ISSN: 2831-6789

Research Article

Journal of Robotics and Automation Research

Strategic Defense in Machine Learning: Assessing the Most Optimal Defense Approach to Mitigate Adversarial Cyber Attacks

Jay Kim*

dslite.ai, Yorba Linda CA 92886, USA

***Corresponding Author**
Jay Kim, dslite.ai, Yorba Linda CA 92886, USA.

Submitted: 2025, Mar 15; **Accepted:** 2025, Apr 11; **Published:** 2025, Apr 21

Citation: Kim, J. (2025). Strategic Defense in Machine Learning: Assessing the Most Optimal Defense Approach to Mitigate Adversarial Cyber Attacks. *J Robot Auto Res, 6*(2), 01-04.

Abstract

In the era of AI proliferation, developing robust defense mechanisms against adversarial cyberattacks is critical. This project focuses on identifying and evaluating the most effective defense strategy to protect AI models from adversarial attacks. To mitigate overfitting, the baseline AI model was constructed with 2 convolutional layers, a dense layer of 256 nodes, pooling, and dropout layers. This foundational model demonstrated exceptional proficiency, achieving a 99.5% accuracy rate on the Modified National Institute of Standards and Technology (MNIST) dataset. The next three defense methodologies: adversarial training (integrating perturbed images into the training regimen), defensive distillation (employing softened probability distributions to enhance data generalization), and gradient masking (nullifying unused gradients to obscure potential attack vectors) were explored. Each method was applied to train distinct defense-augmented versions of the AI control model. The effectiveness of these defense strategies was tested against the Fast Gradient Sign Method (FGSM) attack which manipulates test images to deceive AI models. Each defense-enhanced model was evaluated based on its ability to maintain accuracy in the face of these cyberattacks. This analysis aims to contribute significantly to the field of AI cybersecurity, offering insights into the most viable strategies for safeguarding AI systems against sophisticated adversarial threats.

1. Introduction

As technological advances evolved at a more rapid pace than ever through artificial intelligence, the proliferation of advanced neural networks has brought many groundbreaking advancements across a litany of fields [1]. However, this progress has also unveiled new vulnerabilities that expose models to adversarial attacks. These attacks are techniques that are specifically designed for model manipulation by utilizing minor perturbations within the input data. This creates incorrect predictions which although may seem indifferent to the human eye, can have detrimental effects on the machine-learning process. In addition, the expanding role of machine learning in vital areas like image recognition, natural language processing, and autonomous vehicles has made protecting against adversarial attacks more essential than ever before. For example, university researchers from Washington, and Michigan have demonstrated this by adding slashes, and other images on to stop signs, causing AI-driven cars to misclassify images almost 100% of the time. This is a testament to the profound real-world consequences that these attacks can have, including things such as misclassification of medical images or the manipulation of self-driving car perception systems, being uniquely able to transcend traditional cybersecurity paradigms [2].

Luckily, despite the mounting threat posed by adversarial attacks, researchers and cybersecurity experts have proposed a multitude of solutions all aimed towards mitigating the potential impact of any attacks. These solutions have spanned from input preprocessing and feature denoising to adversarial training and model architecture enhancements [3]. However, the fundamental question persists: which adversarial defense methods offer the most effective safeguard against ever-evolving adversarial attacks? Now through comprehensive exploration and evaluation of the most prominent adversarial defense methods in the current literature, this paper aims to address the answer to this question. Through analyzing the robustness and applicability of each of these defense systems, we hope to shed light on the most promising ones to give insight for practitioners and upcoming researchers who can now make informed decisions based on our research when selecting the most suitable method for their line of work [4].

