Deflecting Adversarial Attacks with Pixel Deflection

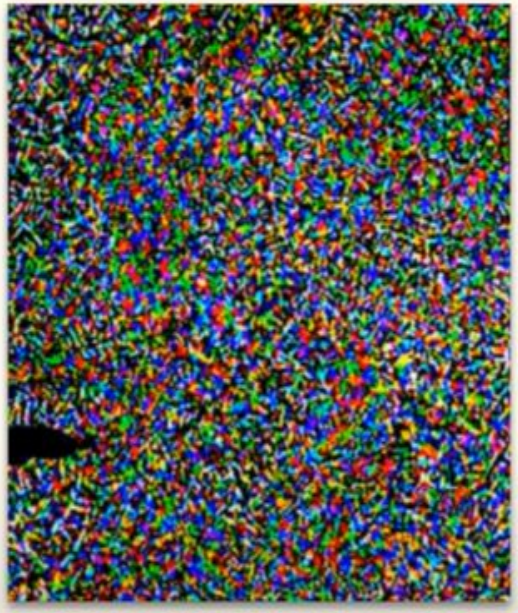
Authors: Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, James Storer

Published: CVPR 2018

Adversarial Attack

Original Image

+



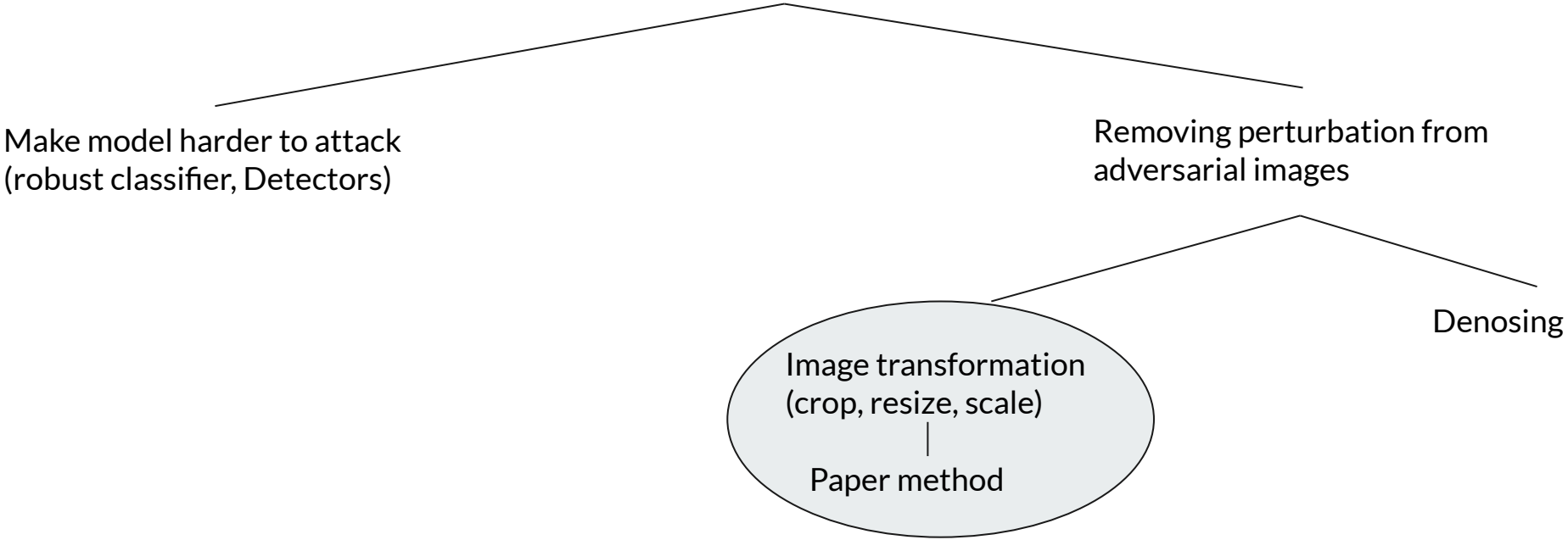
=

Adversarial Image

Predicted: Indian Elephant (99.7%)

Predicted: Guacamole (99.9%)

