Targeted attentional adversarial attack: supplementary material

H. Zeng, K. Yu, and A. Peng

In this supplementary document, we provide results for alternative attentional map generation methods [1, 2]:

1) Comparison of attack ability (Table 1).

2) Comparison of robustness to image processing (Fig. 1 and 2).

When Grad-CAM [1] is adopted for generating attentional maps, the proposed method has more obvious advantages than the baseline. Note the white-box success rate of baseline AI-FGSM is only 87.45% when $\epsilon = 16$. This is because the Grad-CAM method 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 attentional maps associated with high-confidence labels, thus working better under Grad-CAM.

When GradCAM++ [2] is adopted for generating attentional maps, our advantage over the base line shrinks (the lines of the proposed method and those of the baseline overlap with each other in Fig. 2). In [2], only positive gradients are used in producing attentional maps (See Section 7. 1 of [2] for detail discussion). Due to such mechanism, when generating an attention map associated to a low-confidence label, the output of $GradCAM++$ includes the attentional map of the original label, i.e., $GradCAM++_{ori}$. In other words, our idea of combining the attentional map of the original label and that of the target label has been partially internalized in GradCAM++. The results of GradCAM++ show the importance of combining CAM_o and CAM_t in targeted attack from another perspective

Table 1. The attack success rate (%) when alternative methods are adopted for generating attentional maps. * indicates the white-box attacks.

	Attack\Classification model	Resnet18*	Resnet34	Mobilenetv3_large	IncepV3
Grad-CAM [1]	Baseline AI-FGSM, $\epsilon = 8$	82.10	7.85	5.00	3.75
	Proposed AI-FGSM, $\epsilon = 8$	**99.30**	**10.35**	**7.20**	**4.90**
	Baseline AI-FGSM, $\epsilon = 16$	87.45	15.30	9.05	7.50
	Proposed AI-FGSM, $\epsilon = 16$	**99.80**	**22.45**	**13.60**	**11.80**
Grad-CAM++ [2]	Baseline AI-FGSM, $\epsilon = 8$	**99.50**	**11.95**	**7.50**	5.05
	Proposed AI-FGSM, $\epsilon = 8$	**99.50**	11.35	**7.50**	**5.35**
	Baseline AI-FGSM, $\epsilon = 16$	**99.95**	23.85	15.35	12.70
	Proposed AI-FGSM, $\epsilon = 16$	**99.95**	**24.80**	**15.45**	**13.40**

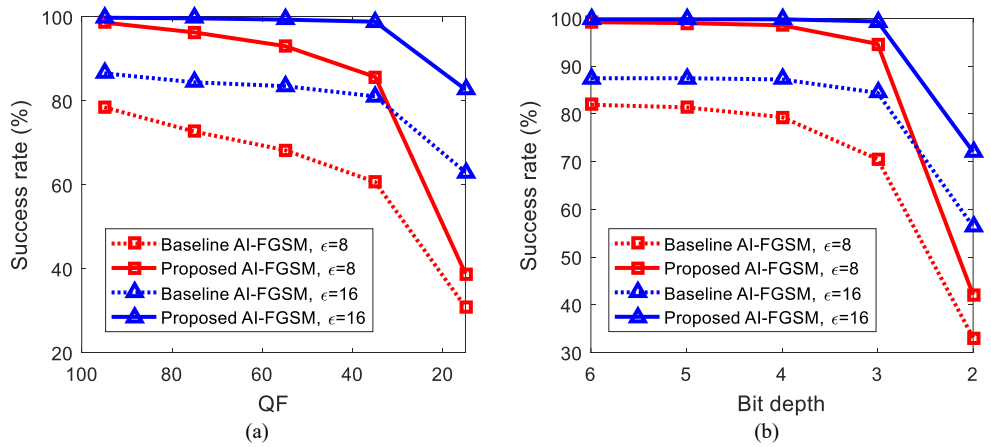

Fig. 1. Comparison of the attack robustness when the attentional map is generated by Grad-CAM, (a) to JPEG compression, (b) to bit depth reduction.

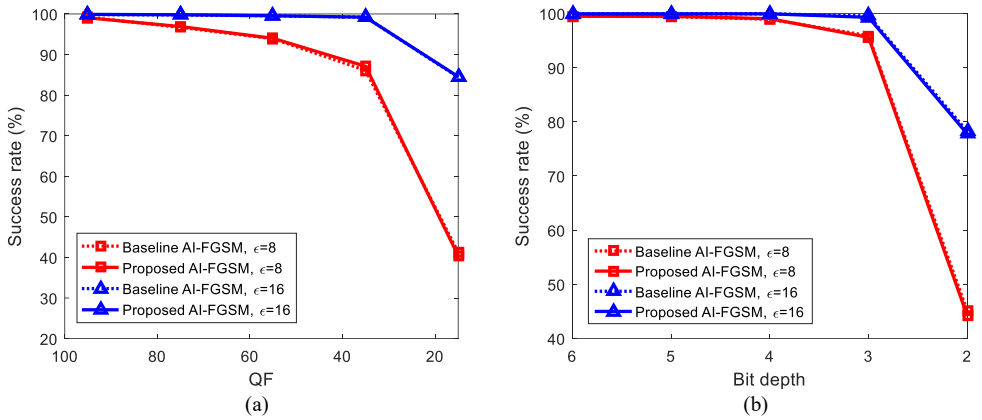

Fig. 2. Comparison of the attack robustness when the attentional map is generated by GradCAM++, (a) to JPEG compression, (b) to bit depth reduction.

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

[2] A. Chattopadhay, A. Sarkar, P. Howlader, et al., Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 839-847