Protecting JPEG images against adversarial attacks

Data Compression Conference
2018

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo and James Storer
Brandeis University

Image consumption

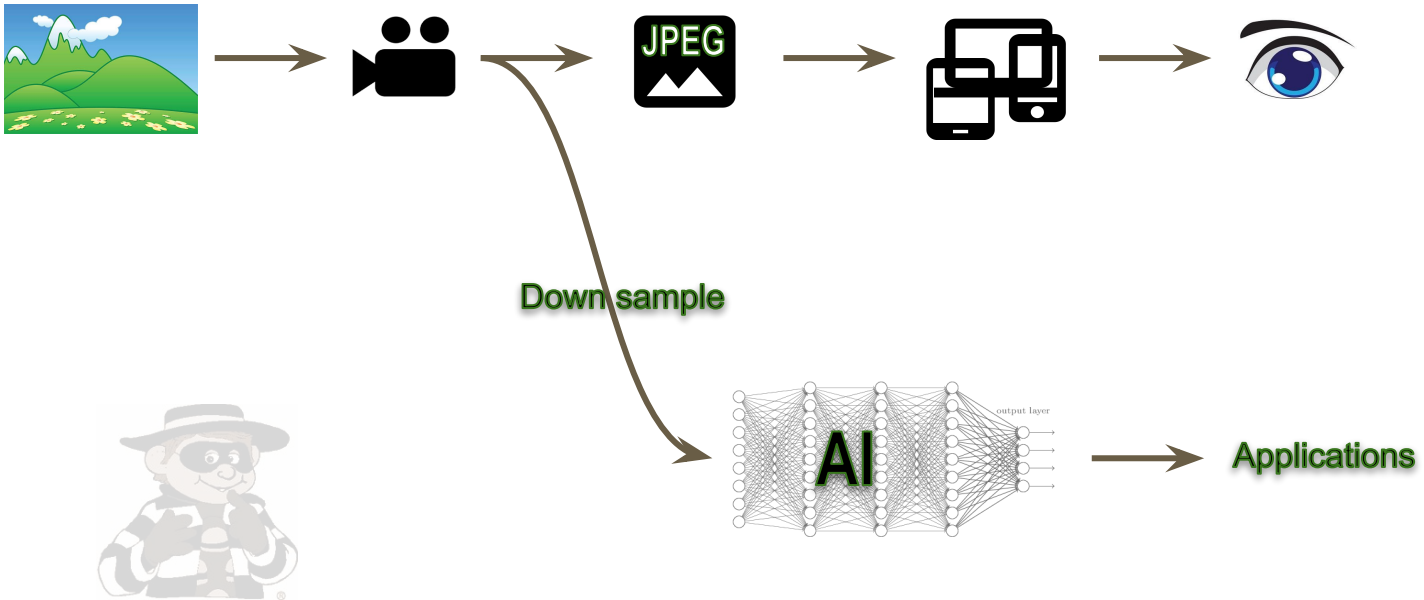
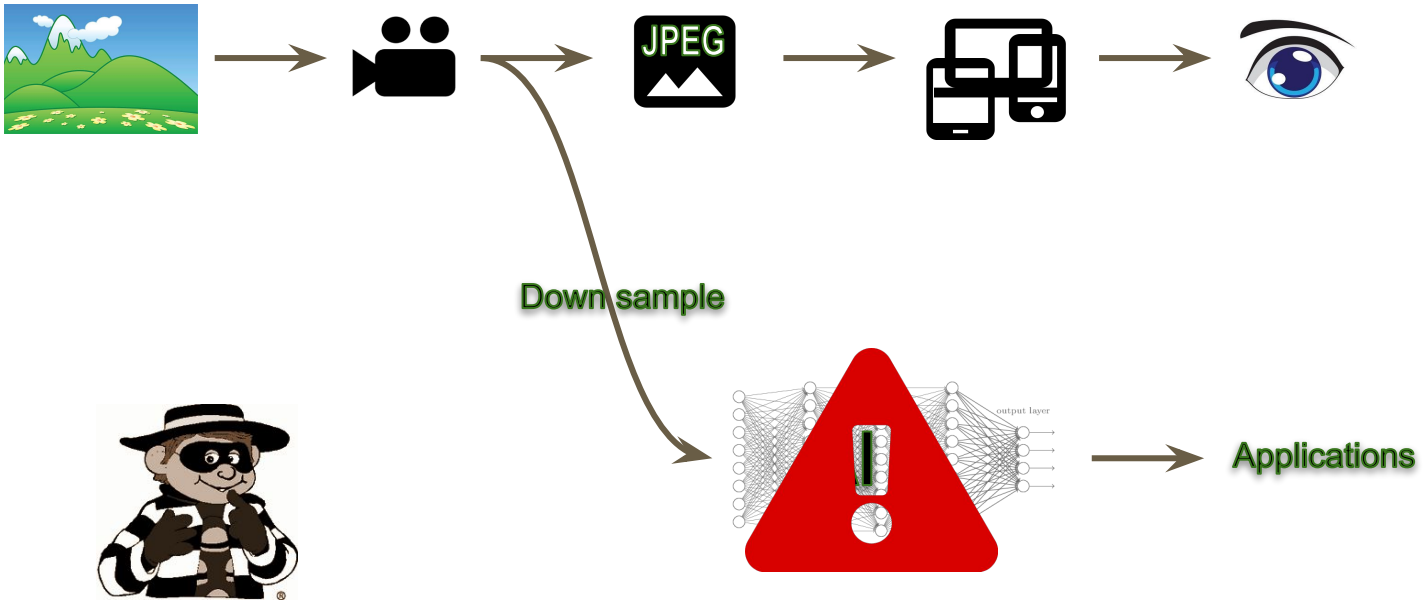
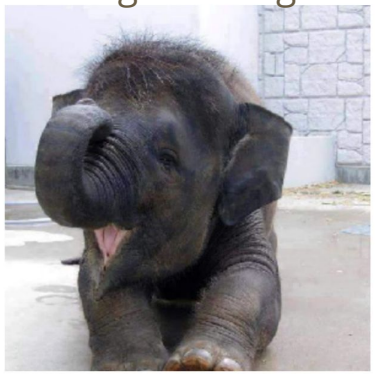


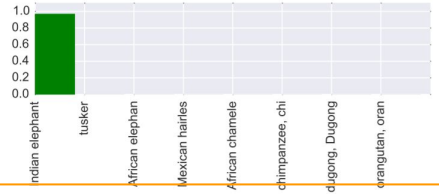
Image consumption



Adversarial Attack

Original Image

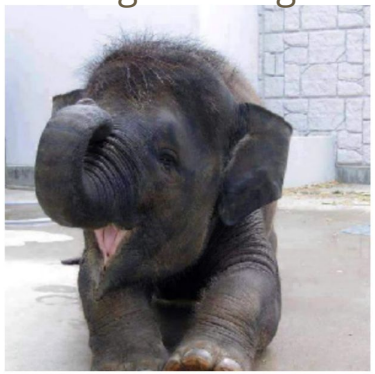




Predicted: Indian Elephant (99.7%)

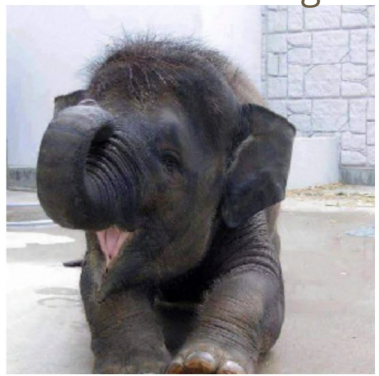
Adversarial Attack

Original Image



Predicted: Indian Elephant (99.7%)

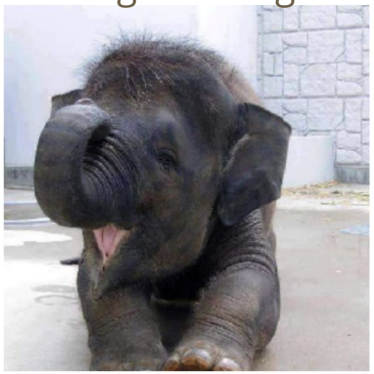
Adversarial Image

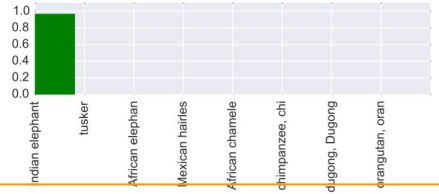


Predicted: Guacamole (99.9%)

Adversarial Attack

Original Image

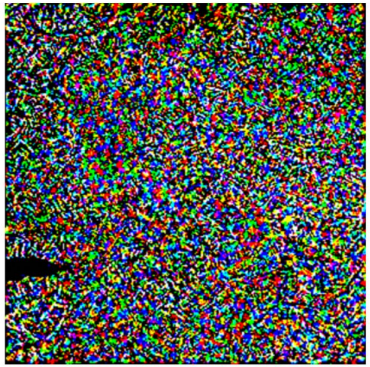


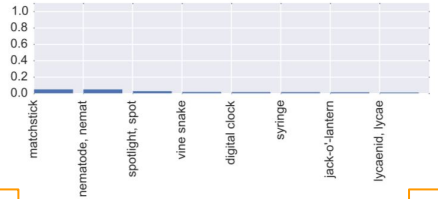


Predicted: Indian Elephant (99.7%)

