ATTACKING DEEP NEURAL NETWORKS WITH ADVERSARIAL IMAGES

Fabrizio Falchi
ISTI, CNR, Pisa, Italy
www.fabriziofalchi.it

COST ACTION CA16101 - Dubrovnik, November 7th
WHAT’S THAT?
WHAT’S THAT?
What's that?
toucan
## Adversarial Examples

<table>
<thead>
<tr>
<th>mushroom</th>
<th>pineapple</th>
<th>toucan</th>
</tr>
</thead>
<tbody>
<tr>
<td>freight car</td>
<td>hummingbird</td>
<td>milk can</td>
</tr>
</tbody>
</table>
Edward H. Adelson
ILLUSIONS

Edward H. Adelson
DUBROVNIK – DEEP DREAM
Goal
Knowledge
Capability
ADVERSARY’S GOAL
GENUINE IMAGES

... Mushrooms
Pineapple
Toucan ...

Stride of 4
Max pooling
Max pooling
Max pooling
Max pooling
Dense
Dense
1000
NON-TARGETED ATTACK

Goal

NON-TARGETED

Mushrooms <whatever>

...
TARGETED ATTACK

Goal

TARGETED

Mushrooms

Toucan
Goal

Knowledge

Capability
Perfect-knowledge (white-box) attacks
- upper bound on the performance degradation under attack

Slide credit: Biggio
ATTACKING DEEP NEURAL NETWORKS

- Attacks are possible:
  - if you have the model [1,2]
  - if you have access to input and output only! [3]

Error Rate:  
- 84.24%  
- 88.94%  
- 96.19%


Black Box Adversarial Example Attacks

0 3
5 8

MNIST Dataset  GTSRD Dataset

Practical Black-Box Attacks against Machine Learning
Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami
ATTACKING FACE RECOGNITION SYSTEMS
ADVERIAL FACES

Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi
ADVERSARIAL FACES

Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi
Figure 5. Examples of the adversarial faces generated using FLM and GFLM. For each subject, five images are shown including the original face image (middle face), the result of GFLM (right face), the result of FLM (right image), displacement field $f$ for GFLM (left field) and displacement field $f$ for FLM (right field). Tags on the bottom left of images show the probability of the true class. Green and red tags denote the correct and incorrect classified samples respectively.
ADVERSARIAL FACES

P(True class) = 0.1054

P(True class) = 0.0135

P(True class) = 0.0151

Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi
Fast Geometrically-Perturbed Adversarial Faces
Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, Nasser M. Nasrabadi
Figure 2. The proposed method optimizes a displacement field $f$ to produce adversarial landmark locations $P^{adv}$. The spatial transformation $T$ transforms the input sample to the corresponding adversarial image $x^{adv}$ such that $\Phi(x^{adv}) = \Phi(x) + f$, and a state-of-the-art face recognition model $g$ mis-classifies the transformed image $x^{adv}$. 
ATTACKING IN REAL WORLD
ADVERSARIAL IMAGE

Photo: labsix
ROTATE ADVERSARIAL IMAGE

Photo: labsix
Adversarial Examples In The Physical World
Fooling Neural Networks in the Real World

labsix
<table>
<thead>
<tr>
<th>Subtle Poster</th>
<th>Subtle Poster Right Turn</th>
<th>Camouflage Graffiti</th>
<th>Camouflage Art (LISA-CNN)</th>
<th>Camouflage Art (GTSRB-CNN)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Robust Physical-World Attacks on Deep Learning Models
Eykholt, Evtimov, Fernandes, Bo Li, Rahmati, Xiao, Prakash, Kohno, Song
Fig. 4: An example of digital dodging. Left: An image of actor Owen Wilson, correctly classified by VGG143 with probability 1.00. Right: Dodging against VGG143 using AGN’s output (probability assigned to the correct class: < 0.01).
Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter
ATTACKING DNN IN REAL WORLD

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter
ATTACKING DNN IN REAL WORLD

