MAY 2023

DEPENDABLE AI

DEFLECTING ADVERSARIAL ATTACKS WITH PIXEL DEFLECTION

PROJECT REPORT

Group Members:

Shubham Kumar (B20AI039)
Tanmay (B20AI047)

GITHUB LINK: HTTPS://GITHUB.COM/TANMAYIITJ/DAI-PROJECT

02

OBJECTIVE

The aim of this project was to demonstrate a straightforward yet effective defense mechanism against adversarial attacks on deep neural networks. We address the vulnerability of deep learning models to adversarial examples, which are carefully crafted inputs designed to deceive the model's predictions.

By introducing the concept of pixel deflection, the paper seeks to disrupt the gradient information exploited by attackers and mitigate the impact of adversarial perturbations.

The primary goal is to enhance the robustness and reliability of deep learning models, allowing them to maintain high accuracy even in the presence of adversarial inputs.

03

INTRODUCTION AND MOTIVATION

Adversarial attacks refer to the deliberate manipulation of input data to deceive the model and induce misclassification. These attacks pose a significant threat to the reliability and security of deep learning models, particularly in critical applications such as autonomous vehicles, medical diagnosis, and malware detection.

While various defense methods have been proposed, they often suffer from high computational costs or limited effectiveness in mitigating adversarial attacks. Therefore, our project aims to address this issue by introducing the concept of pixel deflection as a simple yet powerful defense mechanism.

The motivation behind pixel deflection is to disrupt the gradient information exploited by attackers, thereby impeding their ability to generate effective adversarial examples. By leveraging this transformation-based defense approach, we enhance the robustness and reliability of deep neural networks against adversarial attacks.

04 PROPOSED METHODOLOGY

Below is the methodology we have implemented to get the results:

- **Classify the original image** :
  Take a sample image from dataset and classify it using the model.

- **Generate adversarial image and classify it**:
  Do adversarial attack on original image by any of the method like FGSM, IGSM, DeepFool etc. and then classify the adversarial image using the same model.

- **Do Pixel Deflection on original & adversarial image without CAM** :
  Do pixel deflection neighourhood-wise at a random locations of the image.

- **Analyse total deflection vs classification accuracy on both original image and adversarial image:**
  Increase number of deflected images slowly and anlayse which image orginal or adversarial is more severely affected at any given number of deflections.

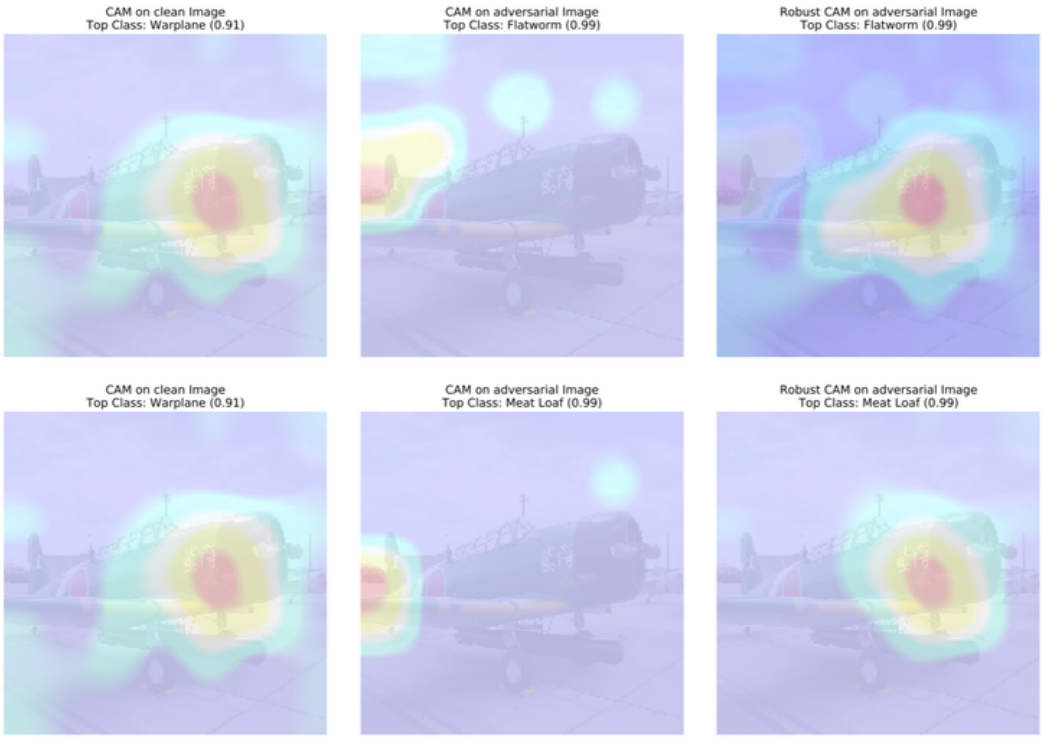
- **Do Pixel deflection on adversarial image using R-CAM** :
  Find most important region of the image for classification using R-CAM and esure pixel deflections are done outside it i.e. in the background.