$+ \epsilon$

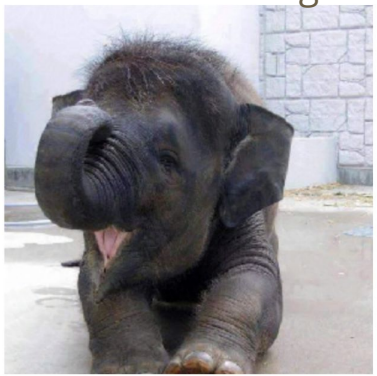
Perturbations

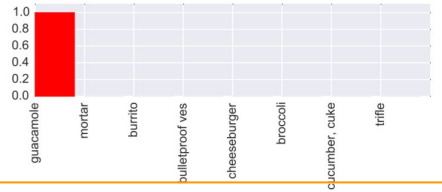




$=$

Adversarial Image

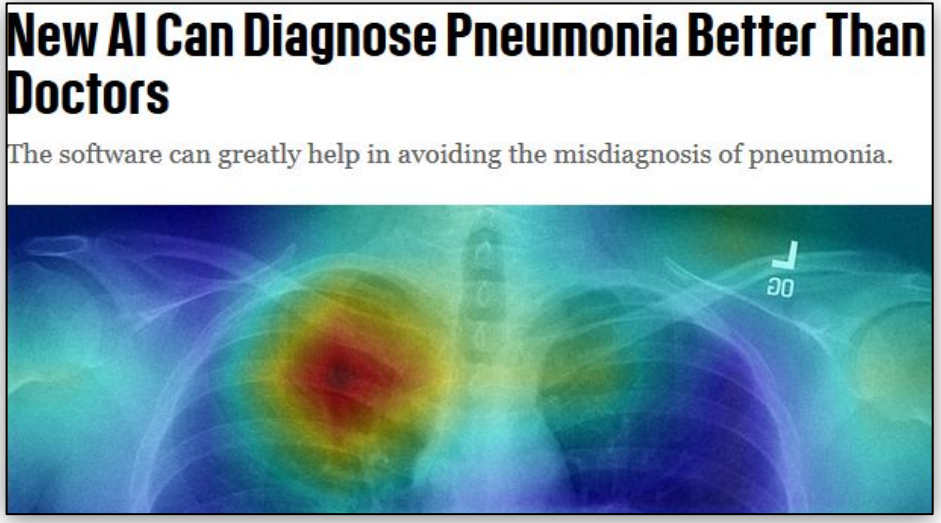


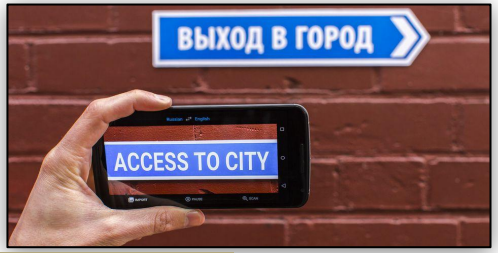


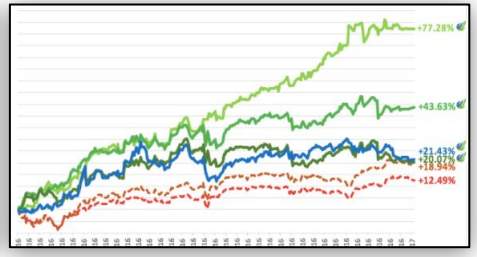
Predicted: Guacamole (99.9%)

- With cleverly designed additive noise the classification of the image can be changed
- It is easy to find noise which makes the classifier predict any given class - Targeted attack