2. Background Research

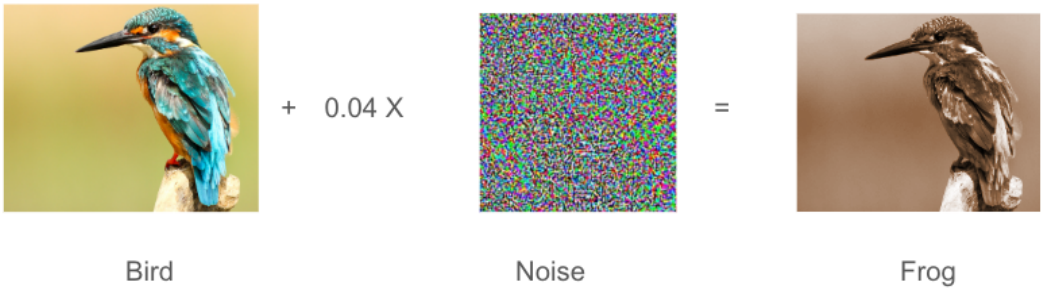

Figure 1: Evasion attack. A model classifies an image as a bird. After adding a small amount of random noise to the image, invisible to the human eye, it is classified as a frog with extremely high confidence |

In the field of Adversarial Machine Learning, there have been numerous comprehensive literature on the topic [5]. However, most only look at the best theoretical method that could be developed to defend against attacks. Moreover, none have compared current strategic methods with one another to determine the most effective one [6]. First, to find this key answer, we must understand how machine learning models work. In traditional machine learning tasks, models are rigorously tested on training data and then tasked with making predictions on previously unseen test data, assigning probabilities to each potential outcome based on their confidence in the prediction. Machine learning then leverages this data as well as sophisticated algorithms to mimic the human learning process, iteratively refining its accuracy. Now, due to the complexity of these models, many vulnerabilities arise [7].

One of these are adversarial attacks. These attacks involve introducing subtle "perturbations" to the training data, changes that often elude human perception but can lead to incorrect classifications and predictions by the model [8]. One such method employed by attackers is the Fast Gradient Sign Method (FGSM), which exploits gradients within the model to manipulate its output [9]. To counter these adversarial threats, we explored three primary defense methods. Adversarial Training, Defensive Distillation, and Gradient Masking. Our first line of defense for testing was Adversarial training. This method involves training the model with a combination of regular and adversarial examples. This comprehensive training approach helps the model generalize its understanding of data, rendering it more resilient to adversarial manipulations [10].

The second method which was tested was defensive distillation. While maintaining the same training processes, this method "distills down" the network's output probabilities, producing softer, less confident predictions. By introducing an element of ambiguity and uncertainty into the training process, the model becomes more vigilant towards subtle changes or perturbations [11].

Finally, we explored Gradient Masking. This defensive method adds an extra layer of protection by injecting "gradient noise" into the data, perturbing the gradients used by the model. This approach introduces complexity to the data, making it arduous for potential attackers to pinpoint the specific areas of training data that need manipulation. As a result, the model's vulnerability to adversarial attacks is reduced. The approach hinges on the notion of enhancing data generalization and injecting unpredictability into the model's learning process [12].

In our pursuit of determining the most optimal methods for detecting and defending against adversarial attacks, we delve deeper into the nuances of these defense strategies. By incorporating them into our project, we aim to bolster the security and dependability of machine learning models in the face of the ever-evolving threat landscape.

3. Methodology

In this research, we began by initializing the MNIST dataset, a collection of 70,000 28x28 color images distributed across ten distinct classes, each containing 6,000 images. This dataset serves as the foundation for our experimental evaluations.

Next, a train-test split was performed on the MNIST dataset to maintain the integrity of our experiments. This division allocates one subset for model training and the other for model evaluation, preventing data leakage and ensuring an unbiased assessment of the models' generalization capabilities. To assess the model's robustness, we employed a custom architecture that included 3 convolutional layers, each with its own Batch Normalization and MaxPooling layer to prevent overfitting. The architecture includes a pooling layer for feature map downsampling, a dense layer with 256 nodes for capturing high-level features, and a final dense layer with 10 nodes representing the 10 MNIST classes.

