Robustness and Generalization via Generative Adversarial Training

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

$\mathcal{M} : \text{Data Manifold}$
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

\[ G_\theta(Z) \approx \mathcal{M} \quad \quad \quad G_\theta : \text{Generator} \]

\[ Z \sim U[0, 1]^k \quad \quad \quad \mathcal{M} : \text{Data Manifold} \]
Unrestricted Adversarial Examples

Disentangled Latent Space

$G_\theta(y, \eta) \simeq M$
Unrestricted Adversarial Examples
Unrestricted Adversarial Examples

Iteratively updating the variables

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

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

Including adversarial images in training the classifier

- Effective as a defense
- Improves performance on clean images

<table>
<thead>
<tr>
<th></th>
<th>Classification (LSUN)</th>
<th>Classification (CelebA-HQ)</th>
<th>Segmentation</th>
<th>Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clean</td>
<td>Adversarial</td>
<td>Clean</td>
<td>Adversarial</td>
</tr>
<tr>
<td>Adv. Trained</td>
<td>89.5%</td>
<td>78.4%</td>
<td>96.2%</td>
<td>83.6%</td>
</tr>
<tr>
<td>Original</td>
<td>88.9%</td>
<td>0.0%</td>
<td>95.7%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
## Adversarial Training

<table>
<thead>
<tr>
<th>Model</th>
<th>Clean</th>
<th>GAT</th>
<th>PGD</th>
<th>Spatial</th>
<th>Recolor</th>
<th>Perceptual</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAT (Ours)</td>
<td>89.5%</td>
<td>78.4%</td>
<td>39.4%</td>
<td>47.8%</td>
<td>52.3%</td>
<td>28.9%</td>
<td>42.1%</td>
</tr>
<tr>
<td>AT PGD [27]</td>
<td>81.2%</td>
<td>6.3%</td>
<td>56.7%</td>
<td>5.1%</td>
<td>37.9%</td>
<td>2.8%</td>
<td>13.0%</td>
</tr>
<tr>
<td>AT AdvProp [37]</td>
<td>89.4%</td>
<td>7.8%</td>
<td>57.6%</td>
<td>6.0%</td>
<td>38.5%</td>
<td>3.5%</td>
<td>22.7%</td>
</tr>
<tr>
<td>AT Spatial [36]</td>
<td>76.3%</td>
<td>5.4%</td>
<td>3.1%</td>
<td>66.0%</td>
<td>4.1%</td>
<td>2.2%</td>
<td>3.7%</td>
</tr>
<tr>
<td>AT Recolor [24]</td>
<td>88.6%</td>
<td>4.7%</td>
<td>7.3%</td>
<td>0.4%</td>
<td>60.7%</td>
<td>1.7%</td>
<td>3.5%</td>
</tr>
<tr>
<td>PAT [25]</td>
<td>72.4%</td>
<td>18.3%</td>
<td>40.1%</td>
<td>46.3%</td>
<td>42.5%</td>
<td>30.1%</td>
<td>36.5%</td>
</tr>
</tbody>
</table>
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%
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.