Adversarial Attacks on Face Detectors using Neural Net based Constrained Optimization

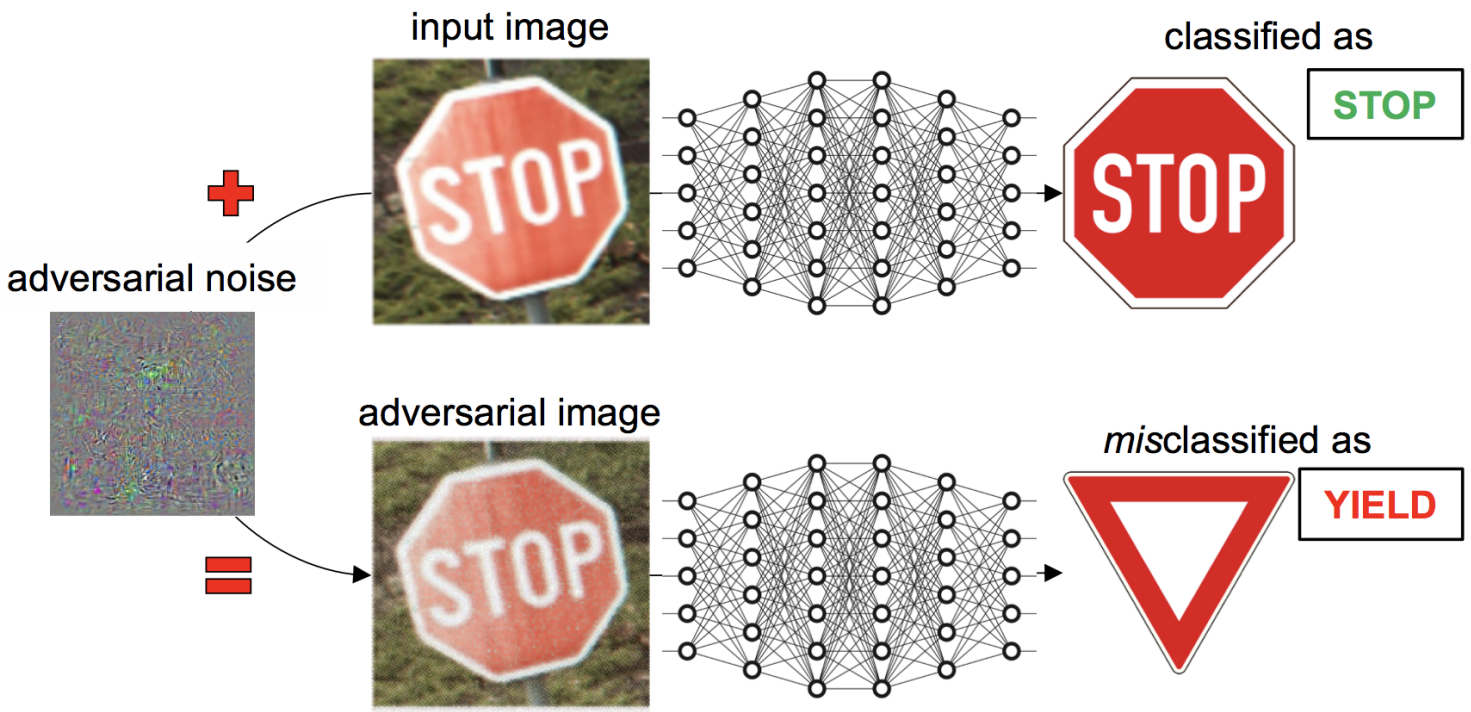
Joey Bose

University of Toronto
joey.bose@mail.utoronto.ca

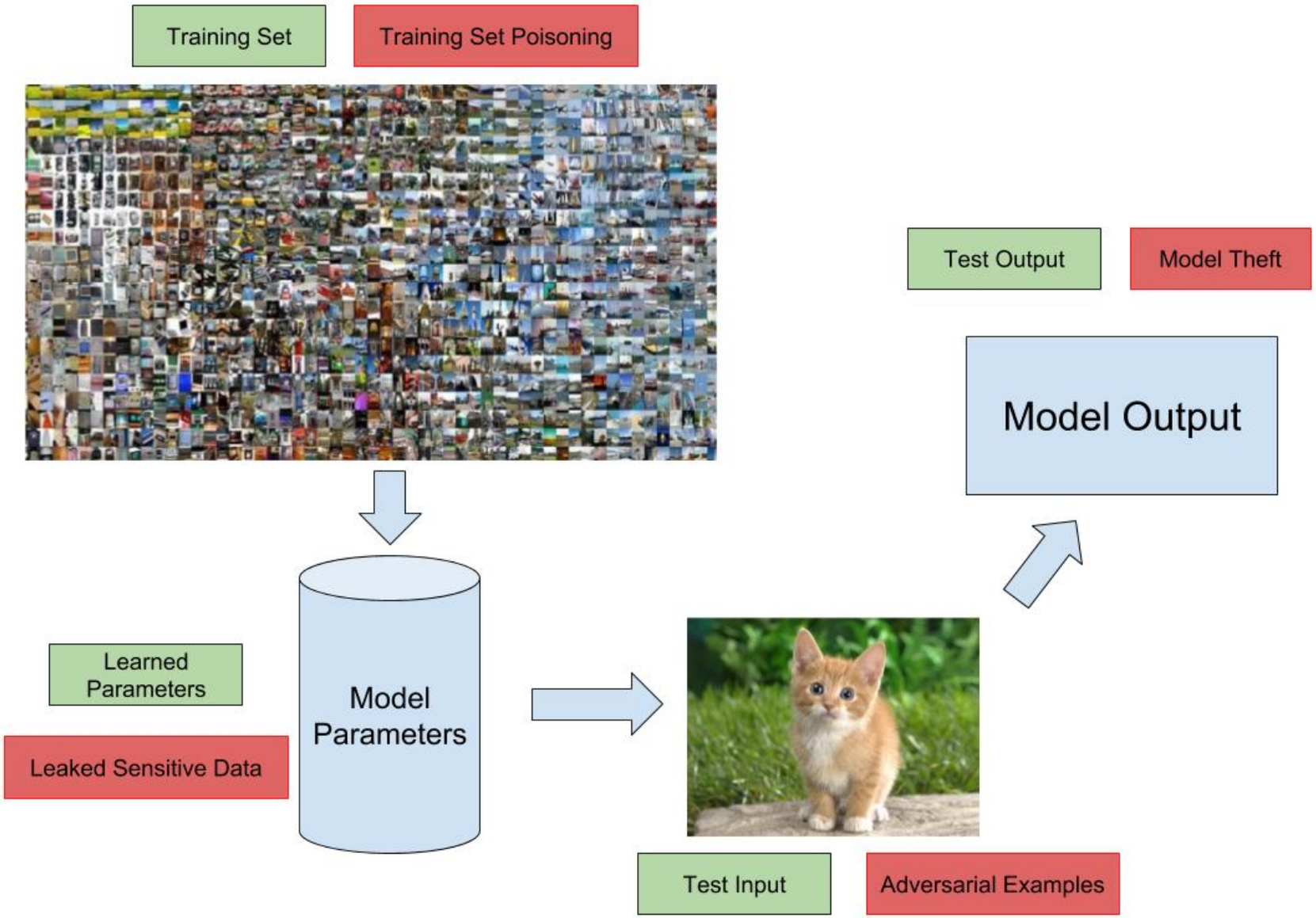
September 26, 2018

Motivation

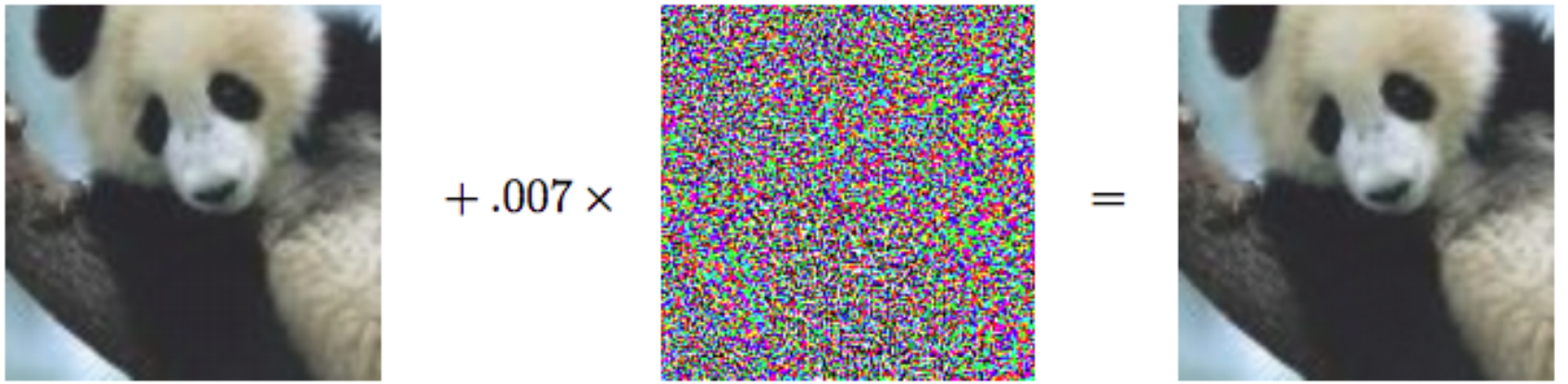
- Machine Learning models are Ubiquitous
- Generalization behavior of Deep Neural Nets is still very poorly understood
- Attacking models reveals weaknesses and drives research towards Robust Models



Attacking the Machine Learning Pipeline



Adversarial Attacks - Basic Phenomena



$$\text{minimize } L(x, x+\delta)$$
$$\text{s.t. } D(x+\delta) = t'$$
$$x + \delta \in [0,1]^n$$

