Delving into Transferable Adversarial Examples and Black-box Attacks

Yanpei Liu, Xinyun Chen ¹ Chang Liu, Dawn Song ²

¹Shanghai JiaoTong University
²University of the California, Berkeley

ICLR 2017/ Presenter: Ji Gao
Outline

1 Motivation
   • Transferability

2 Previous methods

3 Experiment I

4 Method

5 Experiment II

6 Experiment III: On realistic data
Motivation

- Adversarial examples: Samples are close to the normal seeds but are misclassified by the deep model

![Image of panda and gibbon with adversarial noise](image)

Label: Panda + 0.007 = Label: Gibbon

- Adversarial samples can transfer between models: Adversarial samples generated to fool one model can be also applied to another models.
Motivation: Transferability

- Transferability of the adversarial samples: The ability for adversarial samples generated can be transferred to another model.
- Research problem: How to represent transferability? How can we generate samples with more?
- If we can generate transferable sample, it means that we can attack a black-box model easily by generating adversarial sample on any model.
Targeted and non-targeted attack

- Targeted: Have an objective label.
- Non-targeted: Don’t have an objective label, just want to mislead the model to a different label.

Non-targeted samples have been shown easier to transfer.

This paper focus on targeted attack.

What if for the two models, the labels are different?
Formally, adversarial sample can be defined as:

\[
d(x, x^*) \leq B, \quad x^* \in \mathbb{X} \quad F(x) \neq F(x^*)
\]
The Fast Gradient Sign algorithm is one efficient algorithm for creating adversarial samples. Equation:

$$\Delta x = \epsilon \text{sign}(\nabla_x \text{Loss}(F(x^*; \theta), F(x; \theta)))$$ \hspace{1cm} (2)

This is motivated by controlling the $L_\infty$ norm of $\Delta x$, i.e. perturbation in each feature dimension to be the same.

If the sign function is not used, it becomes another method on $L_2$ norm. It is called fast gradient method in this paper.
Previous solutions: Optimization-based approach

L2 attack (Carlini & Wagner (2016))

\[ \arg \min_{x^*} \lambda d(x, x^*) + \text{Loss}(F(x; \theta)) \] (3)

- This approach is more intuitive and accurate, but costly.
Targeted attack

- Previous method is for non-targeted attack, but it’s easy to change to targeted attack.
- Simply use a different loss function:

\[
\arg \min_{x^*} \lambda d(x, x^*) + \text{Loss}_{y^*}(F(x; \theta))
\] (4)
Experiment design:

- **Data**: Imagenet 2012.
- **Metric of transferability**:
  - Non-targeted: Accuracy
  - Targeted: The number of samples predicted exactly the target class (on the transfered model), they called matching rate.
- **Bound of distortion**: RMSD

\[ \sqrt{\sum_i (x_i^* - x_i)^2 / N} \]
Experiment of non-targeted adversarial samples

Transfers well with a large distortion.
FGS result is worse than FG and optimization-based approaches.
Minimal transferable RMSD

(a) Fast Gradient

(b) Fast Gradient Sign
Table 2: The matching rate of targeted adversarial images generated using the optimization-based approach. The first column indicates the average RMSD of the generated adversarial images. Cell $(i, j)$ indicates that matching rate of the targeted adversarial images generated for model $i$ (row) when evaluated on model $j$ (column). The top-5 results can be found in the appendix (Table 12).

Doesn’t transfer well.
• Generate adversarial images for the ensemble of the (five) models.

**Ensemble attack**

\[
\arg\min_{x^*} \lambda d(x, x^*) + \text{Loss} \left( \sum_i \alpha_i F_i(x; \theta) \right) \tag{5}
\]
For each of the five models, we treat it as the black-box model to attack, and generate adversarial images for the ensemble of the rest four, which is considered as white-box.

