Improving Transferability of Adversarial Examples with Input Diversity

by C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille
Table of Contents

1. Introduction
2. Objectives
3. Transformations
4. Related Work
5. Methodology
   • Family of FGSM
   • New Proposed Methods of FGSM
   • Relationships
   • Attacking on Ensemble Networks
6. Experiment
   • Setup
   • Attacking a Single Network
   • Attacking an Ensemble of Networks
   • Ablation Studies
7. For
8. Against
9. Conclusion
Introduction

• **Black-Box setting**
  • Low success rate of adversarial attacks.
  • Single Step attacks perform better.
  • Poor transferability of adversarial examples because of underfitting.

• **White-Box setting**
  • Comparatively better success rate of adversarial attacks.
  • Iterative attacks perform better.
  • Poor transferability of adversarial examples because of overfitting.
Objectives

• Aims to improve transferability of adversarial examples.
• Create diverse input patterns.
• Apply random transformations to input images at each step.
• Test the strategy on several networks under both white-box and black-box settings, single-model and multi-model settings.
• Dataset used is ImageNet.
• Their enhanced attack reached an average success rate of 73%, which outperforms the top-1 attack in NIPS 2017 by 6.6%.
• Code is available at https://github.com/cihangxie/DI-2-FGSM
Transformations

- They also experimented with other image transformations, *e.g.*, rotation or flipping, to create diverse input patterns, and found random resizing & padding yields adversarial examples with the *best* transferability.
Related Work

- Szegedy et al. [36] proposed a box-constrained L-BFGS method.
  - Expensive computation.

  - Less expensive due to a single gradient step.

- Kurakin et al. [16] extended the method above to an iterative version.

- Dong et al. [9] proposed a class of momentum-based iterative algorithms.

- Liu et al. [21] proposed that transferability can also be improved by attacking an ensemble of networks simultaneously.
Methodology

Let $X$ be an image.

Let $y^{true}$ be the corresponding ground-truth label.

Let $\theta$ be network parameters.

$$L(X, y^{true}; \theta) = -1_{y^{true}} \cdot \log(\text{softmax}(l(X; \theta)))$$ is the loss.

**Goal:** to maximize the loss $L(X + r, y^{true}; \theta)$, where $X^{adv} = X + r$. 
Family of FGSM

- Fast Gradient Sign Method (FGSM):
  \[ X_{adv} = X + \epsilon \cdot \text{sign}(\nabla_X L(X, y^{true}; \theta)) \]

- Iterative Fast Gradient Sign Methods (I-FGSM):
  \[ X_{adv}^0 = X \]
  \[ X_{adv}^{n+1} = \text{Clip}_X \{ X_{adv}^n + \alpha \cdot \text{sign}(\nabla_X L(X_{adv}^n, y^{true}; \theta)) \} \]

- Momentum Iterative Fast Gradient Sign Method (MI-FGSM):
  \[ g_{n+1} = \mu \cdot g_n + \frac{\nabla_X L(X_{adv}^n, y^{true}; \theta)}{||\nabla_X L(X_{adv}^n, y^{true}; \theta)||_1} \]
  \[ X_{adv}^{n+1} = \text{Clip}_X \{ X_{adv}^n + \alpha \cdot \text{sign}(g_{n+1}) \} \]
Diverse Inputs Patterns Methods

➢ Diverse Inputs Iterative Fast Gradient Sign Method (DI^2-FGSM):

\[
X_{n+1}^{adv} = \text{Clip}_X \{ X_n^{adv} + \alpha \cdot \text{sign}(\nabla_X L(T(X_n^{adv}; p), y^{true}; \theta)) \}
\]

\[
T(X_n^{adv}; p) = \begin{cases} 
T(X_n^{adv}) & \text{with probability } p \\
X_n^{adv} & \text{with probability } 1 - p
\end{cases}
\]

➢ Momentum Diverse Inputs Iterative Fast Gradient Sign Method (M-DI^2-FGSM):

\[
g_{n+1} = \mu \cdot g_n + \frac{\nabla_X L(T(X_n^{adv}; p), y^{true}; \theta)}{||\nabla_X L(T(X_n^{adv}; p), y^{true}; \theta)||_1}
\]
Relationships between different attacks
Attacking on Ensemble Networks

