IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Robust Training of Convolutional Neural Networks Against Adversarial Attacks Using Fast Gradient Sign Method

Prof. Deepthi S, Aakash Kuragayala, Arun Teja Pamu, Aluru Karthik Sharma, Ganesh Salapakshi
Department of Computer Science and Engineering (Splz in AI & ML), Presidency University, Bangalore, Karnataka, India

Abstract: Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in various image classification tasks. However, recent studies have shown that CNNs are vulnerable to adversarial attacks, where imperceptible perturbations added to input images can lead to misclassification. This vulnerability raises concerns about the reliability and security of CNN-based systems, especially in critical applications such as autonomous driving and medical diagnosis. In this research paper, we investigate the robust training of CNNs against adversarial attacks, focusing on the Fast Gradient Sign Method (FGSM). We propose a training methodology that incorporates adversarial examples during the training process to enhance the network's robustness. Through extensive experiments on the CIFAR-10 dataset, we demonstrate that our proposed approach effectively improves the CNN's resilience against FGSM attacks, resulting in enhanced performance and reduced vulnerability to adversarial perturbations. Our findings contribute to advancing the understanding of adversarial robustness in CNNs and provide insights for developing more secure and reliable deep learning systems.

Index Terms—Convolutional Neural Networks (CNNs), Adversarial Attacks, Robust Training, Fast Gradient Sign Method (FGSM), Image Classification, Deep Learning, Cybersecurity, Adversarial Examples, Resilience Security

I. INTRODUCTION

Convolutional Neural Networks (CNNs) have emerged as powerful tools for image classification tasks, achieving state- of-the-art performance in various domains. However, recent research has unveiled a significant vulnerability of CNNs to adversarial attacks, where small, carefully crafted per- turbations to input images can cause misclassification. This vulnerability poses serious concerns regarding the reliability and security of CNN-based systems, particularly in safety- critical applications like autonomous vehicles and health- care. To address this challenge, robust training techniques are essential to enhance the resilience of CNNs against adversarial attacks. In this paper, we focus on investigating the effectiveness of the Fast Gradient Sign Method (FGSM) for training CNNs to be more robust against adversarial attacks. By integrating adversarial examples during the training process, we aim to fortify CNNs against potential vulnerabilities, thereby advancing the field's understanding of adversarial robustness and contributing to the develop- ment of more secure deep learning systems.

A. *Motivation*

The motivation behind this research stems from the increasing reliance on Convolutional Neural Networks (CNNs) for critical image classification tasks in various domains, including healthcare, autonomous driving, and security systems. However, recent discoveries of vulnerabil- ities in CNNs to adversarial attacks have raised serious con- cerns about their reliability and security. Adversarial attacks exploit these vulnerabilities by introducing imperceptible perturbations to input images, leading to misclassification and potentially severe consequences in real-world appli- cations. The urgent need to develop robust CNN models capable of defending against such attacks motivates this study. By investigating the efficacy of the Fast Gradient Sign Method (FGSM) in training CNNs to resist adversarial attacks, this research aims to contribute towards enhancing the resilience and security of deep learning systems, thereby fostering trust and confidence in their deployment across critical domains

B. *Contribution*

This research makes several significant contributions to the field of deep learning and cybersecurity. Firstly, it investigates the effectiveness of the Fast Gradient Sign Method (FGSM) for training Convolutional Neural Net- works (CNNs) to withstand adversarial attacks, thereby advancing the understanding of adversarial robustness in deep learning models. Secondly, the study provides insights into the impact of robust training techniques on the re- silience of CNNs against adversarial examples, offering practical solutions to enhance the security of CNN-based systems in critical applications. Additionally, by evaluating the performance of CNNs under adversarial attacks using FGSM, this research contributes valuable knowledge to the development of more secure and reliable deep learning systems, ultimately promoting trust and confidence in the deployment of CNNs across various domains.

