Evaluating White-Box Adversarial Attacks on Pre-Trained Deep Neural Networks: Enhancing Inference Robustness with Defensive Mechanism for Image Classification

Project Proposal

September 2024

1 Aim and Background

This project aims to explore and assess the robustness of pre-trained deep neural networks (DNNs) against white-box adversarial attacks, specifically concentrating on inference in image classification tasks. Leveraging popular pre-trained models such as ResNet, VGG, and MobileNet, this study will analyze how different types of white-box attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini-Wagner (C&W) simulate model inference on popular benchmarking datasets CIFAR-100 and SVHN. Adversarial attacks have become a significant challenge in deep learning, where even slight perturbations can drastically degrade a model's performance, leading to incorrect classifications. To mitigate these vulnerabilities, this project will execute defensive distillation as a mechanism to improve the model's robustness during inference. Defensive distillation is performed by softening the output probabilities of the model during training, which smooths the decision boundaries and makes the model less sensitive to adversarial perturbations (Roshan, 2024). By using this technique, the project aims to measure the extent to which defensive distillation enhances the model's resistance to attacks, comparing performance metrics such as classification accuracy under adversarial conditions versus clean input. The project will not involve training new models but will focus completely on inference, using pre-trained models to mimic real-world scenarios where pre-existing systems are subjected to adversarial attacks. Through this work, we will provide insights into the significance of defensive distillation in protecting DNNs during inference and offer recommendations for enhancing model robustness in image classification tasks. Earlier work on adversarial attacks has mainly focused on training robust models(Papernot, 2016), black box attacks, or using adversarial training methods, leaving gaps in understanding how pre-trained models perform under attack during inference alone. Further, while defensive distillation has been explored as a defense mechanism (Netzer, 2011), most studies have not thoroughly investigated its impact on inference robustness across different datasets like CIFAR-100 and SVHN, making this project a novel contribution to the field.

Here are two research questions for this task:

RQ1: How do white-box adversarial attacks (e.g., FGSM, PGD, Carlini-Wagner) affect the inference performance of pre-trained deep neural networks for image classification?

RQ2: How does a defensive mechanism improve the robustness of deep neural networks during inference against white-box adversarial attacks?

1.1 Project Plan

Summary of Major Tasks and Deadline:

1. Adversarial Attack Literature Review and Initial Research: October 10, 2024
   Comprehensive literature review on white-box adversarial attacks and defensive mechanisms, defining the scope of our research.

2. Data Collection and Preprocessing: October 16, 2024
   Collect datasets (CIFAR-100, SVHN) and preprocess them, ensuring consistency for adversarial testing and defensive mechanism evaluation

3. Feature Extraction and Analysis: October 23, 2024
   Extract features from datasets, analyzing the impact of adversarial attacks on different data points.

4. Neural Network Architecture Development: October 30, 2024
   Implement and adapt pre-trained models like ResNet50, VGG19, and MobileNetV2 for image classification and adversarial attack testing.

5. Implementation of White-Box Adversarial Attacks: November 6, 2024
   Implement and run adversarial attacks (FGSM, PGD, Carlini-Wagner) to test inference performance on the DNNs.

6. Defensive Mechanism Implementation: November 12, 2024
   Apply defensive mechanisms (e.g., adversarial training, defensive distillation) to the DNNs, preparing them for robustness testing.

7. Training and Optimization: November 18, 2024
   Train the adversarial models and models with defensive mechanisms. Optimize parameters for effective performance.

8. Model Performance Evaluation and Comparison: November 29, 2024
   Performance measure and compare the performance of the models (attacked and defended) with benchmarks from previous works on the same datasets.

9. Documentation and Final Report Preparation: December 4, 2024

2 Dataset of this Study

CIFAR100: The CIFAR-100 dataset is a popular benchmark for image classification tasks. It consists of 60,000 color images, each with a resolution of 32x32 pixels, split into 50,000 training images and 10,000 test images. CIFAR-100 is more difficult than its counterpart CIFAR-10 as it includes 100 different classes, each describing real-world objects like animals, vehicles, and household items, with 600 images per class. The dataset is further grouped into 20 superclasses, each containing five related subcategories. The diversity and complexity of the dataset make it an excellent choice for estimating the robustness of deep neural networks against adversarial attacks, as models need to differentiate between fine-grained variations in similar-looking images.

SVHN: The Street View House Numbers (SVHN) dataset is another famous dataset used for digit classification tasks, featuring real-world images of house numbers captured by Google Street View. SVHN includes over 600,000 digit images, where the focus is on determining single digits ranging from 0 to 9. The dataset is split into a training set of 73,257 images, a testing set of 26,032 images, and an additional set of 531,131 less difficult samples. SVHN gives a more natural setting compared to CIFAR-100, with images captured in varying conditions, often including distracting elements like nearby numbers or background clutter. This dataset is ideal for evaluating the robustness of DNNs, particularly in handling noisy, real-world data, which is particularly relevant for evaluating defensive mechanisms like distillation under adversarial attacks.

CIFAR100 LINK: Click here to see dataset

SVHN LINK: Click here to see dataset

2.1 Ethical Consideration

Ethical considerations play a vital role, particularly in the context of guaranteeing fairness and transparency in the use of adversarial attacks and defenses on deep neural networks. Consideration is taken to avoid any misuse of adversarial techniques in harmful applications, ensuring that the research is utilized solely for improving model robustness and security. All datasets are publicly available and properly anonymized to protect privacy and confidentiality. Additionally, the research adheres to principles of clarity, reproducibility, and integrity, confirming that the findings contribute positively to the advancement of secure and ethical AI practices.

3 References

1. Papernot, N. and McDaniel, P., 2016. On the effectiveness of defensive distillation. arXiv preprint arXiv:1607.05113.

2. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B. and Ng, A.Y., 2011, December. Reading digits in natural images with unsupervised feature learning. In NIPS workshop on deep learning and unsupervised feature learning (Vol. 2011, No. 2, p. 4).

3. Roshan, K., Zafar, A. and Haque, S.B.U., 2024. Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system. Computer Communications, 218, pp.97-113.