2024

Exploring the Robustness of LoRA-Adapted ResNet18 against Adversarial Attacks

TSAI CHEN HSIUNG, YUNG-HO LEE

Abstract

Recent advances in deep neural networks (DNNs) have significantly enhanced their performance across a variety of tasks. However, these models remain susceptible to adversarial attacks, where subtly modified inputs can lead to incorrect outputs. This study explores the efficacy of Low-Rank Adaptation (LoRA) as a technique to enhance the robustness of ResNet18, a widely used DNN architecture. By adapting the fully-connected and convolutional layers of ResNet18 with LoRA and subjecting the model to Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks, we aim to assess whether LoRA can mitigate the vulnerabilities of DNNs to adversarial examples. Our preliminary results indicate that LoRA adaptations not only preserve the intrinsic performance of ResNet18 but also enhance its resistance to adversarial perturbations, providing a promising avenue for developing more secure DNN architectures.

Introduction

Deep Learning has achieved remarkable success in fields ranging from autonomous driving to medical diagnostics. Yet, the robustness of deep neural networks (DNNs) under adversarial conditions remains a critical challenge, undermining their reliability in safety-critical applications. Adversarial attacks, where input data is perturbed in a way that is often imperceptible to humans, can lead to catastrophic misclassifications, revealing a fundamental vulnerability in current learning algorithms.

In this context, enhancing the adversarial robustness of DNNs without compromising their performance is of paramount importance. Traditional approaches often involve retraining networks with adversarially generated data, a process that is computationally expensive and not always feasible. The Low-Rank Adaptation (LoRA) proposes an innovative approach by injecting trainable low-rank structures into pre-trained networks, thus enabling efficient fine-tuning while maintaining the model's capacity.

This research focuses on the application of LoRA to ResNet18, a model celebrated for its efficiency and accuracy across various tasks. By integrating LoRA into ResNet18's architecture, specifically targeting its fully-connected and convolutional layers, we investigate whether this

adaptation enhances its resilience to FGSM and PGD attacks—common benchmarks for evaluating adversarial robustness.

Related Work

Adversarial Attacks in Deep Learning: Adversarial attacks have emerged as a significant threat to the reliability of machine learning models. Szegedy et al. (2013)[1] first demonstrated that neural networks are vulnerable to imperceptible perturbations of their inputs, prompting a myriad of studies into adversarial examples. Techniques like FGSM (Goodfellow et al., 2014[2]) and PGD (Madry et al., 2017[3]) have been developed to generate these examples, challenging the robustness of deep learning models under adversarial settings.

Resilience Through Network Architecture: Modifications to network architecture can inherently increase robustness. The introduction of architectures like ResNet (He et al., 2016[4]) marked significant progress in performance on visual recognition tasks. However, their resilience to adversarial attacks has not been extensively studied, which is crucial for their application in security-sensitive areas.

Enhancing Robustness via Low-Rank Adaptations: Recent advances have explored the potential of low-rank matrix approximations to enhance model training and inference efficiency (Xu et al., 2020[5]). LoRA specifically extends this concept by adapting large pre-trained models with minimal additional parameters, thus offering a promising solution for improving robustness without extensive retraining.

Parameter-Efficient Tuning of Large Convolutional Models: In the recent research paper titled "Large Convolutional Model Tuning via Filter Subspace," the authors propose an innovative approach to fine-tuning convolutional models, particularly effective for large pre-trained models like ResNet50 and ConvNeXt. The core method focuses on adapting only filter subspace elements, known as "filter atoms," to enhance parameter efficiency significantly. This technique involves a few hundred parameters and leverages recursive decomposition of each filter atom over another set of filter atoms to expand the tunable parameter space, thus facilitating more refined adjustments to the model.

Methodology

ResNet18

ResNet18, a residual network architecture consisting of 18 layers, is a subset of the ResNet family that introduced residual learning to alleviate the vanishing gradient problem in deep networks. It comprises 8 residual blocks, each block containing two convolutional layers with batch normalization and ReLU activation, except for the first convolutional layer which stands alone.