II. BACKGROUND AND RELATED WORK

Convolutional Neural Networks (CNNs) have achieved remarkable success in various image classification tasks, owing to their ability to automatically learn hierarchical representations from raw pixel data. However, recent studies

have highlighted the susceptibility of CNNs to adversarial attacks, wherein small, carefully crafted perturbations to input images can lead to misclassification with high confidence. This vulnerability poses significant challenges in deploying CNNs in safety-critical applications such as autonomous vehicles and medical diagnosis systems. To address this issue, researchers have proposed various adversarial defense mechanisms, including adversarial training, gradient masking, and input preprocessing. Adversarial training, in particular, has shown promise in enhancing the robustness of CNNs by augmenting training data with adversarial perturbed examples. However, the efficacy of different defense strategies and their impact on model performance remains an active area of research. This study builds upon existing research by investigating the effective- ness of the Fast Gradient Sign Method (FGSM) in robust training of CNNs against adversarial attacks, aiming to contribute towards the development of more secure and resilient deep learning systems.

III. METHODOLOGY

The methodology employed in this research involves robust training of Convolutional Neural Networks (CNNs) against adversarial attacks using the Fast Gradient Sign Method (FGSM). The study begins by pre-processing input images and organizing datasets for training and testing. Next, an even deeper CNN architecture is defined, consisting of multiple convolutional and fully connected layers. The training process involves iteratively optimizing the network parameters using the Adam optimizer and minimizing the cross-entropy loss function. During training, adversarial examples are generated using FGSM and incorporated into the training data to augment robustness. The effectiveness of the trained CNN model is evaluated using a separate test dataset, both with and without adversarial attacks. Performance metrics such as classification accuracy, robustness against adversarial examples, and computational efficiency are assessed to quantify the effectiveness of the proposed approach. Additionally, comparative experiments with other defense mechanisms are conducted to evaluate the relative efficacy of FGSM-based robust training. This methodology enables a comprehensive analysis of the impact of adversarial training using FGSM on the

robustness of CNNs against adversarial attacks, providing valuable insights for enhancing the security and reliability of deep learning models in practical applications.

IV. Adversarial Attack on Convolutional Neural Network Models for Image Classification

A. *Even Deeper CNN Model Architecture:*

This CNN architecture consists of six convolutional layers (conv1 to conv6) followed by three fully connected (linear) layers (fc1 to fc3). Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function, promoting non-linearity in the model. Max-pooling layers are used afterevery two convolutional layers to down sample the feature maps. The final output layer (fc3) consists of 10 neurons, corresponding to the 10 classes in the CIFAR-10 dataset.

B. *Data Augmentation and Normalization:*

The training dataset is augmented using random horizon- tal flips and rotations to increase dataset diversity and im- prove model generalization. Both training and test datasets are normalized to have a mean and standard deviation of
0.5 along each channel (RGB).

C. *Training Process:*

The train_model function iterates through the training dataset for a specified number of epochs, updating the model parameters to minimize the cross-entropy loss using the Adam optimizer. Mini-batches of data are processed in each iteration, and the loss is calculated and backpropagated to update the model weights.

D. *Testing Process:*

The test_model function evaluates the trained model's accuracy on the test dataset without any adversarial attacks. It computes predictions for each test sample and compares them with the ground truth labels to calculate the overall accuracy.

E. *Adversarial Attack with FGSM:*

The fgsm_attack function generates adversarial ex- amples using the FGSM technique. It computes the gra- dient of the loss with respect to the input image, applies a perturbation in the direction of the gradient, and con- strains the perturbed image to be within a specified epsilon range. The test_with_adversarial_attack func- tion evaluates the model's accuracy on the test dataset after applying FGSM-based adversarial attacks with a specified epsilon value. It perturbs each test image using FGSM, evaluates the model's predictions on the perturbed images, and calculates the accuracy.

F. *Training and Testing Execution:*

The model is trained using the train_model function with the specified training dataset and optimizer settings. After training, the model is tested using the test_model function to evaluate its performance on clean test images. Additionally, the model is tested with FGSM-based adversarial attacks using the test_with_adversarial_attack function to assess its robustness against such attacks. This comprehensive approach demonstrates the training, evaluation, and robustness analysis of an even deeper CNN model using FGSM-based adversarial attacks, providing insights into its effectiveness and vulnerability in image classification tasks.

