Adversarial Attack

Hung-yi Lee

Source of image: http://www.fafa01.com/post865806
Motivation

• You have trained many neural networks.
• We seek to deploy neural networks in the real world.
• Are networks robust to the inputs that are built to fool them?
  • Useful for spam classification, malware detection, network intrusion detection, etc.
人類不講武德 ...
How to Attack

Example of Attack

Benign Image

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
  \vdots
\end{bmatrix}
+ \begin{bmatrix}
  \Delta x_1 \\
  \Delta x_2 \\
  \Delta x_3 \\
  \vdots
\end{bmatrix}
\]
small

Network (Image Classifier)

Non-targeted
Anything other than “Cat”

Targeted
Misclassified as a specific class (e.g., “Star Fish”)

Something Else

Attacked Image

Tiger Cat
Example of Attack

The target is “Star Fish”

Network = ResNet-50

Benign Image

Tiger Cat

0.64

Attacked Image

Star Fish

1.00
Example of Attack

- Benign Image
  - Tiger Cat
    - 0.64
- Attack Image
  - Star Fish
    - 1.00

50x
Example of Attack

Benign Image

Attacked Image

The target is “Keyboard”
tiger cat

Persian cat

tabby cat

fire screen
How to Attack

Non-targeted

\[ x^* = \arg \min_x L(x) \]

\[ L(x) = -e(y, \hat{y}) \]

Targeted

\[ L(x) = -e(y, \hat{y}) + e(y, y_{target}) \]
Non-perceivable

\[ d(x^0, x) \leq \varepsilon \quad \text{Need to consider human perception} \]

- **L2-norm**

\[ d(x^0, x) = \|\Delta x\|_2 \]
\[ = (\Delta x_1)^2 + (\Delta x_2)^2 + (\Delta x_3)^2 \ldots \]

- **L-infinity**

\[ d(x^0, x) = \|\Delta x\|_\infty \]
\[ = \max\{|\Delta x_1|, |\Delta x_2|, |\Delta x_3|, \ldots\} \]
Attack Approach

\[ w^*, b^* = \arg \min_{w, b} L \quad \text{Difference?} \]

\[ x^* = \arg \min L(x) \]

**Gradient Descent**

Start from original image \( x^0 \)

For \( t = 1 \) to \( T \)

\[ x^t \leftarrow x^{t-1} - \eta g \]

\[ g = \begin{bmatrix} \frac{\partial L}{\partial x_1} \bigg|_{x=x^{t-1}} \\ \frac{\partial L}{\partial x_2} \bigg|_{x=x^{t-1}} \\ \vdots \end{bmatrix} \]
Attack Approach

\[ x^* = \arg \min_{d(x^0, x) \leq \varepsilon} L(x) \]

Gradient Descent

Start from original image \( x^0 \)

For \( t = 1 \) to \( T \)

\[ x^t \leftarrow x^{t-1} - \eta g \]

If \( d(x^0, x) > \varepsilon \)

\[ x^t \leftarrow \text{fix}(x^t) \]

\[ w^*, b^* = \arg \min_{w, b} L \] Difference?

Update input, not parameters

Different optimization methods

Different constraints

L-infinity

\( \varepsilon \)

after update

fixed
Attack Approach

\[ x^* = \arg \min_{d(x^0, x) \leq \varepsilon} L(x) \]

Fast Gradient Sign Method (FGSM)
https://arxiv.org/abs/1412.6572

Start from original image \( x^0 \)
For \( t = 1 \) to \( T \)
\[ x^t \leftarrow x^{t-1} - \eta g \]
**Attack Approach**

\[ x^* = \text{arg min}_{d(x^0, x) \leq \varepsilon} \min L(x) \]

**Fast Gradient Sign Method (FGSM)**

Start from original image \( x^0 \)

For \( t = 1 \) to \( T \)

\[ x^t \leftarrow x^{t-1} - \eta g \]

\[ g = \pm 1 \begin{bmatrix} \text{sign} \left( \frac{\partial L}{\partial x_1} \bigg|_{x=x^t} \right) \\ \text{sign} \left( \frac{\partial L}{\partial x_2} \bigg|_{x=x^t} \right) \\ \vdots \end{bmatrix} \]

\[ \text{if } t > 0, \text{sign}(t) = 1; \text{otherwise, sign}(t) = -1 \]
Attack Approach

\[ x^* = \arg \min_{d(x^0, x) \leq \varepsilon} L(x) \]

Iterative FGSM

Start from original image \( x^0 \)
For \( t = 1 \) to \( T \)
\[ x^t = x^{t-1} - \eta g \]
If \( d(x^0, x) > \varepsilon \)
\[ x^t = \text{fix}(x^t) \]

\[ g = \pm 1 \begin{bmatrix} \text{sign} \left( \frac{\partial L}{\partial x_1} \bigg|_{x=x^{t-1}} \right) \\ \text{sign} \left( \frac{\partial L}{\partial x_2} \bigg|_{x=x^{t-1}} \right) \\ \vdots \end{bmatrix} \]

L-infinity

\[ \varepsilon \]

fixed

after update

https://arxiv.org/abs/1607.02533
White Box v.s. Black Box

• In the previous attack, we know the network parameters $\theta$
  • This is called **White Box Attack**.

• You cannot obtain model parameters in most online API.