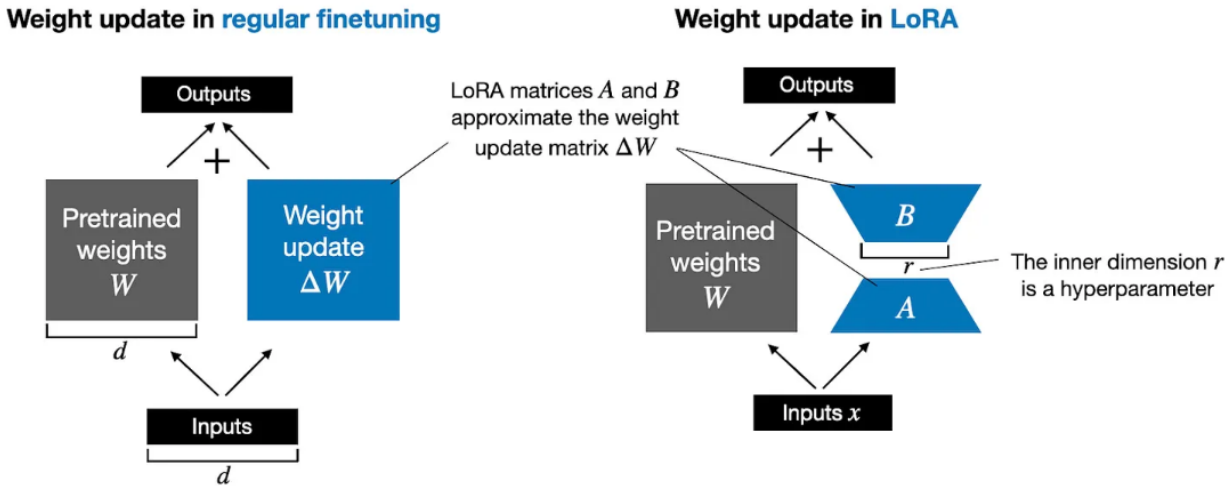

Figure 1. Visualization of LoRA mechanism

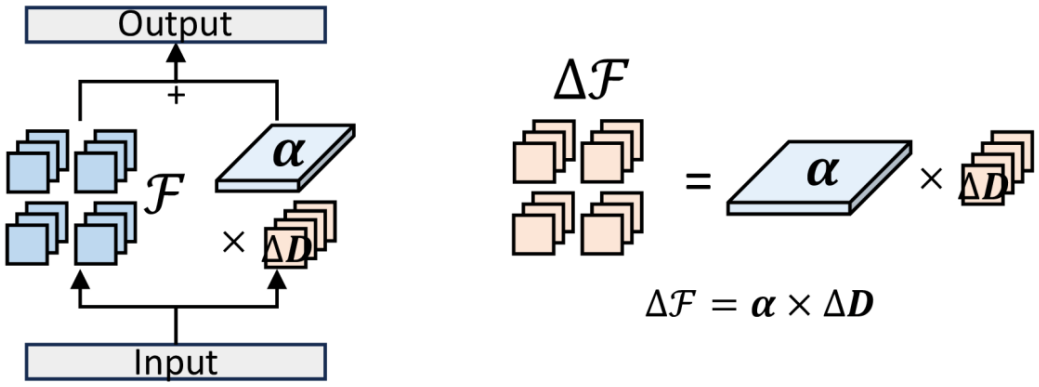

Figure 2. Visualization of LoRA adaptation on Convolutional Neural Network

LoRA Adaptation

Conceptual Overview

LoRA targets specific layers within a pre-trained CNN by introducing two low-rank matrices, A and B, that facilitate an efficient and focused modification of the layer weights. This process retains the original pre-trained weights (denoted as W) and injects an additional update (ΔW) derived from the product of matrices A and B.

