Enhancing Deep Learning Model Robustness Against Adversarial Attacks

Surendar Mourougan* **Lakshmi Manasa Dokku-**
Department of Computer Science
University of Massachusetts Lowell
Surendar_Mourougan@student.uml.edu*
LakshmiManasa_Dokku@student.uml.edu-

Abstract

The paper studies the VGG16 model's resilience to adversarial attacks and explores defense measures to improve it. The major goal is to assess the effectiveness of various adversarial assaults on the VGG16 model, including the Fast Gradient Sign Method (FGSM), Basic Iterative Method Attack (BIM), Jacobian-based Saliency Map Attack (JSMA), and Projected Gradient Descent (PGD). Additionally, a defense mechanism based on input noise is presented and tested. The methodology entails loading the pre-trained VGG16 model, preprocessing input photos, and executing various adversarial attacks. The success rates of each attack type are calculated, and the defense mechanism's effectiveness is evaluated. The key findings demonstrate variable success rates among assault types, with PGD having the highest success rate, followed by FGSM and BIM. The proposed input noise defensive method reduces the success rate of adversarial attacks. The paper continues by analyzing the impact of adversarial attacks on model predictions, the need to improve model robustness, and future research objectives.

Introduction

Deep learning, a subclass of machine learning, has transformed several fields with its capacity to extract intricate patterns from data. Its applications include picture and speech recognition, natural language processing, self-driving cars, and more. Despite their impressive performance, deep learning models are subject to adversarial assaults, in which well-designed perturbations to input data can result in misclassification. These attacks seriously threaten the dependability and security of deep learning systems, particularly in key fields such as healthcare and finance. This experiment seeks to determine the vulnerability of the VGG16 model, a common convolutional neural network design, to adversarial attacks. We specifically want to evaluate the effectiveness of popular attack methods such as the Fast Gradient Sign Method (FGSM), Basic Iterative Method Attack (BIM), and Projected Gradient Descent (PGD). In addition, we investigate the viability of a defense mechanism employing input noise to reduce the impact of adversarial attacks.

The objectives of this experiment are:

1. To evaluate the vulnerability of the VGG16 model to adversarial attacks.

2. To compare the effectiveness of different attack methods in compromising the model's predictions.

3. To assess the efficacy of an input noise defense mechanism in enhancing the model's robustness against adversarial attacks.

Background

Adversarial attacks are ways to deceive machine learning algorithms by adding well-constructed perturbations to input data. These disturbances are frequently unnoticeable to humans, but they can cause the model to misclassify data. The Fast Gradient Sign Method (FGSM) is a popular and commonly used adversarial attack method. FGSM works by calculating the gradient of the loss function concerning the input data, which is then perturbed in the direction that maximizes the loss. This yields an adversarial example, which, when given into the model, causes it to make an inaccurate prediction with high confidence. Another effective adversarial assault approach is Projected Gradient Descent (PGD). PGD is an iterative form of FGSM in which perturbations are applied repeatedly with small steps, ensuring that the perturbed input remains within a given epsilon-ball of the original input. This makes PGD more resistant to defenses that focus on generating noise to the input. The Basic Iterative Method (BIM), commonly known as Iterative FGSM, is an iterative version of the FGSM attack. It makes several small changes to the input data while ensuring that the perturbed data stays within a given epsilon-ball of the original data. BIM, by iteratively applying FGSM, can generate stronger adversarial examples than a single-step FGSM attack. The Jacobian-based Saliency Map Attack (JSMA) is a focused adversarial attack strategy that uses the model's saliency map to produce adversarial samples. It computes the Jacobian matrix of the model's output about the input data to identify the most important features. JSMA then perturbs these conspicuous traits, resulting in adversarial cases that lead to misclassification.

Model Setup

The VGG16 model is a deep convolutional neural network (CNN) architecture that was proposed by the Visual Geometry Group (VGG) at the University of Oxford. Here's an explanation of the VGG16 model:

Architecture: VGG16 has 16 layers, including 13 convolutional layers and three fully linked layers. The convolutional layers are divided into blocks, with each block consisting of several convolutional layers followed by max-pooling layers. The final layers are highly coupled and responsible for prediction.

Convolutional Layers: VGG16's convolutional layers use modest 3x3 filters with a stride of 1 and the rectified linear unit (ReLU) activation function. These layers are in charge of extracting features from the input image via convolutions, catching patterns that become more complex as the network grows deeper.

