack with their respective mechanisms composed of Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini-Wagner (CW), and DeepFool mechanisms. Subsequently, we observe the scope of defenses adversarial training, defensive distillation, input preprocessing, gradient masking. A comprehensive literature review presents historical evolutions and achievements in both attack generation and policies of mitigation. Methodology In the methodology section, an approach to testing adversarial robustness is presented with a standardized framework in terms of reproducibility and benchmark datasets like MNIST, CIFAR-10, ImageNet. We provide the outcomes of the comparative experiments in diverse threat models as well as implications of the research. Lastly, the paper gives a glimpse of the future of adversarial research and the importance of adaptable, strong, and interpretable models. The knowledge generated through our contribution synthesizes the preexisting ones and provides a basis on which strict and safe DNN systems can be developed.*

Keywords - *Adversarial Attacks, Deep Neural Networks, FGSM, PGD, Adversarial Training, Defensive Distillation, Cybersecurity, AI Robustness, Gradient Masking, Transferability.*

1. Introduction

1.1. Rise of Deep Neural Networks

Deep Neural Networks (DNNs) have transformed the sphere of artificial intelligence as it also allowed achieving success in such a great variety of tasks like image recognition, natural language processing, and speech synthesis. Its success is dictated by the knot of the potent computational facilities, [1-3] big-scale data, as well as the development of training regimens.

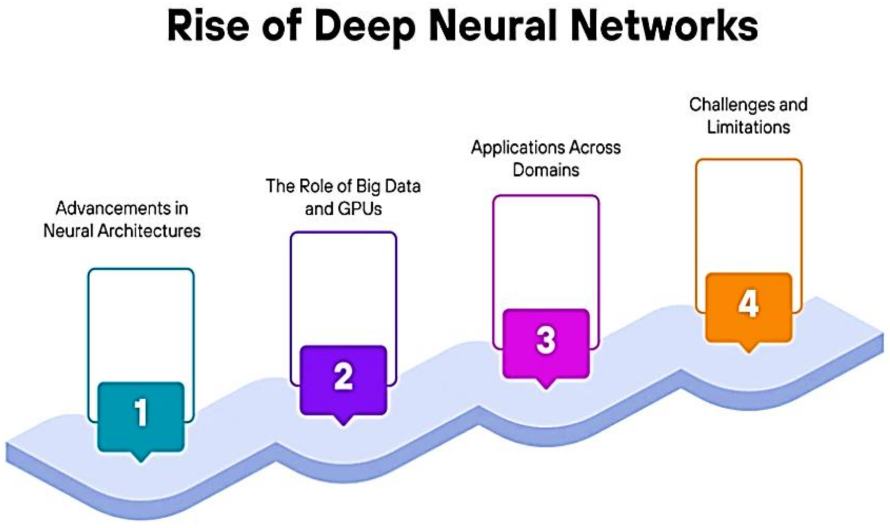
Rise of Deep Neural Networks
Advancements in Neural Architectures
1
The Role of Big Data and GPUs
2
Applications Across Domains
3
Challenges and Limitations
4

Fig 1: Rise of Deep Neural Networks

- **Advancements in Neural Architectures:** Neural networks in early times were shallow and could not be used to model complex patterns. This development gave rise to more powerful architectures like the Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) that observed a tremendous improvement. Networks such as AlexNet, VGGNet, and ResNet showed that when the network is many layers deep, the feature representations and, consequently, the accuracy may increase, especially in visual tasks. The architectural features enabled models to acquire hierarchical patterns via raw data.
- **The Role of Big Data and GPUs:** The availability of big training sets such as ImageNet increased the speed at which DNNs were developed by giving millions of example images that could be labelled to train them. At the same time, the training of deep models which would have been computationally impractical became possible due to the use of the Graphics Processing Units (GPUs). The co-operation between parallel computing infrastructure and big data preconditioned to train more pronounced and convoluted models.
- **Applications across Domains:** Deep neural networks are approximating state-of-the-art in many different applications. They surpass conventional techniques when it comes to object detection, segmentation and classification in computer vision. BERT and GPT are some of the models that have established new standards in sentiment analysis and machine translation in natural language. They can also utilize DNNs in autonomous systems, medical diagnosis, finance, and recommendation systems indicating their flexibility and influence.
- **Challenges and Limitations:** Even though DNNs have very good performance, it encounters multiple issues. They are memory-hungry, are computationally intensive and in most cases they are considered to be black boxes whose decisions can hardly be deciphered. Perhaps most importantly, they are vulnerable to adversarial attacks: with slight manipulation of inputs, we can change output to be wrong. Thus there is a reliability issue when it comes to using them in safety-sensitive applications.