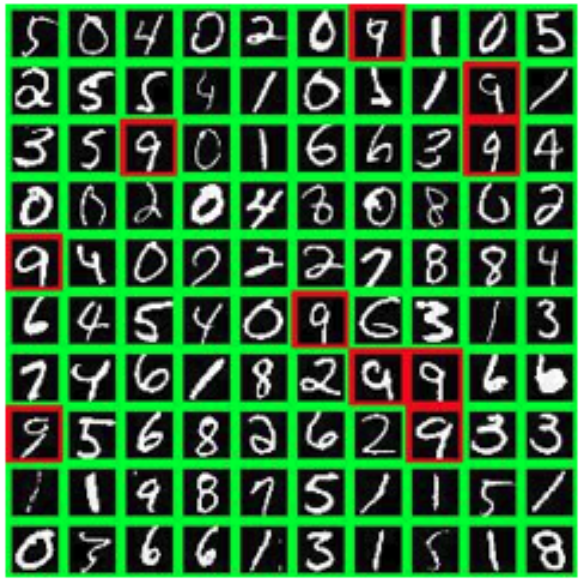
Early Attacks - FGSM (Goodfellow et al. 2014)

Given an image x, the Fast Gradient Sign Method (FGSM) returns a perturbed input x':

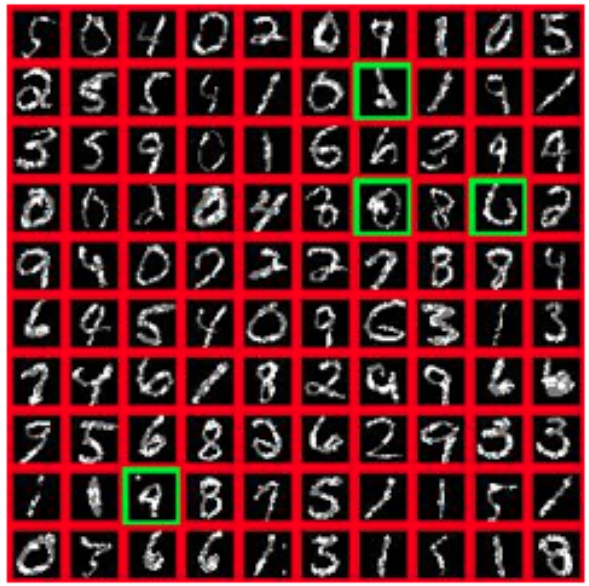
$$x' = x - \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

where J is the loss function for the attacked classifier and ϵ controls the extent of the perturbation.

FGSM on MNIST



Basic Iterative Method on MNIST (Kurakin et. al 2016)



Carlini-Wagner (Carlini and Wagner 2016)

Find some small δ such that $D(x+\delta) = t'$,

$$\begin{aligned} &\text{argmin}_\delta \|\delta\|_p + c \cdot f(x+\delta) \\ &\text{s.t. } x+\delta \in [0,1]^n \end{aligned}$$

where f is an objective function such that $D(x+\delta) = t' \Leftrightarrow f(x+\delta) \leq 0$. The Carlini-Wagner attack is very strong – achieving over 99.8% misclassification on CIFAR-10 – but is slow and computationally expensive

Adversarial Transformation Networks (Baluja et al. 2017)

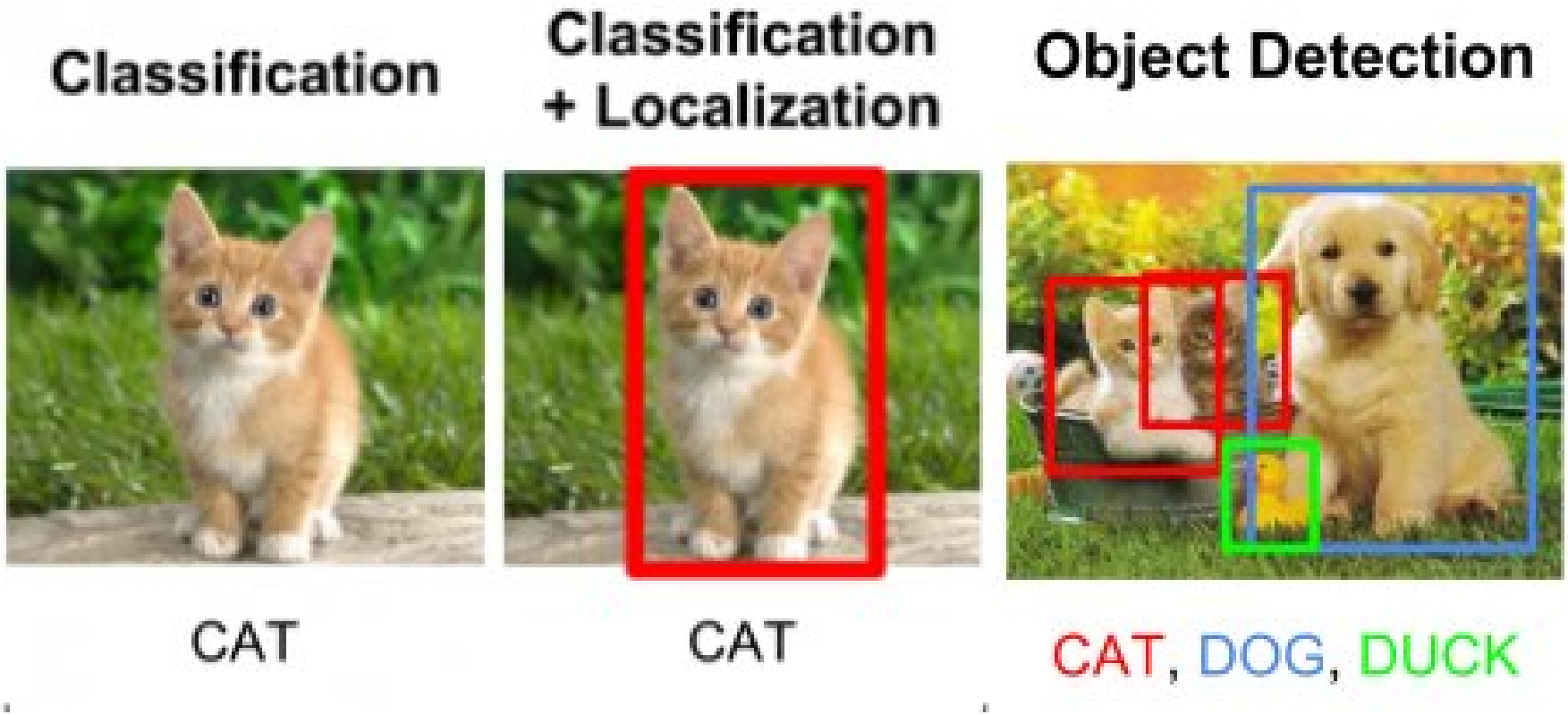
Adversarial Transformative Network (ATN) is any neural network that, given an input image, returns an adversarial image:

