Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks

Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu

Presented by: Qing Feng, Daoyang Song
Overview

- Related Work
- Methodology
- Experiments
- Conclusion
Attention Maps

Normally Trained Models

Defense Models
Related Work

- Adversarial examples
- Black-box attacks
- Attacks for an ensemble of examples
- Defend against adversarial attacks
Methodology

- Adversarial example for a set of images
- Less sensitive
- Higher probability
Algorithms

- TI method
- Optimization
- Kernel matrix
- Attack algorithms
**Translation-Invariant method**

$T_{ij}(x)$: shifts image $x$ by $i$ and $j$ pixels

$W_{ij}$: weight

$J(T_{ij}(x^{adv}), y)$: loss function

$$\begin{aligned}
\arg\max_{x^{adv}} J(x^{adv}, y), \\
\arg\max_{x^{adv}} \sum_{i,j} w_{ij} J(T_{ij}(x^{adv}), y), \\
s.t. \|x^{adv} - x^{real}\|_\infty \leq \epsilon,
\end{aligned}$$
Gradient Calculation

Translation-invariant property

\[ \nabla_x J(x, y) \bigg|_{x = T_{ij}(\hat{x})} \approx \nabla_x J(x, y) \bigg|_{x = \hat{x}}. \]
\[ \nabla_x \left( \sum_{i,j} w_{ij} J(T_{ij}(x), y) \right) \bigg|_{x=\hat{x}} = \sum_{i,j} w_{ij} \nabla_x J(T_{ij}(x), y) \bigg|_{x=\hat{x}} = \nabla_x J(T_{ij}(x), y) \]

\[ = \sum_{i,j} w_{ij} \nabla_{T_{ij}(x)} J(T_{ij}(x), y) \cdot \frac{\partial T_{ij}(x)}{\partial x} \bigg|_{x=\hat{x}} = \left( \nabla_{T_{ij}(x)} J(T_{ij}(x), y) \cdot \frac{\partial T_{ij}(x)}{\partial x} \right) \]

\[ = \sum_{i,j} w_{ij} T_{-i-j} \left( \nabla_x J(x, y) \bigg|_{x=T_{ij}(\hat{x})} \right) \]

\[ \approx \sum_{i,j} w_{ij} T_{-i-j} \left( \nabla_x J(x, y) \bigg|_{x=\hat{x}} \right) \]
Gradient Calculation

\[ \sum_{i,j} w_{ij} T_{-i-j}(\nabla_x J(x, y)|_{x=x}) \leftrightarrow W \ast \nabla_x J(x, y)|_{x=x}, \]

\(W\) is the kernel matrix
Three different types of kernel

Uniform Kernel:

\[ W_{i,j} = \frac{1}{(2k+1)^2}; \]

Linear kernel:

\[ \tilde{W}_{i,j} = (1 - |i|/k+1) \cdot (1 - |j|/k+1) \quad W_{i,j} = \frac{\tilde{W}_{i,j}}{\sum_{i,j} \tilde{W}_{i,j}} \]

A Gaussian kernel:

\[ \tilde{W}_{i,j} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2+j^2}{2\sigma^2}\right) \]

standard deviation

\[ \sigma = \frac{k}{\sqrt{3}} \quad W_{i,j} = \frac{\tilde{W}_{i,j}}{\sum_{i,j} \tilde{W}_{i,j}} \]
Attacking algorithm: TI-BIM

Basic Iterative Method

TI-BIM

\[ x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(\nabla_x J(x_t^{adv}, y)) \]

\[ x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(W \ast \nabla_x J(x_t^{adv}, y)) \]
Attacking algorithm: TI-FGSM

Fast Gradient Sign Method

\[ \mathbf{x}^{adv} = \mathbf{x}^{real} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}^{real}, y)), \]

TI-FGSM

\[ \mathbf{x}^{adv} = \mathbf{x}^{real} + \epsilon \cdot \text{sign}(\mathbf{W} \ast \nabla_{\mathbf{x}} J(\mathbf{x}^{real}, y)). \]
Experiments

- Experimental Settings
- Translation-Invariant Property of CNNs
- The Results of Different Kernels
- The Effect of Kernel Size
- Single-Model Attacks
- Ensemble-based Attacks
Experimental Settings