Mathematical Formulation

The adaptation involves calculating the low-rank update as:

$$\Delta W = \alpha \times (A \times B)$$

Where:

- $A \in R^{Cout \times r}$ and $B \in R^{r \times (Cin \times K^h \times Kw)}$
- α is a scalar that controls the magnitude of the update.
- C_{out}, C_{in}, K_h, and K_w represent the number of output channels, input channels, kernel height, and kernel width, respectively.
- r is the rank, which is significantly smaller than the dimensions of the full weight matrix, thereby ensuring the efficiency of the adaptation.

Implementation Details

1. **Fully-Connected Layers:**
   For layers like the final classification layer in ResNet18, LoRA introduces matrices A and B that directly adjust the layer's weights. The modified weights are computed as $W' = W + \Delta W$, where ΔW is reshaped to match the dimensions of W.
2. **Convolutional Layers:**

In the first convolutional layer of ResNet18, the Low-Rank Adaptation (LoRA) introduces matrices A and B to modify the layer weights. These matrices are carefully reshaped and optimized to align with the dimensionality of the convolutional weights, leveraging the inherent structure of convolution operations. This adaptation enables the modified weights to integrate smoothly into the network, enhancing the model's capacity while retaining the pre-existing learned features. The resulting convolutional operations are thus influenced by both the original weights and the new low-rank adaptations, allowing for improved performance and resilience[6].

The application of LoRA in convolutional layers optimally balances the network's need for specificity and generalization, facilitating focused improvements without disrupting the broader learned patterns.

Adversarial Attacks Implementation

Fast Gradient Sign Method (FGSM)

FGSM is a one-step attack method that utilizes the gradients of the loss with respect to the input data to create new data points (i.e., adversarial examples) with perturbations intended to maximize the error rate under a model. One of the most famous examples of an adversarial image shown below is taken from the aforementioned paper.

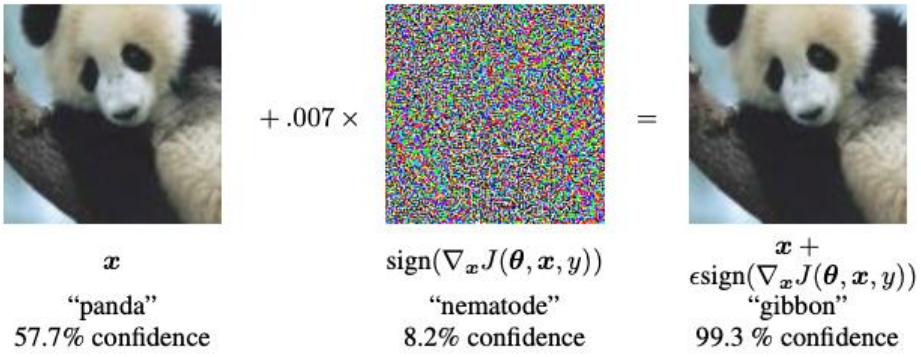

Figure 2. Experiment of FGSM attack on panda image classification

The perturbations are added in the direction of the sign of these gradients:

$X_{adv}=X+\epsilon\cdot\text{sign}(\nabla_X J(\theta,X,y))$

Where:

- X is the input image.
- ϵ is the perturbation magnitude.
- $\nabla_X J$ is the gradient of the loss with respect to the input.

Projected Gradient Descent (PGD)

Projected Gradient Descent (PGD) is a type of adversarial attack that iteratively modifies input data within a defined perturbation range to maximize the prediction error of a neural network. It

is considered one of the most powerful and effective methods for evaluating the robustness of models against adversarial examples.

