Robust Pipeline for Detecting Adversarial Images

Indian Institute of Information Technology, Nagpur

Ankit Vohra (BT17CSE116). Abhibha Gupta (BT17CSE020), Gaurav Agarwal (BT17CSE001)

Dr. Pooja Jain

Introduction

Deep learning has fueled great strides in a variety of computer vision problems, such as object detection, action recognition, human pose estimation, and segmentation. Despite there success, research has shown that DNN's are broadly vulnerable to adversarial examples, carefully chosen inputs that cause the network to change output without a visible change to a human. Yet, for humans these perturbations are often visually imperceptible. In fact, so called adversarial examples are crucially characterized by requiring minimal perturbations that are quasi-imperceptible to a human observer.

Keeping this in mind, we propose a robust pipeline capable of detecting adversarially attacked images

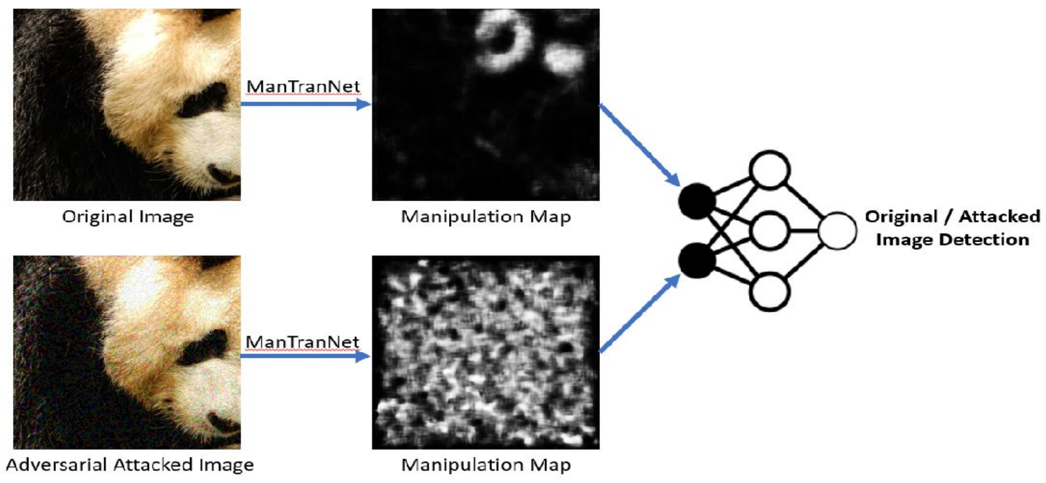
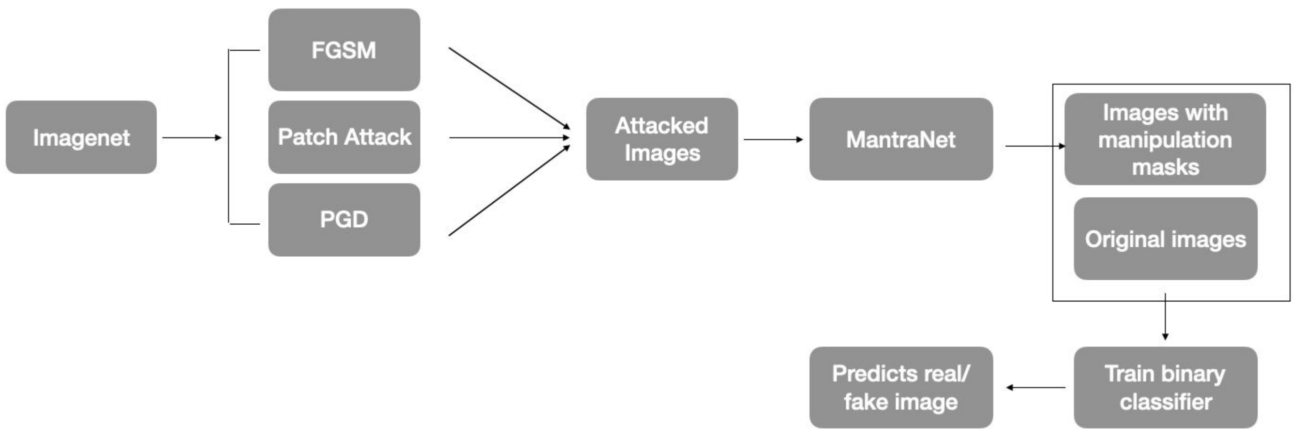


Fig 1: Proposed Model

Proposed Methodology

The pipeline consists of 3 steps. First is generating adversarially attacked images using different attacking methods (FGSM, PGD and Patch based attack) from the ILSVRC 2012 dataset . Next, we pass the generated samples through Mantranet that identifies perturbations and stores manipulation masks of the image. Finally a binary classifier is trained on the manipulation masks to differentiate between real and fake images.