Eight defense models
- Inc-v3ens3, Inc-v3ens4, IncRes-v2ens
- HGD, R&P, JPEG, TVM, NIPS-r3

Four normally trained models
- Inc-v3, Inc-v4, IncRes-v2, Res-v2

Maximum perturbation per pixel: 16
Step size: 1.6
Number of iterations: 10

For MI-FGSM and TI-MI-FGSM
Decay factor: 1.0

For DIM and TI-DIM
Transformation probability: 0.7

ImageNet-compatible dataset: 1,000 images
Translation-Invariant Property of CNNs

k: Maximal number of pixels to shift

Calculate the gradients for \((2k+1)^2\) images?

The loss surfaces
## The Results of Different Kernels

<table>
<thead>
<tr>
<th>Attack</th>
<th>Inc-v3\textsubscript{ens3}</th>
<th>Inc-v3\textsubscript{ens4}</th>
<th>IncRes-v2\textsubscript{ens}</th>
<th>HGD</th>
<th>R&amp;P</th>
<th>JPEG</th>
<th>TVM</th>
<th>NIPS-r3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI-FGSM</td>
<td>Uniform</td>
<td>25.0</td>
<td>27.9</td>
<td>21.1</td>
<td>15.7</td>
<td>19.1</td>
<td>24.8</td>
<td>32.3</td>
</tr>
<tr>
<td></td>
<td>Linear</td>
<td>30.7</td>
<td>32.4</td>
<td>24.2</td>
<td>20.9</td>
<td>23.3</td>
<td>28.1</td>
<td>34.6</td>
</tr>
<tr>
<td></td>
<td>Gaussian</td>
<td>28.2</td>
<td>28.9</td>
<td>22.3</td>
<td>18.4</td>
<td>19.8</td>
<td>25.5</td>
<td>30.7</td>
</tr>
<tr>
<td>TI-MI-FGSM</td>
<td>Uniform</td>
<td>30.0</td>
<td>32.2</td>
<td>22.8</td>
<td>21.7</td>
<td>22.8</td>
<td>26.4</td>
<td>32.7</td>
</tr>
<tr>
<td></td>
<td>Linear</td>
<td>35.8</td>
<td>35.0</td>
<td>26.8</td>
<td>25.5</td>
<td>23.4</td>
<td>29.0</td>
<td>35.8</td>
</tr>
<tr>
<td></td>
<td>Gaussian</td>
<td>35.8</td>
<td>35.1</td>
<td>25.8</td>
<td>25.7</td>
<td>23.9</td>
<td>28.2</td>
<td>34.9</td>
</tr>
<tr>
<td>TI-DIM</td>
<td>Uniform</td>
<td>32.6</td>
<td>34.6</td>
<td>25.6</td>
<td>24.1</td>
<td>27.2</td>
<td>30.2</td>
<td>34.9</td>
</tr>
<tr>
<td></td>
<td>Linear</td>
<td>45.2</td>
<td>47.0</td>
<td>34.9</td>
<td>35.6</td>
<td>35.2</td>
<td>38.5</td>
<td>43.6</td>
</tr>
<tr>
<td></td>
<td>Gaussian</td>
<td>46.9</td>
<td>47.1</td>
<td>37.4</td>
<td>38.3</td>
<td>36.8</td>
<td>37.0</td>
<td>44.2</td>
</tr>
</tbody>
</table>

Table 1. The success rates (%) of black-box attacks against eight defenses with different choices of kernels
The Effect of Kernel Size

The success rates (%) of black-box attacks against IncRes-v2-ens, HGD, R&P, TVM, and NIPS-r3.
The Effect of Kernel Size

The adversarial examples generated for Inc-v3 by TI-FGSM with different kernel sizes.
Table 2. The success rates (%) of black-box attacks against eight defenses.
## Single-Model Attacks