Mathematical Formulation: Given an input xx, a target model ff with parameters $\theta\theta$, and a loss function LL, the PGD attack iteratively updates the adversarial example $xadvxadv$ using the formula:

$$x_{t+1}^{adv} = \prod_{x+S}(x_t^{adv} + \alpha \cdot sign(\nabla_x L(\theta, x_t^{adv}, y)))$$

where:

- x_t^{adv} is the adversarial example at iteration t,
- α is the step size,
- $\nabla_x L$ is the gradient of the loss with respect to the input,
- y is the true label for x,
- S is the allowed perturbation set, typically defined as $\{\delta: \|\delta\|_\infty \leq \epsilon\}$,
- $\prod$ denotes the projection operation that keeps the perturbations within the ϵ-ball around the original image x.

Experimental Setup

Dataset Preparation

Dataset Description

For this study, we utilized the ImageNet-tiny[7] dataset, a subset of the larger ImageNet database known for its diversity and complexity. This dataset offers a robust testing ground for evaluating model performance under normal and adversarial conditions due to its wide variety of image classes.

Preprocessing

All images in the ImageNet-tiny dataset were preprocessed to conform to the input requirements of the ResNet18 architecture. This preprocessing involved resizing the images to the appropriate dimensions, followed by standard normalization to scale pixel values. Additionally, data augmentation techniques such as random cropping and horizontal flipping were employed to

enhance model generalizability and prevent overfitting.

Training Strategy

Training was carried out using stochastic gradient descent with a momentum of 0.9 and a learning rate of 0.01, adjusted during training with a learning rate decay factor as per standard practice. The models were trained for a total of 50 epochs, with early stopping implemented to prevent overtraining.

Hardware and Software Configuration

All experiments are conducted on NVIDIA Tesla V100 GPUs, with the models implemented and tested using PyTorch 1.8.0 and CUDA 11.1 to ensure reproducibility. The adversarial attack implementations are verified using the Foolbox 3.3.0 library.

Evaluation Metrics

Accuracy

Model performance was primarily evaluated using accuracy, defined as the ratio of correctly predicted observations to the total observations:

$$Accuracy = \frac{True\ Positive + True\ Negative}{All\ data\ points}$$

Model Certainty

We measured model certainty through the average probability of class predictions, providing insights into the confidence of the model's predictions across different classes.

Attack Success Rate

Adversarial robustness was assessed using the attack success rate metric, specifically designed to measure the proportion of previously correct predictions that were misled by an attack:

$$Attack\ Success\ Rate = \frac{Number\ of\ success\ attack\ for\ correct\ prediction}{Success\ Prediction\ before\ attack}$$

Results

This section presents the outcomes of our experiments with the ResNet18 architecture adapted using Low-Rank Adaptation (LoRA), evaluated on the ImageNet-tiny dataset under both standard and adversarial conditions.

Model Performance

The application of LoRA significantly impacted the validation accuracy of the models. The baseline ResNet18 model, without any adaptations, achieved an accuracy of 90.00% on the validation set. When LoRA was applied to the fully connected layers (LoRA-FC), there was a notable increase in accuracy to 96.00%. Further adaptation of both the fully connected and the first convolutional layer (LoRA-FC+Conv1) resulted in the highest accuracy of 98.00%, indicating a substantial enhancement in the model's ability to generalize to new data.

	Baseline Resnet-18	LoRA-FC	LoRA-FC/Conv1
Validation Accuracy	90.00%	96.00%	98.00%

Table 1. Model Performance

Model Certainty

The introduction of adversarial attacks using the Fast Gradient Sign Method (FGSM) revealed differences in model certainty, a measure of the confidence in the model's predictions. Before the attack, all model configurations displayed high certainty in their predictions. However, post-attack, the LoRA-FC model showed a significant drop in certainty, suggesting that while the accuracy was high, the model's resilience to adversarial inputs was limited. In contrast, the LoRA-FC+Conv1 model maintained a higher level of certainty, indicating a better retention of confidence despite adversarial perturbations.