1.2. Security and Trustworthiness Challenges

Although Deep Neural Networks (DNNs) have notably snatched in a large number of tasks, their adversarial robustness is still an issue of high concern. Among the most alarming weaknesses is the fact that they are vulnerable to adversarial examples, inputs that have the model acceptably modified in subtle, strategic ways, but are nearly the same to human observers. [4-6] Such perturbations, which may be hard to notice to the naked eye, can induce high-confidence incorrect classification, e.g., a stop sign as a speed limit sign or a cat as a dog. This activity is hazardous to security especially in applications that demand high security such as autonomous cars, medical diagnosis and verification or authentication where the associated cost of misclassifying may be huge. The fact that adversarial examples exist makes trust in machine learning highly questionable. In contrast to traditional software applications, failure conditions in DNNs are unintuitive and hard to be detected, especially when compared to the environments where failure conditions can be predicted and possible to test. The unpredictability undermines user trust and constrains the use of AI systems in applications such that reliability and interpretability are not negotiable. Moreover, the adversarial examples can transfer between various models and architectures, making the situation even worse, as such black-box attacks are possible even when the attacker knows nothing about the internal structure of the attacked model. The search to reduce these vulnerabilities has resulted in the formulation of a number of defense mechanisms; many of which have not been found effective against adaptive attacks where the attackers have taken into consideration the very defense mechanism when generating the attack. What is seen in the arms race of attack and defense is more of an even greater challenge that involves comprehension of the theoretical basis of the vulnerability of adversariality. Since DNNs become an increasingly essential part of both infrastructure and decision-making systems, it suggests that the issue of security and reliability is not a purely technical matter that one choses or does not choose to invest in, but rather a primary prerequisite of an ethical and responsible use of artificial intelligence.

1.3. Adversarial Attacks and Defenses in Deep Neural Networks

One of the red flags in the use of deep neural networks (DNNs) has been the adversarial attacks that have brought into light the vulnerabilities linked to their robustness. This kind of attack targets the careful construction of inputs that are virtually indistinguishable with clean data, except that they are used to induce a model to make erroneous predictions. Adversarial examples attack using the high-dimensional and intricate decision borders that neural networks can study, due to frequently interfering with input information with the direction of the gradient of the model loss. Such minor control is enough to mislead even cutting-edge models, which means their applicability to reality is questionable. Several mechanisms of attack have been formulated, but the main of them are quick and single-pass (such as the Fast Gradient Sign Method FGSM) and more powerful iterative attacks (Projected Gradient Descent PGD) and optimization-based attacks (the Carlini Wagner CW attack). These are the techniques that may be used both in white-box and black-box scenarios, depending on the access an attacker would have to a target model. As a reaction to these threats, numerous defense measures have been suggested. Perhaps the most well-known attempt is adversarial training, in which a model is retrained to make use of adversarial examples in order to be made more robust. Other strategies have been defensive distillation, input preprocessing methods (e.g., denoising or compression), and gradient masking, which tries to blur the gradients that can be used to create adversarial inputs. Nevertheless, most of these defenses have been found to be prone to adaptive attacks that are specially designed to overcome them. This recurrent game between attack and defense underscores the importance of developing a better theoretical model of the reasons behind DNNs susceptibility as well as what can be done to create models that are more robust to the attack. Overcoming these difficulties is fundamental to the promotion of comfortable and secure use of AI technologies in real practice.

2. Literature Survey

2.1. Early Discoveries of Adversarial Vulnerabilities

In 2013, Szegedy et al. first reported the phenomenon of adversarial vulnerabilities in deep neural networks (DNNs). They proved that one can apply tiny, well-designed adjustments to images that are quite imperceptible by the human eye but which trick DNNs into classifying an input with near-certainty. [7-10] These unforeseen conducts indicated a blind area in the manner neural networks make meaning and systematize patterns. The discoveries led to an impressive surge of subsequent studies about reasons behind these vulnerabilities and their implications to deploy machine learning models in real world, safety-sensitive applications like autonomous driving and facial recognition.