05

- **Do Wavelet Denoising with RCAM after pixel deflection**
  Apply a wavelet denoising transform to lessen the noise effect added due to pixel deflection and adversarial perturbations.

- **Analyse the accuracies in various cases:**
  Comapre and contrast the accuraccies in the case of clean image, adversarial image, pixel deflected image, pixel deflected image with RCAM and wavelength denoising.

CAM VS ROBUST CAM



06

CAM VS ROBUST CAM

Class Activation Maps are heatmaps which highlight the areas most discriminative for a given image and a given class. This works well when the image contains the object given as the class but not so much if the given class has no presence in the image. In a given adversarial image by definition, the ‘class’ of the image is not the same as the true class. Thus, generating CAMs for adversarial images is difficult.
In order to overcome this, we propose a robust version of CAM. If we see the Top-5 predictions for the adversarial image img_adversary other than the ‘adversarial class’, it is no surprise that most of these classes are closely related to the true class even if not the same; skunk, polecat, weasel, and mink are all similar looking animals. This is a well-known side effect of ImageNet because a thousand classes of ImageNet have a lot of fine-grained classes of similar objects/species.

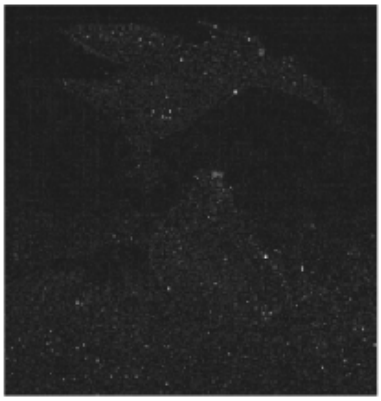
Robust Activation Map is geometric mean of CAM obtained using the top-K classes. By taking the top-K classes, we average out the impact that a single bad adversarial class may have on the activation map.
In above page we have a visual comparison of CAM and robust version of CAM.

07 RESULTS

EFFECT OF WAVELET DENOISER

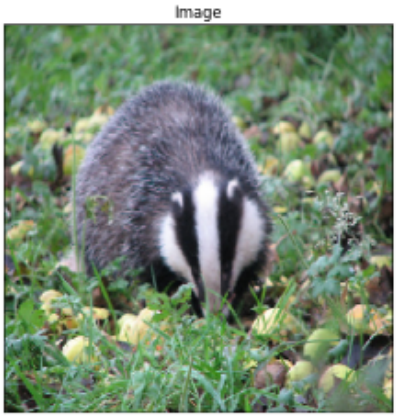
Image with 500 deflections

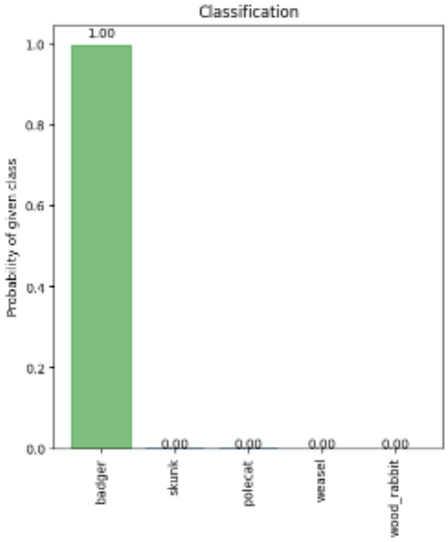
Difference between clean and deflected image