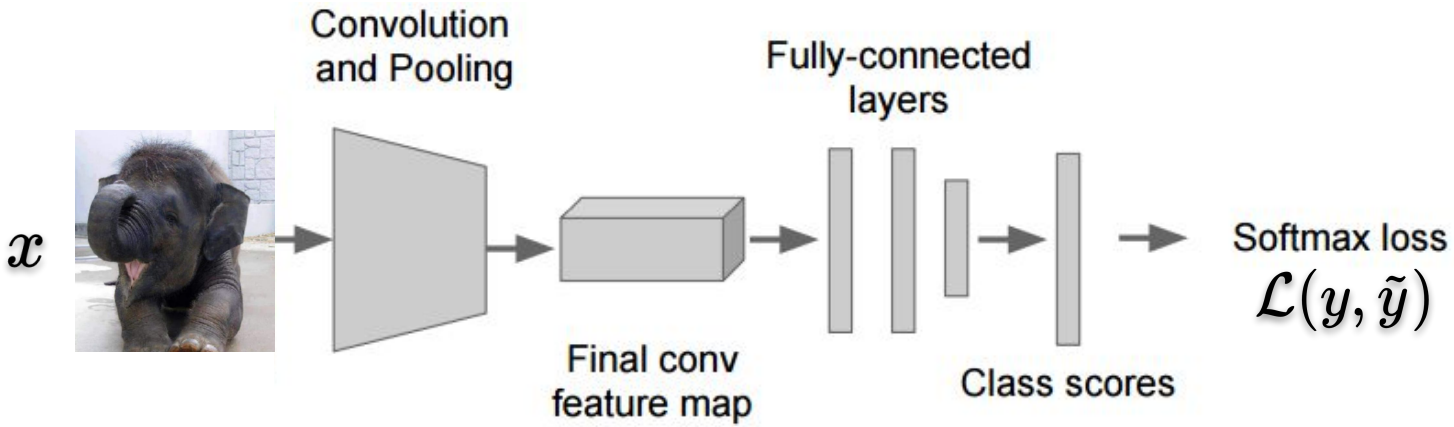
CNNs are used in critical applications



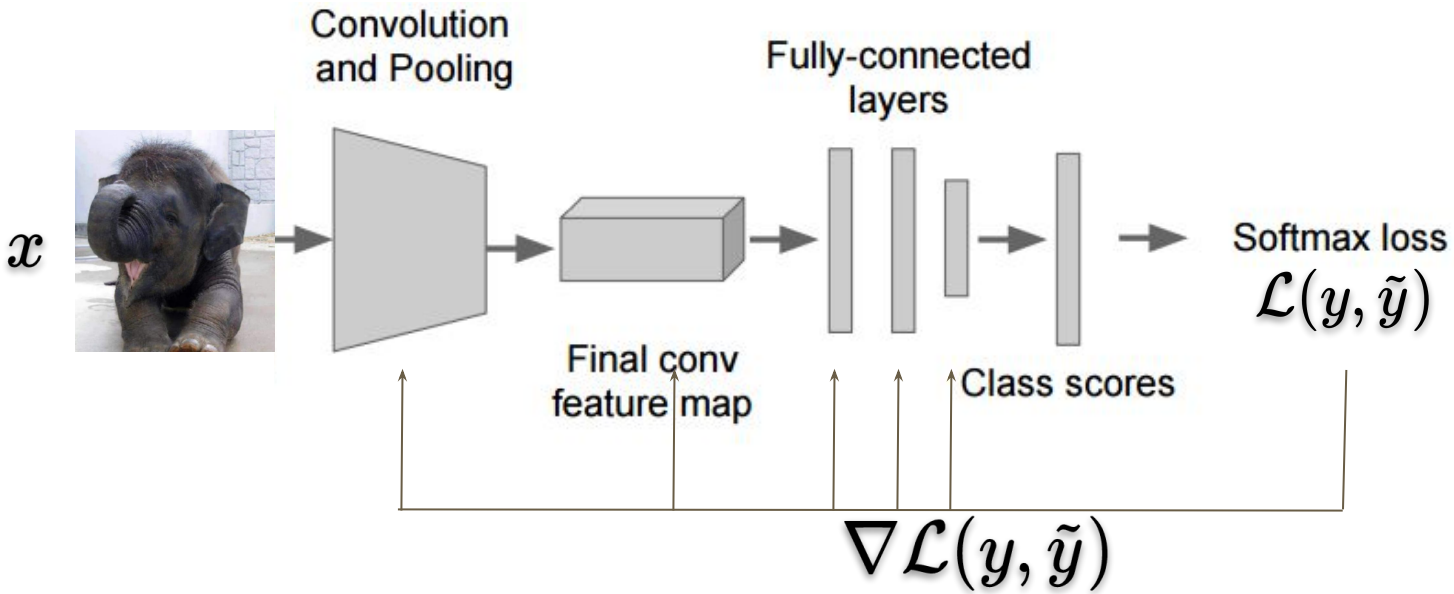




Standard Convolutional Neural Network

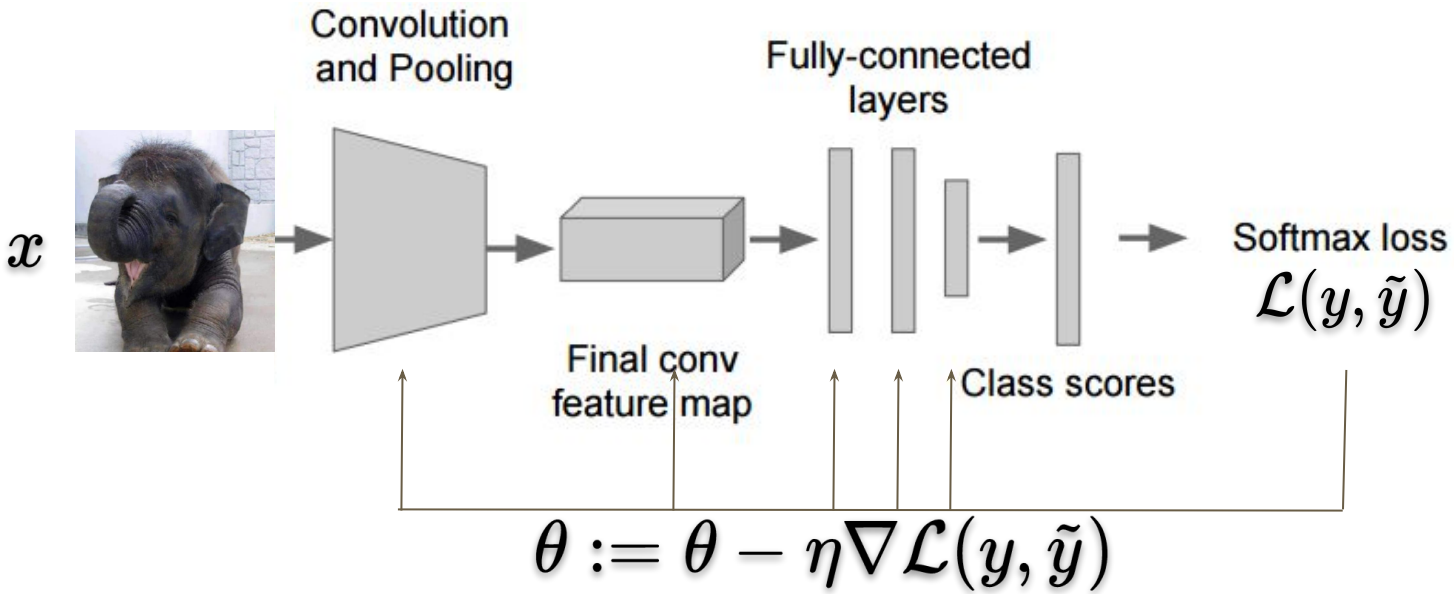


Backpropagation



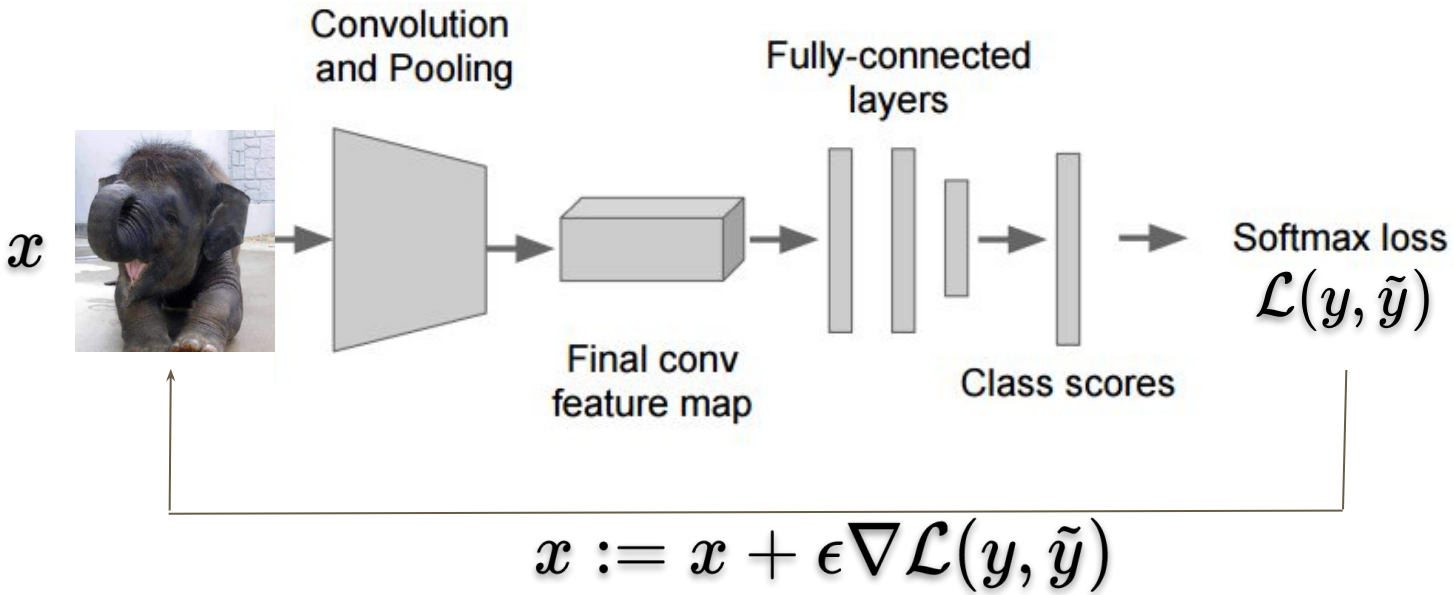
- Find the gradient (∇) with respect to the loss function, which is a measure of true vs predicted values.

Gradient Descent



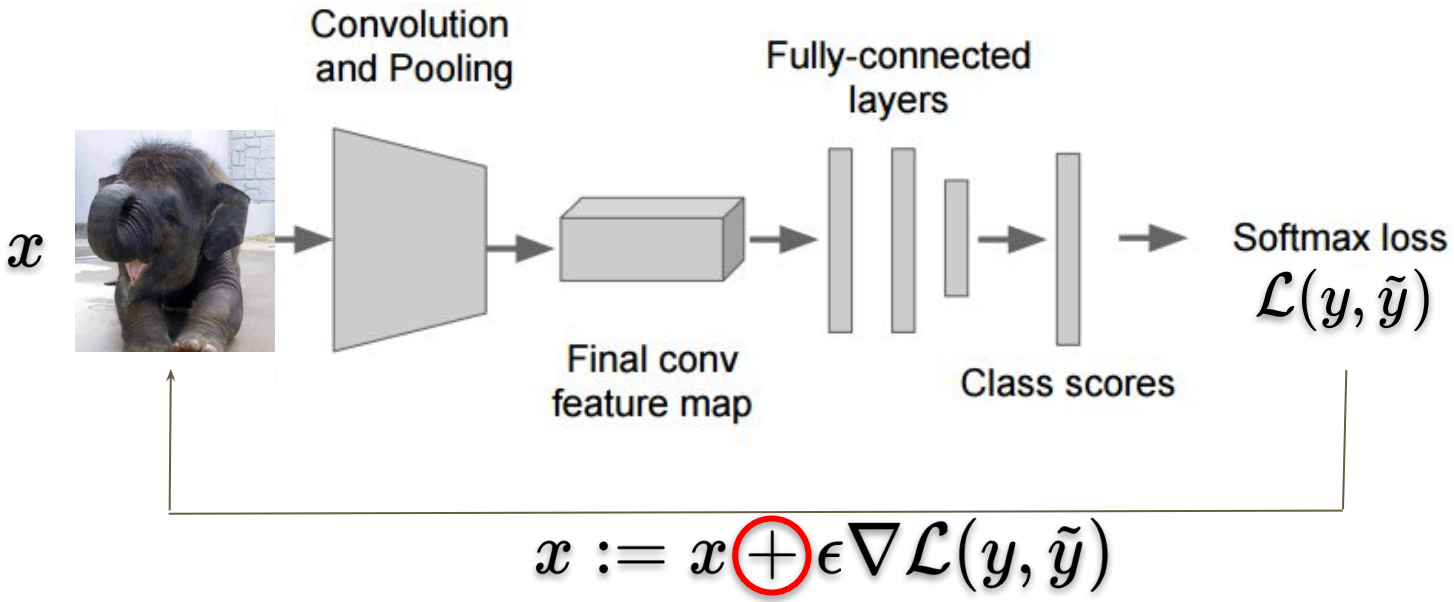
- Update the parameters with small value (η) in the direction which decreases the loss

Gradient ascent attack



- Update the image with small value (ϵ) but in the direction which increases the loss

Gradient ascent attack



- Update the image with small value (ϵ) but in the direction which increases the loss

Various attack models

- Gradient Attack

- Fast Gradient Sign Method

- Iterated Gradient Sign Method

- Deep Fool

- JSMA

- L-BFGS

(and many more since we wrote our paper)

Various types of adversarial perturbations

One of the most popular attack is - **F**ast **G**radient **S**ign **M**ethod (**FGSM**)

$$x' = x + \epsilon \times \text{sign}(\nabla \mathcal{L}(y, \tilde{y}))$$

It is very efficient and but requires higher ϵ value compared to other techniques.

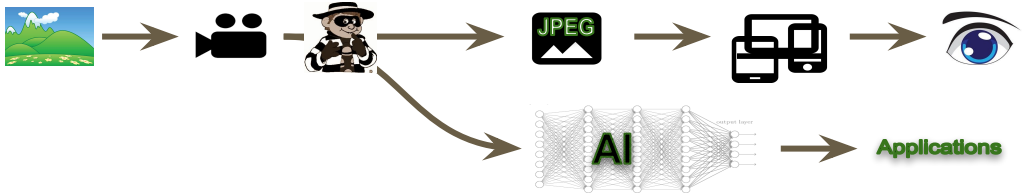
Iterative version of this method is called **I**terated **G**radient **S**ign **M**ethod (**IGSM**)

$$x'_0 = x, \qquad x'_{N+1} = \text{Clip}_{x,\epsilon}\left\{x'_N + \alpha \times \text{sign}(\nabla \mathcal{L}(y, \tilde{y}))\right\}$$

For every iteration the image is clipped to be within ±ϵ, thus the process is slow but generates more robust adversarial images.

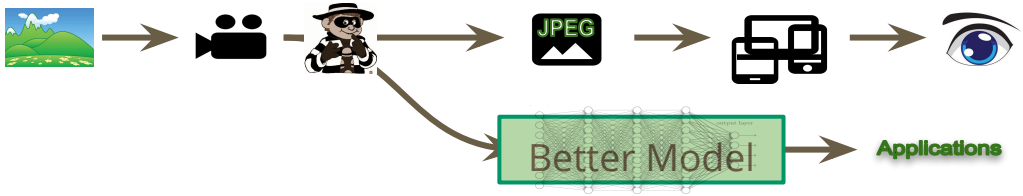
Details of other attacks like **JSMA** and **DeepFool** are included in our paper.

Defending against adversarial attacks



Defending against adversarial attacks

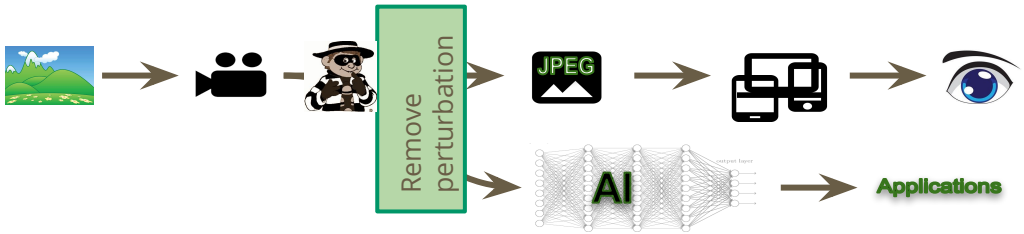
Make models harder to attack
(robust classifiers, detectors, adversarial-training)



Defending against adversarial attacks

Make models harder to attack
(robust classifiers, detectors, adversarial-training)

Remove perturbations from adversarial images



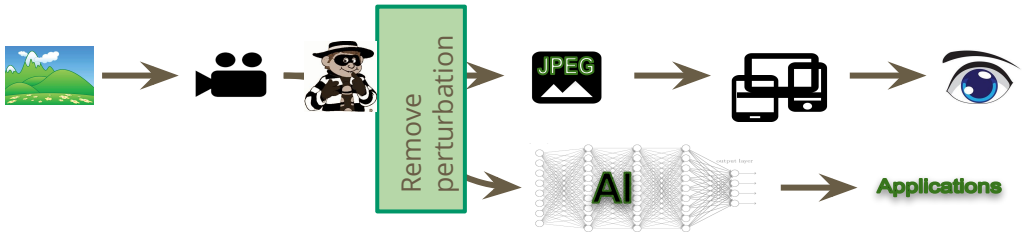
Defending against adversarial attacks

Make models harder to attack
(robust classifiers, detectors, adversarial-training)

Remove perturbations from adversarial images

Image transformations
(crop,reconstruct, pixel-deflection)

Denoising
(quantization, smoothing, shrinkage)



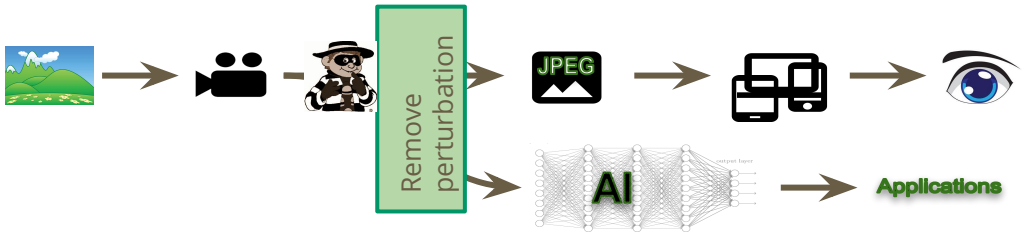
Defending against adversarial attacks

Make models harder to attack
(robust classifiers, detectors, adversarial-training)