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows $S_A$ (top) and $S_B$ (bottom) dodging against $DNN_B$. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows $S_A$ impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsW1C); (c) $S_B$ impersonating $S_C$; and (d) $S_C$ impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter
ATTACKING FACE VERIFICATION SYSTEMS
FACE RECOGNITION

ID1
ID2
ID3
...
ID10
...
IDn
FACE VERIFICATION
FACE VERIFICATION

Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks
Goswami, Ratha, Agarwal, Singh, Vatsa
Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks

Goswami, Ratha, Agarwal, Singh, Vatsa
ADVERSARY-AWARE MACHINE LEARNING
Machine learning system should be aware of the *arms race* with the adversary

*Security evaluation of pattern classifiers under attack*

Biggio, Fumera, Roli
NIPS 2018 : Adversarial Vision Challenge (Robust Model Track)

Pitting machine vision models against adversarial attacks.

Completed

41429  327  1953
Views  Participants  Submissions

Competition tracks

There will be three tracks in which you and your team can compete:

- Robust Model Track
- Untargeted Attacks Track
- Targeted Attacks Track
ADVERSARIAL EXAMPLE DETECTION
NON-TARGETED ATTACK

Mushrooms <whatever>
Increase robustness
**Detection**

Attack detection
Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter
OUR APPROACH
DEEP LEARNING (FROM NATURE)

Yann LeCun, Yoshua Bengio & Geoffrey Hinton
Representation learning methods that allow a machine to be fed with raw data and to automatically discover the representations needed for detection or classification.

Deep-learning are representation learning methods

- with multiple levels of representation, obtained by
- composing simple but non-linear modules that each
- transform the representation at one level into a representation at a higher, slightly more abstract level.
MULTIPLE LEVELS OF ABSTRACTION
MULTIPLE LEVELS OF ABSTRACTION
Our Approach

A detection scheme for adversarial images based on internal representation (aka deep features) of the neural network classifier.

• **Main intuition**: look at the evolution of features, i.e. the path formed by their positions in the feature spaces, during the forward pass of the network.

• **Claim**: The trajectories traced by authentic inputs and adversarial examples differ and can be used to discern them.

Adversarial examples detection in features distance spaces
F. Carrara, R. Becarelli, R. Caldelli, F. Falchi, G. Amato
ECCV WOCM Workshop 2018
MULTIPLE LEVELS OF ABSTRACTION
## Our Approach: Results

<table>
<thead>
<tr>
<th>Attacked Model</th>
<th>ResNet-50 pretrained on ILSVRC’12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crafting Algorithms</td>
<td>L-BFGS, FGSM, BIM, PGD, MI-FGSM</td>
</tr>
<tr>
<td>Emb. Pivots</td>
<td>1000 Class (C)entroids / (M)edoids from ILSVRC train set</td>
</tr>
<tr>
<td>Emb. Distance Function</td>
<td>L2 / cosine similarity (cos)</td>
</tr>
<tr>
<td>Emb. Size</td>
<td>16-length 1000-dim sequences, TRAIN / VAL / TEST = 12k / 1k / 3k</td>
</tr>
<tr>
<td>Detector</td>
<td>MLP (2-layer, 100 and 1 neurons) / LSTM (100-dim)</td>
</tr>
<tr>
<td>Threat Model</td>
<td>zero-knowledge (attacker not aware of detector)</td>
</tr>
<tr>
<td>Method</td>
<td>L-BFGS</td>
</tr>
<tr>
<td>LSTM + M + cos</td>
<td>.854</td>
</tr>
<tr>
<td>LSTM + M + L2</td>
<td>.743</td>
</tr>
<tr>
<td>MLP + M + cos</td>
<td>.551</td>
</tr>
<tr>
<td>MLP + M + L2</td>
<td>.681</td>
</tr>
<tr>
<td>LSTM + C + cos</td>
<td>.709</td>
</tr>
<tr>
<td>LSTM + C + L2</td>
<td>.482</td>
</tr>
<tr>
<td>MLP + C + cos</td>
<td>.388</td>
</tr>
<tr>
<td>MLP + C + L2</td>
<td>.626</td>
</tr>
</tbody>
</table>
## Easy to Identify Adversarial Images