To attack an ensemble of $K$ models, the logits are fused by

$$l(X; \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} w_k l_k(X; \theta_k)$$

$$w_k \geq 0$$

$$\sum_{k=1}^{K} w_k = 1$$
Experiment - Setup

- Dataset: ImageNet validation set (5000 images).
- Networks:
  - Inception-v3 (Inc-v3)
  - Inception-v4 (Inc-v4)
  - Resnet-v2-152 (Res-152)
  - Inception-Resnet-v2 (IncRes-v2)
- 3 adversarial trained networks:
  - ens3-adv-Inception-v3 (Inc-v3ens3)
  - ens4-adv-Inception-v3 (Inc-v3ens4)
  - ens-adv-Inception-ResNet-v2 (IncRes-v2ens)
- Step size: $\alpha = 1$ and $N = \min(\varepsilon + 4, 1.25\varepsilon)$
- Maximum perturbation $\varepsilon = 15$
- $\mu = 1$
- $\rho = 0.5$
- Input $X$ is randomly resized $\text{rnd} \times \text{rnd} \times 3$ image with $\text{rnd} \varepsilon [299; 330)$ and padded to the size $330 \times 330 \times 3$ in a random manner.
# Attacking on Single Networks

The success rates on seven networks (single network attack)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc-v3</td>
<td>FGSM</td>
<td>64.6%</td>
<td>23.5%</td>
<td>21.7%</td>
<td>21.7%</td>
<td>8.0%</td>
<td>7.5%</td>
<td>3.6%</td>
</tr>
<tr>
<td></td>
<td>I-FGSM</td>
<td>99.9%</td>
<td>14.8%</td>
<td>11.6%</td>
<td>8.9%</td>
<td>3.3%</td>
<td>2.9%</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>99.9%</td>
<td>35.5%</td>
<td>27.8%</td>
<td>21.4%</td>
<td>5.5%</td>
<td>5.2%</td>
<td>2.8%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>99.9%</td>
<td>36.6%</td>
<td>34.5%</td>
<td>27.5%</td>
<td>8.9%</td>
<td>8.4%</td>
<td>4.7%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>99.9%</td>
<td>63.9%</td>
<td>59.4%</td>
<td>47.9%</td>
<td>14.3%</td>
<td>14.0%</td>
<td>7.0%</td>
</tr>
<tr>
<td>Inc-v4</td>
<td>FGSM</td>
<td>26.4%</td>
<td>49.6%</td>
<td>19.7%</td>
<td>20.4%</td>
<td>8.4%</td>
<td>7.7%</td>
<td>4.1%</td>
</tr>
<tr>
<td></td>
<td>I-FGSM</td>
<td>22.0%</td>
<td>99.9%</td>
<td>13.2%</td>
<td>10.9%</td>
<td>3.2%</td>
<td>3.0%</td>
<td>1.7%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>43.3%</td>
<td>99.7%</td>
<td>28.9%</td>
<td>23.1%</td>
<td>5.9%</td>
<td>5.5%</td>
<td>3.2%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>51.1%</td>
<td>99.9%</td>
<td>39.4%</td>
<td>33.7%</td>
<td>11.2%</td>
<td>10.7%</td>
<td>5.3%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>72.4%</td>
<td>99.5%</td>
<td>62.2%</td>
<td>52.1%</td>
<td>17.6%</td>
<td>15.6%</td>
<td>8.8%</td>
</tr>
<tr>
<td>IncRes-v2</td>
<td>FGSM</td>
<td>24.3%</td>
<td>19.3%</td>
<td>39.6%</td>
<td>19.4%</td>
<td>8.5%</td>
<td>7.3%</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>I-FGSM</td>
<td>22.2%</td>
<td>17.7%</td>
<td>97.9%</td>
<td>12.6%</td>
<td>4.6%</td>
<td>3.7%</td>
<td>2.5%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>46.5%</td>
<td>40.5%</td>
<td>95.8%</td>
<td>28.6%</td>
<td>8.2%</td>
<td>6.6%</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>53.5%</td>
<td>45.9%</td>
<td>98.4%</td>
<td>37.8%</td>
<td>15.3%</td>
<td>13.0%</td>
<td>8.8%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>71.2%</td>
<td>67.4%</td>
<td>96.1%</td>
<td>57.4%</td>
<td>25.1%</td>
<td>20.7%</td>
<td>14.9%</td>
</tr>
<tr>
<td>Res-152</td>
<td>FGSM</td>
<td>34.4%</td>
<td>28.5%</td>
<td>27.1%</td>
<td>75.2%</td>
<td>12.4%</td>
<td>11.0%</td>
<td>6.0%</td>
</tr>
<tr>
<td></td>
<td>I-FGSM</td>
<td>20.8%</td>
<td>17.2%</td>
<td>14.9%</td>
<td>99.1%</td>
<td>5.4%</td>
<td>4.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>53.8%</td>
<td>49.0%</td>
<td>44.8%</td>
<td>99.2%</td>
<td>13.0%</td>
<td>11.1%</td>
<td>6.9%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>50.1%</td>
<td>44.1%</td>
<td>42.2%</td>
<td>99.0%</td>
<td>18.2%</td>
<td>15.2%</td>
<td>9.0%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>78.9%</td>
<td>76.5%</td>
<td>74.8%</td>
<td>99.2%</td>
<td>35.2%</td>
<td>29.4%</td>
<td>19.0%</td>
</tr>
</tbody>
</table>
Attacking a Single Network