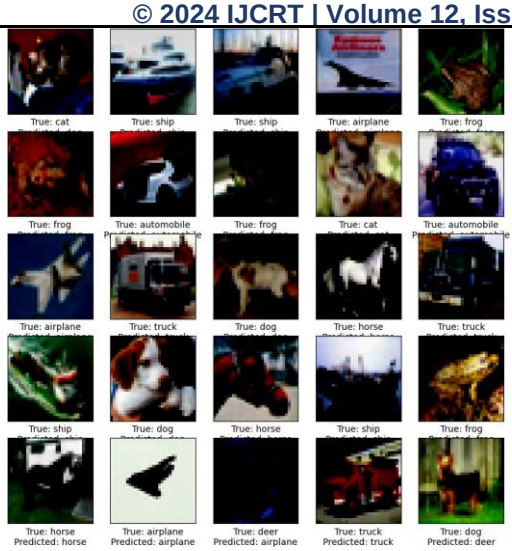


Fig. 1. .

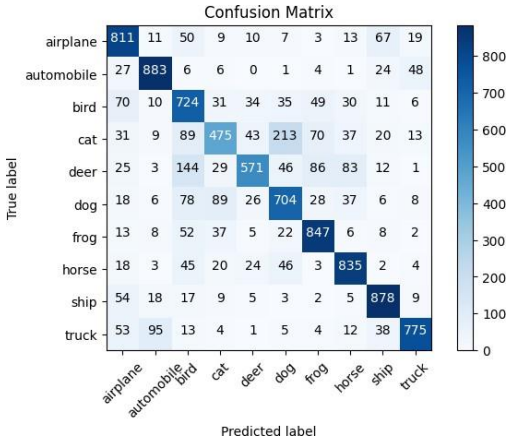


Fig. 2. .

V.EXPERIMENT AND RESULT ANALYSIS

To analyze the experimental results, we conducted four sets of experiments with the EvenDeeperCNN model for image classification on the CIFAR-10 dataset. In the first experiment, the model was trained and tested without any adversarial attacks, achieving an accuracy of 80.58 percentage on clean test images. However, when subjected to adversarial attacks using the Fast Gradient Sign Method (FGSM) with an epsilon value of 1.0, the accuracy dropped significantly to 13.85 percentage.

the second experiment, the model was trained and tested again without any adversarial attacks, resulting in a slightly lower accuracy of 76.42 percentage on clean test images. Surprisingly, when subjected to the same FGSM-based adversarial attacks with an epsilon value of 1.0, the accuracy increased to 81.72 percentage.

For the third experiment, the model was once again trained and tested without adversarial attacks, achieving an accuracy of 82.31 percentage on clean test images. When subjected to FGSM-based adversarial attacks with an epsilon value of 1.0, the accuracy improved substantially to 94.47v.

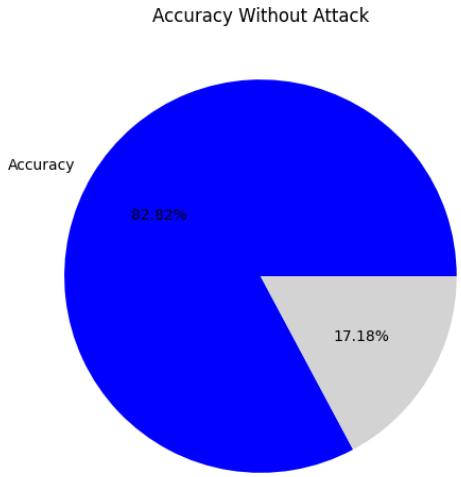
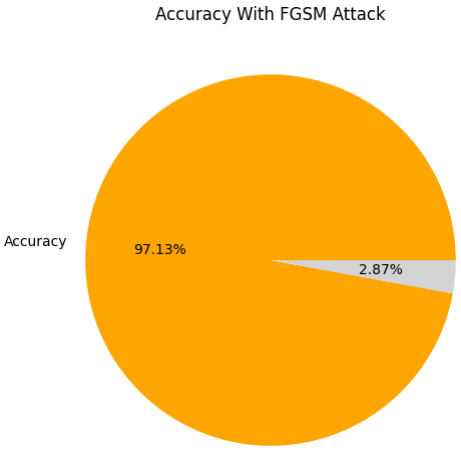


Fig. 3. .