$$\text{argmin}_\theta \sum_{x_i \in \mathcal{X}} \beta \cdot L_{\mathcal{X}}(g_{f,\theta}(x_i), x_i) + L_{\mathcal{Y}}(f(g_{f,\theta}(x_i)), f(x_i))$$

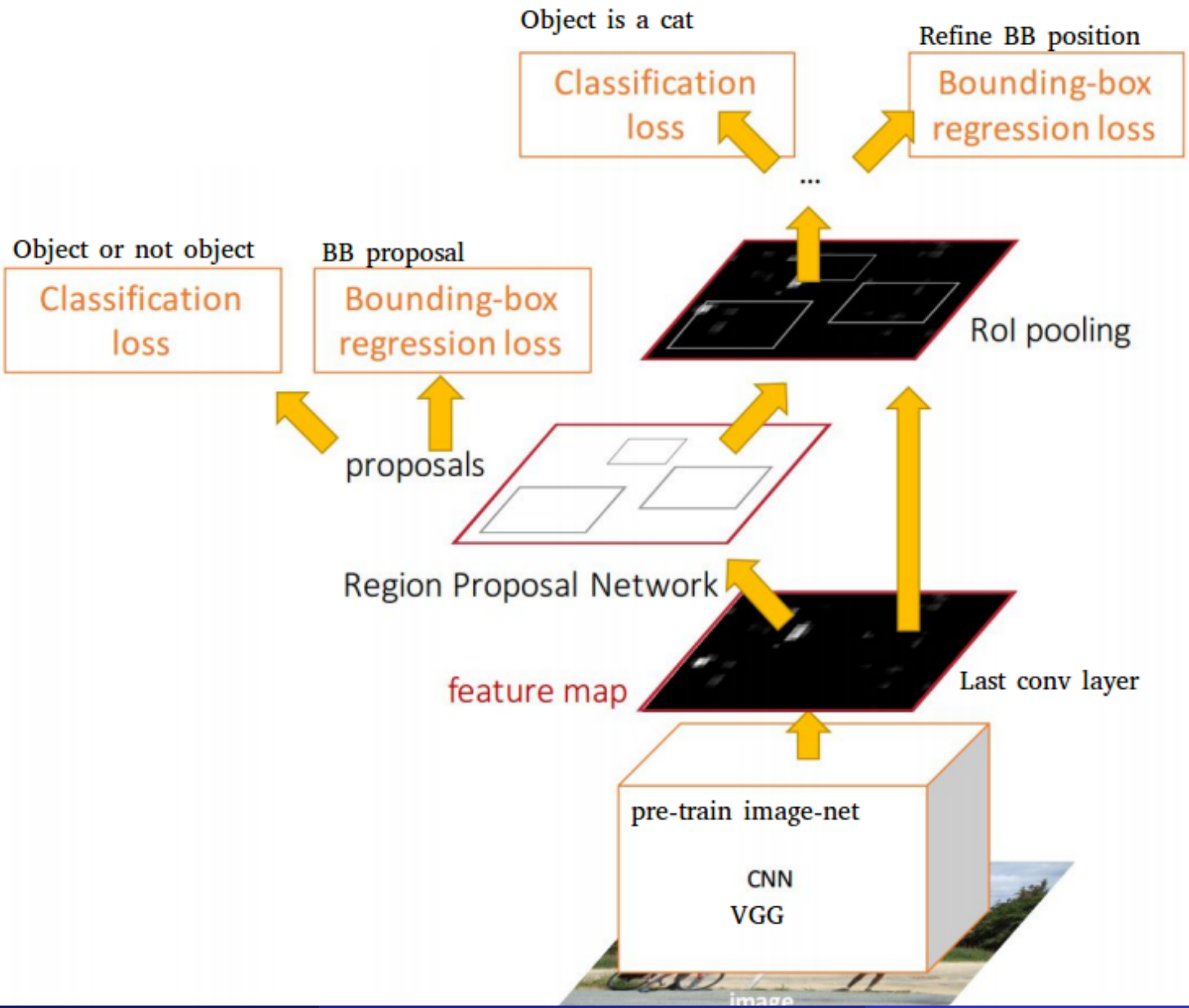
where β is a scalar, $L_{\mathcal{X}}$ is a perceptual loss (e.g., the L_2 distance) between the original and perturbed inputs and $L_{\mathcal{Y}}$ is the loss between the classifier's predictions on the original inputs and the perturbed inputs.