Visualization of randomly selected clean images and their corresponding adversarial examples

<table>
<thead>
<tr>
<th>Model</th>
<th>Attack</th>
<th>Inc-v3</th>
<th>Inc-v4</th>
<th>IncRes-v2</th>
<th>Res-152</th>
<th>Inc-v3_{ens3}</th>
<th>Inc-v3_{ens4}</th>
<th>IncRes-v2_{ens}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc-v3</td>
<td>C&amp;W</td>
<td>100.0%</td>
<td>5.7%</td>
<td>5.3%</td>
<td>5.1%</td>
<td>3.0%</td>
<td>2.5%</td>
<td>1.1%</td>
</tr>
<tr>
<td></td>
<td>D-C&amp;W (Ours)</td>
<td>100.0%</td>
<td>16.8%</td>
<td>13.0%</td>
<td>11.2%</td>
<td>5.8%</td>
<td>3.9%</td>
<td>2.1%</td>
</tr>
<tr>
<td>Inc-v4</td>
<td>C&amp;W</td>
<td>15.1%</td>
<td>100.0%</td>
<td>9.2%</td>
<td>7.8%</td>
<td>4.4%</td>
<td>3.5%</td>
<td>1.9%</td>
</tr>
<tr>
<td></td>
<td>D-C&amp;W (Ours)</td>
<td>29.3%</td>
<td>100.0%</td>
<td>20.1%</td>
<td>15.4%</td>
<td>7.1%</td>
<td>5.3%</td>
<td>3.1%</td>
</tr>
<tr>
<td>IncRes-v2</td>
<td>C&amp;W</td>
<td>15.8%</td>
<td>11.2%</td>
<td>99.9%</td>
<td>8.6%</td>
<td>6.3%</td>
<td>3.6%</td>
<td>3.4%</td>
</tr>
<tr>
<td></td>
<td>D-C&amp;W (Ours)</td>
<td>33.9%</td>
<td>25.6%</td>
<td>100.0%</td>
<td>19.4%</td>
<td>11.2%</td>
<td>7.3%</td>
<td>4.0%</td>
</tr>
<tr>
<td>Res-152</td>
<td>C&amp;W</td>
<td>11.4%</td>
<td>6.9%</td>
<td>6.1%</td>
<td>100.0%</td>
<td>4.4%</td>
<td>4.1%</td>
<td>2.3%</td>
</tr>
<tr>
<td></td>
<td>D-C&amp;W (Ours)</td>
<td>33.0%</td>
<td>27.7%</td>
<td>24.4%</td>
<td>100.0%</td>
<td>13.1%</td>
<td>9.3%</td>
<td>5.7%</td>
</tr>
</tbody>
</table>

The success rates on seven networks where we attack a single network using C&W attack.
Attacking a Ensemble of Network

- IncRes-V2-ens reaches max success of only 19.0%.

- Attacking group of networks simultaneously to improve transferability.

- Adversarial examples are generated on ensemble of 6 networks.

- Tested on ensembled network (white-box setting) and hold-out network (black-box setting).

- FGSM attack is not used due to its low success rates on white-box models.

