Robust Discriminative Localization Maps

A. Prakash, N. Moran, S. Garber, A. DiLillo & J. Storer

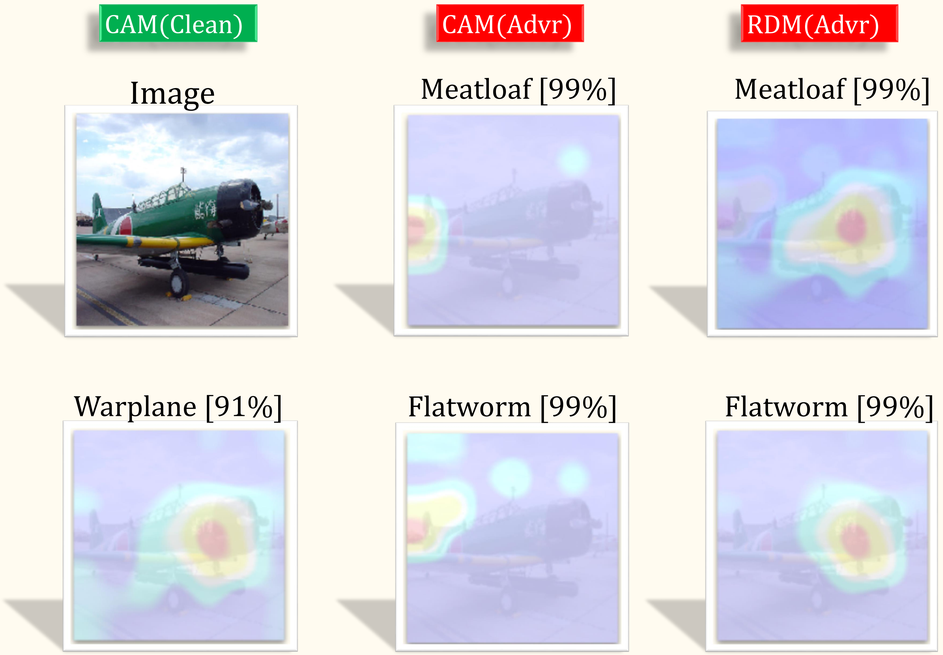
aprakash@brandeis.edu, iamaaditya.github.io

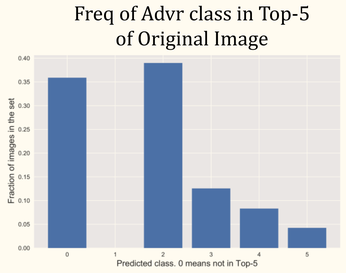
github.com/iamaaditya/robust-activation-maps

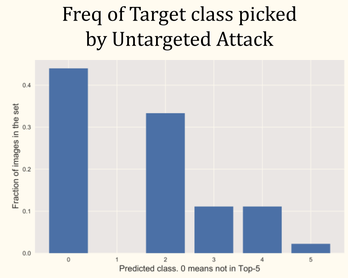
CVPR 2018
Salt Lake City

Class Activation Maps + **Adversarial Systems** = **Bad ! Bad !! FAKE Activation Maps !!**

Visual Question Answering	Image Classification	Image Compression	Image Retrieval	Guided Attention	Adversarial Defense
[Zhou, 2015]	[Zhou, 2016]	[Prakash, 2017]	[Jimenez, 2017]	[Li, 2018]	[Prakash, 2018]







Class Activation Map

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

Robust Discriminative Map

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

Defense	FGSM	IGSM	DFool	C&W
Feature Squeezing (Xu et al [49])				
(a) Bit Depth (2 bit)	0.132	0.511	0.286	0.170
(b) Bit Depth (5 bit)	0.057	0.022	0.310	0.957
(c) Median Smoothing (2x2)	0.358	0.422	0.714	0.894
Quilting + TVM (Guo *et al.* [19])				
Quilting	0.611	0.862	0.858	0.843
TVM + Quilting	0.619	0.866	0.866	0.841
Cropping + TVM + Quilting	0.629	0.882	0.883	0.859
Our work: PD - Pixel Deflection, R-CAM: Robust CAM				
PD	0.735	0.880	0.914	0.931
PD + R-CAM	0.746	0.912	0.911	0.952