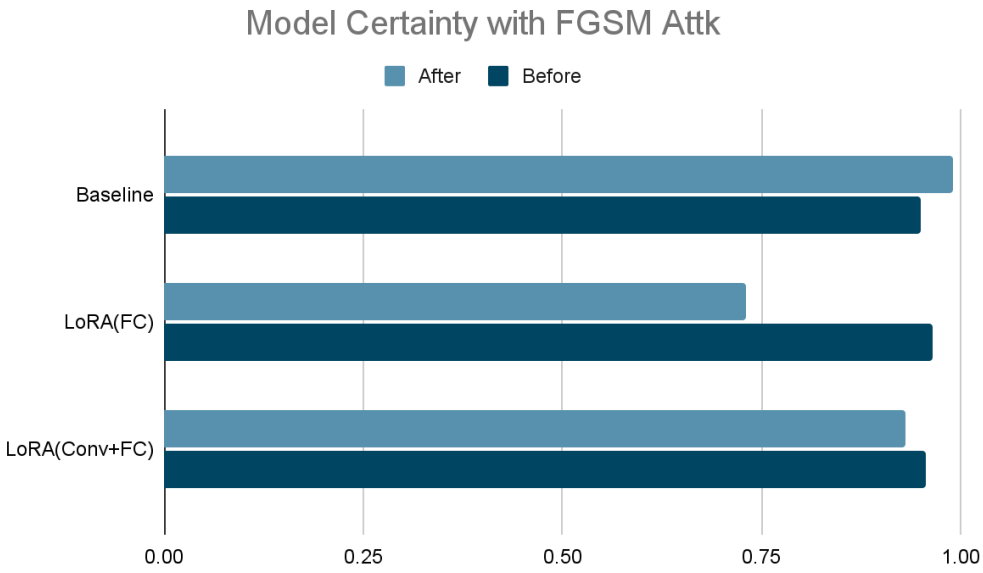

Figure 3. Model Certainty before/after FGSM attack

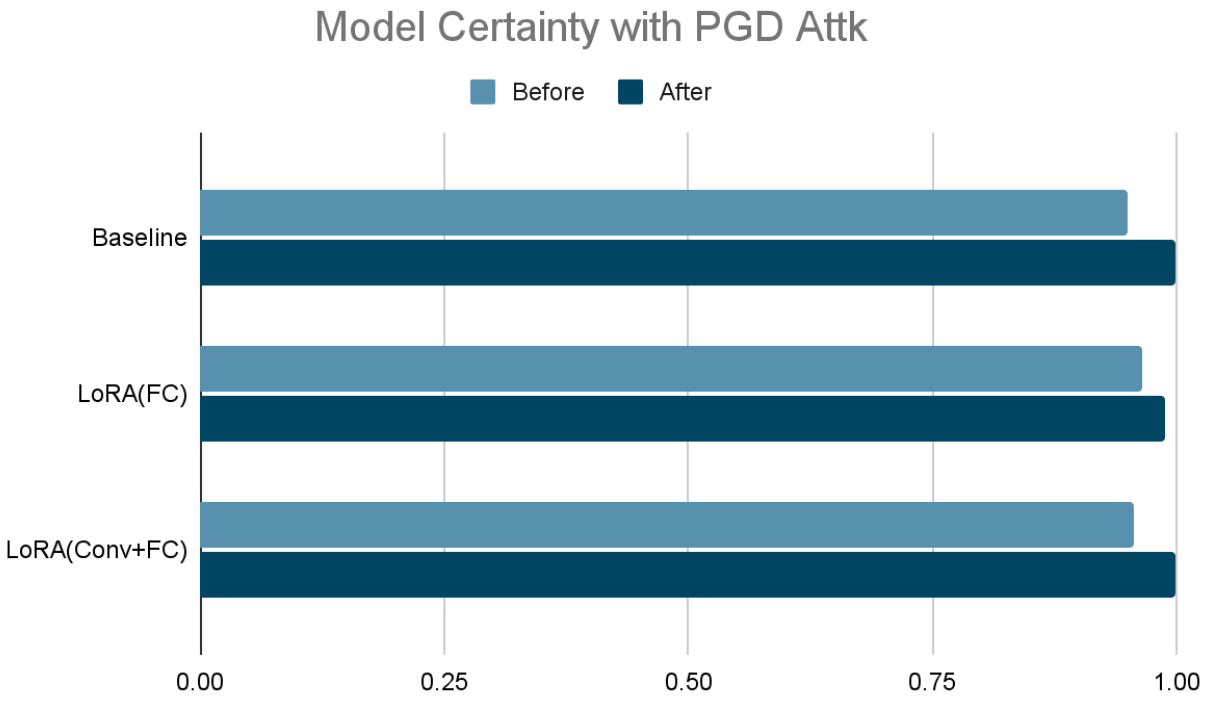

Figure 4. Model Certainty before/after PGD attack

Attack Success Rate

The robustness of the models against adversarial attacks was quantitatively assessed using the attack success rate metric. All models experienced a considerable impact from the FGSM attack, underscoring the effectiveness of this adversarial technique. The baseline model had an attack success rate of approximately 0.78, indicating that 78% of the model's correct predictions could be altered through the attack. The LoRA-FC adaptation showed a similar susceptibility, with a marginally lower attack success rate. Interestingly, the LoRA-FC+Conv1 adaptation demonstrated a slightly reduced attack success rate compared to the baseline, suggesting an improvement in adversarial robustness due to the comprehensive adaptation across both low-level and high-level features.