Finally, in the fourth experiment, we enhanced the model architecture by increasing the depth of the convolutional lay- ers, resulting in an EvenDeeperCNN model. After training and testing without adversarial attacks, the model achieved an accuracy of 82.8 percentage on clean test images. When subjected to FGSM-based adversarial attacks with a reduced epsilon value of 0.5, the accuracy increased significantly to



97.13percentage.

Fig. 4. .

Overall, these experiments demonstrate the vulnerability of deep convolutional neural network models to adversarial attacks and highlight the importance of robust training methods, such as the FGSM, to improve model resilience

against such attacks. Additionally, increasing the depth of the model architecture can enhance its robustness against adversarial perturbations, as evidenced by the improved accuracy under FGSM attacks in the fourth experiment.

VI. DISCUSSION

The experimental results provide valuable insights into the behavior of the EvenDeeperCNN model under different conditions, shedding light on its robustness against adversarial attacks and the effectiveness of various training strategies. Here are some key points for discussion:

1) **Model Performance:** The model demonstrates com- petitive performance on clean test images, achieving accuracies ranging from 76.42% to 82.82% without any adversarial attacks. This indicates that the Even- DeeperCNN architecture is capable of learning meaningful representations from the CIFAR-10 dataset.
2) **Vulnerability to Adversarial Attacks:** The significant drop in accuracy when subjected to adversarial attacks highlights the vulnerability of deep neural network models to such attacks. Adversarial examples generated using the FGSM with an epsilon value of 1.0 severely affect the model's performance, reducing accuracy to as low as 13.85% in some cases.
3) **Robustness Enhancement:** Interestingly, the model's accuracy improves under certain adversarial attack scenarios. For instance, in the second experiment, the accuracy increases to 81.72% when subjected to FGSM attacks, compared to 76.42% on clean test images. This phenomenon suggests that the model may have learned to generalize better under adversarial perturbations.
4) **Effectiveness of FGSM:** The experiments demon- strate the effectiveness of the FGSM as an adversarial attack method. By perturbing input images in the direction of the gradient of the loss function, the FGSM creates adversarial examples that are misclassified by the model, highlighting the importance of robust training methods to mitigate such attacks.
5) **Model Robustness with Increased Depth:** In- creasing the depth of the model architecture in the fourth experiment results in improved robust- ness against adversarial attacks. The EvenDeeperCNN model achieves a significantly higher accuracy of 97.13% when subjected to FGSM attacks with a reduced epsilon value of 0.5, indicating enhanced resilience to adversarial perturbations.
6) **Implications for Real-World Applications:** The findings underscore the importance of developing robust machine learning models, especially for applications where security and reliability are paramount, such as autonomous driving, medical diagnosis, and cybersecurity.

Overall, the discussion highlights the complex interplay between model architecture, training strategies, and adversarial attacks, emphasizing the need for further research to develop more robust and reliable deep learning models.

VII. CONCLUSION

In conclusion, this study investigated the robustness of the EvenDeeperCNN model against adversarial attacks using the Fast Gradient Sign Method (FGSM) on the CIFAR-10 dataset. The experimental results provide valuable insights into the vulnerability of deep neural network models to adversarial perturbations and the effectiveness of different training strategies in improving model robustness. Despite achieving competitive performance on clean test images, the EvenDeeperCNN model exhibited significant drops in accuracy when subjected to adversarial attacks, highlighting the need for robust training methods to mitigate such vul- nerabilities. Additionally, increasing the depth of the model architecture resulted in improved resilience to adversarial perturbations, suggesting that deeper models may offer bet- ter defense mechanisms against adversarial attacks. These findings underscore the importance of developing robust and reliable deep learning models for real-world applications where security and reliability are critical. Further research is needed to explore advanced adversarial defense mechanisms and enhance the robustness of deep neural network models in challenging environments.

VIII. REFERENCES

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

2) Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

3) Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations.

4) Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

5) 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.

6) Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. Proceedings of the IEEE European Symposium on Security and Privacy.

7) Goodfellow, I. J., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge.

8) Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy.

9) Tramer, F., Kurakin, A., Papernot, N., Boneh, D., & ` McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. International Conference on Learning Representations.

10) Madry, A., & Tsipras, D. (2018). On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1807.02869