Defending against adversarial attack



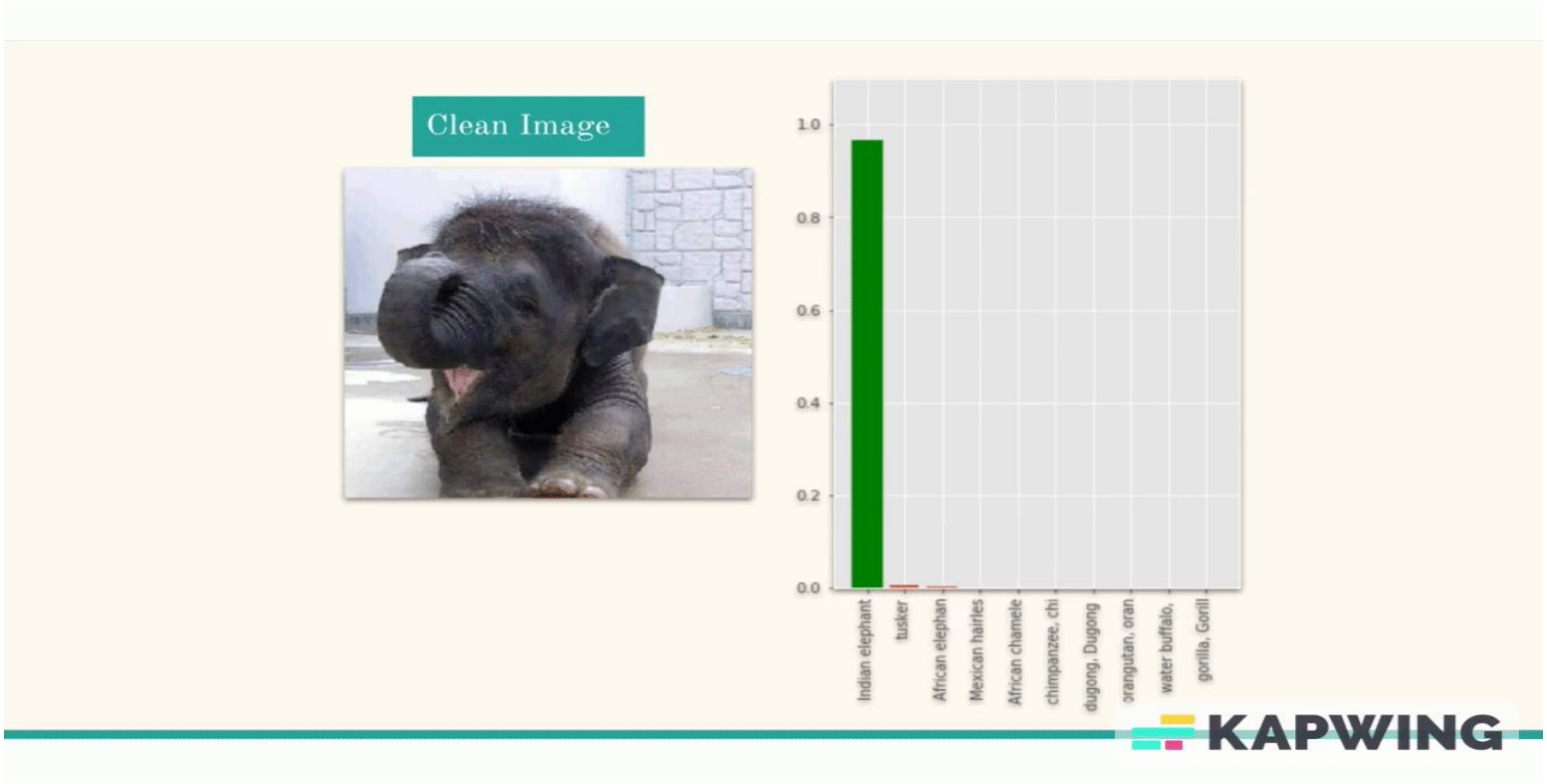
KEY INSIGHT #1

CLASSIFIER are robust to **noise**

