Towards undetectable adversarial examples: a steganographic perspective

Hui Zeng, Biwei Chen, Rongsong Yang, Chenggang Li, and Anjie Peng

No Institute Given

The supplementary document consists of three parts of content: A) More results of attack ability comparison; B) Additional examples; C) Ablation study on the distortion function.

1 More results of attack ability comparison

We launch untargeted attacks against a Resnet50 model. These attacked images are then evaluated with the source model (Fig. 1(a, b)) and four target models: Dense121 (Fig. 1(c, d)), VGG16 (Fig. 1(e, f)), Resnet50adv (Fig. 1(g, h)), and InceptionV3adv (Fig. 1(i, j)).

2 Additional examples

We display several AEs crafted by different methods with ϵ=16 in Fig. 2. The proposed 'W-' and 'WA-' schemes tend to add adversarial perturbations in rich texture regions that are hard for humans to perceive.

3 Ablation study on the distortion function

Is the proposed method sensitive to the embedding suitability map? To answer this question, in Fig. 3, we draw the undetectability-attack ability curves for the proposed WA-IFGSM attack when different embedding suitability maps are used. To facilitate comparison, the curve of the baseline IFGSM is redrawn. 'WA-IFGSM (WOW)' and 'WA-IFGSM (MiPOD)' correspond to the curves of WOW [1] and MiPOD [2], respectively, when they are used to calculate embedding suitability maps. The proposed WA-IFGSM attack performs consistently for various embedding suitability maps, and is way superior to the baseline IFGSM.

References

1. Holub V., Fridrich J.: Designing steganographic distortion using directional filters. In: IEEE Workshop on Information Forensic and Security, pp. 234–239 (2012).
2. Sedighi V., Cogranne R. and Fridrich J.: Content-adaptive steganography by minimizing statistical detectability. IEEE Trans. Info. Forensics and Security, 11(2): 221–234 (2016).