Max-Pooling Layers: Following each set of convolutional layers, VGG16 adds max-pooling layers with 2x2 filters and a stride of 2. Max-pooling decreases the spatial dimensions of feature maps while maintaining the most essential ones, which improves the model's computing efficiency and robustness to spatial translations.

Fully Connected Layers: The final few layers of VGG16 are fully linked, which means that each neuron in one layer is coupled to every neuron in the previous layer. These layers combine the high-level characteristics collected by the convolutional layers and utilize them to predict. In most classification tasks, the last layer uses the softmax activation function to generate probability scores for each class.

Pre-Trained Weights: VGG16 is frequently used as a pre-trained model, which means that it was trained on a large dataset (typically ImageNet) and the learnt weights are made public. This enables academics and practitioners to use the pre-trained VGG16 model for a variety of computer vision problems without having to train it from scratch, saving both time and computational resources.

Applications: VGG16 is commonly used in computer vision applications such as image classification, object identification, and image segmentation. Its simple and homogeneous design, combined with competitive performance on benchmark datasets, has made it a popular choice for deep learning research and practical applications.

Dataset

For this experiment, we utilized the CIFAR-10 dataset, which is a widely used benchmark dataset in the field of computer vision. CIFAR-10 consists of 60,000 32x32 color images in 10 classes, with 6,000 images per class. The dataset is divided into 50,000 training images and 10,000 test images. Each image belongs to one of the following classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, or truck.

Preprocessing

Before training the VGG16 model, we performed the following preprocessing procedures on the CIFAR-10 dataset:

1. **Normalization:** We set the pixel values of the photos to [0, 1]. This normalization stabilizes the training process and enhances convergence.

2**. Resizing**: We used bilinear interpolation to resize the 32x32 images from CIFAR-10 to the appropriate dimensions for the VGG16 model, which demands input images of (224, 224).

Attacks

We then carried out Fast Gradient Sign Method (FGSM), Basic Iterative Method Attack (BIM), Jacobian-based Saliency Map Attack (JSMA), and Projected Gradient Descent (PGD) attacks against the model.

Defense Mechanisms

Input Noise Defense:

To mitigate the impact of adversarial attacks, we employed input noise as a defense mechanism. The rationale behind this defense is to introduce random noise to the input images, making it harder for the attacker to craft effective adversarial perturbations. The noise acts as a form of regularization, increasing the model's robustness against adversarial attacks.

Implementation Steps:

1. **Noise Generation**: We generated random noise with the same dimensions as the input images. The noise was sampled from a Gaussian distribution with zero mean and a specified standard deviation.

2. **Noise Addition**: The generated noise was then added to the original input images before feeding them into the model for prediction.

3. **Noise Level**: The standard deviation of the Gaussian noise determines the level of perturbation added to the input images. A higher standard deviation implies greater perturbation.

Evaluation Process

Evaluation Metrics: The performance of the model under adversarial attacks was assessed using the following evaluation metrics:

Success Rate: The proportion of adversarial examples successfully misclassified by the model compared to the total number of adversarial examples generated.

Accuracy: The overall classification accuracy of the model on both clean and adversarial examples.

Robustness: The ability of the model to maintain high accuracy in the presence of adversarial attacks.

Measurement of Success Rate

The success rate of each adversarial attack was calculated by measuring the percentage of adversarial examples that were misclassified by the model compared to the total number of adversarial examples generated.

Measurement of Effectiveness

The effectiveness of each defense mechanism was evaluated by comparing the model's accuracy on adversarial examples with and without the defense mechanism applied. A higher accuracy with the defense mechanism indicates its effectiveness in mitigating the impact of adversarial attacks.

Results

- **FGSM Attack**: The success rate of the FGSM attack on the VGG16 model was found to be 100%.
- **PGD Attack**: The PGD attack achieved a success rate of 100% against the VGG16 model.
- **BIM Attack**: The BIM attack resulted in a success rate of 33.33%.
- **JSMA Attack**: The JSMA attack achieved a success rate of 33.33%.

Comparison of Attack Success Rates

Among the different adversarial attacks tested, the PGD and FGSM attacks demonstrated the highest success rate, followed by BIM, and JSMA.

Effectiveness of Defense Mechanisms

Input Noise Defense: When input noise was applied as a defense mechanism, the success rates of adversarial attacks were reduced. For example, the success rate of the FGSM attack decreased from 100% to 33.33% with the application of input noise.