2.2. White-box and Black-box Attacks

The adversarial attacks can be classified in two broad categories: white-box and black-box. In white-box attack, the adversary knows the architecture and parameter of the model. A seminal white-box attack that operates by making changes in the direction of the gradient of the loss function with respect to the input features was proposed by Goodfellow et al. (2014) and is known as Fast Gradient Sign Method (FGSM). This method is effective and uses only one step to come up with adversarial examples Conversely, black-box attacks do not require the internal working of the target model. Papernot et al. (2016) were the first who tried similar attacks by training surrogate models that resemble to target. They also exploited the transferability property of adversarial examples and demonstrated that the examples created to deceive one model are likely to deceive other models as well, such as ones with different architecture or different training on data.

2.3. Defense Strategies

Several defense strategies have been suggested as a reaction to the increasing number of adversarial attacks to make DNNs even more robust. Out of this, adversarial training especially the one suggested by Madry et al. (2018) has been one of the most successful ones. It includes learning adversarial examples and augmenting the training data by them, therefore, instructing the model to properly label noisy inputs. The other strategies are defensive distillation, which attempts to make the model less sensitive due to the training using softened labels and input preprocessing, such as denoising and feature squeezing. Subsequent work has however revealed that most of these defenses can be broken by adaptive attacks, carefully designed opponents who take into consideration the very defense mechanism itself, and with the only remaining swinging challenge being designing defense mechanisms that we can state are undoubtedly robust.

2.4. Benchmark Datasets and Evaluation Metrics

To maintain consistency and comparability in the research on adversarial robustness, a number of benchmark datasets and evaluation protocols are defined. Most often, they work with such datasets as MNIST, CIFAR-10, ImageNet because of the diversity of their complexity and their popularity in the machine learning community. Among the metrics that are most likely used to assess the performance of the system is the attack success rate, i.e. the ability of adversarial examples to induce misclassification, the classification accuracy of the model under various attacks. The more complete description of model performance at a range of adversarial intensties can be obtained using robustness curves averaging accuracy over perturbation magnitude (usually a L 2 norm). These metrics and measures are essential to gauge and evaluate the outcome of the attacks and defenses in uniform and regular manner.

3. Methodology

3.1. Threat Model Definition

During an adversarial machine learning, the threat model specifies the capabilities and assumptions of the attacker, and it is paramount in determining the robustness of a system. [11-13] There are two broad categories of attacks; white-box, black-box and transfer-based.

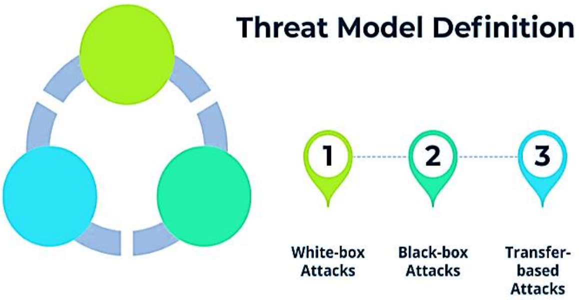

Fig 2: Threat Model Definition

- **White-box Attacks:** In a white-box environment, the adversary knows everything about the target e.g. architecture, parameters, and gradients. This lets create adversarial examples precisely and efficiently using techniques like optimization on the gradient. The white-box attacks are the worst-case scenario concerning model evaluation frequently due to the fact that because of access to the inner workings of the model, the adversary can tailor the attack to exploit it as directly as possible.
- **Black-box Attacks:** Black-box attacks postulate that the attacker does not have access to the internal parameters or structure of the model, only observing the model output (e.g., the labels of classes or confidence scores) to queries it provides. These attacks are normally based on the querying the model in order to evaluate the approaches such as finite-difference approximations or training alternative models to approximate the decision boundaries. Many practical applications have more realistic black-box scenarios in which the details of an internal model are proprietary or confidential.
- **Transfer-based Attacks:** The transferability of adversarial examples can also be used as an attack technique: that is, transfer-based attacks generate adversarial examples with respect to one model and transfer them to another. Here, the attacker uses the surrogate model to train the surrogate model to resemble the target and then uses adversarial examples to attack the target. Such examples are subsequently used to attack the target model. The importance of transfer based attacks is that they are especially valuable in the black-box settings and illustrate more general weaknesses that pervade models.