• Are we safe if we do not release model? ☺

• No, because **Black Box Attack** is possible. 😞
Black Box Attack

If you have the training data of the target network
Train a proxy network yourself
Using the proxy network to generate attacked objects

What if we do not know the training data?
Black Box Attack

Be Attacked

<table>
<thead>
<tr>
<th></th>
<th>ResNet-152</th>
<th>ResNet-101</th>
<th>ResNet-50</th>
<th>VGG-16</th>
<th>GoogLeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-152</td>
<td>0%</td>
<td>13%</td>
<td>18%</td>
<td>19%</td>
<td>11%</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>19%</td>
<td>0%</td>
<td>21%</td>
<td>21%</td>
<td>12%</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>23%</td>
<td>20%</td>
<td>0%</td>
<td>21%</td>
<td>18%</td>
</tr>
<tr>
<td>VGG-16</td>
<td>22%</td>
<td>17%</td>
<td>17%</td>
<td>0%</td>
<td>5%</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>39%</td>
<td>38%</td>
<td>34%</td>
<td>19%</td>
<td>0%</td>
</tr>
</tbody>
</table>

(lower accuracy → more successful attack)

Ensemble Attack

<table>
<thead>
<tr>
<th></th>
<th>ResNet-152</th>
<th>ResNet-101</th>
<th>ResNet-50</th>
<th>VGG-16</th>
<th>GoogLeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ResNet-152</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-ResNet-101</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-ResNet-50</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-VGG-16</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>-GoogLeNet</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
</tr>
</tbody>
</table>
The attack is so easy! Why?

To learn more:

Adversarial Examples Are Not Bugs, They Are Features


One pixel attack


Video: https://youtu.be/tfpKIZIWidA
Universal Adversarial Attack

https://arxiv.org/abs/1610.08401

Black Box Attack is also possible!
Beyond Images

- Speech processing
  Detect synthesized speech

- Natural language processing

---

**Question:** Why did he walk?
For *exercise*, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells.

**Question:** Why did the university see a drop in applicants?
In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . .

https://arxiv.org/abs/1908.07125
Attack in the Physical World

- An attacker would need to find perturbations that generalize beyond a single image.
- Extreme differences between adjacent pixels in the perturbation are unlikely to be accurately captured by cameras.
- It is desirable to craft perturbations that are comprised mostly of colors reproducible by the printer.
<table>
<thead>
<tr>
<th>Distance/Angle</th>
<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>5° 0°</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>5° 15°</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>10° 0°</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>10° 30°</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>40° 0°</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
</tbody>
</table>


Targeted-Attack Success: 100% 73.33% 66.67% 100% 80%
Attack in the Physical World

read as an 85-mph sign


https://youtu.be/4uGV_fRj0UA
Adversarial Reprogramming

(a) **counting ImageNet**

- $y_{adv}$
- 1 square: tench
- 2 squares: goldfish
- 3 squares: white shark
- 4 squares: tiger shark
- 5 squares: hammerhead
- 6 squares: electric ray
- 7 squares: stingray
- 8 squares: cock
- 9 squares: hen
- 10 squares: ostrich

(b) **Adversarial Program**

(c) **ImageNet Classifier**

- tiger shark, ostrich
- 4 squares, 10 squares

https://arxiv.org/abs/1806.11146
“Backdoor” in Model

- Attack happens at the training phase

Goal: misclassified as “dog”

be careful of unknown dataset ……

https://arxiv.org/abs/1804.00792
Defense
Passive v.s. Proactive

http://3png.com/a-27051273.html
Passive Defense

Original

Do not influence classification

Filter

Network

e.g. Smoothing

Attack signal

Less harmful

Tiger Cat

Keyboard
Keyboard
0.98

Smoothing

Smoothing

tiger cat
0.64

tiger cat
0.37

Side Effect!
Passive Defense

Image Compression

https://arxiv.org/abs/1704.01155

Generator

https://arxiv.org/abs/1805.06605
Passive Defense - Randomization

https://arxiv.org/abs/1711.01991
Proactive Defense

**Adversarial Training**

Training a model that is robust to adversarial attack.

Given training set $\mathcal{X} = \{(x^1, \hat{y}^1), (x^2, \hat{y}^2), \ldots, (x^N, \hat{y}^y)\}$

Using $\mathcal{X}$ to train your model

For $n = 1$ to $N$

Find adversarial input $\tilde{x}^n$ given $x^n$ by an attack algorithm

We have new training data

$\mathcal{X}' = \{ (\tilde{x}^1, \hat{y}^1), (\tilde{x}^2, \hat{y}^2), \ldots, (\tilde{x}^N, \hat{y}^y) \}$

Using both $\mathcal{X}$ and $\mathcal{X}'$ to update your model

Data Augmentation

Can it deal with new algorithm?

Fix it!
Concluding Remarks

- Attack: given the network parameters, attack is very easy.
- Even black box attack is possible
- Defense: Passive & Proactive
- Attack / Defense are still evolving.

https://www.gotrip.hk/179304/weekend_lifestyle/pokemon-go-%E7%B2%BE%E9%9D%88%E9%80%B2%E5%8C%96/
Acknowledgement

• 感謝作業十助教團隊林毓宸同學、黃啟斌同學幫忙蒐集參考
Attack Approaches

- ...... only list a few
What happened?

$\gamma_{Egyptian\ cat}$  $\gamma_{tiger\ cat}$  $\gamma_{Persian\ cat}$

Random

$\gamma_{tiger\ cat}$  $\gamma_{key\ board}$

Specific Direction

$x^0$