- ATNs were less effective than strong attacks like Carlini-Wagner
- ATN's are fast, adversarial image can be created with just a forward pass through the ATN
- ATN's adversarial images are not transferable

Object Detection in Pictures



Faster RCNN


Object is a cat
Classification loss
Refine BB position
Bounding-box regression loss
...
Object or not object
Classification loss
BB proposal
Bounding-box regression loss
RoI pooling
proposals
Region Proposal Network
feature map
Last conv layer
pre-train image-net
CNN
VGG

Adversarial Attacks on Object Detection

Object Detectors are much harder to attack than classification models due to:

- Number of Targets in an Image are much higher
- A successful attack must fool **ALL** Proposed Bounding Boxes
- Older Detectors are not always end to end differentiable

Problem Setup

Constructing adversarial examples for face detectors can be framed as a constrained optimization problem similar to the Carlini-Wagner attack.

$$\begin{aligned} &\text{minimize } L(x, x+\delta) \\ &\text{s.t. } D(x+\delta) = t' \\ &x+\delta \in [-1,1]^n \end{aligned}$$

This optimization problem is typically very difficult as the constraint $D(x+\delta) = t'$ is highly non-linear due to D being a neural network.

Relaxation

The constraint can be moved to the objective function as a penalty term for violating the original constraint.

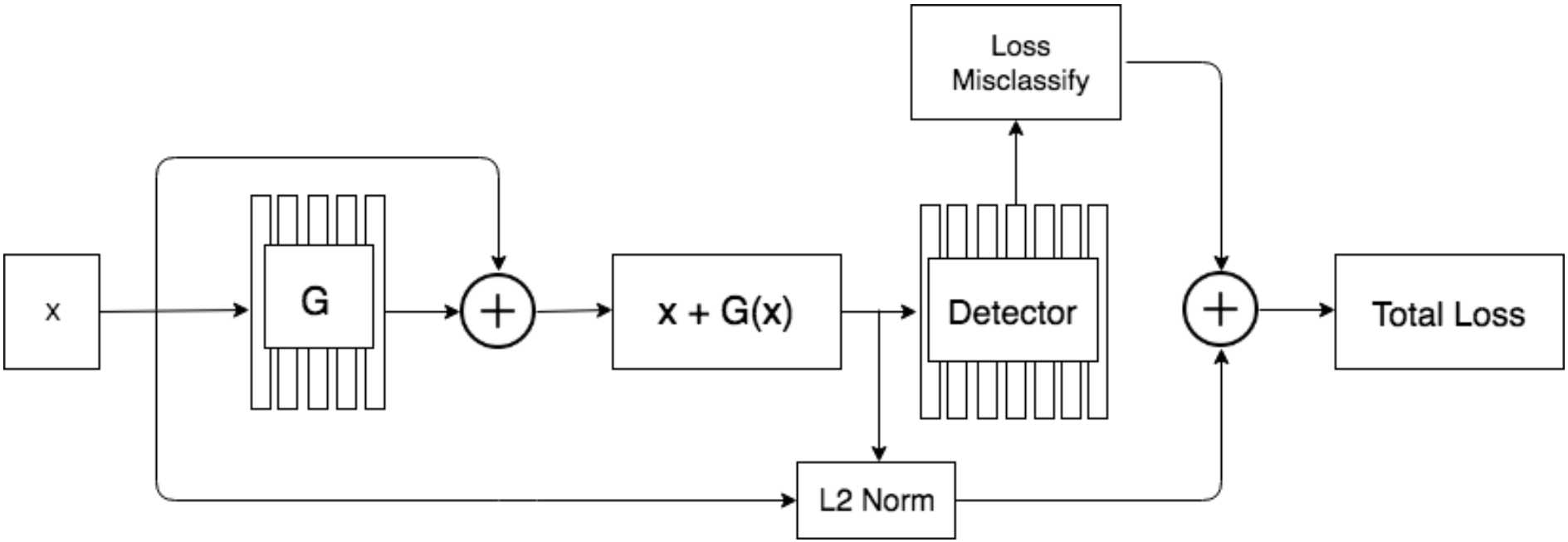
$$\begin{aligned} &\text{minimize } L(x, x+\delta) + \lambda L_{\text{misclassify}}(x+\delta) \\ &\text{s.t. } x+\delta \in [-1,1]^n \end{aligned}$$

The constant $\lambda > 0$ balances the magnitude of the perturbation generated to the actual adversarial goal.