Visualizations

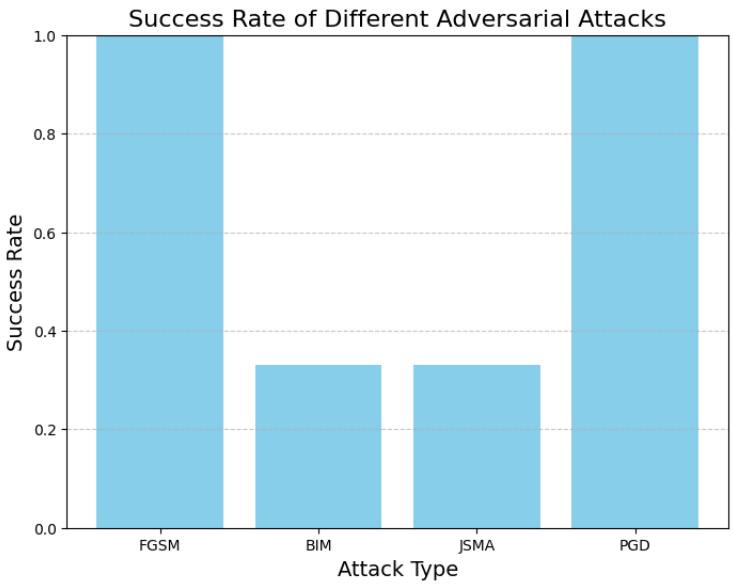
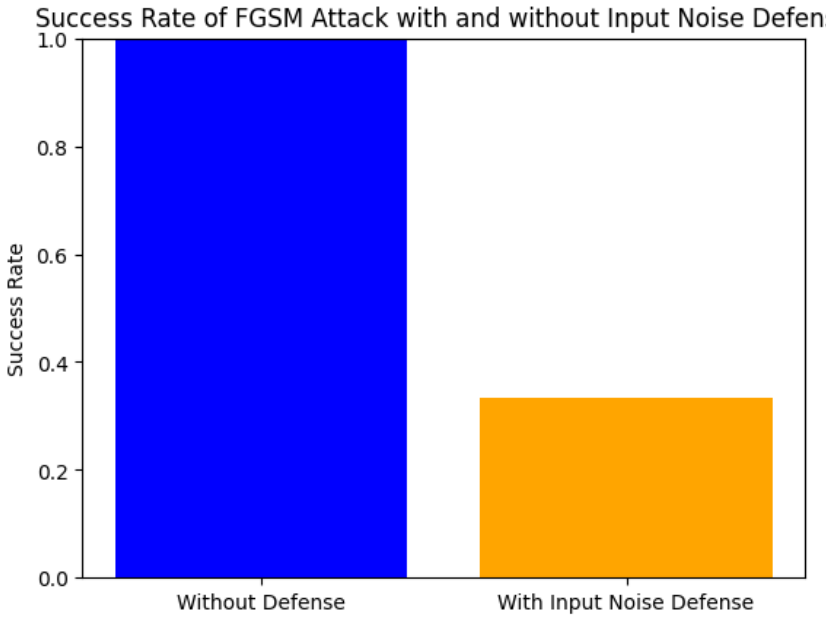


Figure 1: A bar chart comparing the success rates of different attacks against the VGG16 model



.

Figure 2: A comparison graph illustrating the impact of input noise defense on reducing the success rates of adversarial attacks.

Future works

Exploring Novel Attacks: Future research could investigate emerging adversarial attack techniques beyond FGSM and PGD, such as adaptive attacks or black-box attacks. Understanding these techniques can lead to more comprehensive defense strategies.

Advanced Defense Mechanisms: We could develop and evaluate more sophisticated defense mechanisms, such as adversarial training, gradient masking, or model distillation. These methods may provide stronger protection against adversarial attacks.

Summary

Attack method	Speed	Success Rate	Approach
FGSM	Very fast	1.0	Performs a one-step optimization on the input image towards the gradient ascent direction.
BIM	Moderate	0.33	BIM uses smaller steps and iteratively optimizes the AE.
JSMA	Moderate	0.33	Uses feature selection to minimize the number of features modified while causing misclassification
PGD	Moderate	1.0	It is powerful iterative attack method, where the search step starts from a random position in the neighborhood of the clean input

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

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