- All ensembled models are assigned with equal weight.
## Attacking a Ensemble of Network

<table>
<thead>
<tr>
<th>Model</th>
<th>Attack</th>
<th>-Inc-v3</th>
<th>-Inc-v4</th>
<th>-IncRes-v2</th>
<th>-Res-152</th>
<th>-Inc-v3_ens3</th>
<th>-Inc-v3_ens4</th>
<th>-IncRes-v2_ens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td>I-FGSM</td>
<td>96.6%</td>
<td>96.9%</td>
<td>98.7%</td>
<td>96.2%</td>
<td>97.0%</td>
<td>97.3%</td>
<td>94.3%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>88.9%</td>
<td>89.6%</td>
<td>93.2%</td>
<td>87.7%</td>
<td>91.7%</td>
<td>91.7%</td>
<td>93.2%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>96.9%</td>
<td>96.9%</td>
<td>98.8%</td>
<td>96.8%</td>
<td>96.8%</td>
<td>97.0%</td>
<td>94.6%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>90.1%</td>
<td>91.1%</td>
<td>94.0%</td>
<td>89.3%</td>
<td>92.8%</td>
<td>92.7%</td>
<td>94.9%</td>
</tr>
<tr>
<td>Hold-out</td>
<td>I-FGSM</td>
<td>43.7%</td>
<td>36.4%</td>
<td>33.3%</td>
<td>25.4%</td>
<td>12.9%</td>
<td>15.1%</td>
<td>8.8%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>69.9%</td>
<td>67.9%</td>
<td>64.1%</td>
<td>51.7%</td>
<td>36.3%</td>
<td>35.0%</td>
<td>30.4%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>71.4%</td>
<td>65.9%</td>
<td>64.6%</td>
<td>55.6%</td>
<td>22.8%</td>
<td>26.1%</td>
<td>15.8%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>80.7%</td>
<td>80.6%</td>
<td>80.7%</td>
<td>70.9%</td>
<td>44.6%</td>
<td>44.5%</td>
<td>39.4%</td>
</tr>
</tbody>
</table>

The success rates of ensemble attacks

<table>
<thead>
<tr>
<th>IncRes-v2</th>
<th>FGSM</th>
<th>24.3%</th>
<th>19.3%</th>
<th>39.6%</th>
<th>19.4%</th>
<th>8.5%</th>
<th>7.3%</th>
<th>4.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I-FGSM</td>
<td>22.2%</td>
<td>17.7%</td>
<td>97.9%</td>
<td>12.6%</td>
<td>4.6%</td>
<td>3.7%</td>
<td>2.5%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>46.5%</td>
<td>40.5%</td>
<td>95.8%</td>
<td>28.6%</td>
<td>8.2%</td>
<td>6.6%</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>53.5%</td>
<td>45.9%</td>
<td>98.4%</td>
<td>37.8%</td>
<td>15.3%</td>
<td>13.0%</td>
<td>8.8%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>71.2%</td>
<td>67.4%</td>
<td>96.1%</td>
<td>57.4%</td>
<td>25.1%</td>
<td>20.7%</td>
<td>14.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Res-152</th>
<th>FGSM</th>
<th>34.4%</th>
<th>28.5%</th>
<th>27.1%</th>
<th>75.2%</th>
<th>12.4%</th>
<th>11.0%</th>
<th>6.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I-FGSM</td>
<td>20.8%</td>
<td>17.2%</td>
<td>14.9%</td>
<td>99.1%</td>
<td>5.4%</td>
<td>4.6%</td>
<td>2.8%</td>
</tr>
<tr>
<td></td>
<td>DI²-FGSM (Ours)</td>
<td>53.8%</td>
<td>49.0%</td>
<td>44.8%</td>
<td>99.2%</td>
<td>13.0%</td>
<td>11.1%</td>
<td>6.9%</td>
</tr>
<tr>
<td></td>
<td>MI-FGSM</td>
<td>50.1%</td>
<td>44.1%</td>
<td>42.2%</td>
<td>99.0%</td>
<td>18.2%</td>
<td>15.2%</td>
<td>9.0%</td>
</tr>
<tr>
<td></td>
<td>M-DI²-FGSM (Ours)</td>
<td>78.9%</td>
<td>76.5%</td>
<td>74.8%</td>
<td>99.2%</td>
<td>35.2%</td>
<td>29.4%</td>
<td>19.0%</td>
</tr>
</tbody>
</table>

Best with single network attack
Ablation Studies

• Claim: White-Box success rates will improve if:
  • Transformation probability is smaller.
  • Increase total number of iterations, or
  • Using a smaller step size.