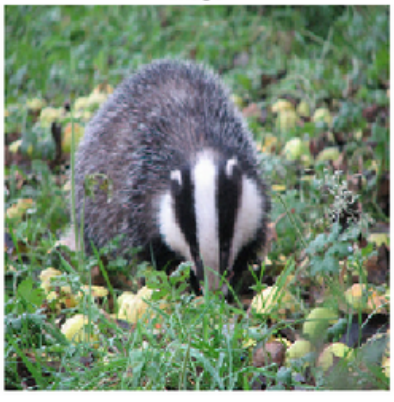
Difference between clean and deflected after denoising

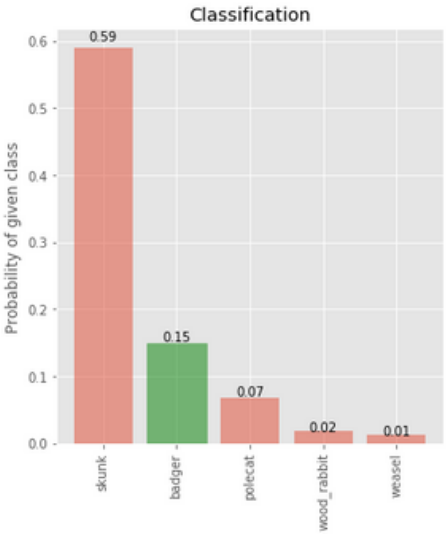
Performance Across Classifier with Only Pixel Deflections





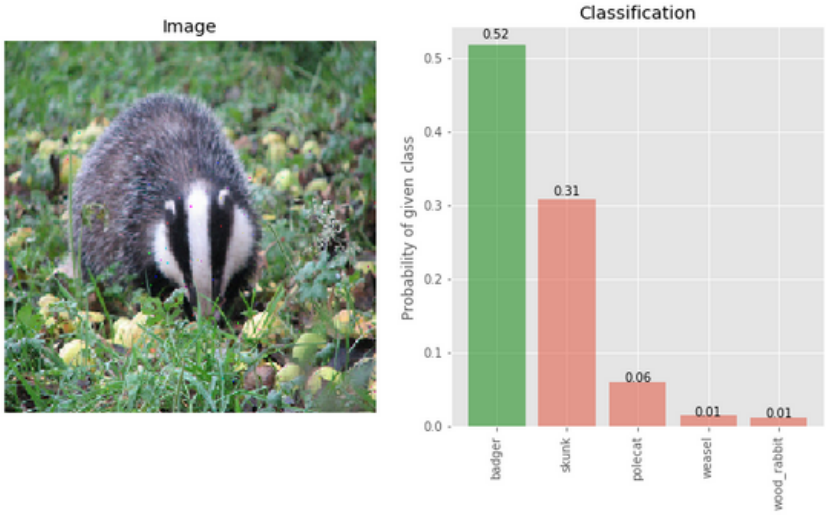
CLEAN IMAGE





ADVERSARIAL IMAGE

08



PIXEL DEFLECTED ADVERSARIAL IMAGE

ACCURACIES COMPARISION

Class	Clean	Adversary	PD	PD+WD	PD+RCAM	PD+RCAM+WD
True class - Badger	100	15	75	92	82	97
Adversary - Skunk	0.0	59	16	07	12	02

REFERENCES

- https://github.com/iamaaditya/pixel-deflection
- https://github.com/mashaan14/MNIST-M
- https://arxiv.org/abs/1801.08926