3.2. Attack Algorithms

The algorithm of adversarial attacks is designed to make minor (and usually non-noticeable) changes to data passed to an input of a machine learning model that will make it give the wrong outputs. [14-16] The following are some of the most manipulative and popular attack mechanisms:

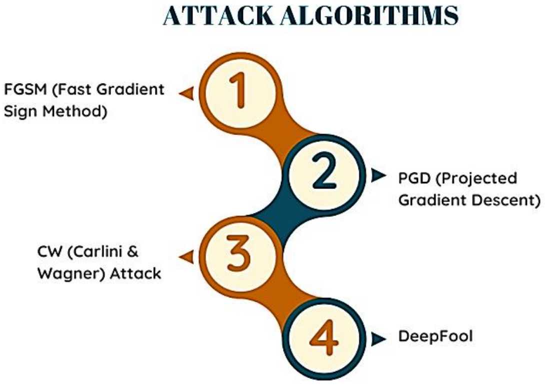
ATTACK ALGORITHMS
1
FGSM (Fast Gradient Sign Method)
2
PGD (Projected Gradient Descent)
3
CW (Carlini & Wagner) Attack
4
DeepFool

Fig 3: Attack Algorithms

- **FGSM (Fast Gradient Sign Method):** One of the commonly used fast and simple one-step attacks has been introduced by Goodfellow et al. (2014), and it is called FGSM, meaning FGM+the recently introduced gradient sign method (Goodfellow et al., 2014), which is a perturbation in the direction of the gradient of the loss function with respect to the input. The perturbation is magnified by a very small value that makes it notionally unnoticeable to humans but is also enough to trick the model. FGSM running time is linear in the dimensionality and many adversarial robustness papers consider FGSM as a baseline attack.
- **PGD (Projected Gradient Descent):** PGD is a multi-shot variant of FGSM developed by Madry et al (2018). It repeatedly applies tiny FGSM-like steps to incrementally increase the adversarial perturbation, and projects the outcome back into an Lp After every step, it -norm balls the original input to make sure the perturbation can be completed within a fixed perturbation budget. PGD has been viewed as a robust first-order attack and been extensively used to compare advances in adversarial training and defense.
- **CW (Carlini & Wagner) Attack:** The Carlini Wagner attack is one of the optimization-based approaches where the generation of adversarial examples is framed as a constrained optimization problem. It aims at identifying the minimum perturbation that leads to misclassification, and strikes a balance between the used perturbation scaling and the classification loss. CW attack is very powerful and capable of evading most of the previous defenses and this is the reason why it has been one of the most powerful and researched attacks.
- **DeepFool:** DeepFool by Moosavi-Dezfooli et al. (2016) is an iterative untargeted attack designed to determine the smallest perturbation necessary to modify the classification performed by the model. It operates through the linearization of decision boundaries of the model and corresponding projection of the input along lines towards the closest boundary until misclassification is achieved. DeepFool has a reputation of giving smaller perturbations than e.g. FGSM and PGD.

3.3. Defense Mechanisms

Countering adversarial attacks on neural networks is an open research area and a number of methods have been developed. [17-20] either these two defenses seek to strengthen the robustness of the model or ensure that the adversarial examples crafted by attackers are incapable of succeeding against the model.