<table>
<thead>
<tr>
<th>Adversarial Image</th>
<th>Generation Algorithm</th>
<th>Actual Class</th>
<th>Fooled Class</th>
<th>Nearest Neighbor</th>
<th>kNN score</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td>L-BFGS</td>
<td>bikini, two-piece</td>
<td>pomegranate</td>
<td><img src="image2.png" alt="Image" /></td>
<td>0.01</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td>FGS</td>
<td>brassiere, bra, bandeau</td>
<td>Chihuahua</td>
<td><img src="image4.png" alt="Image" /></td>
<td>0.01</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td>FGS</td>
<td>revolver, six-gun, six-shooter</td>
<td>mousetrap</td>
<td><img src="image6.png" alt="Image" /></td>
<td>0.00</td>
</tr>
<tr>
<td><img src="image7.png" alt="Image" /></td>
<td>L-BFGS</td>
<td>assault rifle, assault gun</td>
<td>Border terrier</td>
<td><img src="image8.png" alt="Image" /></td>
<td>0.00</td>
</tr>
<tr>
<td>Adversarial Image</td>
<td>Generation Algorithm</td>
<td>Actual Class</td>
<td>Fooled Class</td>
<td>Nearest Neighbor</td>
<td>kNN score</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------</td>
<td>--------------</td>
<td>--------------</td>
<td>-----------------</td>
<td>-----------</td>
</tr>
<tr>
<td><img src="image1.png" alt="Adversarial Image" /></td>
<td>FGS</td>
<td>chime, bell, gong</td>
<td>barometer</td>
<td><img src="image2.png" alt="Nearest Neighbor" /></td>
<td>0.13</td>
</tr>
<tr>
<td><img src="image3.png" alt="Adversarial Image" /></td>
<td>L-BFGS</td>
<td>basenji</td>
<td>Arctic fox, white fox, Alopex lagopus</td>
<td><img src="image4.png" alt="Nearest Neighbor" /></td>
<td>0.13</td>
</tr>
<tr>
<td><img src="image5.png" alt="Adversarial Image" /></td>
<td>FGS</td>
<td>Greater Swiss Mountain dog</td>
<td>Bernese mountain dog</td>
<td><img src="image6.png" alt="Nearest Neighbor" /></td>
<td>0.11</td>
</tr>
<tr>
<td><img src="image7.png" alt="Adversarial Image" /></td>
<td>FGS</td>
<td>jeep, landrover</td>
<td>pickup, pickup truck</td>
<td><img src="image8.png" alt="Nearest Neighbor" /></td>
<td>0.11</td>
</tr>
</tbody>
</table>
OTHER DETECTION APPROACHES

• **Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods** [2017]
  Nicholas Carlini, David Wagner

• **On Detecting Adversarial Perturbations** [2017]
  Jan Hendrik Metzen, Tim Genewein, Volker Fischer, Bastian Bischoff

• **Trace and detect adversarial attacks on CNNs using feature response maps** [2018]
  Mohammadreza, Friedhelm, Thilo

• **Adversarial examples detection in features distance spaces** [2018]
  F. Carrara, R. Becarelli, R. Caldelli, F. Falchi, G. Amato
Detection of Face Morphing Attacks by Deep Learning
C. Seibold, W. Samek, A. Hilsmann, P. Eisert
ADVERSARIAL EXAMPLES DETECTION

Cover Image + HiDDeN Perturbation = “Copyright ID: 1337”

HiDDeN: Hiding Data With Deep Networks
Jiren Zhu, Russell Kaplan, Justin Johnson, Li Fei-Fei
Questions are welcomed

Fabrizio Falchi
fabrizio.falchi@cnr.it
CONCLUSIONS

• Machine Learning and Deep Learning in particular can be attacked
  o Slightly modifying images but also in real world
  o Even if our neural network is a black box for the enemy

• Many approaches have been proposed to make DL more robust

• Adversarial examples detection is its early stages

• We need adversary-aware machine learning