Defending against Physically Realizable Attacks on Image Classification

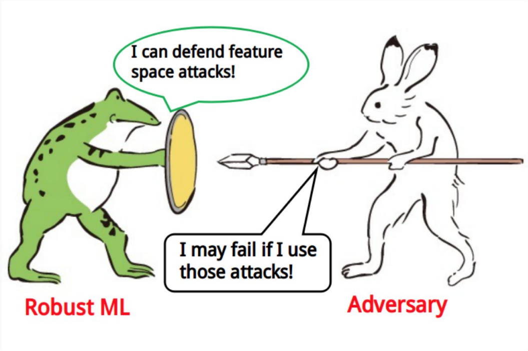
Tong Wu, Liang Tong, Yevgeniy Vorobeychik

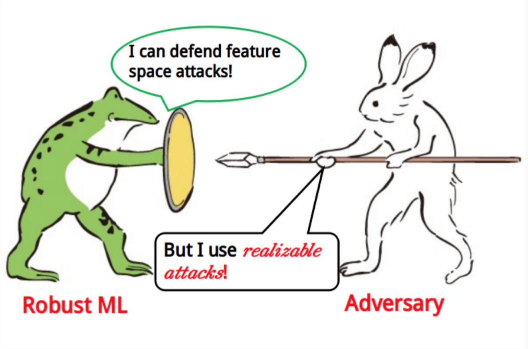
Washington University in St. Louis

Motivation

A large literature has emerged on defending deep neural networks against adversarial examples on the feature space, namely l_2, l_infty etc.

However,there seem no effective methods specifically to defend against physically realizable attacks (major concern in real life).


I can defend feature space attacks!
I may fail if I use those attacks!
Robust ML
Adversary


I can defend feature space attacks!
But I use *realizable attacks*!
Robust ML
Adversary

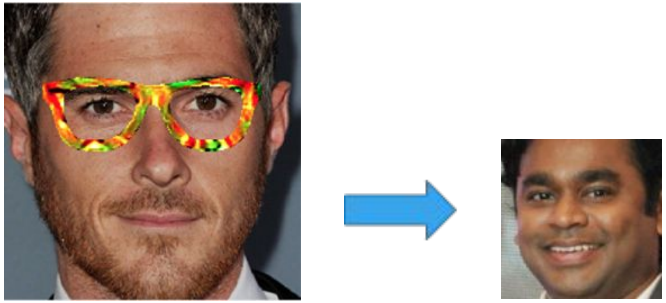
Done by Liang Tong,
USENIX 2020

What is Physically Realizable Attack

Dave Annable

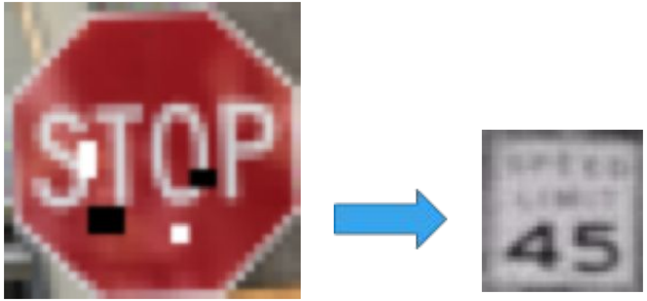
Stop Sign

What is Physically Realizable Attack



A.R. Rahman A.R. Rahman

Sharif et al. CCS



Speed Limit Speed Limit

Eykholt et al., CVPR

Physically Realizable Attack

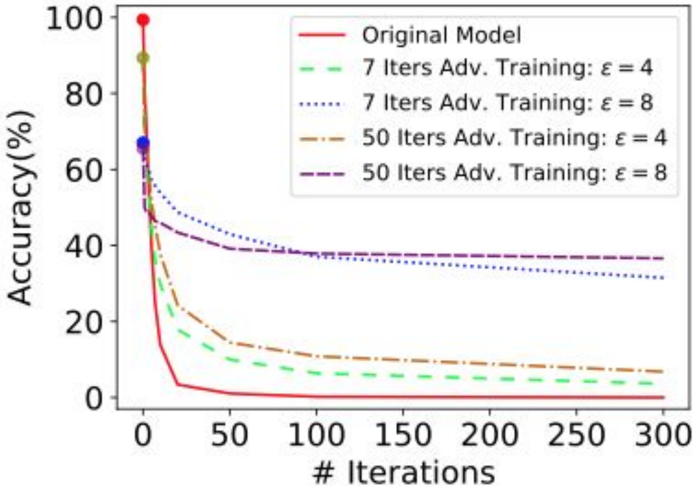
1. The attack can be implemented in the physical space (e.g., modifying the stop sign);