We then defined a function for Fast Gradient Sign Method (FGSM) attacks, a white-box attack that perturbs the input data to maximize prediction errors. Using the function, tf.GradientTape(), tape.watch, and categorical cross entropy, the attack calculated the gradient direction of maximum loss. Upon knowing the direction which would create maximum loss, the fgsm_attack function would apply that perturbation to the input data's pixel values and effectively increase the error from the model, because it has never encountered this new perturbed image before. To test each defense method's resilience against adversarial attacks, we created 3 functions to implement the three distinct defense strategies.

The first strategy employed was adversarial training, incorporating adversarial examples generated by the FGSM attack during model training. The second strategy was Defensive Distillation. This function was created by 2 separate models, one student and one teacher model. The teacher model would have a higher "temperature" which meant that it would produce softened probabilities using the softmax cross entropy function. Next, the student model would learn off of the gradients that the teacher model produced to achieve a balance between hard and softened probability output. This should help the model be less confident in its probability outputs to reduce the impact of adversarial perturbations, as well as increase the generalization of the data. The third method was gradient masking, which involved feature extraction, where unused nodes were changed to 0 to obscure the correct direction of perturbation from the attacker. To gauge the effectiveness of these defense mechanisms, we assess the robustness of each model by subjecting them to an FGSM attack.

This was performed by creating adversarial images using the FGSM function and recording the resulting accuracy of each defense method through the categorical cross-entropy function. Next every possible permutation of the defense methods was tested. This was performed by training 4 epochs of the first model, saving it into a .H5 file, and training that model for another 3 epochs on the second method. Once all of these were tested, the best defense method was determined.

Consecutive testing on different model complexities was tested to determine the optimal version of the best performing defense method. This evaluation enables a comparative analysis of the performance of the three defense methods in the presence of adversarial attacks, to determine efficacy in bolstering model security. The aim is to maximize the valuable insights gained into the enhancement of model robustness against adversarial threats, ultimately advancing the development of more secure deep learning models.

4. Results

This study analyzed three defense techniques against FGSM attacks conducted over many epochs. We found that all defense model accuracies were significantly different from each other as well as compared to the Control Model without a defense method. The best-performing model was a purely adversarially trained model, with an accuracy of 88.5% after a FGSM attack. This was significantly higher than the accuracy of the Control Model without the defense of 29.15%.

This approach has proven to increase model robustness by exposing it to specifically crafted examples designed to deceive it. Moreover, the combination of training with regular and adversarial examples allows the model to better generalize and defend against unseen adversarial perturbations, thus accounting for its high accuracy.

Next, Defensive Distillation method was tested and showed an average accuracy of 33.12%. This demonstrates that there is no significant difference found between utilizing Defensive Distillation versus a regular defenseless model. This technique introduces a level of uncertainty and variability in training by softening output probabilities, to mitigate the impact of adversarial perturbations. Despite this approach, there are potential trade-offs associated with determining the optimal level of softening, and gradients lost through the student-teacher model transfer, proving it may be necessary to conduct further fine-tuning and optimization in order to maintain competitive levels of accuracy.

Finally, Gradient Masking method was tested and had an average accuracy of 47.52%, performing better than Defensive Distillation however falling short of Adversarial Training and was significantly less. This effectively added complexity to the data, making it difficult for potential attackers to pinpoint vulnerable areas. However, the accuracy fluctuated across epochs, indicating potential sensitivity to this method. It's likely that forcibly altering already trained gradients played a role in its overall performance. It is also important to note that the Defenseless Control Model as expected, performed the worst over 16 test runs, exhibiting an average accuracy of 29.15%.