Approach Motivation

- Optimizing over a single parameter per image is still difficult for a detection network.
- Adversarial attacks against face detectors should perturb pixels mostly on face regions
- Learning abstract representations of a face should help constructing attacks on new faces
- Fast generation of adversarial images enables Adversarial Training

Threat Model



Choice of Misclassification Loss

There are many possible choices for Misclassification Loss

- Likelihood of perturbed images under D
- $\sum_{i=1}^{N} \max(0, Z(x_i')_{\text{face}} - Z(x_i')_{\text{background}})$
- $\sum_{i=1}^{N} \max(0, D(x_i')_{\text{face}} - D(x_i')_{\text{background}})$

Empirically, some loss functions are better than others as the constant λ is either too small or too large during different phases in training.

Learning the Generator

$$L_{\text{total}}(x, x') = \|x - x'\|_2^2 + \lambda \cdot \sum_{i=1}^{N} \max(0, Z(x_i')_{\text{face}} - Z(x_i')_{\text{background}}) \quad (1)$$

- Conditional generator G is trained using a pretrained detector over **ALL** targets proposed by the detector.
- Spending more time on a given example allows greatly stabilizes training
- Choosing the same misclassification loss as the Carlini Wagner attack is more robust to the choice of λ. Training was not successful otherwise.

Implementation Details

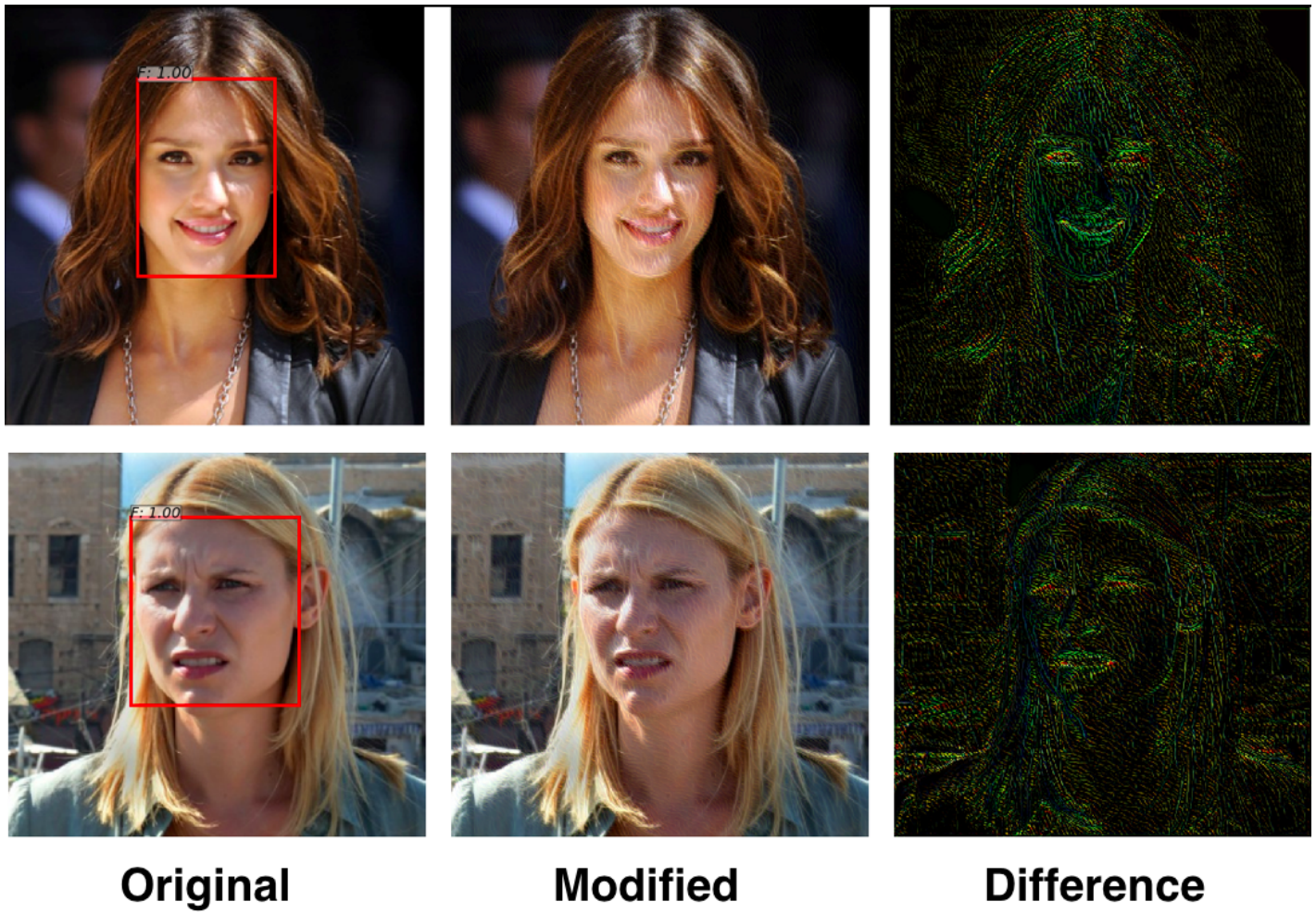
- Input are resized to a resolution of 600 by 800 pixels
- The number of object proposals are restricted to a maximum of 2000 during training and 300 during test
- Only Object proposals with probability greater than $\alpha = 0.7$ are considered
- We pre train our Faster R-CNN face detector on the WIDER face dataset for 14 epochs with the ADAM optimizer