Remove perturbations from adversarial images

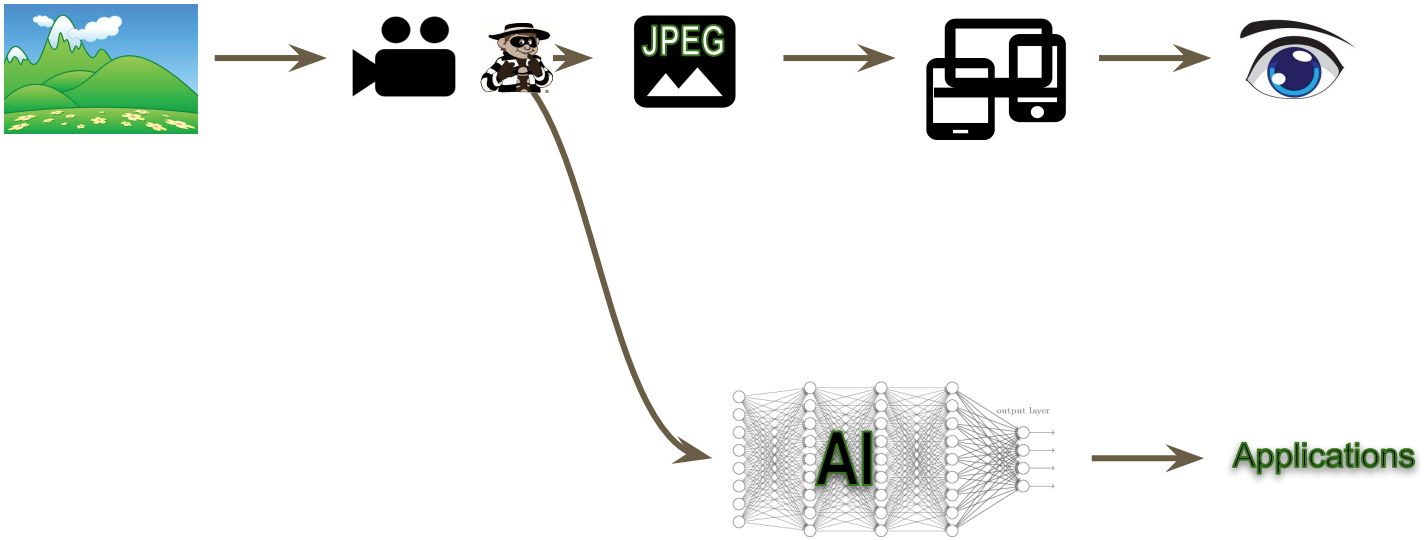
Image transformations
(crop,reconstruct, pixel-deflection)

Denoising
(quantization, smoothing, shrinkage)

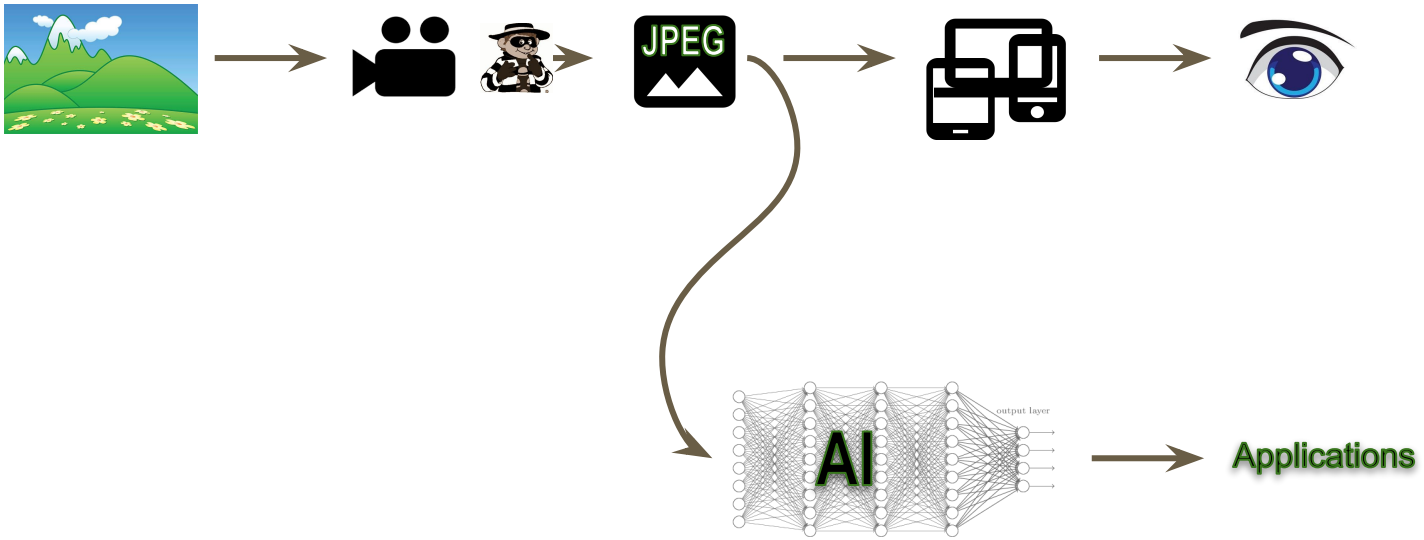


JPEG

JPEG as a possible defense



JPEG as a possible defense



JPEG as a possible defense

Keeping the Bad Guys Out:
Protecting and Vaccinating Deep Learning with JPEG Compression

Nilaksh Das[1], Madhuri Shanbhogue[1], Shang-Tse Chen[1], Fred Hohman[1], Li Chen[2], Michael E. Kounavis[2], and Duen Horng Chau[1]

[1]Georgia Institute of Technology
[2]Intel Corporation

A study of the effect of JPG compression on adversarial images

Gintare Karolina Dziugaite
Department of Engineering
University of Cambridge

Zoubin Ghahramani
Department of Engineering
University of Cambridge

Daniel M. Roy
Department of Statistical Sciences
University of Toronto

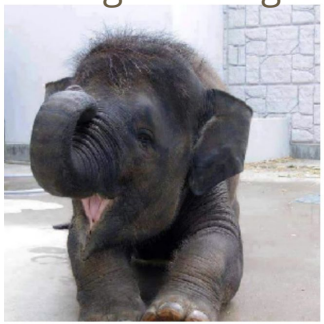
ADVERSARIAL EXAMPLES IN THE PHYSICAL WORLD

Alexey Kurakin
Google Brain
kurakin@google.com

Ian J. Goodfellow
OpenAI
ian@openai.com

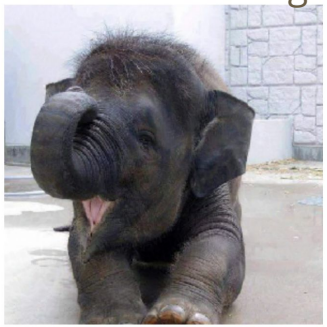
Samy Bengio
Google Brain
bengio@google.com

Original Image



Predicted: Indian Elephant (99.7%)

Adversarial Image

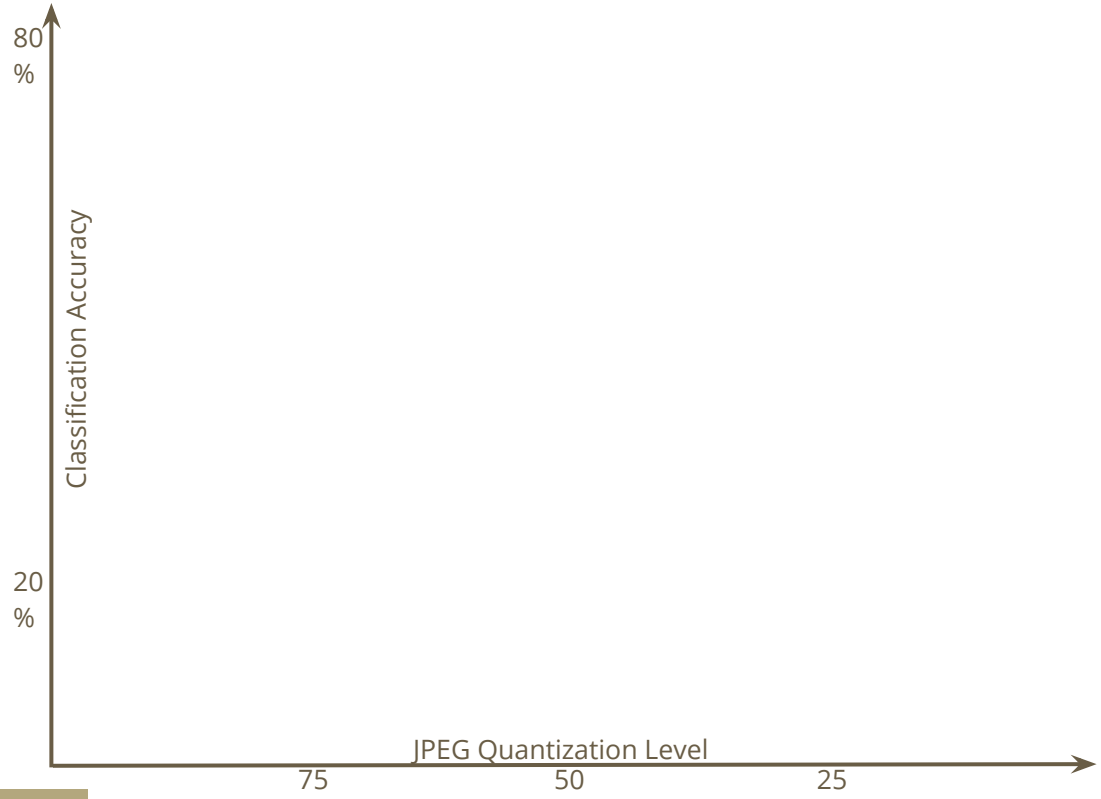


Guacamole (99.9%)

Adversarial Image after JPEG

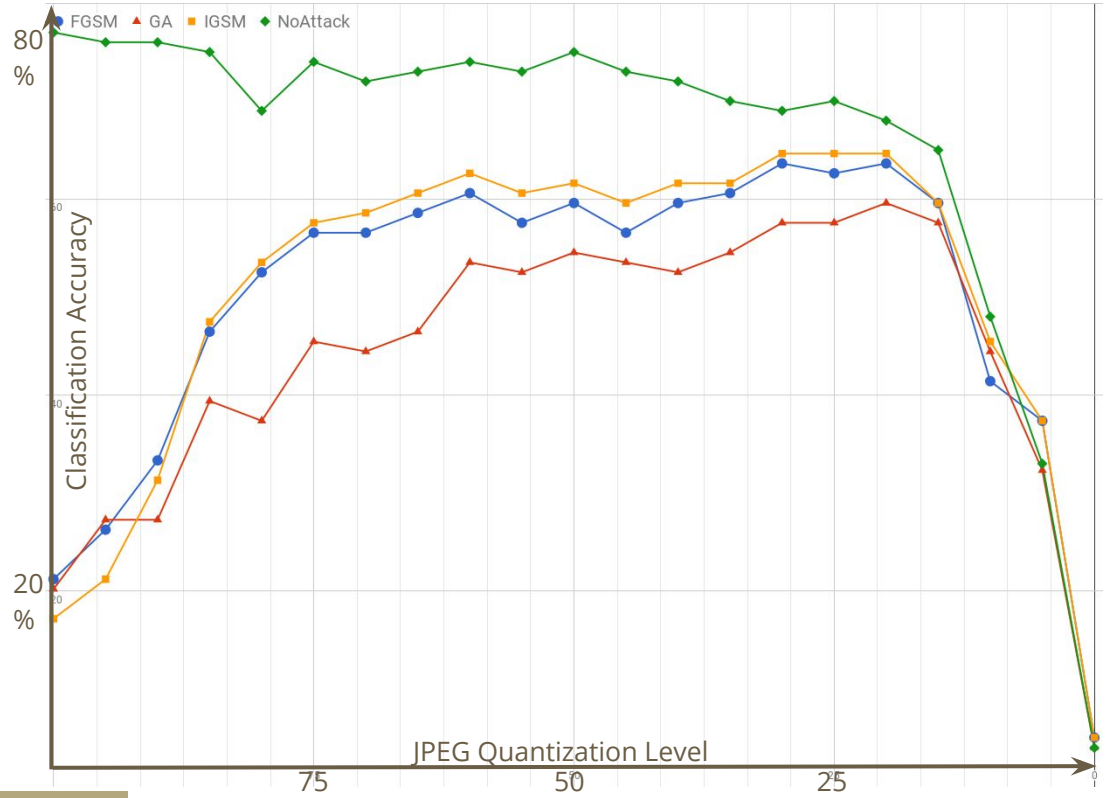
Indian Elephant (21.1%)

But JPEG has its limitations



But JPEG has its limitations

- Recovery of classification accuracy is limited.
- Negatively impacts clean images - images which have not been perturbed.
- Optimum quantization value is not universal across attacks.
- Recovery is worse on stronger attacks.
- Works well with small perturbations only.