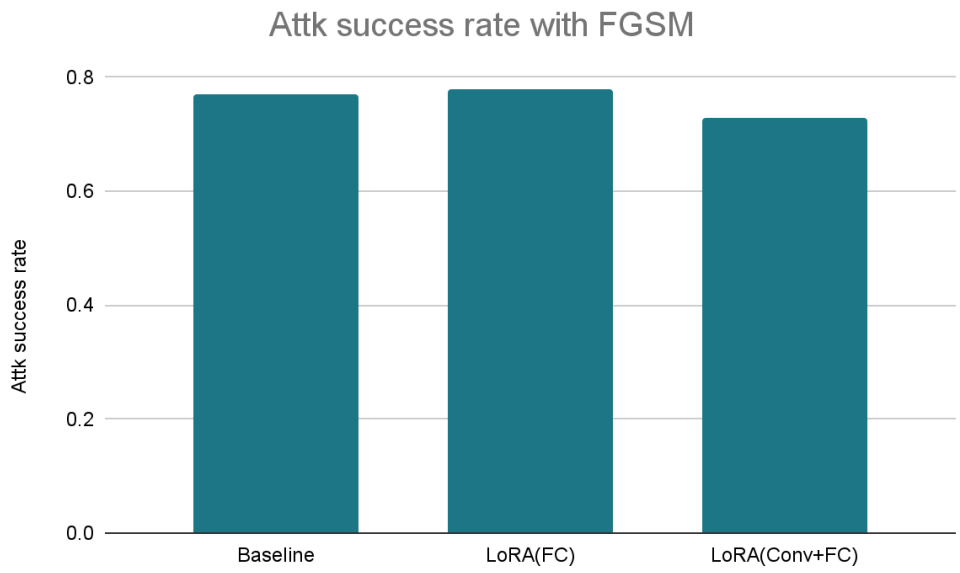

Figure 5. FGSM Attack Success Rate across three adaptation setting

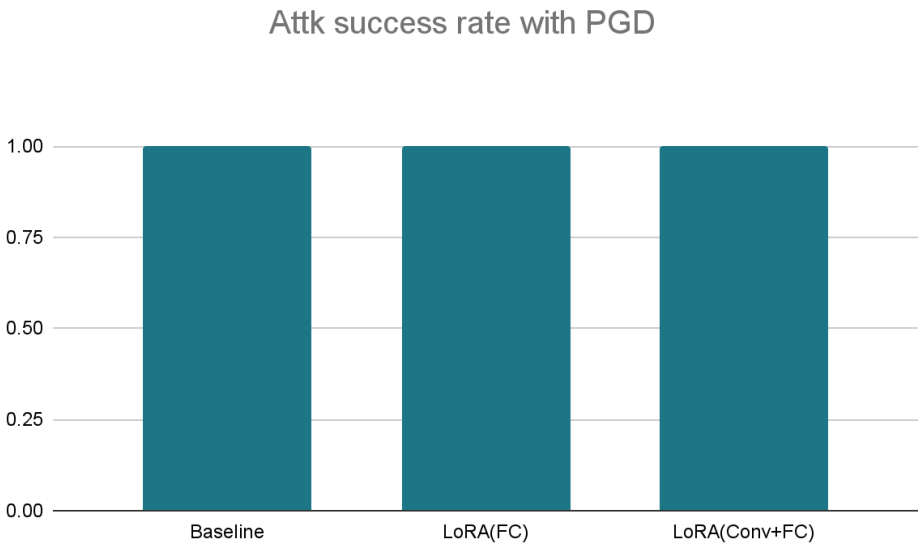

Figure 6. PGD Attack Success Rate cross three adaptation setting

Discussion

The results indicate that LoRA can enhance not only the accuracy but also the robustness of CNNs to adversarial attacks, particularly when applied across multiple layers. The improvement in attack success rates and retention of model certainty post-attack provides empirical evidence supporting the hypothesis that layer-specific adaptations can confer additional resilience against adversarial conditions. However, the increased performance also underscores the need for more nuanced adaptations, as the simple application of LoRA to the fully connected layers did not substantially enhance adversarial robustness compared to more comprehensive adaptations.

Future Work

Expansion to Black Box Attacks

Given the explorations into adversarial robustness under white box settings, a natural extension is to assess the model's performance against black box attacks. Notably, the Zeroth Order Optimization (ZOO) Attack and Boundary Attack offer avenues to understand the external robustness of the model without direct access to its gradients or architecture. Future work will involve employing these black box methods to evaluate if the observed drop in decision certainty

and vulnerability profiles noted under FGSM persist. This will help in establishing a more comprehensive understanding of the model’s behavior in real-world scenarios where attackers might not have inside knowledge of the system.

Dynamic Model Switching

