Targeted attentional adversarial attack: supplementary material

Hui Zeng, Kun Yu, Biwei Chen, and Anjie Peng

In this supplementary document, we provide:

1) Our used subset of the ImageNet (Table 1).

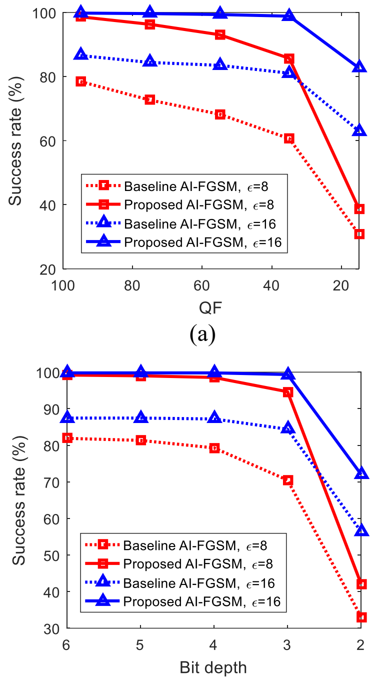
2) Comparison of attack ability (Table 2) and robustness (Fig. 1) when Grad-CAM [1] is adopted for generating attentional maps.

As observed from Table 2, the proposed method triumphs the baseline in all cases. Note the white-box success rate of baseline AI-FGSM is only 87.45% when ϵ=16. This is because the Grad-CAM may generate an all-zero attentional map for a low-confidence label. In this case, the baseline attack is doomed to fail. On the contrary, the proposed attack is always guided with high-confidence attentional maps, thus free of such dilemma. Fig. 1 shows that the proposed AI-FGSM is consistently more robust than the baseline for both JPEG compression and bit depth reduction.

The results above verify the superiority of the proposed scheme for Grad-CAM. According to our preliminary experiments, the proposed scheme can be easily integrated with a more recent CAM [2] as well.

[1] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," 2017CVPR, pp. 618-626.

[2] A. Chattopadhay, A. Sarkar, P. Howlader, et al., "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," 2018WCACV, pp. 839-847.



(a)

(b)

Fig. 1. Comparison of the attack robustness, (a) to JPEG compression, (b) to bit depth reduction.

Table 1. Image categories used in the experiments.

folder name	category	folder name	category	folder name	category	folder name	category
n01440764	tench	n01698640	alligator	n01860187	black swan	n02077923	sea lion
n01530575	brambling	n01740131	night snake	n01924916	flatworm	n02088094	Afghan
n01601694	water ouzel	n01770081	harvestman	n01980166	fiddler crab	n02090721	wolfhound
n01641577	bullfrog	n01795545	black grouse	n02007558	flamingo	n02093428	pit bull terrier
n01682714	anole	n01820546	lorikeet	n02027492	dunlin	n02095889	Sealyham

Table 2. Targeted attack success rates (%) of the baseline/proposed scheme when Grad-CAM [1] is adopted for generating attentional maps. * indicates the white-box attacks. Higher success rates are in **bold**.

Attack\Classification model	Resnet18*	Resnet34	Mobilenet_v3_large	IncepV3
AI-FGSM ($\epsilon = 8$)	82.10/**99.30**	7.85/**10.35**	5.00/**7.20**	3.75/**4.90**
AI-FGSM ($\epsilon = 16$)	87.45/**99.80**	15.30/**22.45**	9.05/**13.60**	7.50/**11.80**