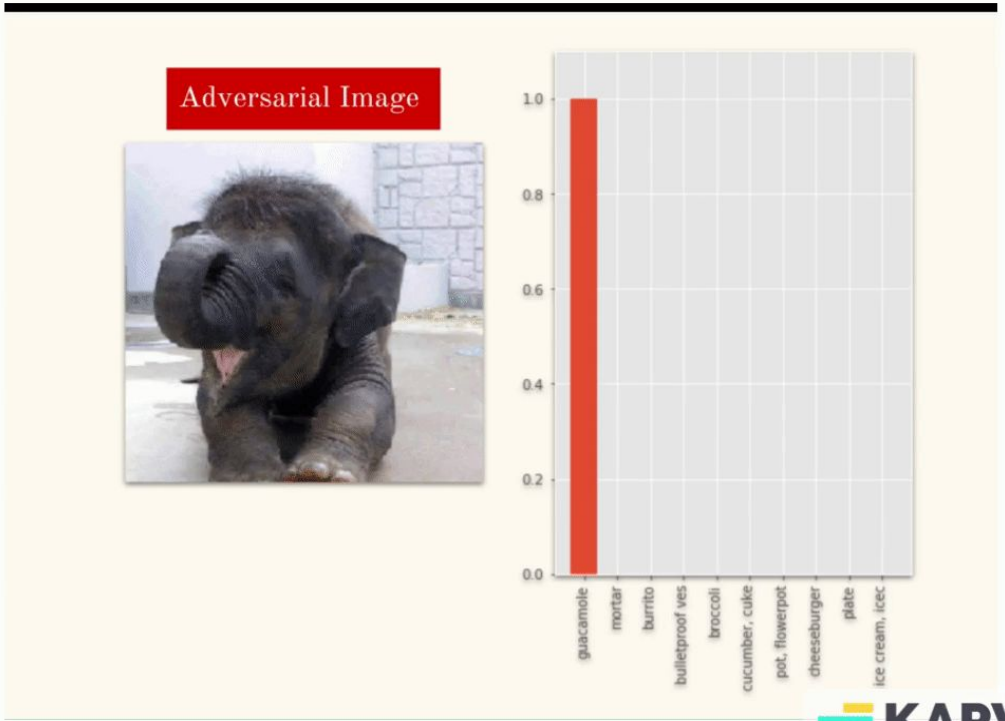
But

Adversarial Systems are not.

EXAMPLE (Pixel replacement in Original Image)



