Cost-Effective Adversarial Attacks against Scene Text Recognition

Mingkun Yang[1], Haitian Zheng[2], Xiang Bai[1], Jiebo Luo[2]

[1]Huazhong University of Science and Technology, [2]University of Rochester

Introduction

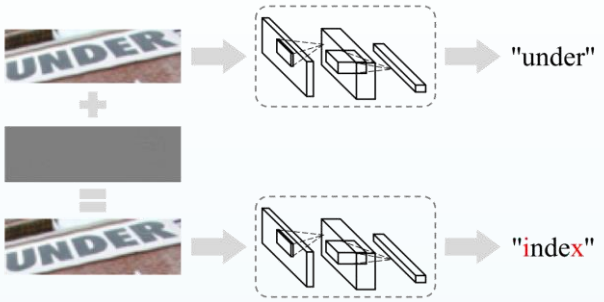

Fig. 1: Illustration of adversarial attack on scene text recognition: given an image, adding an imperceptible perturbation generated by our method makes the prediction different from the original one. Red characters are mis-recognized characters.

- 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

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

Where δ 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

$$\mathcal{L}(I+\delta, y') = S \cdot H \cdot \sum_{t=1}^{T} \max(q_t^{y_t} - \max_{y'_t \neq y_t} q_t^{y'_t}, 0)$$

$$H = \text{sigmoid}(k \cdot \min_{t}\{d_t\}),$$

$$S = \prod_{t=1}^{T} p(y_t),$$

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.

Methods	IIIT5k		SVT		IC03	
	SR	ℓ_2 Dist	SR	ℓ_2 Dist	SR	ℓ_2 Dist
PGD	83.10	1.77	93.97	1.66	80.05	1.76
MI-FGSM	17.42	3.55	80.98	2.83	57.20	3.98
FGSM	25.10	75.45	52.70	37.60	33.21	59.33
none	99.93	1.89	100	1.43	100	1.87
with S	99.87	1.77	100	1.27	100	1.77
with H	99.93	1.75	100	1.26	100	1.71
Ours (with S and H)	**99.93**	**1.66**	**100**	**1.17**	**100**	**1.69**

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

Experiments

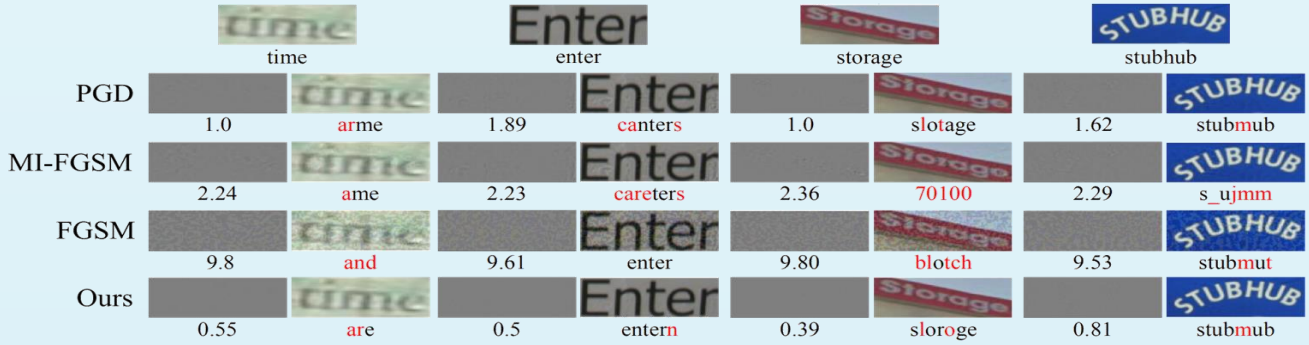

Fig. 2: Adversarial examples generated by various attack methods.

References

[1] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," IEEE Symposium on Security and Privacy, 2017.
[2] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," IEEE TPAMI, 2018