Attacks on Cropped 300-W Dataset

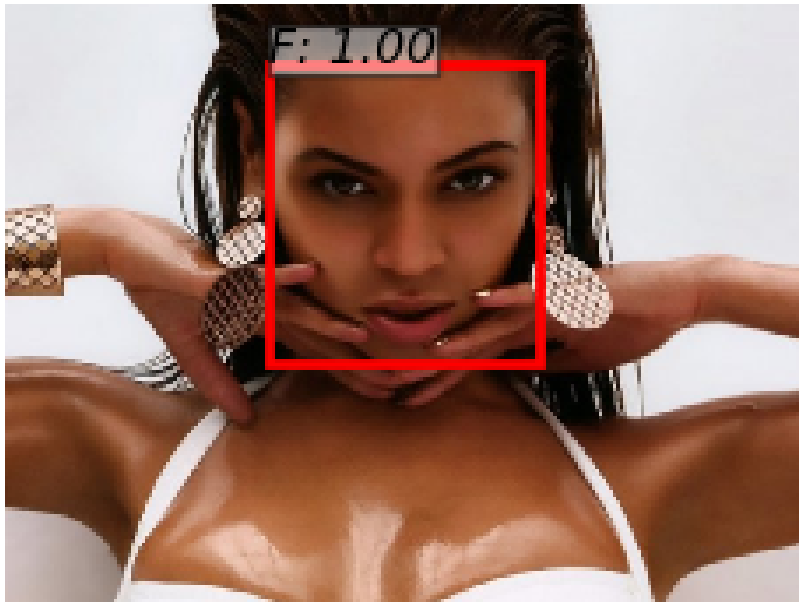
	Faster R-CNN	Our Attack
$\alpha = 0.5$	599	8
$\alpha = 0.6$	599	4
$\alpha = 0.7$	597	3
$\alpha = 0.8$	595	2
$\alpha = 0.9$	593	1
$\alpha = 0.99$	563	0

