Cost-Effective Adversarial Attacks against Scene Text Recognition
Mingkun Yang¹, Haitian Zheng², Xiang Bai¹, Jiebo Luo²
¹Huazhong University of Science and Technology, ²University of Rochester

Introduction

Fig. 1: Illustration of adversarial attack on scene text recognition: given an image, adding an imperceptible perturbation generated by our method makes the prediction different from the original one. Red characters are mis-recognized characters.

➢ Deep neural networks are vulnerable to adversarial attacks.
➢ But there has been no work on adversarial attacks against scene text recognition.

Contribution

➢ The first attempt to investigate the adversarial effects in scene text recognition with attention-based methods.
➢ A novel and effective optimization-based adversarial attack objective function specifically designed for untargeted adversarial attacks against sequential tasks.
➢ Generating adversarial examples with less perceptible perturbation and an even higher attack success rate.

Methodology

Adversarial Attacks for Non-sequential Tasks

\[
\min_\delta \mathcal{D}(I, I + \delta) + c \cdot \mathcal{L}(I + \delta, y') \\
\text{s.t. } I + \delta \in [-1, 1],
\]

Where \( \delta \) is the adversarial perturbation, \( y' \) is the predicted label which is different from the true label \( y \), and \( c \) is the weight to trade off the relative importance of being adversarial and remaining close to the original example. Generally, \( \mathcal{L}(I + \delta, y') = \max(q^y - \max_{y' \neq y} q^{y'}, 0) \).

Adversarial Attacks for Scene Text Recognition

\[
\mathcal{L}(I + \delta, y') = S \cdot H \cdot \sum_{t=1}^{T} \max(q^y_i - \max_{y' \neq y} q^{y'}_i, 0) \\
H = \text{sigmoid}(k \cdot \min_{t} \{d_t\}), \\
S = \prod_{t=1}^{T} p(y_t),
\]

Where \( S \) is the recognition score. The higher \( S \) is, the more difficult to attack, and the larger \( L \) should be. When \( k \) is large enough, \( H \) is a step function and ensures that once any one character is successfully attacked.

Experiments

Fig. 2: Adversarial examples generated by various attack methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>HITTk</th>
<th>SR</th>
<th>e2 Dist</th>
<th>SVT</th>
<th>e2 Dist</th>
<th>ICO3</th>
<th>e2 Dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGD</td>
<td>83.10</td>
<td>1.77</td>
<td>93.97</td>
<td>1.66</td>
<td>80.05</td>
<td>1.76</td>
<td></td>
</tr>
<tr>
<td>MI-FGSM</td>
<td>17.42</td>
<td>3.55</td>
<td>80.98</td>
<td>2.83</td>
<td>57.29</td>
<td>3.98</td>
<td></td>
</tr>
<tr>
<td>FGSM</td>
<td>25.10</td>
<td>75.45</td>
<td>52.70</td>
<td>37.60</td>
<td>33.21</td>
<td>59.33</td>
<td></td>
</tr>
<tr>
<td>o GMC</td>
<td>99.93</td>
<td>1.89</td>
<td>100.00</td>
<td>1.43</td>
<td>100.00</td>
<td>1.87</td>
<td></td>
</tr>
<tr>
<td>with ( S )</td>
<td>99.87</td>
<td>1.77</td>
<td>100.00</td>
<td>1.27</td>
<td>100.00</td>
<td>1.77</td>
<td></td>
</tr>
<tr>
<td>with ( H )</td>
<td>99.93</td>
<td>1.75</td>
<td>100.00</td>
<td>1.26</td>
<td>100.00</td>
<td>1.71</td>
<td></td>
</tr>
<tr>
<td>Ours (with ( S ) and ( H ))</td>
<td>99.93</td>
<td>1.66</td>
<td>100.00</td>
<td>1.17</td>
<td>100.00</td>
<td>1.69</td>
<td></td>
</tr>
</tbody>
</table>

Tab. I: Performance of adversarial attacks over several methods and variants.

References