# Evaluating the Resilience of White-Box Defenses Against Adversarial Examples


## Abstract

It is well-established that neural networks exhibit susceptibility to adversarial examples. This paper assesses two defenses designed to counter white-box attacks and demonstrates their lack of effectiveness. Through the implementation of established methodologies, we successfully diminish the accuracy of these protected models to zero percent.


## 1 Introduction

A significant hurdle in the field is the development of neural networks that are resistant to adversarial examples. This paper shows that defenses created to address this issue are inadequate when faced with a white box scenario. Adversarial examples are generated that diminish classifier accuracy to zero percent on a well known dataset, while adhering to a minimal perturbation constraint of 4/255, a more stringent limit than what was taken into account in the initial studies. The proposed attacks effectively generate targeted adversarial examples, achieving a success rate exceeding 97

## 2 Background

This paper assumes prior knowledge of neural networks and the methods for creating potent attacks against adversarial examples, alongside calculating such examples for neural networks possessing non-differentiable layers. A concise review of essential details and notation will be provided.

Adversarial examples are defined as inputs that closely resemble a given input with regard to a certain distance metric (ŏ0a3, in this instance), yet their classification differs from that of the original input. Targeted adversarial examples are instances engineered to be classified as a predetermined target label.

Two defenses are scrutinized: Pixel Deflection and High-level Representation Guided Denoiser. The authors of these defenses are thanked for making their source code and pre-trained models accessible.

Pixel Deflection introduces a non-differentiable preprocessing step for inputs. A subset of pixels, determined by an adjustable parameter, is substituted with adjacent pixels. The resultant image often exhibits noise. To mitigate this, a denoising procedure is employed.

High-level Representation Guided Denoiser (HGR) employs a trained neural network to denoise inputs prior to their classification by a standard classifier. This denoiser is a differentiable, non-randomized neural network.

## 3 Methodology

The defenses are evaluated under the white-box threat model, generating adversarial examples using Projected Gradient Descent (PGD) to maximize cross-entropy loss, with the ŏ0a3, distortion limited to 4/255.

.

Many studies assert that white-box security is only applicable against attackers who are entirely ignorant of the defense mechanism in use. HGD, for example, states that the white-box attacks described in their research should be classified as oblivious attacks, according to previous research work's definition.

Protection against oblivious attacks proves to be ineffective. The concept of the oblivious threat model was introduced in prior work to examine the scenario involving an exceptionally weak attacker, highlighting that certain defenses fail to provide robustness even under such lenient conditions. Moreover, numerous previously disclosed systems already demonstrate security against oblivious attacks. A determined attacker would undoubtedly explore the potential presence of a defense and devise strategies to bypass it, should a viable method exist.

Consequently, security against oblivious attacks falls considerably short of being either intriguing or practical in real-world scenarios. Even the black-box threat model permits an attacker to recognize the implementation of a defense, while keeping the precise parameters of the defense confidential. Furthermore, it has been observed that systems vulnerable to white-box attacks are frequently susceptible to black-box attacks as well. Hence, this paper concentrates on evaluating systems against white-box attacks.

### 3.1 Pixel Deflection

It is demonstrated that Pixel Deflection lacks robustness. The defense, as implemented by the original authors, is analyzed and the code used for this evaluation is accessible to the public.

BPDA is applied to Pixel Deflection to address its non-differentiable replacement operation. This attack successfully diminishes the defended classifier's accuracy to 0

### 3.2 High-Level Representation Guided Denoiser

It is shown that employing a High-level representation Guided Denoiser is not resilient in the white-box threat model. The defense, as implemented by its developers, has been analyzed, and the code for this evaluation is openly accessible.

PGD is utilized in an end-to-end fashion without any alterations. This method reduces the accuracy of the defended classifier to 0

## 4 Conclusion

This paper shows that Pixel Deflection and High-level representation Guided Denoiser (HGD) are vulnerable to adversarial examples.