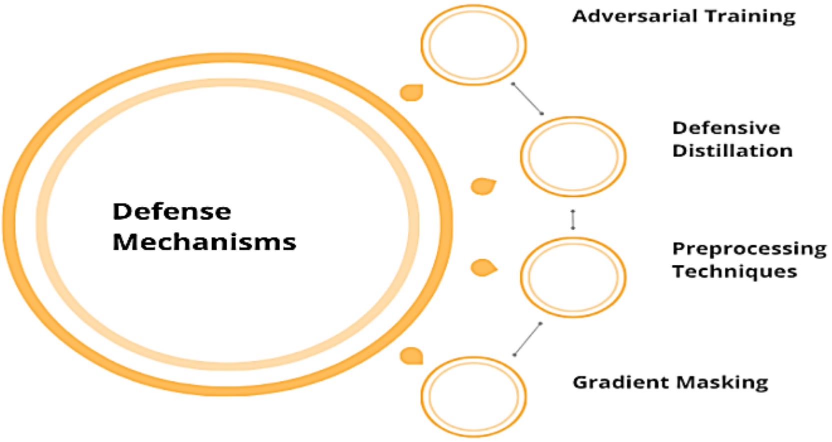

Fig 4: Defense Mechanisms

- **Adversarial Training:** One of the most used defense strategies is adversarial training. It takes the training data and adds adversarial examples to it and retrains the model in such a way that it can classify both unperturbed and perturbed data correctly. This procedure assists the model to acquire stronger decision boundaries. Particularly, the same researchers (Madry et al., 2018) showed that the adversarial training with a strong attack strategy such as PGD can make models much stronger under white-box attack.
- **Defensive Distillation:** Papernot et al. (2016) proposed a defense technique called defensive distillation based on the concept of knowledge distillation. A model is initially trained with softmax outputs of higher temperature, and then another model is trained to agree with the outputs of the first model. This operation helps to increase the robustness of the model against perturbations made on the inputs as it smooths the gradient landscape thus making it more difficult to create an effective perturbation. But afterwards, it was discovered that it can be evaded using adaptive attacks.
- **Preprocessing Techniques:** The protections offered by preprocessing defenses are aimed at diminishing adversarial perturbations prior to reaching the model. Such techniques consist of denoising input, applying JPEG compression, or feature-squeezing the input, i.e., downsizing the precision of variations shown in the input to clear out those that are unnecessary. Many of these techniques are easy to apply or can be integrated to existing models, but protection may be partial only and can hinder clean accuracy or could be beaten by well-crafted attacks.
- **Gradient Masking:** Gradient masking implements concrete approaches that deliberately conceal or otherwise manipulate the gradients of the model to make it impossible to exploit it effectively by an attacker aimed at constructing adversarial examples. Such methods of evasion as non-differentiable preprocessing layers or randomization may fool gradient-based attacks. Gradient masking is however regarded as a flimsy defense since it creates a false sense of security and in many cases, it is susceptible to transfer-based or black-box attacks.

3.4. Experimental Setup

In order to compare the capabilities of adversarial attacks and defenses, we make up an experimental framework that incorporates standard datasets, widely used architectures of neural networks, and broadly accepted evaluation measures. With the datasets, we consider MNIST and CIFAR-10 as they are the most frequently employed datasets in publications that study the claims on adversarial machine learning. MNIST is a collection of 70,000 images of handwritten digits (0-9), each is 28x28 pixels and represents a rather easy type of classification task. CIFAR-10, in turn, has a much more complicated and demanding evaluation of 60,000 color images of 10 categories (32x32 pixels) each. The latter datasets will enable us to answer the question of how generalizable attacks and defenses are to multiple levels of input complexity. We will use LeNet and ResNet-18 in the case of model structures. LeNet is a small convolutional neural network originally trained on digit recognition tasks such as MNIST, and thus it makes a good example of a baseline model to use with low-dimensional data. To experiment on CIFAR-10, the deeper and more advanced model that uses residual connection ResNet-18 is chosen since it makes very strong performance on complex image classification problems. These model selections guarantee that our experiments will model both low and high capacity networks to answer the question of whether the complexity of a given model influences the adversarial robustness. Evaluation wise, we will be basing on robust accurate where accumulated accuracy of the model is meant on adversary example contained within a specified budget of perturbation. We also evaluate performance under a range of norm-induced perturbation levels (e.g., at $L\infty$ bounded perturbation) in order to know how the model is sensitive to different degrees of adversarial noise. Also, a visual inspection of the adversarial examples is conducted so that it could be seen that the perturbations in question do not affect the human eye and that the attack success could be qualitatively evaluated. Collectively, they form an overall picture of the analysis of adversarial robustness in both attack and defence scenarios.

4. Results and Discussion

4.1. Robust Accuracy Comparison

We compare the models where we consider output classification accuracy on both clean data and data attacked in three state-of-the-art methods: FGSM, PGD and CW. The outcomes point to adversarial performance in decline, which makes it clear the weakness of each architecture in protection.