Pixel replacement in adversarial Image



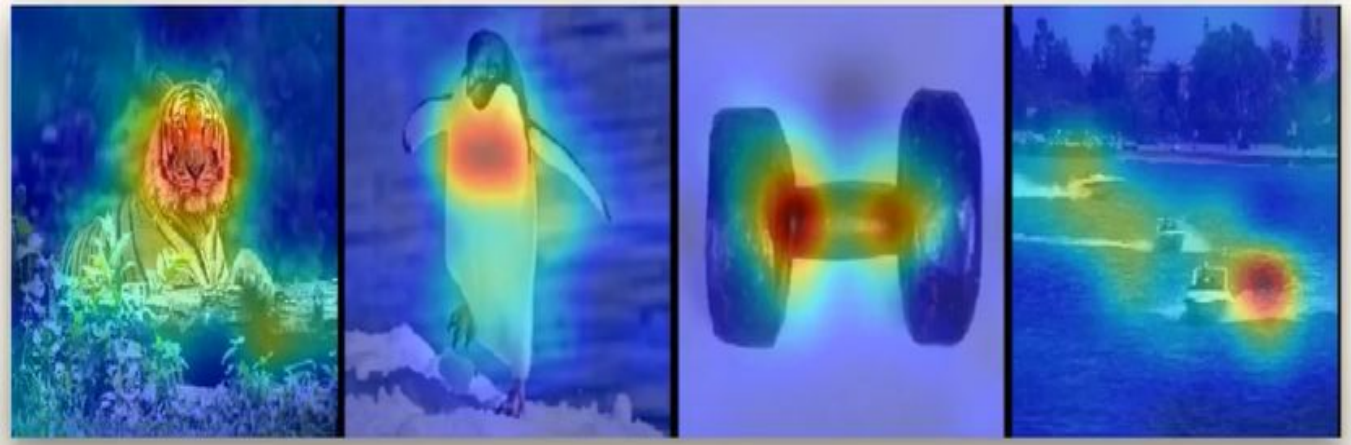
KEY INSIGHT #2

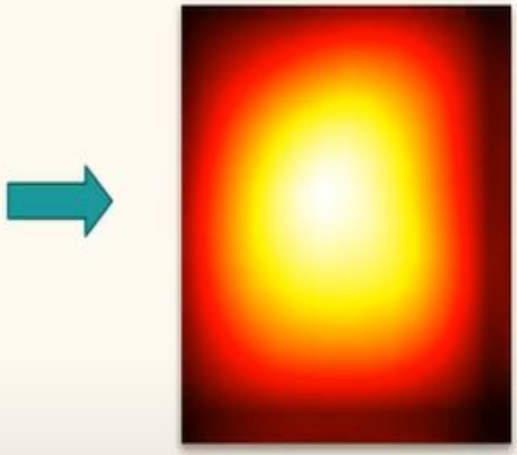
Classifiers look for **semantic regions**

but

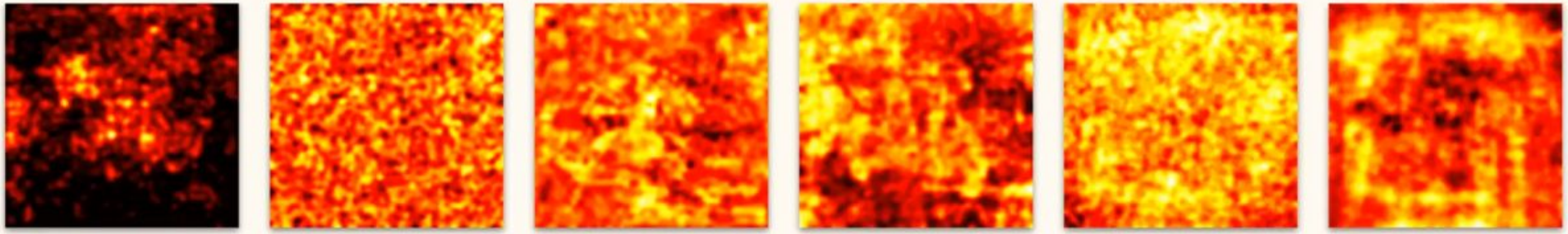
Adversarial Systems are content agnostic.

CLASS ACTIVATION MAP





Average Location of Adversarial Perturbation



JSMA C&W IGSM LBFGS FGSM Deep Fool

Average location of objects

PIXEL DEFLECTION

Algorithm 1: Pixel deflection transform

Input : Image I, neighborhood size r
Output: Image I' of the same dimensions as I

for $i \leftarrow 0$ **to** K **do**
  Let $p_i \sim \mathcal{U}(I)$
  Let $n_i \sim \mathcal{U}(R_p^r \cap I)$
  $I'[p_i] = I[n_i]$
end