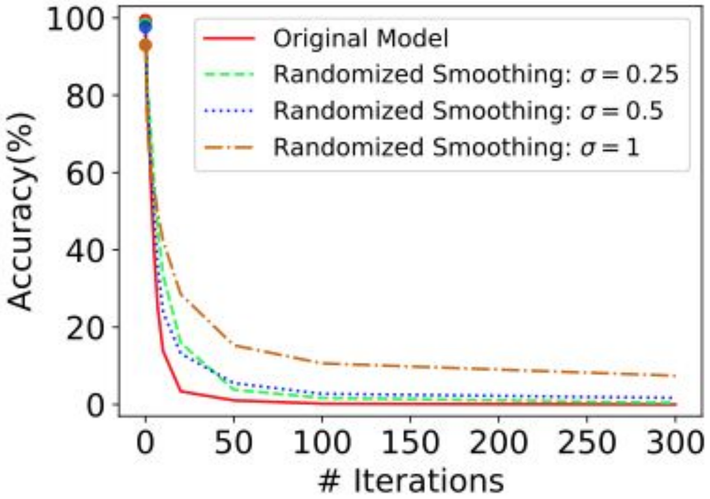
2. the attack has low suspiciousness; this is operationalized by modifying only a small part of the object, with the modification similar to common "noise" that obtains in the real world;

3. the attack causes misclassification by state-of-the-art deep neural network

In the experiment, we did not use real glass frame, but we defend the physically realizable attacks in digital domain. This seems to be stronger attacks.

Failure of Robust Learning & Randomized Smoothing against Physically Realizable Attacks





Is It even Possible to Build a Robust Model against Physically Realizable Attacks ?

The evidence suggests that the conventional models which succeed in lp-bounded attacks are not particularly useful when facing such physical attacks.

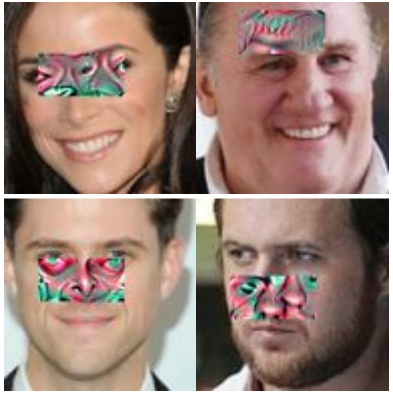
The common constraint of such attacks is to limit the size of the adversarial occlusion (for the purpose of avoiding suspicious), but not its shape or location.

Abstract Attack Model: Rectangular Occlusion Attacks (ROA)

This rectangle can be placed by the adversary anywhere in the image.

Attacker can furthermore introduce l_infty noise inside the rectangle with epsilon = 255.

This attack is untargeted attack



Determine the Location

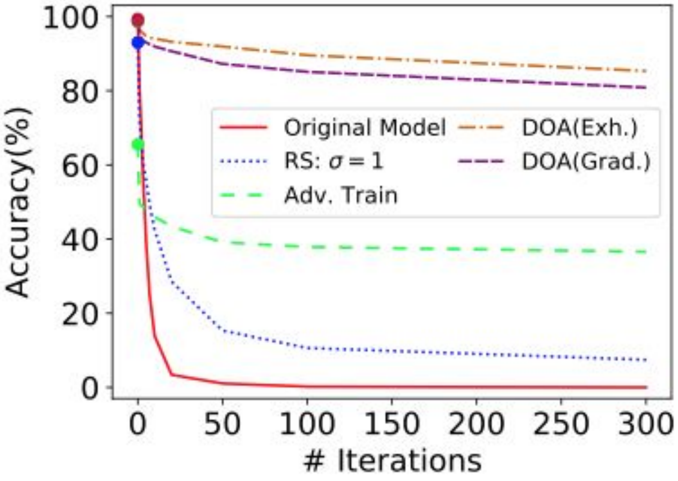
Exhaustive Searching : Adding a grey rectangular sticker to image, considering all possible locations and choosing the worst-case attack

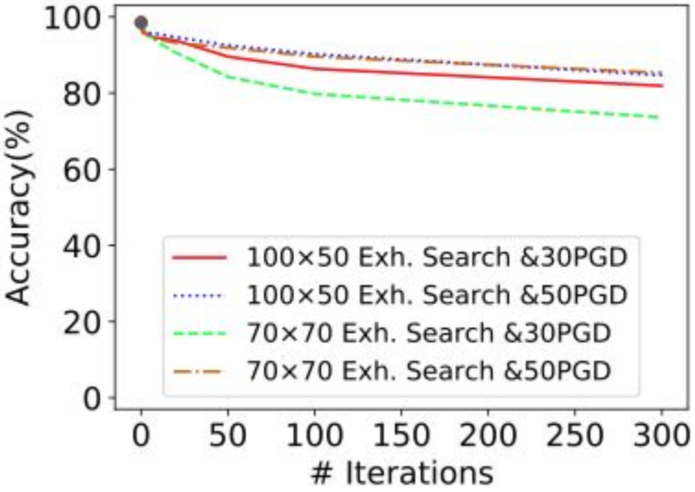
Gradient Based Searching : Computing the magnitude of the gradient w.r.t each pixel, considering all possible locations and choosing C locations with largest magnitude. Exhaustively searching among these C locations.

Applying PGD attacks to the sticker with epsilon of 255.

Defense against Occlusion Attacks (DOA)

We apply the Standard Adversarial training approach for ROA





Thank you