3. Private Fast Gradient Sign Method (P-FGSM)

adversarial image

\[ \hat{x} = x + \delta^* \]

original image

adversarial perturbation

privacy protection \[ M(x) \neq M(x) \]

unnoticeability \[ ||\hat{x} - x|| \to 0 \]

irreversibility the true class or \( M(x) \) cannot be deduced from \( M(x) \)

P-FGSM: iterative adversarial perturbation generation

\[ \hat{x}_0 = x \]

cost function of \( M \) parameters of \( M \)

\[ \hat{x}_N = \hat{x}_{N-1} - \varepsilon \text{sign}(\nabla_{\theta} M(\theta, \hat{x}_{N-1}, \hat{y})) \]

magnitude of perturbation target class

\[ \hat{y} = \hat{y}\left(\sum_{i=1}^{N} p_i > \sigma\right) \]

random selection function set of target candidate classes

Proposed target class \( \hat{y} \) selection

from classes with cumulative probability > threshold \( \sigma \)

avoid targeting true class even when \( M \) is incorrect

4. Experiments

Dataset: Mediaeval 2018 Pixel Privacy Challenge [3]

Classifier: ResNet50 365-class classifier

Preprocessing: resize to 224×224 pixels with bilinear interpolation

Parameters: \( \sigma = 0.99; \varepsilon = 0.007 \)

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%) \dagger</th>
<th>PSNR</th>
<th>BRISQUE \S</th>
<th>Euclidean distance\S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1</td>
<td>56.40</td>
<td>86.47</td>
<td>-</td>
<td>26.72</td>
</tr>
<tr>
<td>Top-5</td>
<td>8.83</td>
<td>23.00</td>
<td>40.62</td>
<td>4.75</td>
</tr>
<tr>
<td>N-FGSM</td>
<td>0.17</td>
<td>7.00</td>
<td>40.24</td>
<td>2.87</td>
</tr>
<tr>
<td>R-FGSM</td>
<td>0.00</td>
<td>5.60</td>
<td>39.99</td>
<td>2.30</td>
</tr>
<tr>
<td>L-FGSM</td>
<td>0.00</td>
<td>5.60</td>
<td>39.99</td>
<td>2.30</td>
</tr>
<tr>
<td>P-FGSM</td>
<td>0.00</td>
<td>5.60</td>
<td>39.99</td>
<td>2.30</td>
</tr>
</tbody>
</table>

\* between discrete uniform distribution and average discrete distribution of target class

\dagger the smaller the better; \S the larger the better

5. Conclusions

P-FGSM: protects privacy against automatic inference

by generating corresponding adversarial images

misleads ResNet50 (always in its top-1 and 94.40% of the times in its top-5)

higher degree of irreversibility compared to N-FGSM and L-FGSM

comparable visual quality with other FGSMs

References