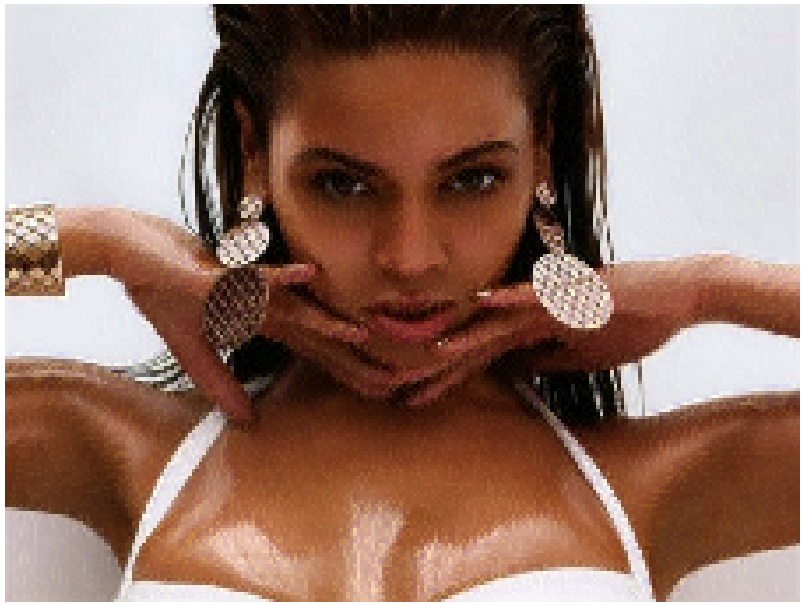
Results

	FGSM	C-W	Ours
Runtime	2.21s	>6300s	**1.21s**



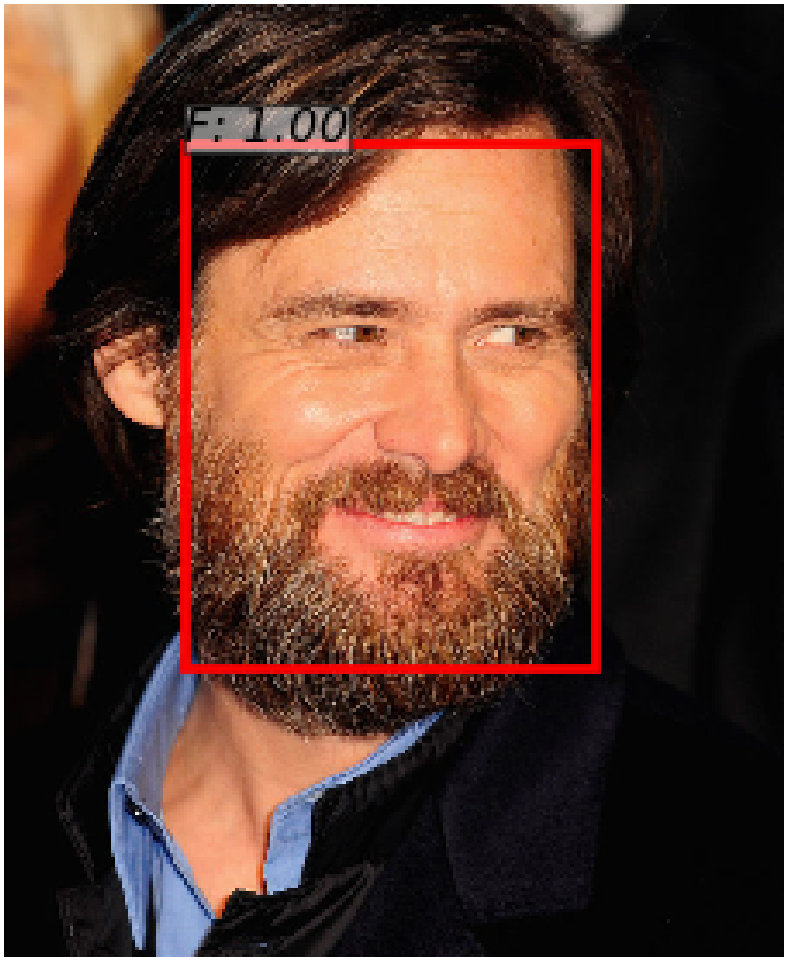
More Results





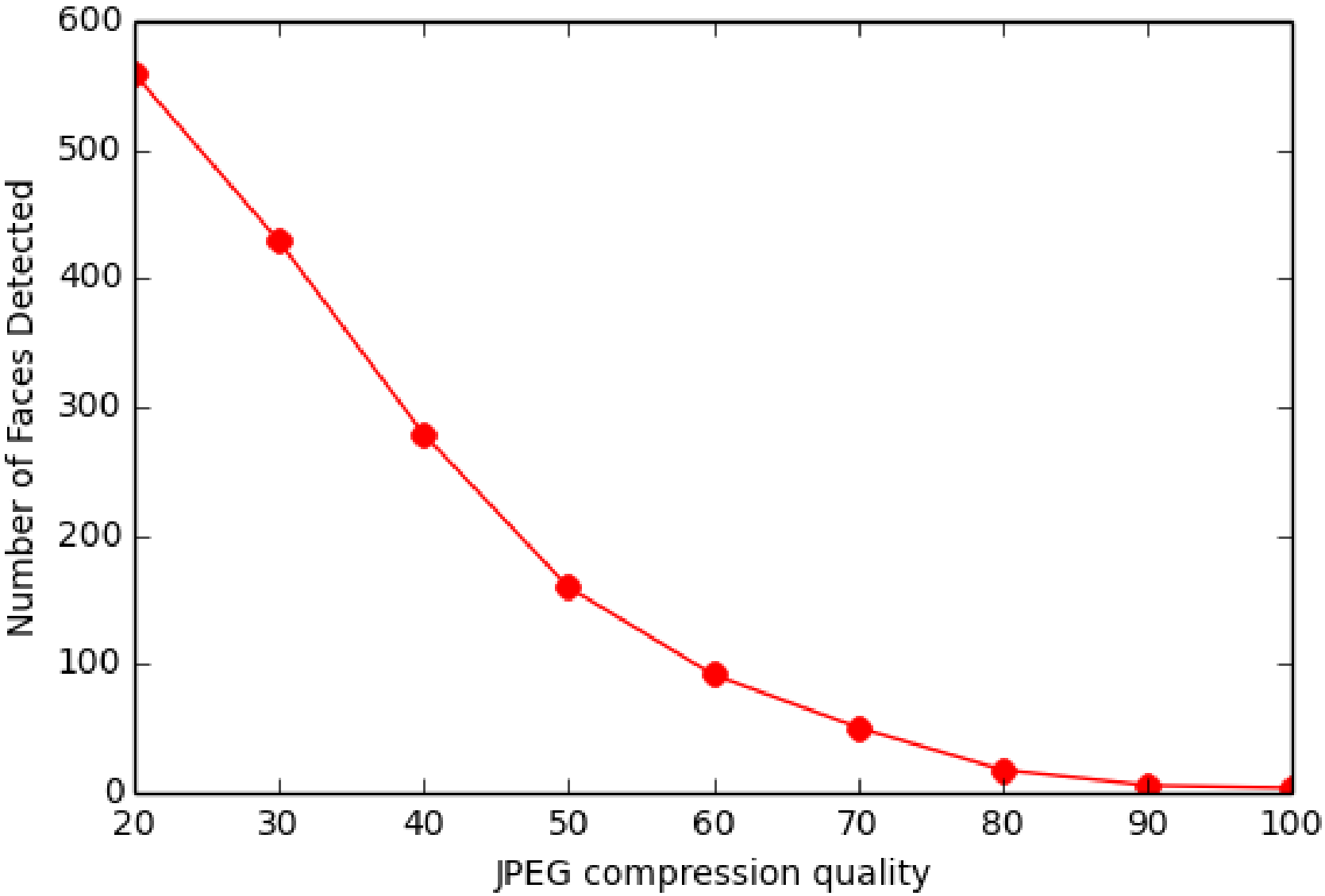
More Results

More Results



Video Demo

Attacks under JPEG compression



Ongoing and Future Research Directions

- Extend attack to multiple detectors
- Construct a Black-box variation of this attack using Policy Gradients
- Characterize the space of adversarial examples between two detectors.