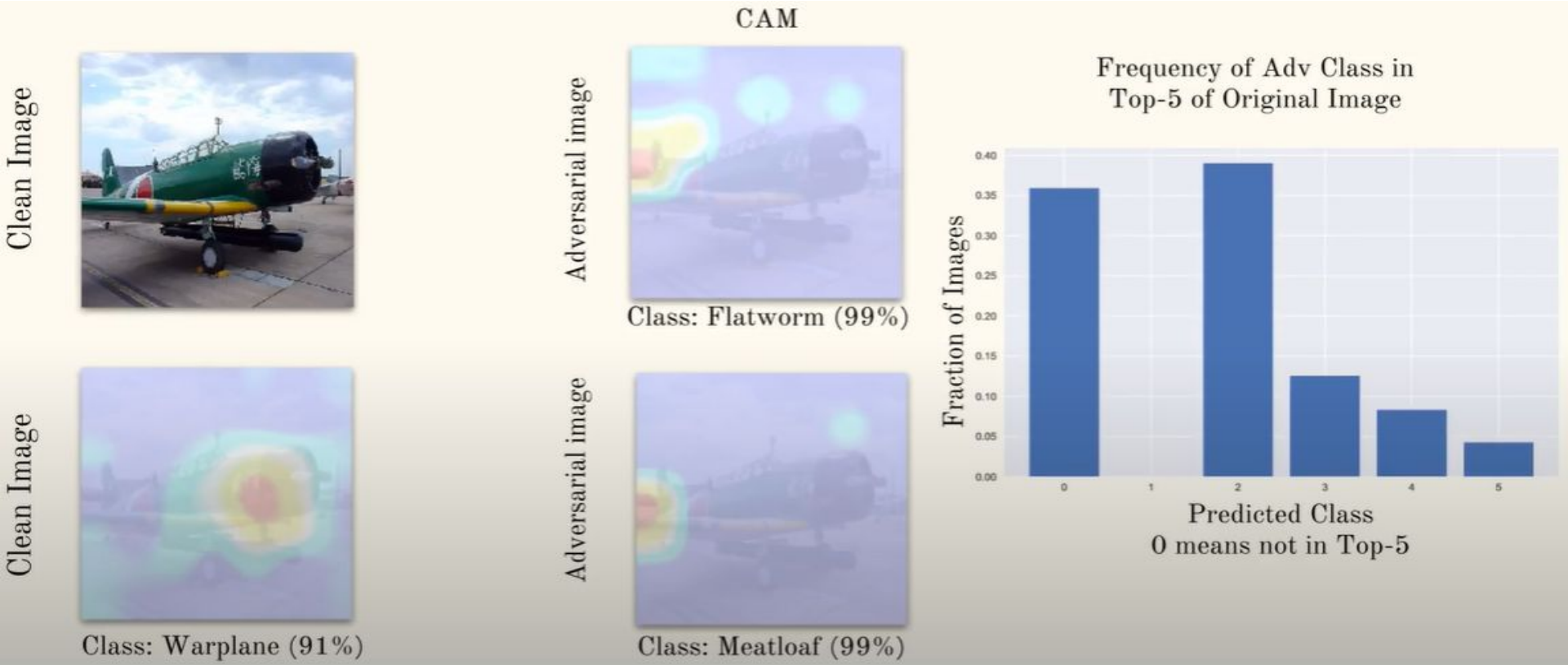
PIXEL DEFLECTION

Algorithm 1: Pixel deflection transform

Input : Image I, neighborhood size r
Output: Image I' of the same dimensions as I

for $i \leftarrow 0$ **to** K **do**
  Let $p_i \sim \mathcal{U}(I)$
  Let $n_i \sim \mathcal{U}(R_p^r \cap I)$
  $I'[p_i] = I[n_i]$
end

Class Activation Map is Fooled too !



Class Activation Map is Fooled too !

CAM

Clean Image

Adversarial image

Class: Flatworm (99%)

CAM

$$M_c(x,y) = \sum_k w_c^k f_k(x,y)$$

Clean Image

Class: Warplane (91%)

Adversarial image

Class: Meatloaf (99%)

Robust CAM

$$\hat{M}(x,y) = \sum_c \frac{M_c(x,y)}{2^i}$$

WAVELET DENOISER

- Since both pixel deflection and adversarial attacks add noise to the image, it is desirable to apply a denoising transform to lessen these effects.

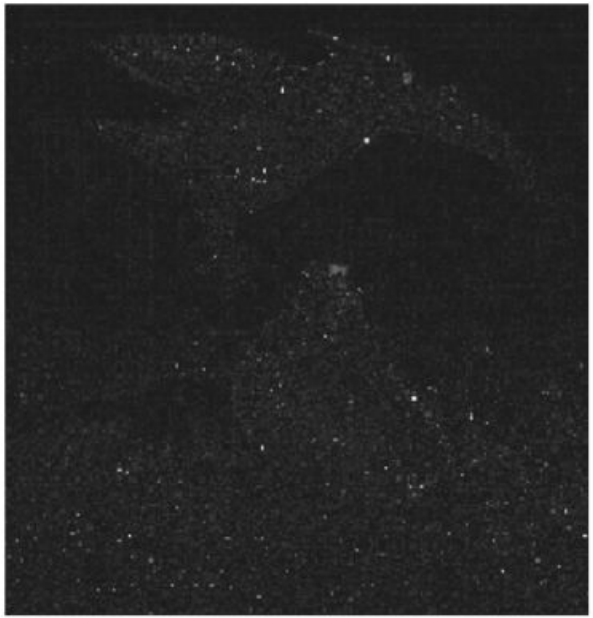
- It involves decomposing the image into its wavelet coefficients, thresholding these coefficients to remove noise, and then reconstructing the image from the modified coefficients.