JPEG is not the best defense

PIXELDEFEND: LEVERAGING GENERATIVE MODELS TO UNDERSTAND AND DEFEND AGAINST ADVERSARIAL EXAMPLES

Yang Song
Stanford University
yangsong@cs.stanford.edu

Taesup Kim
Université de Montréal
taesup.kim@umontreal.ca

Sebastian Nowozin
Microsoft Research
nowozin@microsoft.com

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Nate Kushman
Microsoft Research
nkushman@microsoft.com

MITIGATING ADVERSARIAL EFFECTS THROUGH RANDOMIZATION

Cihang Xie, Zhishuai Zhang & Alan L. Yuille
Department of Computer Science
The Johns Hopkins University
Baltimore, MD 21218 USA
{cihangxie306, zhshuai.zhang, alan.l.yuille}@gmail.com

Jianyu Wang
Baidu Research USA
Sunnyvale, CA 94089 USA
wjyouch@gmail.com

Zhou Ren
Snap Inc.
Venice, CA 90291 USA
pchat.com

Better than JPEG

COUNTERING ADVERSARIAL IMAGES USING INPUT TRANSFORMATIONS

Chuan Guo*
Cornell University

Mayank Rana & Moustapha Cissé & Laurens van der Maaten
Facebook AI Research

Deflecting Adversarial Attacks with Pixel Deflection

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, James Storer
Brandeis University
{aprakash,nemtiax,solomongarber,dilant,storer}@brandeis.edu

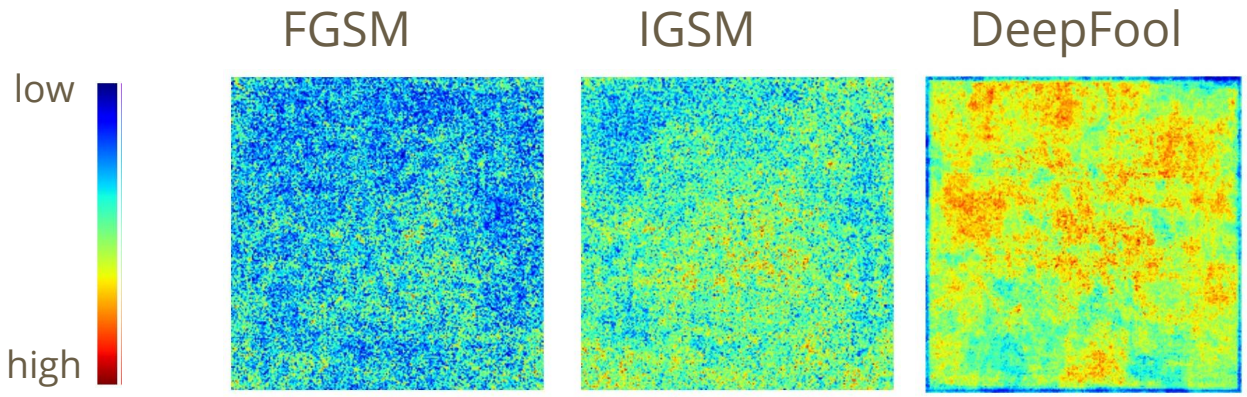
Why do we care about JPEG?

- Fast decoding available in virtually all devices

- Deep Learning datasets already store images as JPEG

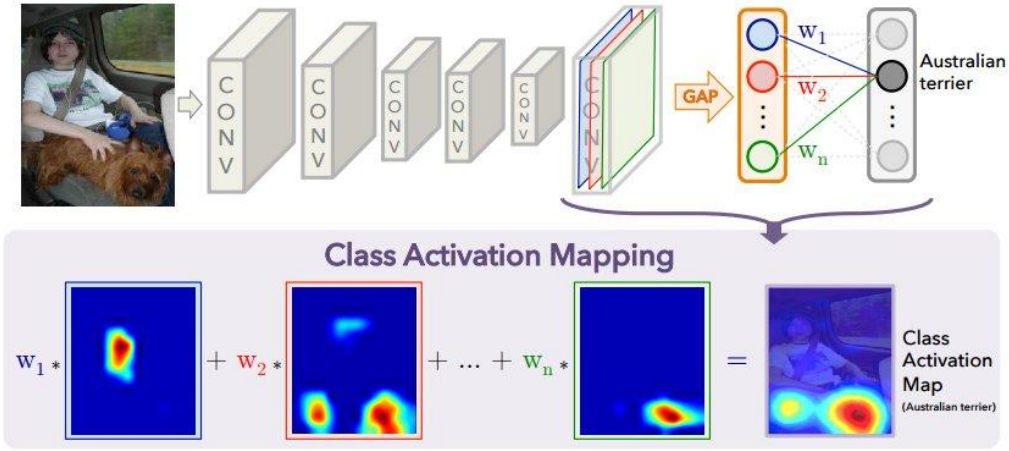
- JPEG defenses provide adequate visual quality

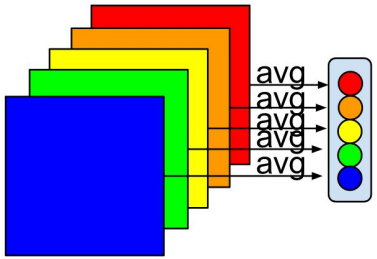
Attacks are agnostic to object location



Average location of adversarial perturbations

Object localization

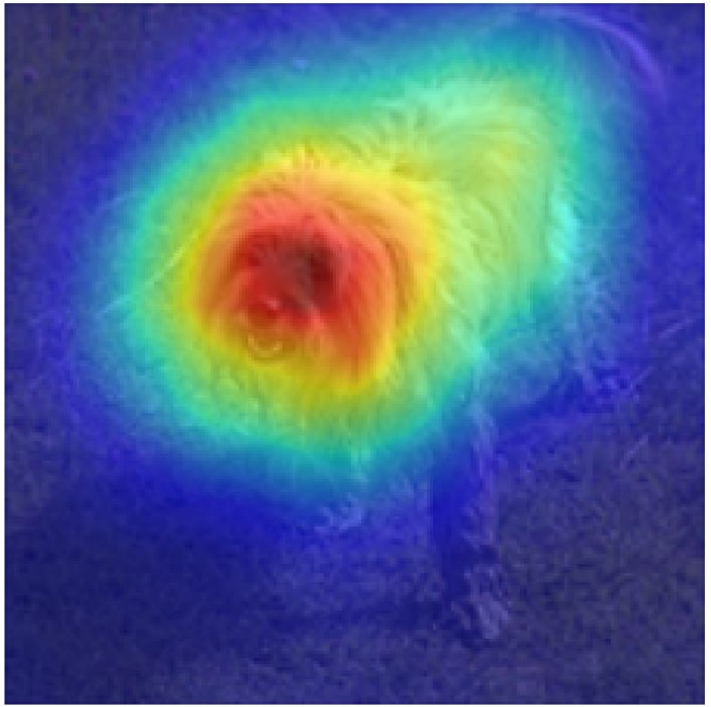


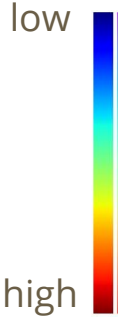


GAP
Global Activation Pooling

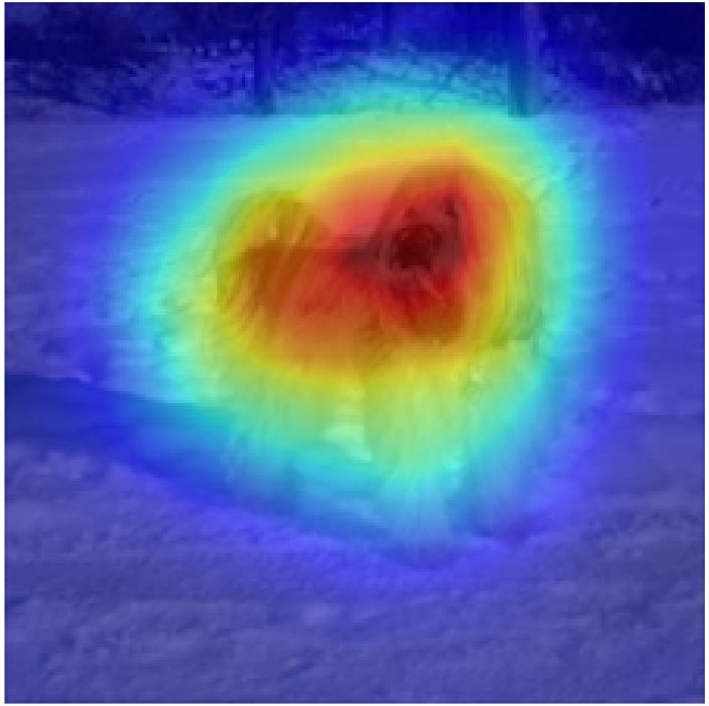
- Class activation map is the obtained by taking the output of GAP and learning weights that maximize the discriminative activations for a given class.
- Widely used to for discriminative localization
- Is good only for one object - the class label of the image