• Ablation Studies is explaining of how different parameters affect the success rates.

• Only considering attacking an ensemble of networks.

• Max Perturbation is still set to 15.
Ablation Studies

Transformation probability:

The success rates of DI^2-FGSM

The success rates of M-DI^2-FGSM
Ablation Studies

Total iteration number:

The success rates of DI^2-FGSM

The success rates of M-DI^2-FGSM
Ablation Studies

Step Size:

The success rates of DI^2-FGSM

The success rates of M-DI^2-FGSM
NIPS 2017 adversarial competition

<table>
<thead>
<tr>
<th>Attack</th>
<th>TsAIL</th>
<th>iyswim</th>
<th>Anil Thomas</th>
<th>Inc-v3_{adv}</th>
<th>IncRes-v2_{ens}</th>
<th>Inc-v3</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-FGSM</td>
<td>14.0%</td>
<td>35.6%</td>
<td>30.9%</td>
<td>98.2%</td>
<td>96.4%</td>
<td>99.0%</td>
<td>62.4%</td>
</tr>
<tr>
<td>D1^2-FGSM (Ours)</td>
<td>22.7%</td>
<td>58.4%</td>
<td>48.0%</td>
<td>91.5%</td>
<td>90.7%</td>
<td>97.3%</td>
<td>68.1%</td>
</tr>
<tr>
<td>MI-FGSM</td>
<td>14.9%</td>
<td>45.7%</td>
<td>46.6%</td>
<td>97.3%</td>
<td>95.4%</td>
<td>98.7%</td>
<td>66.4%</td>
</tr>
<tr>
<td>MI-FGSM*</td>
<td>13.6%</td>
<td>43.2%</td>
<td>43.9%</td>
<td>94.4%</td>
<td>93.0%</td>
<td>97.3%</td>
<td>64.2%</td>
</tr>
<tr>
<td>M-D1^2-FGSM (Ours)</td>
<td>20.0%</td>
<td>69.8%</td>
<td>64.4%</td>
<td>93.3%</td>
<td>92.4%</td>
<td>97.9%</td>
<td>73.0%</td>
</tr>
</tbody>
</table>

The comparison of success rates using three different attacks

The success rates on top defense solutions and official baselines from NIPS 2017 adversarial competition
For

- Proposed methods improve the transferability of adversarial examples attacking a single network.
  - DI^2-FGSM improves the success rates of I-FGSM on black-box models and maintains high success rates on white-box models.
  - M-DI^2-FGSM outperforms all attacks on all black-box models and maintains high success rates on all white-box models.

- Proposed methods improve the transferability of adversarial examples attacking an ensemble of networks.
  - DI^2-FGSM improves the success rates of I-FGSM on black-box models and maintains high success rates on white-box models.
  - M-DI^2-FGSM outperforms all attacks on all black-box models and maintains high success rates on all white-box models.

- Proposed method M-DI^2-FGSM reached an average success rate of 73% which outperforms top-1 attack at NIPS 2017 competition.
Against

- Same dataset ImageNet used for all experimental procedures.
- Would be interesting to see the same methodology over more than just C&W and FGSM.
- Tabular data could be represented in better sub-tables.
- Claimed after looking at general trend that their white-box attack can perform better with altered parameters.
- Proposed solution is computationally expensive.
- No data provided about the count of ensemble networks and how that figure could alter the results.
Conclusion

• Aims to improve the transferability of adversarial examples with input diversity.

• Applies random transformations to the input images at each iteration in the attack process.

• Compared with traditional iterative attacks, the results on ImageNet show that proposed attack method gets significantly higher success rates for black-box models and maintains similar success rates for white-box models.

• This enhanced attack reaches an average success rate of 73.0%.

• Proposed attack strategy can serve as a benchmark for evaluating the robustness of networks to adversaries and the effectiveness of different defense methods in future.

• Code is publicly available at https://github.com/cihangxie/DI-2-FGSM.
Thank You!

Questions?