I-PT-RED backdoor defense

Course Instructor:
Professor Goerge Kesidis
Student:
Samira Malek

Pennsylvania State University

Contents

1 | Introduction

In this project, we aimed to utilize the I-PT-RED [Miller et al., 2023] method to determine whether a deep neural network (DNN) has been compromised by an attack. Specifically, we implemented the I-PT-RED method on the ResNet18 model, which was targeted in an attack designed to alter the classification of class 8 to class 4 whenever an image was subjected to a 4-pixel perturbation with a poisoning rate of 1%. This type of attack introduces subtle changes to the image data that may not be easily perceptible but can significantly impact the network's output.
In the "Estimate Perturbation" section of our study, we detailed the methodology used to optimize the perturbation (v_{st}) for each source-target (s,t) class pair. This step is crucial for understanding the specific characteristics of the attack and designing effective countermeasures.
Following this, in the "Detection Inference" section, we explained the I-Pt-RED approach for detecting whether a DNN has been attacked. Our detection method assumes that there is at most one target class for any given attack, simplifying the inference process and focusing on identifying significant deviations in the classification behavior of the network.
Finally, in the concluding section of our report, we presented a ROC plot for our detection algorithm. This plot illustrates the trade-offs between the true positive rate and false positive rate of the detection method, providing a visual representation of its effectiveness in identifying attacks on the DNN.

2 | Estimate Perturbation

We utilized a ResNet18 model that had been trained to misclassify instances of class label 8 as class label 4 with a poisoning rate of 1%. The perturbation pattern applied was 4-pixel perturbation and conformed to an L2 norm of 0.4. To identify the optimal perturbation v_{st} for each source-target class pair (s, t), we employed the perturbation optimization algorithm detailed in Section 6.1 of [Miller et al., 2023] on adversarial attacks. This methodology allowed us to systematically explore and implement effective perturbations for each targeted misclassification scenario. Figure 2.1 displays v_{84}, which represents the smallest perturbation in terms of the L2 norm.

Figure 2.1: Estimated Perturbation

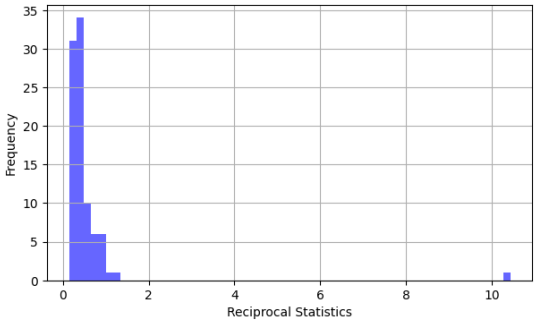


Figure 2.2: Histogram of reciprocal statistics $(\frac{1}{||v_{st}||_2})$

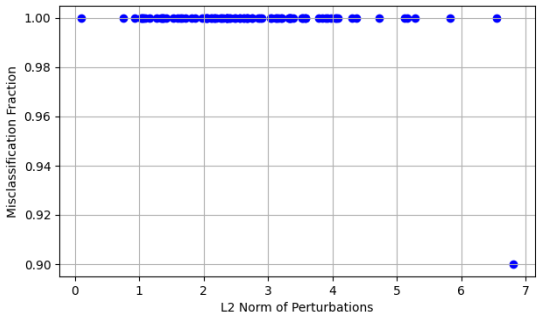


Figure 2.3: Sequences of $\left(\|\underline{v}_{st}\|_2, \rho_{st}\right)$ for all (s,t) pairs.

