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

- Training Set
- Training Set Poisoning
- Test Output
- Model Theft
- Model Output
- Learned Parameters
- Leaked Sensitive Data
- Test Input
- Adversarial Examples
minimize $L(x, x + \delta)$

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 $x'$:

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

where $J$ is the loss function for the attacked classifier and $\epsilon$ controls the extent of the perturbation.
FGSM on MNIST
Basic Iterative Method on MNIST (Kurakin et. al 2016)
Find some small $\delta$ such that $D(x + \delta) = t'$,

$$\arg\min_{\delta} \|\delta\|_p + c \cdot f(x + \delta)$$

s.t. $x + \delta \in [0, 1]^n$

where $f$ is an objective function such that $D(x + \delta) = t' \iff f(x + \delta) \leq 0$.

The Carlini-Wagner attack is very strong – achieving over 99.8% misclassification on CIFAR-10 – but is slow and computationally expensive.
Adversarial Transformative Network (ATN) is any neural network that, given an input image, returns an adversarial image:

\[
\arg\min_{\theta} \sum_{x_i \in X} \beta \cdot L_X(g_{f,\theta}(x_i), x_i) + L_Y(f(g_{f,\theta}(x_i)), f(x_i))
\]

where \( \beta \) is a scalar, \( L_X \) is a perceptual loss (e.g., the \( L_2 \) distance) between the original and perturbed inputs and \( L_Y \) 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

Classification

Classification + Localization

Object Detection

CAT
CAT
CAT, DOG, DUCK
Faster RCNN

Object is a cat

Classification loss

Bounding-box regression loss

Refine BB position

Object or not object

Classification loss

Bounding-box regression loss

Region Proposal Network

proposals

feature map

pre-train image-net

CNN

VGG

Rol pooling

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

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

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

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

The constant \( \lambda > 0 \) balances the magnitude of the perturbation generated to the actual adversarial goal.
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.
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 $\lambda$ is either too small or too large during different phases in training.
Learning the Generator

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

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Faster R-CNN</th>
<th>Our Attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>599</td>
<td>8</td>
</tr>
<tr>
<td>0.6</td>
<td>599</td>
<td>4</td>
</tr>
<tr>
<td>0.7</td>
<td>597</td>
<td>3</td>
</tr>
<tr>
<td>0.8</td>
<td>595</td>
<td>2</td>
</tr>
<tr>
<td>0.9</td>
<td>593</td>
<td>1</td>
</tr>
<tr>
<td>0.99</td>
<td>563</td>
<td>0</td>
</tr>
</tbody>
</table>

Joey Bose (UofT)  
GeekPwn Las Vegas  
September 26, 2018  
20 / 27
<table>
<thead>
<tr>
<th></th>
<th>FGSM</th>
<th>C-W</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runtime</td>
<td>2.21s</td>
<td>&gt;6300s</td>
<td>1.21s</td>
</tr>
</tbody>
</table>

Results

Original

Modified

Difference
More Results
More Results
Attacks under JPEG compression

![Graph showing the relationship between the number of faces detected and JPEG compression quality. As compression quality increases, the number of faces detected decreases.]
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.