Imagenet

FGSM

Patch Attack

PGD

Attacked Images

MantraNet

Images with manipulation masks

Original images

Train binary classifier

Predicts real/ fake image

Fig 2: Flowchart of the Training process

Results

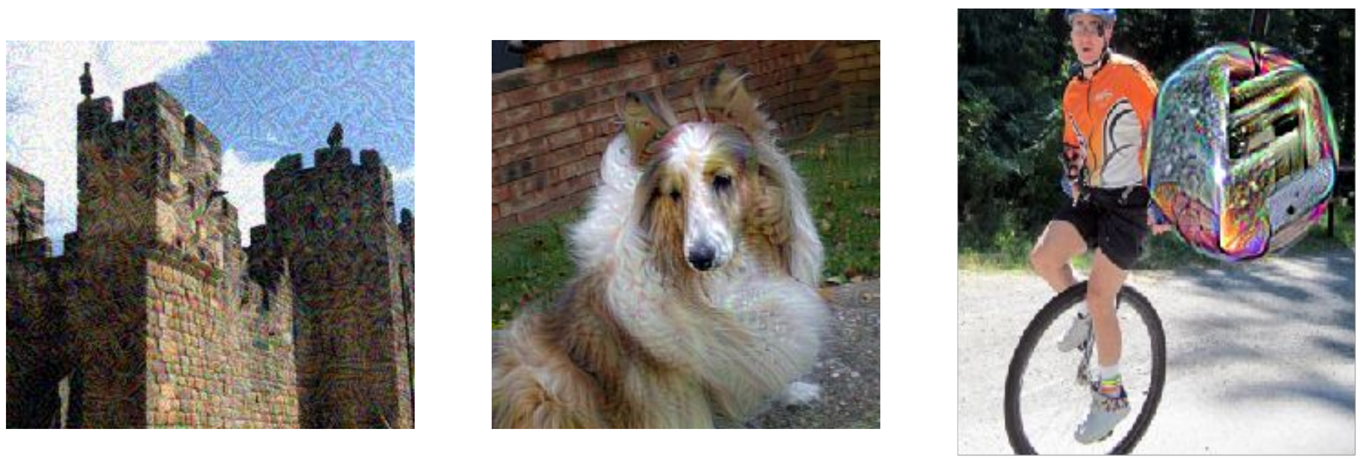


Fig 3: i. FGSM Attack, ii. PGD Attack iii. Patch Attack

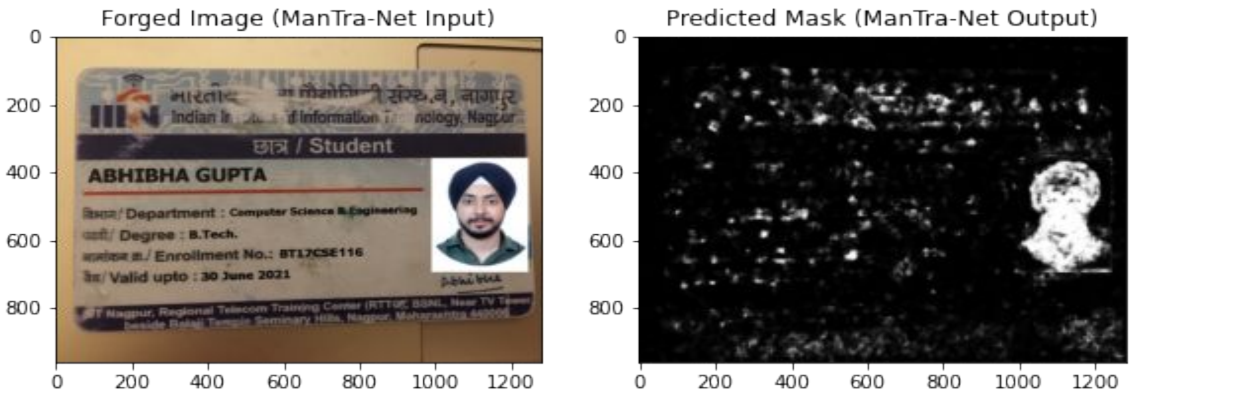

Forged Image (ManTra-Net Input)
Predicted Mask (ManTra-Net Output)
0
200
400
600
800
1000
1200

Fig 4: Photo Editing Attack and Corresponding manipulation mask output from ManTraNet

Accuracy: It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

	Previous SOTA	Basic Binary Classifier	Binary Classifier with ManTraNet
Accuracy of Detecting Adversarial Sample	99.9% for ϵ>=0.03 0.05% for ϵ<0.03	68%	98.6%

The pipeline provides 99.6% accuracy on adversarial image detection. It handles a variety of edge cases

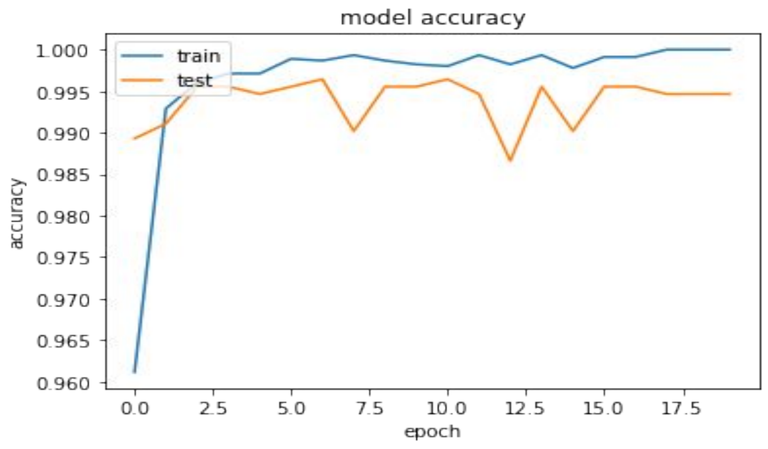


Fig 5: Classifier Accuracy vs Epch

Conclusions and Future Work

Images detection systems can be deceived by creating Adv. Images that appear to be similar to original image. In this paper, we propose a robust pipeline that focuses on avoiding adversarial attacks. We introduce a simple method to detect spoofed images, irrespective of the attack method used. We obtain a good accuracy score of 98.6. To further improve results ManTra-Net can be trained for the specific task of detecting manipulation masks. Although the pipeline is robust but morphological operations on manipulation masks can be tried out to improve the results.

Team No.- CSE17_2_5