<table>
<thead>
<tr>
<th></th>
<th>RMSD</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>30.68</td>
<td>38%</td>
<td>76%</td>
<td>70%</td>
<td>97%</td>
<td>76%</td>
</tr>
<tr>
<td>-ResNet-101</td>
<td>30.76</td>
<td>75%</td>
<td>43%</td>
<td>69%</td>
<td>98%</td>
<td>73%</td>
</tr>
<tr>
<td>-ResNet-50</td>
<td>30.26</td>
<td>84%</td>
<td>81%</td>
<td>46%</td>
<td>99%</td>
<td>77%</td>
</tr>
<tr>
<td>-VGG-16</td>
<td>31.13</td>
<td>74%</td>
<td>78%</td>
<td>68%</td>
<td>24%</td>
<td>63%</td>
</tr>
<tr>
<td>-GoogLeNet</td>
<td>29.70</td>
<td>90%</td>
<td>87%</td>
<td>83%</td>
<td>99%</td>
<td>11%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>RMSD</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>17.17</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-ResNet-101</td>
<td>17.25</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-ResNet-50</td>
<td>17.25</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>-VGG-16</td>
<td>17.80</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>-GoogLeNet</td>
<td>17.41</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>5%</td>
</tr>
</tbody>
</table>
Observations

- Several observations found:
- The gradient directions of different models in our evaluation are almost orthogonal to each other.
- Decision boundaries of the non-targeted approaches using different models align well.

Figure 4: The decision boundary to separate the region within which all points are classified as the ground truth label (encircled by each closed curve) from others. The plane is the same one described in Figure 3. The origin of the coordinate plane corresponds to the original image. The units of both axises are 1 pixel values.

Figure 5: The decision boundary to separate the region within which all points are classified as the target label (encircled by each closed curve) from others. The plane is spanned by the targeted adversarial direction and a random orthogonal direction. The targeted adversarial direction is computed as the difference between the original image in Figure 2 and the adversarial image generated by the optimization-based approach for an ensemble. The ensemble contains all models except ResNet-101. The origin of the coordinate plane corresponds to the original image. The units of both axises are 1 pixel values.
On realistic data

- Clarifai.com: A commercial company providing state-of-the-art image classification services
- Submit adversarial samples to the website.
- Result:
  - Non-targeted: 57% of the targeted adversarial examples generated using VGG-16, and 76% of the ones generated using the ensemble can mislead Clarifai.com to predict labels irrelevant to the ground truth
  - Targeted: 18% of those generated using the ensemble model can be predicted as labels close to the target label by Clarifai.com. The corresponding number for the targeted adversarial examples generated using VGG-16 is 2%.
## Some adversarial samples

<table>
<thead>
<tr>
<th>original image</th>
<th>true label</th>
<th>Clarifai.com results of original image</th>
<th>target label</th>
<th>targeted adversarial example</th>
<th>Clarifai.com results of targeted adversarial example</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="viaduct" /></td>
<td>viaduct</td>
<td>bridge, sight, arch, river, sky</td>
<td>window screen</td>
<td><img src="image" alt="window screen" /></td>
<td>window, wall, old, decoration, design</td>
</tr>
<tr>
<td><img src="image" alt="hip, rose hip, rosehip" /></td>
<td>hip, rose hip, rosehip</td>
<td>fruit, fall, food, little, wildlife</td>
<td>stupa, tope</td>
<td><img src="image" alt="stupa, tope" /></td>
<td>Buddha, gold, temple, celebration, artistic</td>
</tr>
<tr>
<td><img src="image" alt="dogsled, dog sled, dog sleigh" /></td>
<td>dogsled, dog sled, dog sleigh</td>
<td>group together, four, sledge, sled, enjoyment</td>
<td>hip, rose hip, rosehip</td>
<td><img src="image" alt="hip, rose hip, rosehip" /></td>
<td>cherry, branch, fruit, food, season</td>
</tr>
<tr>
<td><img src="image" alt="pug, pug-dog" /></td>
<td>pug, pug-dog</td>
<td>pug, friendship, adorable, purebred, sit</td>
<td>sea lion</td>
<td><img src="image" alt="sea lion" /></td>
<td>sea seal, ocean, head, sea, cute</td>
</tr>
</tbody>
</table>