Fine-grained Generation of Unrestricted Adversarial Examples

Omid Poursaeed
Adversarial Image Manipulation

- Look like real images
- Misclassified by the model
Similarity of Images

$L_p$ similarity: $||x - \hat{x}||_p < \epsilon$

$p = 0, 2, \infty, \ldots$
Manifold of Natural Images
Manifold of Natural Images

\[ \mathcal{M} : \text{Data Manifold} \]

\[ D : \text{Classifier's decision boundary} \]
Manifold of Natural Images

Classifiers equipped with defense

Larger perturbation norms
Unrestricted Adversarial Examples

Can we move on the manifold?
Unrestricted Adversarial Examples

Using a generative model to approximate the manifold

\[ G_\theta(Z) \simeq \mathcal{M} \]

\[ Z \sim U[0, 1]^k \]

\( G_\theta \) : Generator

\( \mathcal{M} \) : Data Manifold
Unrestricted Adversarial Examples

Disentangled Latent Space

\[ \mathcal{G}_\theta(y, \eta) \simeq \mathcal{M} \]
Unrestricted Adversarial Examples
Unrestricted Adversarial Examples

Iteratively updating the variables

\[ y_{\text{adv}}^{(t+1)} = y_{\text{adv}}^{(t)} - \epsilon \cdot \text{sign}(\nabla_{y_{\text{adv}}^{(t)}} J(F(g(y_{\text{adv}}^{(t)}, \eta_{\text{adv}}^{(t)}), T)) \]

\[ \eta_{\text{adv}}^{(t+1)} = \eta_{\text{adv}}^{(t)} - \delta \cdot \text{sign}(\nabla_{\eta_{\text{adv}}^{(t)}} J(F(g(y_{\text{adv}}^{(t)}, \eta_{\text{adv}}^{(t)}), T)) \]
Fine-grained Unrestricted Adversarial Examples

Only manipulating specific layers

Top layers: high-level changes

Bottom layers: low-level changes
Fine-grained Unrestricted Adversarial Examples

Results on LSUN: Non-targeted
Fine-grained Unrestricted Adversarial Examples

Results on LSUN: Targeted
Fine-grained Unrestricted Adversarial Examples

Results on CelebA-HQ Gender Classification
Evaluation on Certified Defenses

Certified defenses exist on norm-bounded attacks

- Vulnerable to our unrestricted attack

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>63.1%</td>
</tr>
<tr>
<td>Adversarial (style)</td>
<td>21.7%</td>
</tr>
<tr>
<td>Adversarial (noise)</td>
<td>37.8%</td>
</tr>
</tbody>
</table>

Table 1: Accuracy of a certified classifier equipped with randomized smoothing on adversarial images.
Adversarial Training

Including adversarial images in training the classifier

- Effective as a defense
- Does not decrease performance on clean images

<table>
<thead>
<tr>
<th></th>
<th>Adv. Trained</th>
<th>Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>87.6%</td>
<td>87.7%</td>
</tr>
<tr>
<td>Adversarial (noise)</td>
<td>81.2%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Adversarial (style)</td>
<td>76.9%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 2: Accuracy of adversarially trained and original classifiers on clean and adversarial test images.
User Study

Real or Fake?

- Accuracy on un-adversarial generated images: 74.7%
- Accuracy on style-based adversarial images: 70.8%
- Accuracy on noise-based adversarial images: 74.3%

Correct category?

- Accuracy on style-based images: 98.7%
- Accuracy on noise-based images: 99.2%