Considering the results that show different adaptations of LoRA affecting attack success rates variably while maintaining similar accuracy levels, another promising area of investigation is dynamic model switching. This approach would involve switching between models with different LoRA adaptations based on the attack context or perceived threat level, potentially increasing system resilience by not relying on a single model's strengths and weaknesses. Analyzing the feasibility, efficiency, and effectiveness of such a strategy will provide insights into adaptive defenses in cybersecurity.

Enhanced Adaptation Techniques

Further research could also explore more granular adaptations using LoRA, such as varying the rank or exploring different layers for adaptation to optimize both performance and security. Additionally, integrating other defensive techniques like adversarial training or feature squeezing with LoRA adaptations could yield models that are robust across a broader spectrum of adversarial attacks.

References

1. Szegedy, C., et al. "Intriguing properties of neural networks." ICLR, 2013.
2. Goodfellow, I.J., Shlens, J., and Szegedy, C. "Explaining and harnessing adversarial examples." ICLR, 2014.
3. Madry, A., et al. "Towards deep learning models resistant to adversarial attacks." ICLR, 2017.
4. He, K., Zhang, X., Ren, S., and Sun, J. "Deep residual learning for image recognition." CVPR, 2016.
5. Xu, Y., et al. "Training behavior of deep neural network in frequency domain." NeurIPS, 2020.

6. Wei, C., et al. “Parameter-Efficient Tuning of Large Convolutional Models” arXiv, 2024
7. https://www.kaggle.com/c/tiny-imagenet