Table 1: Robust Accuracy Comparison

Model	Clean Accuracy	FGSM	PGD	CW
LeNet	99.2%	65%	45%	50%
ResNet18	93.5%	60%	40%	38%

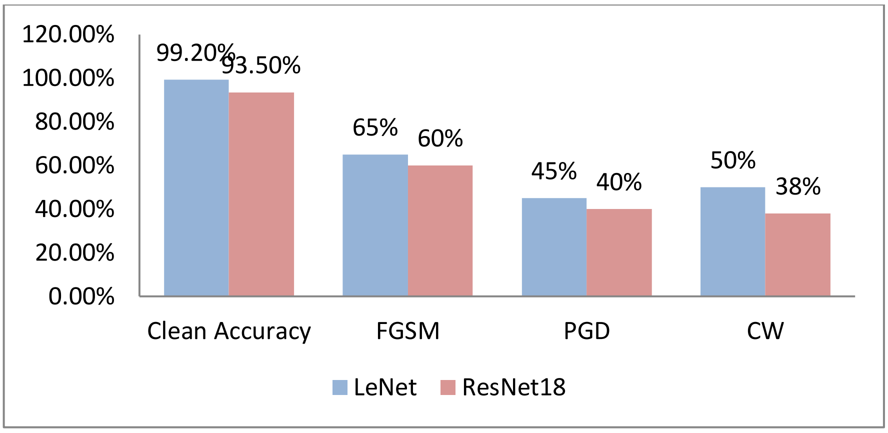

Fig 5: Graph Representing Robust Accuracy Comparison

- **LeNet:** LeNet model has a high clean accuracy of 99.2 on the MNIST making it have a highly performing model on clean data. But when the adversarial attacks happen, its resilience greatly reduces. When attacked using FGSM, the accuracy drops to 65, meaning that it is mildly resistant to basic one-step attacks. It is even worse under PGD where accuracy reaches 45% because this type of attack is more potent. The CW attack, which produces low perturbed but extremely effective examples, sets an accuracy of 50 percent, meaning that LeNET is found to be especially susceptible to an optimization-based attack even though it performs well on clean data.
- **ResNet-18:** ResNet-18 has a good clean accuracy of 93.5% on CIFAR-10, which shows that it is able to deal with complex image information. But just like LeNet, it has a steep decrease in performance because of adversarial effects. Its accuracy drops to 60 percent under FGSM and down to 40 percent when its adversary is the more powerful PGD attack. The CW attack is the best and it drops the accuracy to 38% accuracy. Such findings mean that even contemporary deep networks such as ResNet-18 are incredibly prone to adversarial noise and it is vital to have defense mechanisms incorporated in the training and evaluation phase.

4.2. Effectiveness of Defenses

In order to evaluate the effectiveness of different defense mechanisms in defending against adversarial attacks we compare the performance of various models in terms of adversarial attacks by FGSM, PGD and CW under the three settings (1) no defense, (2) adversarial training and (3) defensive distillation. As indicated in the results, it is evident that defense strategies have influence on model robustness.

Table 2: Effectiveness of Defenses

Defense	FGSM	PGD	CW
No Defense	60%	45%	50%
Adversarial Training	85%	70%	65%
Defensive Distillation	78%	60%	55%

- **No Defense:** When there is no protection, the accuracy of the model reduces drastically on adversarial attacks. The model has an accuracy of about 60 percent against FGSM, however, with PGD and the CW attack, the accuracy drops to 45 percent and 50 percent respectively. These findings enforce the concept of the explicit defense mechanism as unprotected models by far are complex models and are susceptible, particularly to more powerful, repeated or optimization-reliant attacks.

- **Adversarial Training:** The best defense in this domain is adversarial training. The adversarial training so far outperforms the trained model in withstanding attacks and maintaining 85 percent accuracy in FGSM, 70 percent accuracy in PGD, and 65 percent against the CW attack. This large enhancement on all types of attack shows that despite being exposed to carefully-designed perturbations, adversarial training helps the model to better generalize in case of the adversary during training in the form of strong ones, such as PGD.