WAVELET DENOISER

Deflections=500

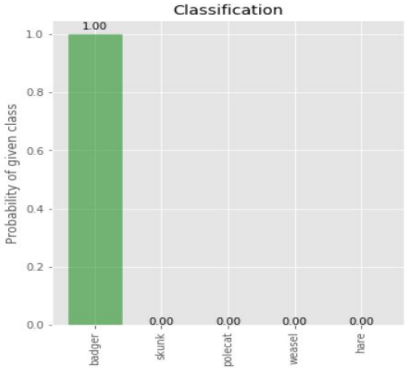
Image with 500 deflections

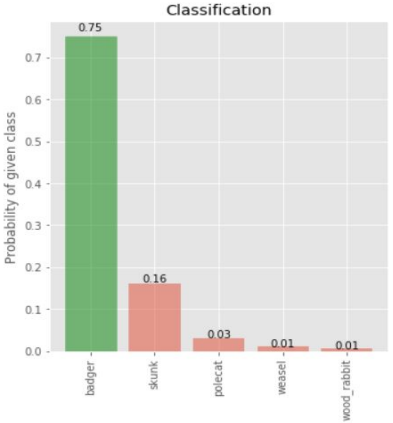
Difference between clean and deflected image

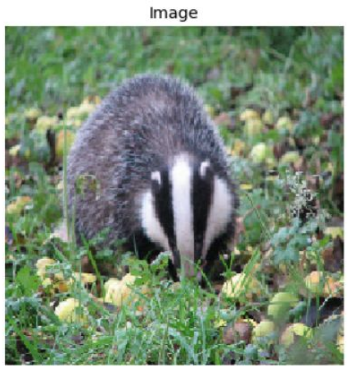


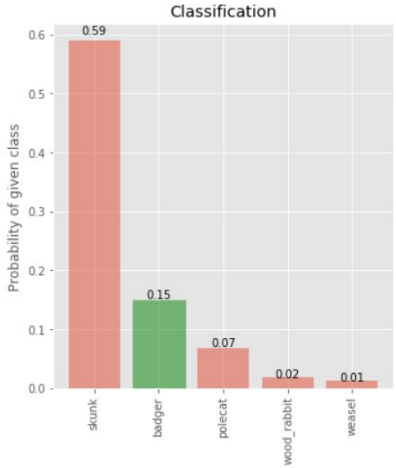
Difference between clean and deflected after denoising

Results - Performance Across Classifier with Only PD







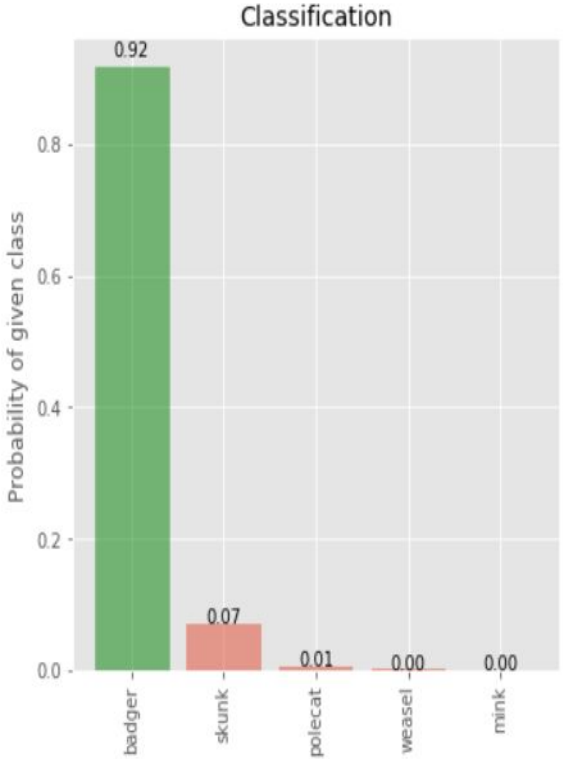


Class	Clean	Adversary	Pixel Deflection
True class - Badger	100	15	75
Adversary - Skunk	0.0	59	16

(numbers denote confidence in each class)

Results - Performance Across Classifier with PD+WD

Image



Class	Clean	Adversary	PD + WD
True class - Badger	100	15	92
Adversary - Skunk	0.0	59	07

(numbers denote confidence in each class)

THANK YOU

Tanmay B20AI047

Shubham Kumar B20AI039