3 | Detection Inference

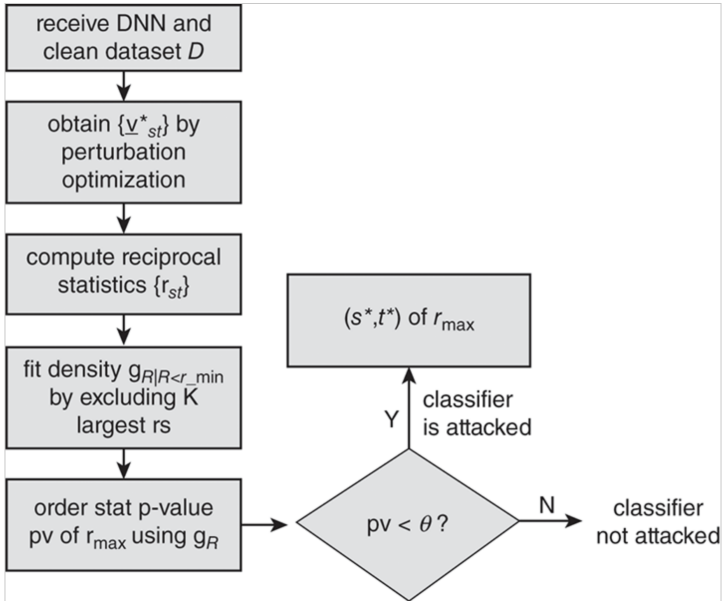


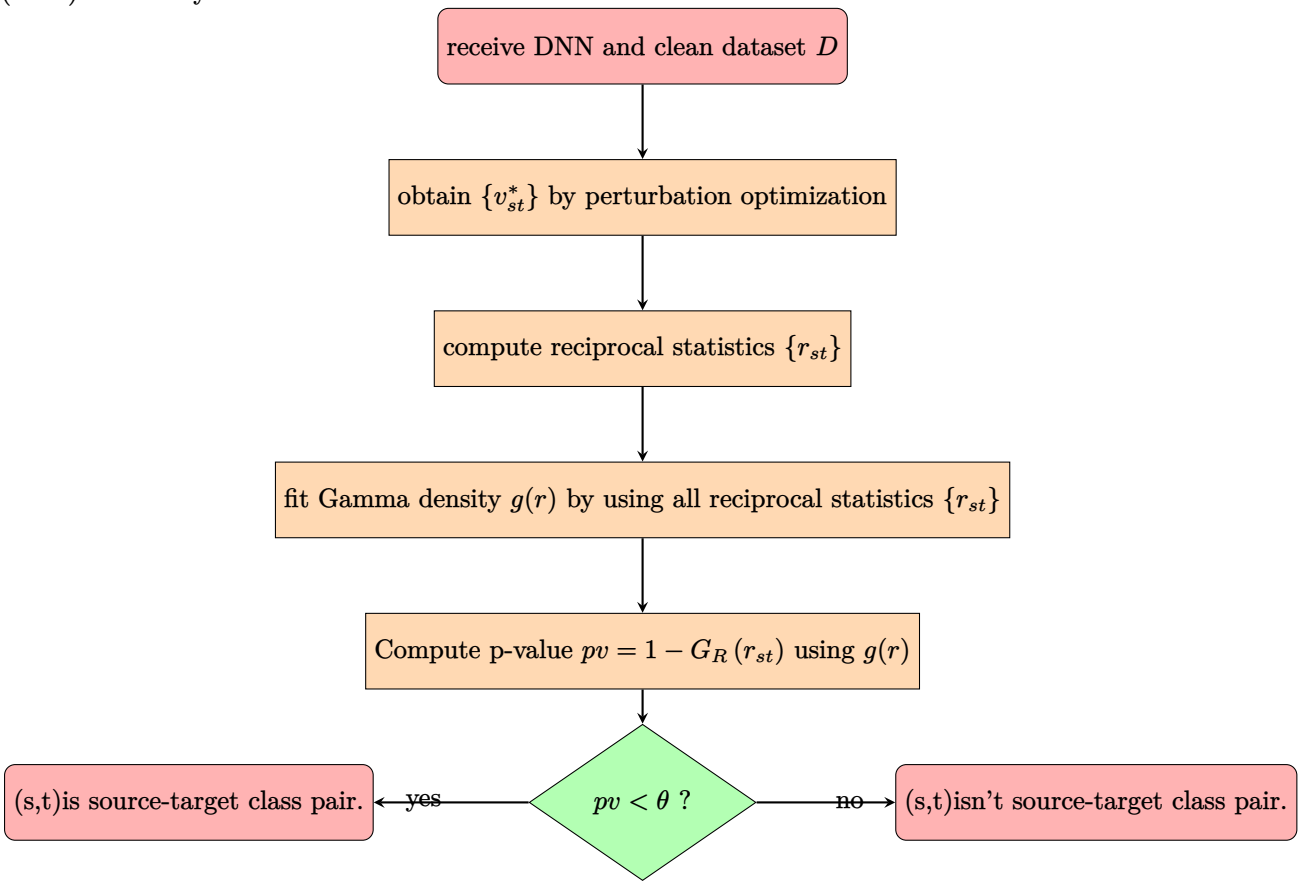
Figure 3.1: I-PT-RED flow chart

By applying the I-PT-RED method (see Figure 3.1), we determined that the DNN had been compromised, identifying (8,4) as the source-target class pair. Additionally, we modified the π values, optimized v_{st}, and reapplied the I-PT-RED method to assess whether the DNN had been attacked again. Our findings suggest that the choice of π does not significantly impact detection performance. This is because the required perturbation size for a true attack pair (s^*, t^*) is consistently observed to be anomalously small compared to that for non-attack class pairs across a broad range of π values. Then we computed p-value for all (s,t) pairs reciprocal r_{st}

4 | ROC Curve