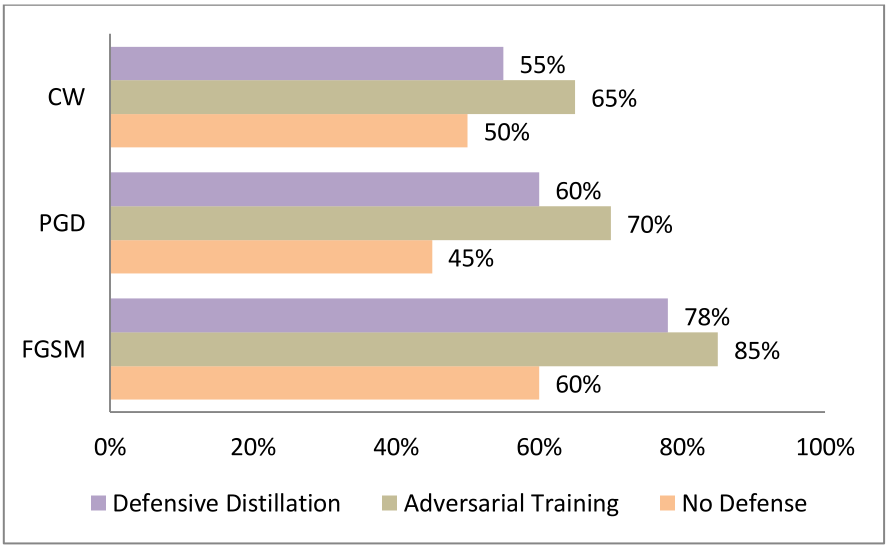

Fig 6: Graph Representing Effectiveness of Defenses

- **Defensive Distillation:** Like adversarial training, defensive distillation enhances robustness too, although to a lesser extent. The accuracy by the model on FGSM is 78 percent, 60 percent under PGD and 55 percent on the attack by CW. Although this approach smoothes the gradients of the model and causes problems to attackers because they are less able to create a perturbation, this trick has been displayed as less robust against adaptive and optimization attacks. Nevertheless, it is a suitable, although not the strongest, albeit a weaker defense, as it poses little to no effect on clean accuracy.

4.3. Insights

Several valuable conclusions about the essence of adversarial robustness and defense strategy efficiency can be drawn based on the results of the conducted experiment. To begin with, it is adversarial training which is the best possible way of resisting a wide variety of attacks, both gradient-based (e.g., FGSM, PGD) and optimization-based (e.g., CW). This is achieved by introducing adversarial examples into the training stage itself that will, in turn, result in a much more robust framework. This enhancement however, is at a high computational cost because the generation of adversarial instances in training particularly with powerful iterative approaches such as PGD is resource-demanding and time-consuming. Such a trade off between robustness and efficiency should certainly be taken into consideration particularly in real time or in resource limited systems. The other interesting phenomenon is the so-called transferability that adversarial examples created against a given model tend to be effective against another model with a similar architecture or decision boundary. This is an illustration of another inherent weakness of neural networks: models similarly trained on the same input are prone to having similar internal representations, and are therefore subject to cross-model attacks. The consequences are of special concern in black-box applications, where adversaries are able to use this property even without access to the target model, by attacking a surrogate. Finally, gradient masking based on obscuring or misleading gradient information to make it harder to generate an attack can give a false sense of security. These may only serve to temporarily decrease the effectiveness of some gradient-based attacks, but they do not make a system truly robust. Such defenses are commonly evaded by black-box or transfer-based tactics by adaptive attackers. Gradient masking in most of the instances just destroys the attack algorithm instead of making the model stronger. It is thus essential to make this distinction and understand that apparent robustness does not necessarily mean that a model is robust in all forms of attack and that a model with high values of measured robustness may not be robust against other harder and more adaptive attacks.

5. Conclusion

The paper covered a general overview of adversarial attacks and defense strategies of deep neural networks (DNNs) that slowly gain prominence as machine learning systems are already introduced to the real world. We started with an overview of the historical progress of adversarial machine learning, starting with the first discovery that small, distortions that were imperceptible to the human eye, could significantly confuse well-trained models. The main reasonable attack algorithms, namely, FGSM, PGD, CW, and DeepFool were outlined and examined, each of which forms a different level of adversarial input designing complexity. Under the defense column, we have discussed several defensive approaches, such as adversarial training, defensive distillation, preprocessing of the inputs and gradient masking. Our analysis was conducted in experimenting on the resilience of various models, namely LeNet and ResNet-18 against these attacks with and without the defensive counter-measures. The outcomes showed the weaknesses of even the most advanced architectures and the comparative effectiveness of various defenses, with adversarial training being the most successful one, however, it was also the most computationally demanding method.