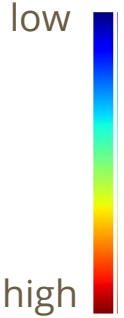
Semantic contents are localized



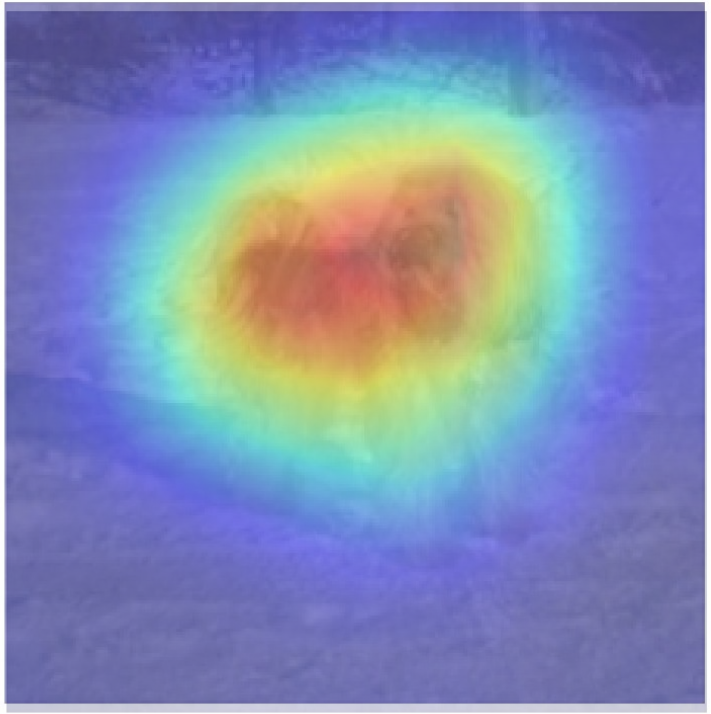


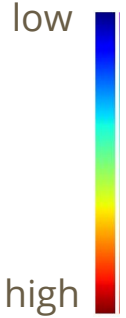
Semantic contents are localized



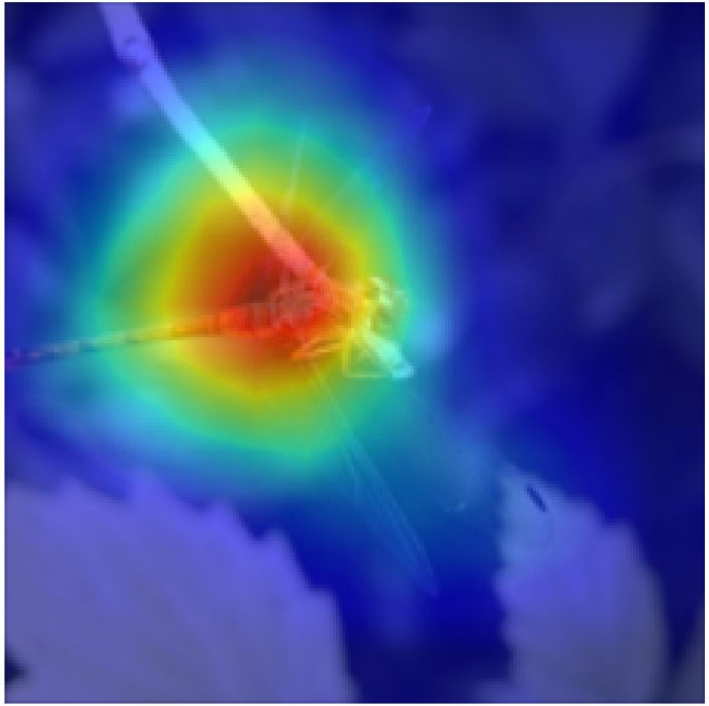


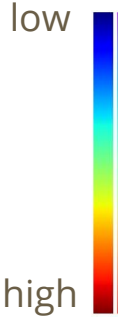
Semantic contents are localized



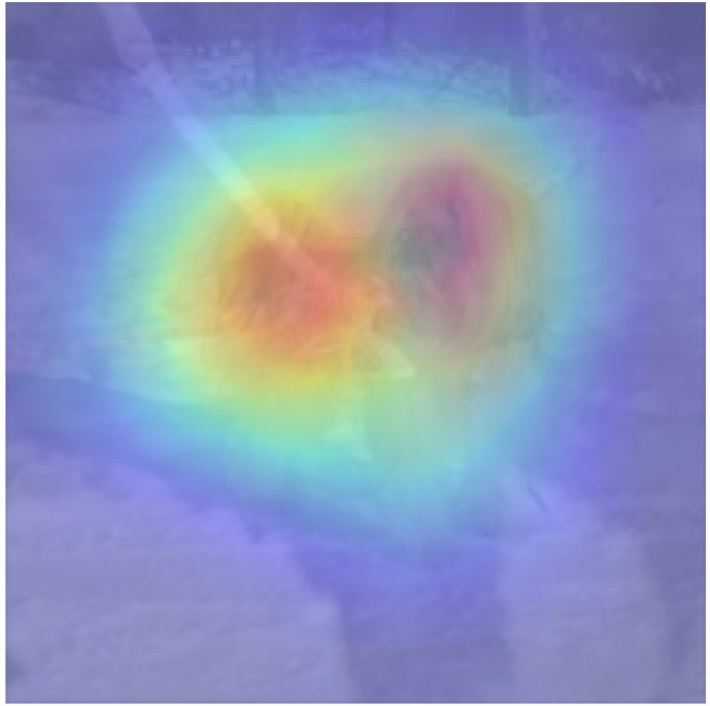


Semantic contents are localized





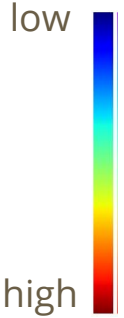
Semantic contents are localized



low

high

Semantic contents are localized



Semantic contents are localized

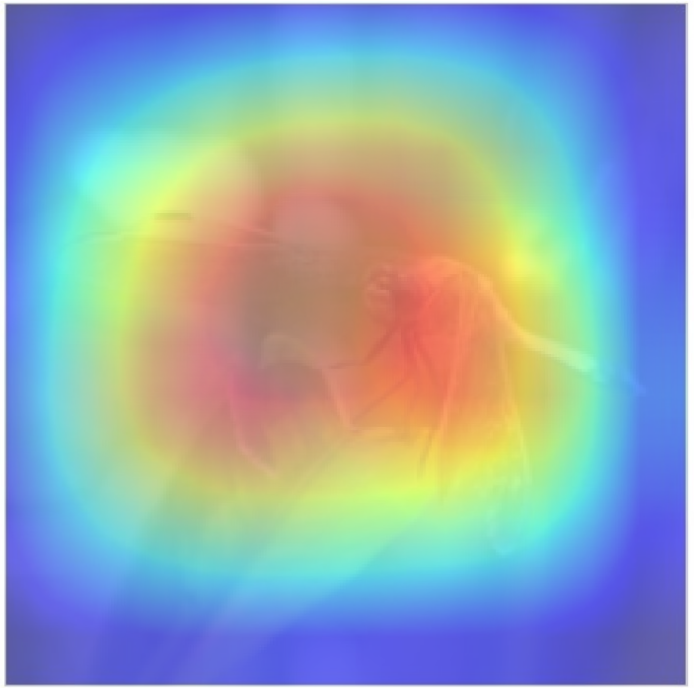


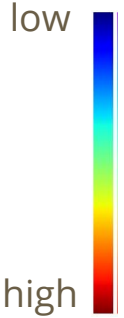


low

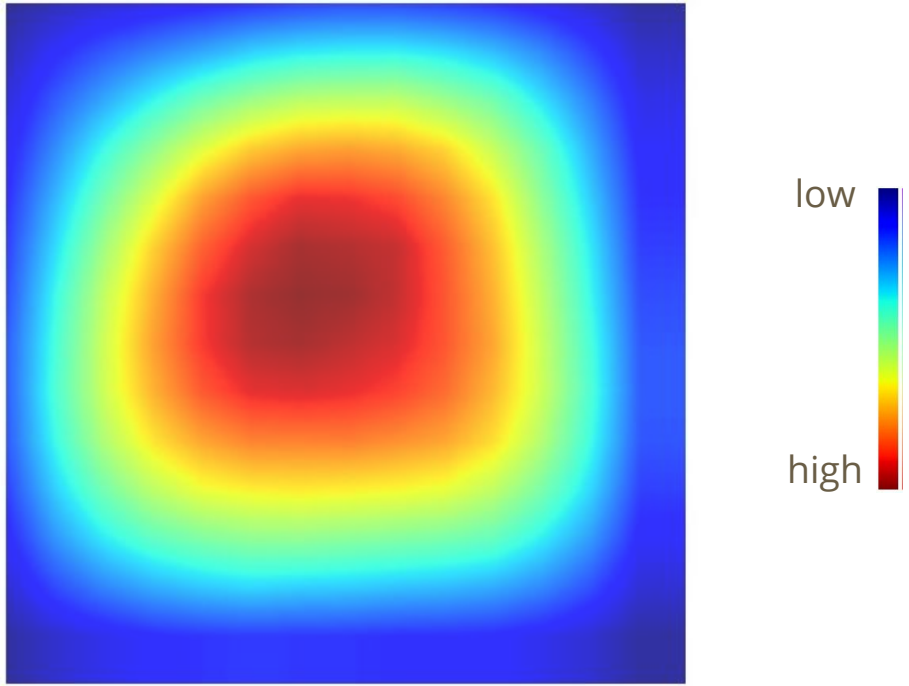
high

Semantic contents are localized



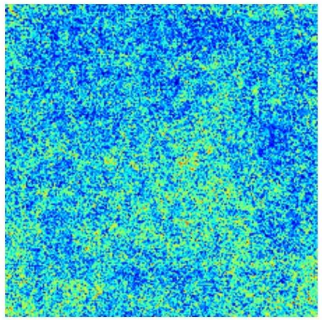


Semantic contents are localized

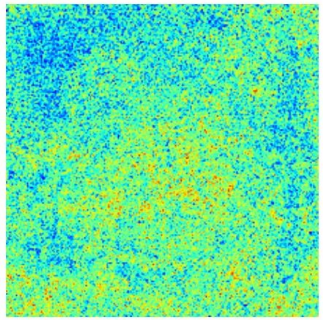


Where are the attacks located?