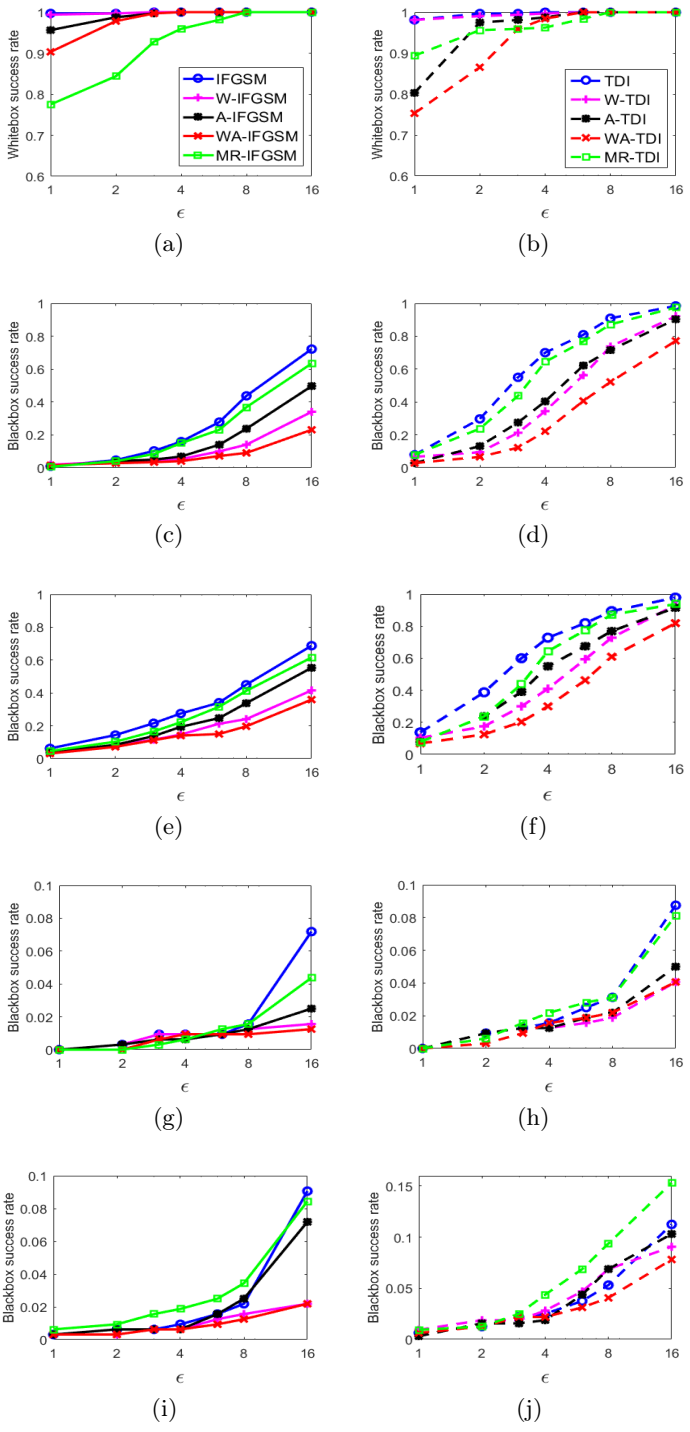

Fig. 1. ASR against different models. (a, b) Resnet50 (white-box), (c, d) Densenet121, (e, f) VGG16, (g, h) Resnet50adv, (i, j) InceptionV3adv. Fig. 5(a), (c), (e), (g) and (i) share a legend, and (b), (d), (f), (h) and (j) share another legend for better visualization.

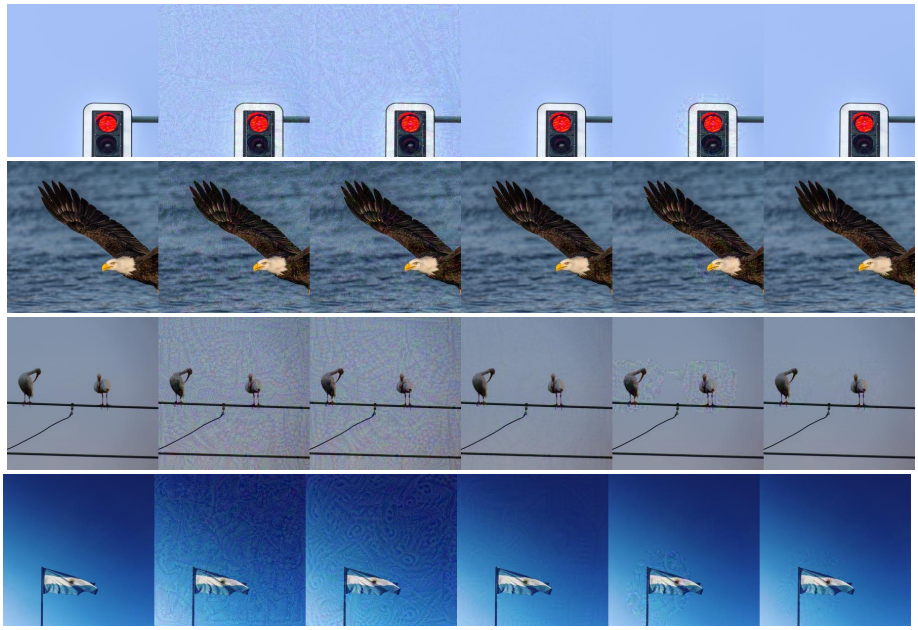

Fig. 2. Visual comparison of AEs, ϵ=16. The baseline attack is IFGSM for the top two rows, and TDI for the bottom two. For each row, from left to right, we show in order the original image, the adversarial images generated by the baseline attack, the 'MR-' attack, the 'W-' attack, the 'A-' attack, and the 'WA-' attack.

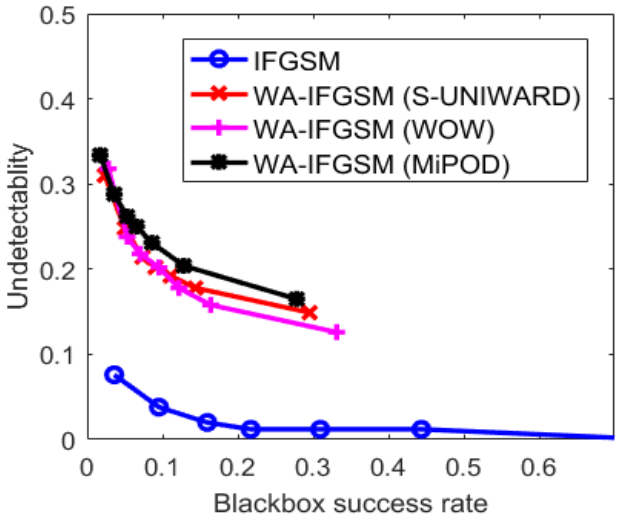

Fig. 3. Undetectability-attack ability curves for different embedding suitability maps.