Additionally testing revealed an increase in model accuracy after adding additional convolutional layers. Gradient Masking and Adversarial Training combined was the best-performing combination of defense methods with an average accuracy of 85.56. Furthermore, the Adversarial Training Defense Model accuracy increased from 88.5% with 2 convolutional layers to a maximum accuracy of 97.9% with 5 convolutional layers. The model's inability to withstand FGSM attacks highlights the necessity of incorporating sophisticated defense mechanisms to bolster its security.

5. Conclusion

As it is shown, Adversarial Training performed the best comparatively performing much better then Gradient Masking and Defensive Distillation. The control model included convolutional layers and dense layers, as well as pooling and drop out layers. It performed flawlessly with an average accuracy of 99.5% on the MNIST dataset without an attack. However correctly classified only 29% of the Adversarially Perturbed images demonstrating how crucial developing defense methods are. Adversarial training, which includes the addition of perturbed images to the training process, performed exceptionally well with an average of 90.69%. This technique proved to be unrivaled in fortifying model robustness by exposing the model to carefully crafted adversarially attacks. Defensive Distillation, which utilizes softened probability distributions, performed decently with an average of 33.12%. However, it was not statistically significant compared to if the model did not have any defenses. This highlights how challenging it is to balance softening with competitive accuracy simultaneously. Finally was introducing "gradient noise" into data as an effective technique that uses gradient masking to increase complexity and deter attackers. However, its performance can vary across different epochs, suggesting sensitivity and potential challenges associated with manipulating trained gradients. This highlights the importance

of sophisticated defense mechanisms in the face of ever-evolving adversarial threats. Adversarial training has shown its ability to improve model generalization and robustness against unexpected adversarial disturbances. These findings provide valuable guidance for practitioners and researchers to choose defense strategies based on the specific requirements of their applications.

Future Direction

Through more detailed research, and projects such as these researchers and cybersecurity experts can offer promising prospects for fortifying artificial intelligence systems against evolving security threats. During our testing, it was discovered that a model with only 1 extra convolutional layer developed a much higher accuracy. With Adversarial Training, it achieved accuracies of 96% instead of 90%, Defensive Distillation went from 30% to 70%, and Gradient Masking went from 50% to 90%. More detailed research could greatly determine the effects of model complexity on robustness. Future research could include testing other defense methods (Adversarial Logit Pairing, Feature Denoising, Gaussian Data Augmentation) as well as testing for the most optimal combination of defense methods. i.e. Defensive Distillation + Adversarial Training or Gradient Masking + Adversarial Training. In conclusion, this study advances our understanding of machine learning in the context of adversarial defense. It provides a strong basis for developing AI systems that are safer and more resilient in the rapidly evolving environment of cyber threats.

Going forward, it is essential to fine-tune and optimise the defense strategies outlined in this study, as well as explore novel approaches to remain ahead and hope to save many lives.

References

1. Rozenwald, M. B., Galitsyna, A. A., Sapunov, G. V., Khrameeva, E. E., & Gelfand, M. S. (2020). A machine learning framework for the prediction of chromatin folding in Drosophila using epigenetic features. *PeerJ Computer Science, 6*, e307.
2. Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. *In Artificial intelligence safety and security* (pp. 99-112). Chapman and Hall/CRC.
3. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
4. 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.
5. Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765-1773).
6. Addepalli, S., & Jain, S. (2022). Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems, 35*, 1488-1501.
7. Croce, F., & Hein, M. (2020, November). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning* (pp. 2206-2216). PMLR.
8. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185-9193).
9. Sen, J., & Dasgupta, S. (2023). Adversarial attacks on image classification models: FGSM and patch attacks and their impact. *arXiv preprint arXiv:2307.02055*.
10. Huang, J., Wen, M., Wei, M., & Bi, Y. (2024). Enhancing the transferability of adversarial samples with random noise techniques. *Computers & Security, 136*, 103541.
11. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
12. Boenisch, F., Sperl, P., & Böttinger, K. (2021). Gradient masking and the underestimated robustness threats of differential privacy in deep learning. *arXiv preprint arXiv:2105.07985*.