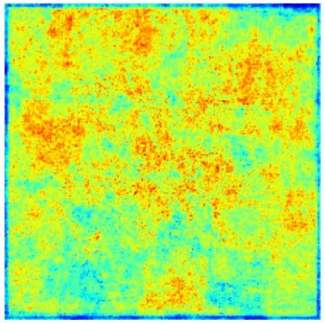
FGSM



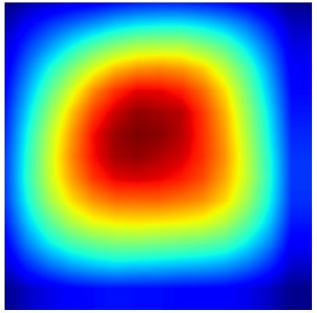
IGSM

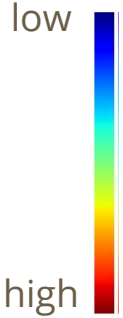


DeepFool

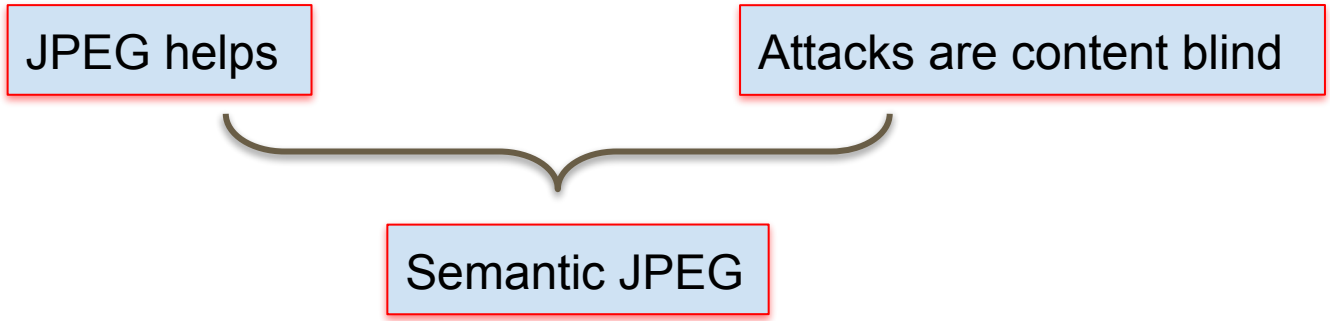


Natural Images



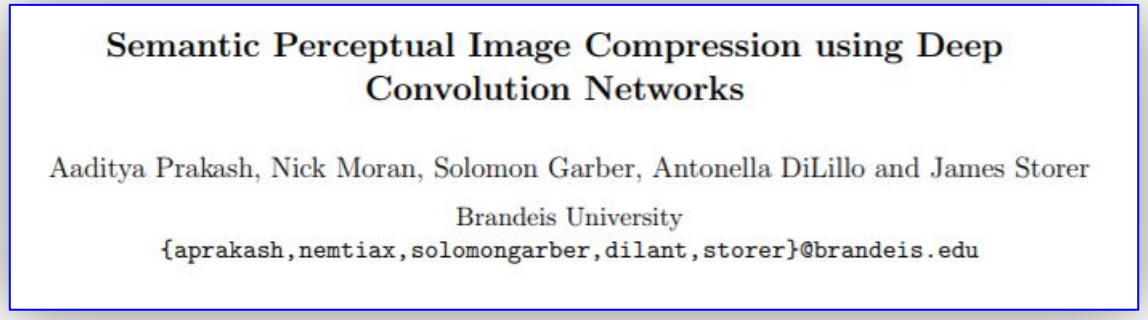


Key to improving JPEG's ability to defend



Key to improving JPEG's ability to defend

JPEG with semantic quantization - DCC 2017



Semantic Perceptual Image Compression using Deep Convolution Networks

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo and James Storer

Brandeis University

{aprakash,nemtiax,solomongarber,dilant,storer}@brandeis.edu

Soon to be available in Firefox browser

MSROI - DCC 2017

(a) Original (b) Human Fixation (c) CAM (d) MS-ROI

Boy's face and hand (only captured by our method)

CAM

$$M_c(x,y) = \sum_{d \,\epsilon\, \mathbf{D}} w_d^c \, f_d(x,y)$$

where w_d^c is learned for every class c and for layer 'd'

MSROI Map

$$Z_l^c = \sum_{d \,\epsilon\, \mathbf{D}} \sum_{x,y} f_d^c(x,y)$$

$$\widehat{M}(x,y) = \sum_{c \,\epsilon\, \mathbf{c}} \begin{cases} \sum_d f_d^c(x,y), & \text{if } Z_l^c > T \\ 0 & \text{otherwise} \end{cases}$$

Advantage

- MSROI - extracts all salient objects in the given image
- Thus, more useful for image compression

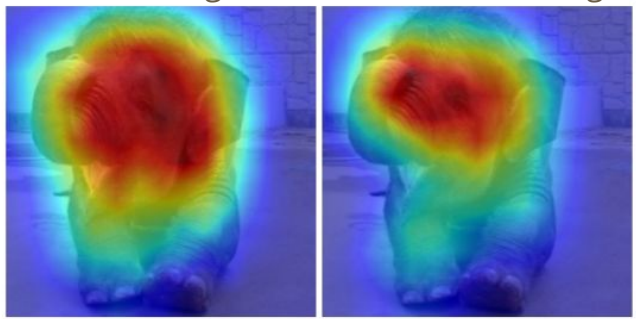
Disadvantage

- Uses classes of the object to get the map
- Thus susceptible to adversarial attack

Making MSROI secure against adversary

MSROI on Clean Image

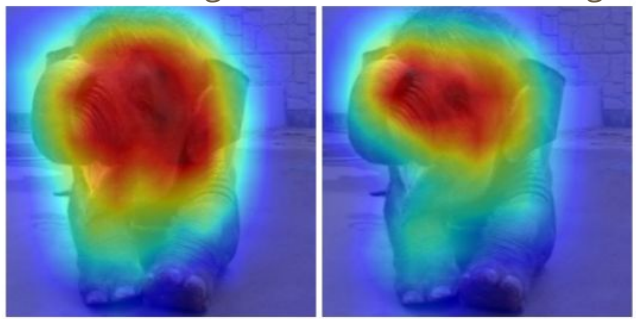
MSROI on Adversarial Image

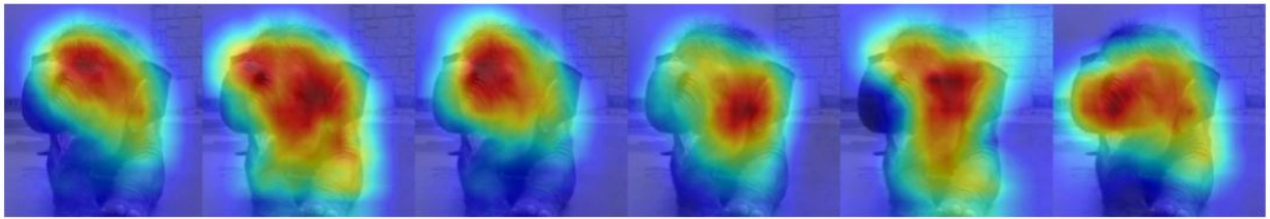


Making MSROI secure against adversary

MSROI on Clean Image

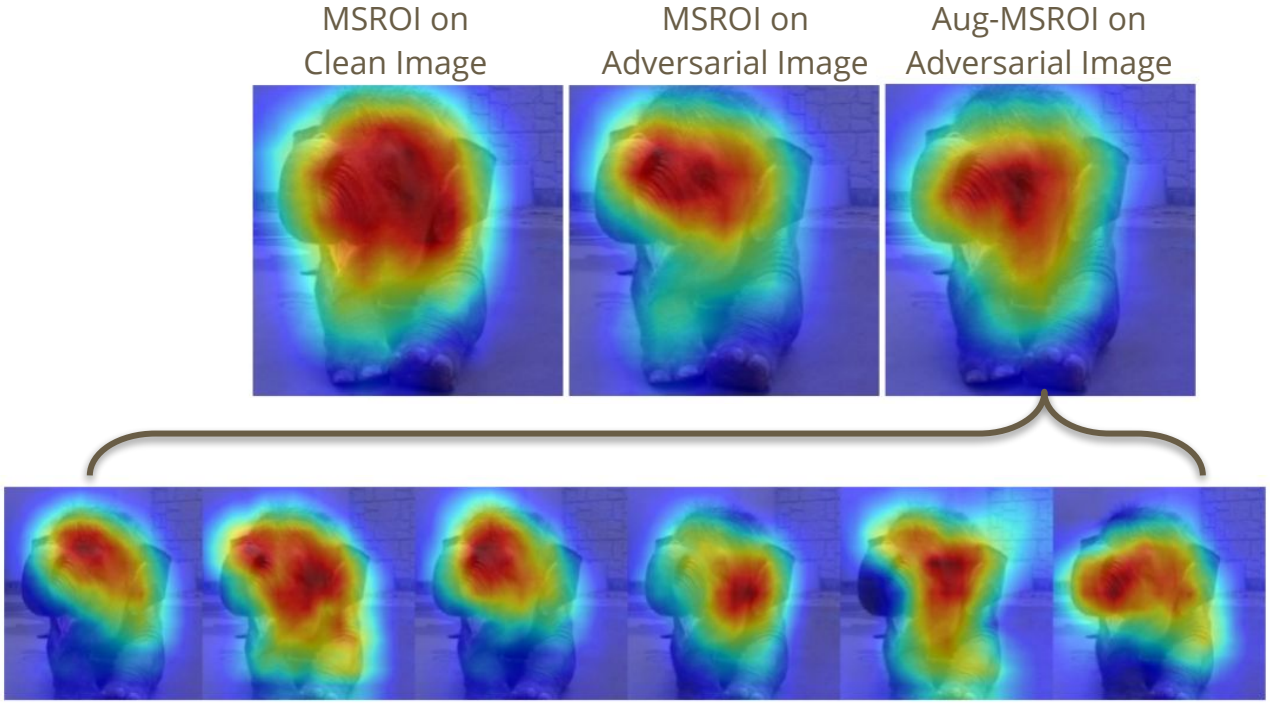
MSROI on Adversarial Image





Various perturbations with Augmented MSROI

Making MSROI secure against adversary



Various perturbations with Augmented MSROI

Making MSROI secure against adversary

MSROI

$$M(i,j) = \sum_{c \in C} \begin{cases} \sum_l f_l^c(i,j) & \text{if } \sum_l \sum_{x,y} f_d^c(i,j) > \mathcal{T} \\ 0 & \text{otherwise} \end{cases}$$

Augmented MSROI

$$\widetilde{M}(i,j) = \sum_{c \in C} \sum_l (a_l^c(i,j)) \text{and } a_{l+1}^c = f_l(a_l^c(i,j) + \Delta)$$

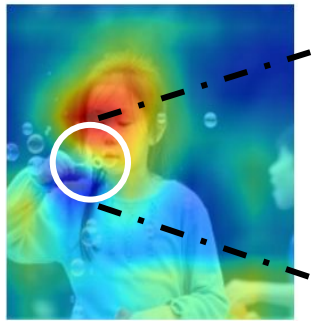
where, Δ is random perturbation

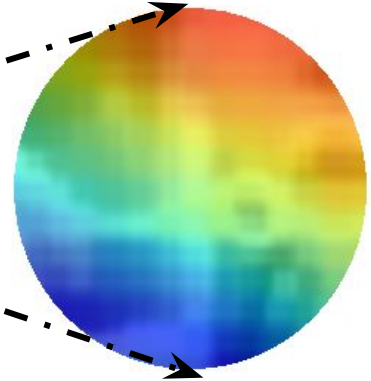
- Aug MSROI is average of maps over several perturbations.