<table>
<thead>
<tr>
<th>Attack</th>
<th>Inc-v3 ens3</th>
<th>Inc-v3 ens4</th>
<th>IncRes-v2 ens</th>
<th>HGD</th>
<th>R&amp;P</th>
<th>JPEG</th>
<th>TVM</th>
<th>NIPS-r3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc-v3</td>
<td>MI-FGSM</td>
<td>20.5</td>
<td>17.4</td>
<td>9.5</td>
<td>6.9</td>
<td>8.7</td>
<td>20.3</td>
<td>19.4</td>
</tr>
<tr>
<td></td>
<td>TI-MI-FGSM</td>
<td><strong>35.8</strong></td>
<td><strong>35.1</strong></td>
<td><strong>25.8</strong></td>
<td><strong>25.7</strong></td>
<td><strong>23.9</strong></td>
<td><strong>28.2</strong></td>
<td><strong>34.9</strong></td>
</tr>
<tr>
<td>Inc-v4</td>
<td>MI-FGSM</td>
<td>22.1</td>
<td>20.1</td>
<td>12.1</td>
<td>9.6</td>
<td>12.1</td>
<td>26.0</td>
<td>24.8</td>
</tr>
<tr>
<td></td>
<td>TI-MI-FGSM</td>
<td><strong>36.7</strong></td>
<td><strong>39.2</strong></td>
<td><strong>28.7</strong></td>
<td><strong>27.8</strong></td>
<td><strong>28.0</strong></td>
<td><strong>31.6</strong></td>
<td><strong>38.4</strong></td>
</tr>
<tr>
<td>IncRes-v2</td>
<td>MI-FGSM</td>
<td>31.3</td>
<td>27.2</td>
<td>19.7</td>
<td>19.6</td>
<td>18.6</td>
<td>31.6</td>
<td>34.4</td>
</tr>
<tr>
<td></td>
<td>TI-MI-FGSM</td>
<td><strong>50.7</strong></td>
<td><strong>51.7</strong></td>
<td><strong>49.3</strong></td>
<td><strong>45.1</strong></td>
<td><strong>45.2</strong></td>
<td><strong>45.9</strong></td>
<td><strong>55.4</strong></td>
</tr>
<tr>
<td>Res-v2-152</td>
<td>MI-FGSM</td>
<td>25.1</td>
<td>23.7</td>
<td>13.3</td>
<td>15.1</td>
<td>14.6</td>
<td>31.2</td>
<td>24.5</td>
</tr>
<tr>
<td></td>
<td>TI-MI-FGSM</td>
<td><strong>39.9</strong></td>
<td><strong>37.7</strong></td>
<td><strong>32.8</strong></td>
<td><strong>31.8</strong></td>
<td><strong>31.1</strong></td>
<td><strong>38.3</strong></td>
<td><strong>41.2</strong></td>
</tr>
</tbody>
</table>

Table 3. The success rates (%) of black-box attacks against eight defenses.
## Single-Model Attacks

Table 4. The success rates (%) of black-box attacks against eight defenses.

<table>
<thead>
<tr>
<th>Attack</th>
<th>Inc-v3&lt;sub&gt;ens3&lt;/sub&gt;</th>
<th>Inc-v3&lt;sub&gt;ens4&lt;/sub&gt;</th>
<th>IncRes-v2&lt;sub&gt;ens&lt;/sub&gt;</th>
<th>HGD</th>
<th>R&amp;P</th>
<th>JPEG</th>
<th>TVM</th>
<th>NIPS-r3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc-v3</td>
<td>DIM</td>
<td>24.2</td>
<td>24.3</td>
<td>13.0</td>
<td>9.7</td>
<td>13.3</td>
<td>30.7</td>
<td>24.4</td>
</tr>
<tr>
<td></td>
<td>TI-DIM</td>
<td>46.9</td>
<td>47.1</td>
<td>37.4</td>
<td>38.3</td>
<td>36.8</td>
<td>37.0</td>
<td>44.2</td>
</tr>
<tr>
<td>Inc-v4</td>
<td>DIM</td>
<td>28.3</td>
<td>27.5</td>
<td>15.6</td>
<td>14.6</td>
<td>17.2</td>
<td>38.6</td>
<td>29.1</td>
</tr>
<tr>
<td></td>
<td>TI-DIM</td>
<td>48.6</td>
<td>47.5</td>
<td>38.7</td>
<td>40.3</td>
<td>39.3</td>
<td>43.5</td>
<td>45.6</td>
</tr>
<tr>
<td>IncRes-v2</td>
<td>DIM</td>
<td>41.2</td>
<td>40.0</td>
<td>27.9</td>
<td>32.4</td>
<td>30.2</td>
<td>47.2</td>
<td>41.7</td>
</tr>
<tr>
<td></td>
<td>TI-DIM</td>
<td>61.3</td>
<td>60.1</td>
<td>59.5</td>
<td>58.7</td>
<td>61.4</td>
<td>55.7</td>
<td>66.2</td>
</tr>
<tr>
<td>Res-v2-152</td>
<td>DIM</td>
<td>40.5</td>
<td>36.0</td>
<td>24.1</td>
<td>32.6</td>
<td>26.4</td>
<td>42.4</td>
<td>36.8</td>
</tr>
<tr>
<td></td>
<td>TI-DIM</td>
<td>56.1</td>
<td>55.5</td>
<td>49.5</td>
<td>51.8</td>
<td>50.4</td>
<td>50.8</td>
<td>55.7</td>
</tr>
</tbody>
</table>
Ensemble-based Attacks