In the future there are a number of research directions that look promising. Such avenues include developing certified defenses, which provide mathematical assurances that the predictions of a model will not change within a specified area of an input. Although existing certified procedures are generally very computationally costly and scale poorly, they are principled vehicle of strong robustness. The other promising line of research is adaptive learning, where models would not only protect against known attack tactics, but would adapt against novel or changing attacks in the field. This would reflect how biological systems respond to competitors, and would make defenses more resilient in the long term. Also, there is the objective of improving the explainability of the vulnerabilities being adversarial. Knowing which perturbations fail can also inform the development of inherently robust architectures and can be used to discern failure modes, occurring through interpretable model behaviors (or visualization techniques).

Finally, with deep learning models being heavily incorporated into life-sustaining fields like healthcare and finance as well as into fully autonomous systems, it is no longer a matter of choice whether they can be trusted and secure but rather a necessity. Adversarial threats provide a real challenge to the trustworthiness of AI, but by further studying strategies of attacks, the quality evaluation standards, and principled defenses, the community can strive to create not just accurate models but also robust, interpretable, and safe ones. The way ahead involves a degree of interdisciplinary effort and a long-term effort to comprehend and counteract the danger of overpowering learning systems that accompanies them.

References

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[2] Ozdag, M. (2018). Adversarial attacks and defenses against deep neural networks: a survey. Procedia Computer Science, 140, 152-161.

[3] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[4] Tirumala, S. S., Ali, S., & Ramesh, C. P. (2016, August). Evolving deep neural networks: A new prospect. In 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD) (pp. 69-74). IEEE.

[5] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In 2016 IEEE European symposium on security and privacy (EuroS&P) (pp. 372-387). IEEE.

[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

[7] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE symposium on security and privacy (SP) (pp. 582-597). IEEE.

[8] Zhang, Z., & Gupta, B. B. (2018). Social media security and trustworthiness: overview and new direction. Future Generation Computer Systems, 86, 914-925.

[9] Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155.

[10] Gu, S., & Rigazio, L. (2014). Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068.

[11] Carlini, N., & Wagner, D. (2017, November). Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM workshop on artificial intelligence and security (pp. 3-14).

[12] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204.

[13] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

[14] Moosavi-Dezfooli, S. M., Fawzi, A., & 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).

[15] Miller, D. J., Xiang, Z., & Kesidis, G. (2020). Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. Proceedings of the IEEE, 108(3), 402-433.

[16] Athalye, A., Carlini, N., & Wagner, D. (2018, July). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In International conference on machine learning (pp. 274-283). PMLR.

[17] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

[18] Bhambri, S., Muku, S., Tulasi, A., & Buduru, A. B. (2019). A survey of black-box adversarial attacks on computer vision models. arXiv preprint arXiv:1912.01667.

[19] Dodge, S., & Karam, L. (2017, July). A study and comparison of human and deep learning recognition performance under visual distortions. In 2017 26th international conference on computer communication and networks (ICCCN) (pp. 1-7). IEEE.

[20] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[21] Pappula, K. K., & Anasuri, S. (2020). A Domain-Specific Language for Automating Feature-Based Part Creation in Parametric CAD. International Journal of Emerging Research in Engineering and Technology, 1(3), 35-44. https://doi.org/10.63282/3050-922X.IJERET-V1I3P105

[22] Rahul, N. (2020). Optimizing Claims Reserves and Payments with AI: Predictive Models for Financial Accuracy. *International Journal of Emerging Trends in Computer Science and Information Technology*, *1*(3), 46-55. https://doi.org/10.63282/3050-9246.IJETCSIT-V1I3P106

[23] Enjam, G. R. (2020). Ransomware Resilience and Recovery Planning for Insurance Infrastructure. *International Journal of AI, BigData, Computational and Management Studies*, *1*(4), 29-37. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V1I4P104

[24] Pappula, K. K., Anasuri, S., & Rusum, G. P. (2021). Building Observability into Full-Stack Systems: Metrics That Matter. *International Journal of Emerging Research in Engineering and Technology*, *2*(4), 48-58. https://doi.org/10.63282/3050-922X.IJERET-V2I4P106

[25] Rahul, N. (2021). Strengthening Fraud Prevention with AI in P&C Insurance: Enhancing Cyber Resilience. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 43-53. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I1P106

[26] Enjam, G. R. (2021). Data Privacy & Encryption Practices in Cloud-Based Guidewire Deployments. *International Journal of AI, BigData, Computational and Management Studies*, *2*(3), 64-73. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I3P108