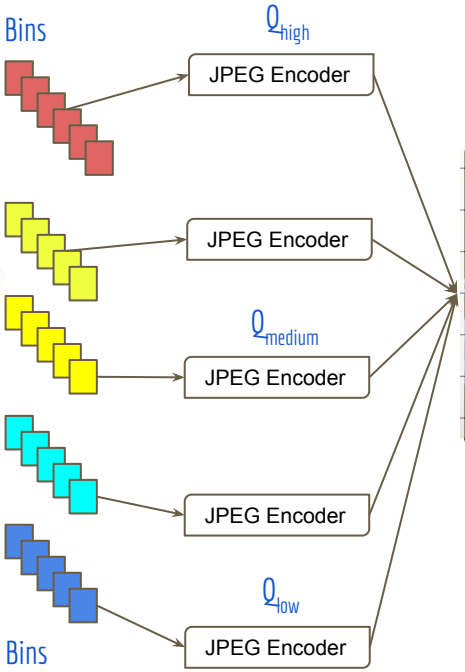
- Random perturbation helps overcome changes in the activation due to adversarial input

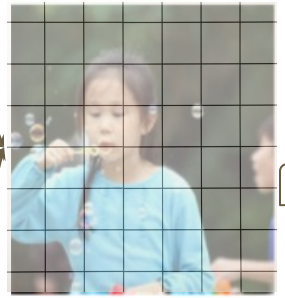
- Classifiers are robust enough that small perturbations do not change the overall class of the image

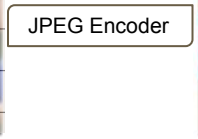
Variable ‘Q’ JPEG





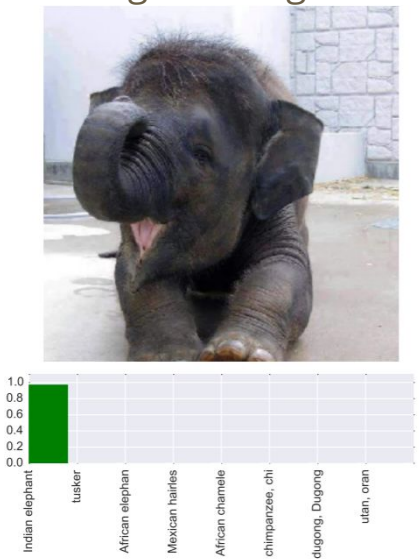






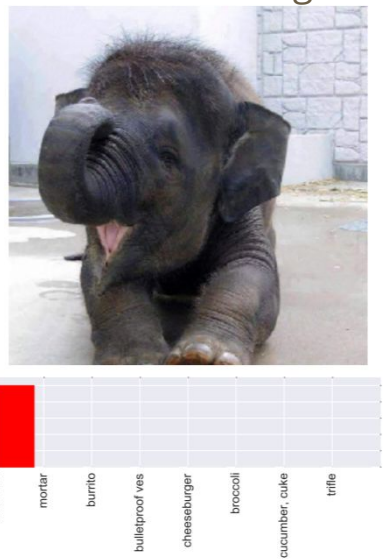
Aug MSROI - Accuracy

Original Image



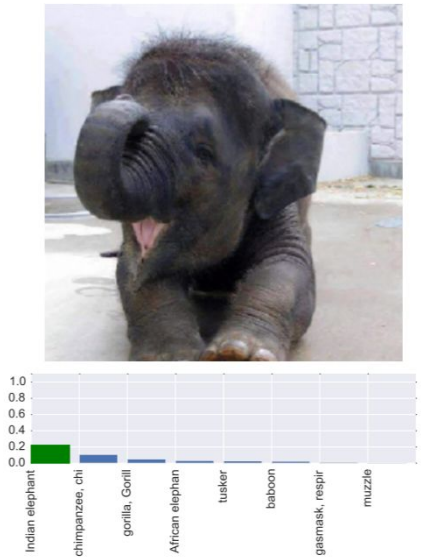
Predicted: Indian Elephant (99.7%)

Adversarial Image



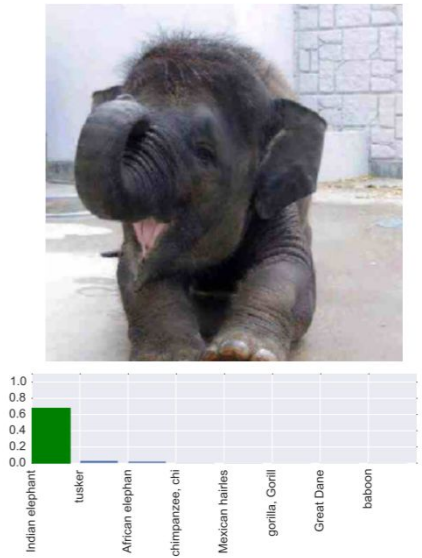
Guacamole (99.9%)

Adversarial Image after JPEG



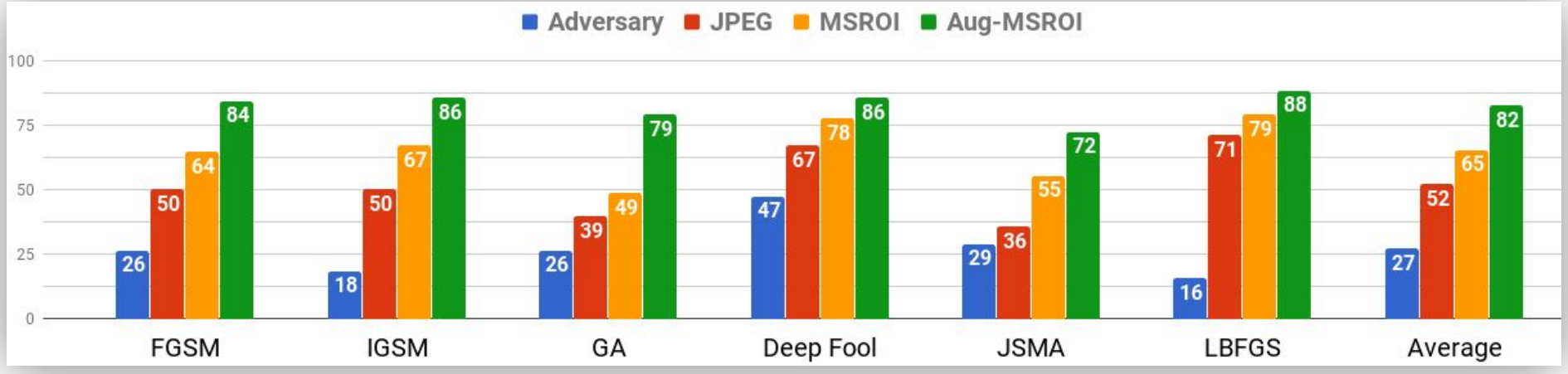
Indian Elephant (21.1%)

Adversarial Image after Aug-MSROI

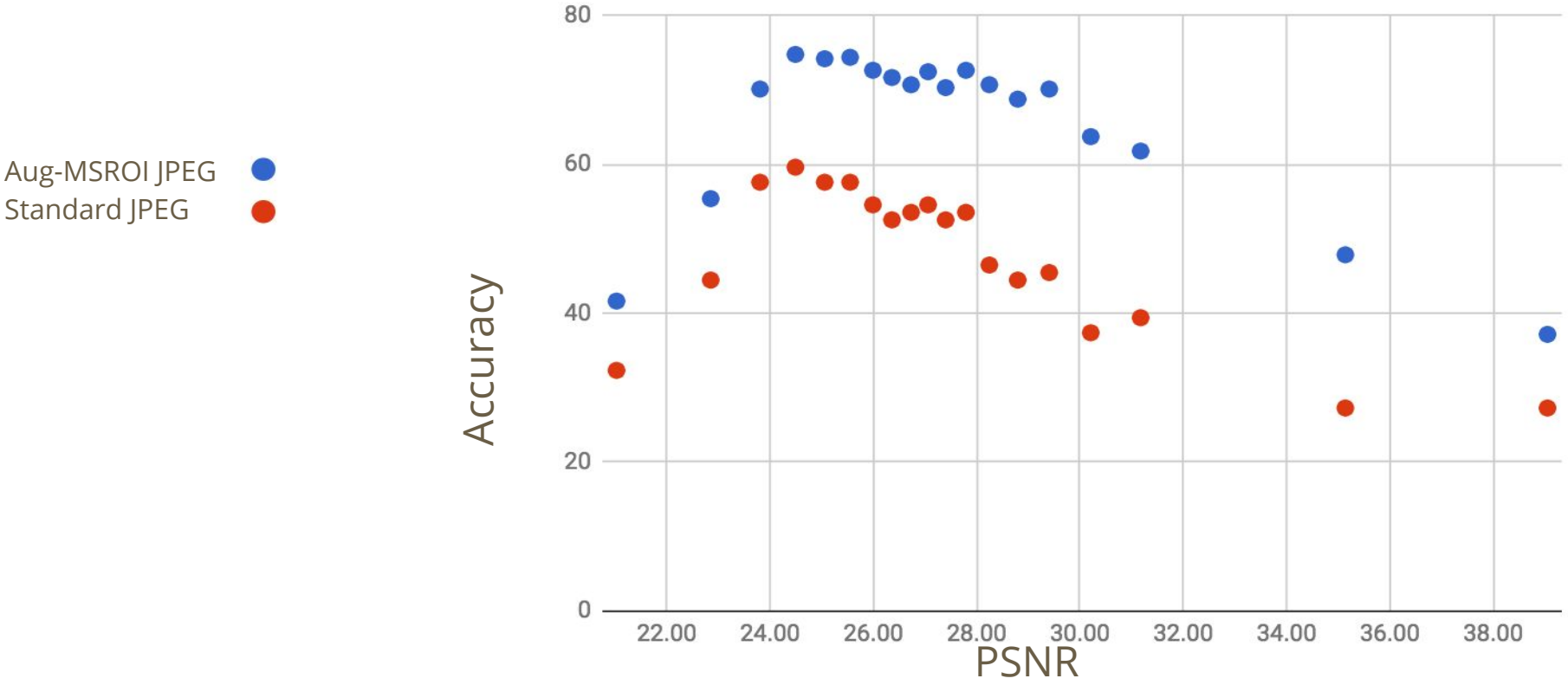


Indian Elephant (70.3%)

Aug MSROI - Accuracy



Aug MSROI - Image quality vs Accuracy



Summary

Code: github.com/iamaaditya/protecting-jpeg

- JPEG can provide some level of protection from small perturbations
- Semantic JPEG can do better than standard JPEG
  but using MSROI to get semantic JPEG is also susceptible to adversary
- Aug-MSROI leverages the localization capabilities of standard MSROI for robust localization in the presence of adversarial attacks
- Final image is decodable using any off-the-shelf JPEG decoder

Thank You

Contact: aprakash@brandeis.edu