<table>
<thead>
<tr>
<th>Attack</th>
<th>Inc-v3&lt;sub&gt;ens3&lt;/sub&gt;</th>
<th>Inc-v3&lt;sub&gt;ens4&lt;/sub&gt;</th>
<th>IncRes-v2&lt;sub&gt;ens&lt;/sub&gt;</th>
<th>HGD</th>
<th>R&amp;P</th>
<th>JPEG</th>
<th>TVM</th>
<th>NIPS-r3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGSM</td>
<td>27.5</td>
<td>23.7</td>
<td>13.4</td>
<td>4.9</td>
<td>13.8</td>
<td>38.1</td>
<td>30.0</td>
<td>19.8</td>
</tr>
<tr>
<td>TI-FGSM</td>
<td>39.1</td>
<td>38.8</td>
<td>31.6</td>
<td>29.9</td>
<td>31.2</td>
<td>43.3</td>
<td>39.8</td>
<td>33.9</td>
</tr>
<tr>
<td>MI-FGSM</td>
<td>50.5</td>
<td>48.3</td>
<td>32.8</td>
<td>38.6</td>
<td>32.8</td>
<td>67.7</td>
<td>50.1</td>
<td>43.9</td>
</tr>
<tr>
<td>TI-MI-FGSM</td>
<td>76.4</td>
<td>74.4</td>
<td>69.6</td>
<td>73.3</td>
<td>68.3</td>
<td>77.2</td>
<td>72.1</td>
<td>71.4</td>
</tr>
<tr>
<td>DIM</td>
<td>66.0</td>
<td>63.3</td>
<td>45.9</td>
<td>57.7</td>
<td>51.7</td>
<td>82.5</td>
<td>64.1</td>
<td>63.7</td>
</tr>
<tr>
<td>TI-DIM</td>
<td>84.8</td>
<td>82.7</td>
<td>78.0</td>
<td>82.6</td>
<td>81.4</td>
<td>83.4</td>
<td>79.8</td>
<td>83.1</td>
</tr>
</tbody>
</table>

Table 5. The success rates (%) of black-box attacks against eight defenses.
Conclusion

- The translation-invariant attack method are less sensitive to the discriminative regions of the white-box model.
- Have higher transferability against the defense models.
- Experiment results identify the vulnerability of the current defenses.
For

- The authors apply the kernel matrix to calculate the gradient at the untranslated image instead of computing all the shifted gradients. It can reduce a lot of calculation.
- Using attention maps to describe how the defense models against transferable adversarial examples which is visualized and easy to understand.
- The translation-invariant attack method proposed in this paper can be combined with other baseline methods and can improve the success rate significantly. It is a strong black-box attack against the defenses.
- The authors did extensive experiments to show the effectiveness of their methods.
Against

- The idea of attention maps is interesting. But the reasons that defense models against transferable adversarial examples due to the different discriminative regions are not explained.
- When calculating the gradient, they did not explain how they get the average of all the shifted gradients.
- The integrated attack algorithm TI-DIM and TI-MI-FGSM are not written in their paper.
- They did not do experiments to compare BIM and its integrated method TI-BIM during experiments and did not specify why BIM is excluded from the experiment.
Thank you!

Questions?