Since we had only one attacked DNN, we could not use the exact same algorithm as depicted in Figure 3.1 to plot the ROC curve. Therefore, we made some modifications. Unlike the original method, we did not assume that there was at most one target class; instead, we considered each (s,t) pair as a potential source-target attack pair. Consequently, we used all of the reciprocal statistics $(K(K-1))$ to estimate the null density using a Gamma distribution. We then computed the p-value for all (s,t) pairs' reciprocal r_{st} to determine whether (s,t) was a source-target attack pair. In Figure **??**, we plotted the ROC curve for various threshold values θ ranging from 0 to 1 at increments [0, 0.05, 0.1, 0.15, 0.2, 0.4, 0.6, 1]. It is

important to note that, since (8,4) is the only identified source-target attack pair, the true positive rate (TPR) could only be 1 or 0.



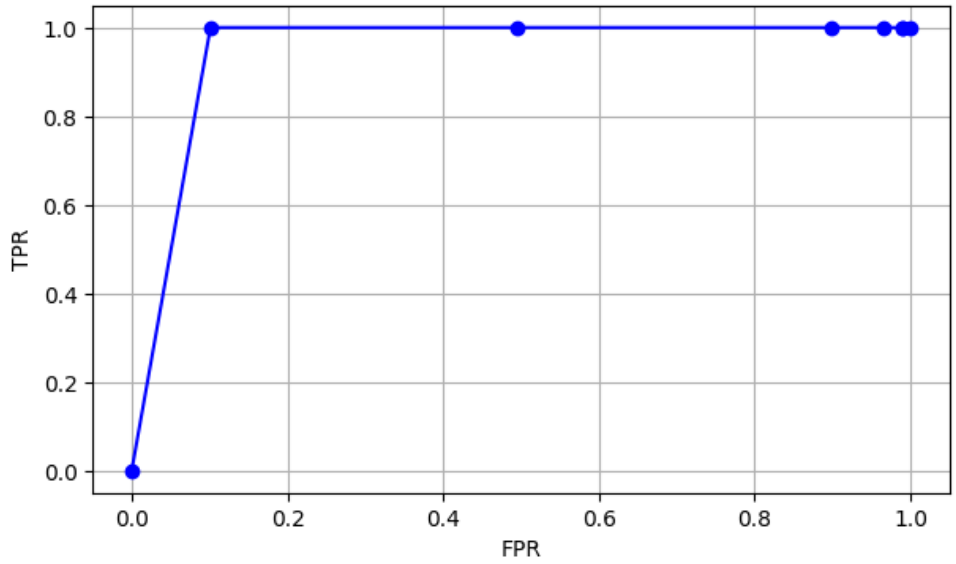


Figure 4.1: ROC curve

5 | References

[Miller et al., 2023] Miller, D. J., Xiang, Z., and Kesidis, G. (2023). *Adversarial Learning and Secure AI*. Cambridge University Press.