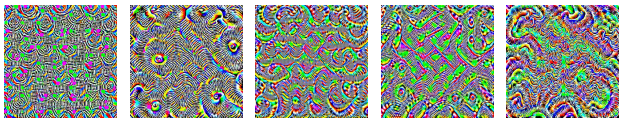
Supplementary: Enhancing Generalization of Universal Adversarial Perturbation through Gradient Aggregation

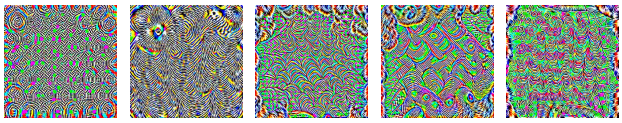
Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, Weihong Deng*
Beijing University of Posts and Telecommunications
{liuxuannan, zhongyaoyao, zyhzyh, lxqin, whdeng}@bupt.edu.cn

A. Visualization results



(a) AlexNet (b) GoogleNet (c) VGG16 (d) VGG19 (e) ResNet152

Figure A1. The visualization of universal adversarial perturbations generated by the proposed SGA with the maximum perturbation $\epsilon = 10$. The UAPs are crafted on the five normally trained models respectively by implementing the clipped cross-entropy loss.



(a) AlexNet (b) GoogleNet (c) VGG16 (d) VGG19 (e) ResNet152

Figure A2. The visualization of universal adversarial perturbations generated by the proposed SGA with the maximum perturbation $\epsilon = 10$. The UAPs are crafted on the five normally trained models respectively by implementing the logit loss.

Various-UAPs Setting. Fig A1 and Fig A2 respectively visualize the various UAPs generated by our methods under different loss function settings, *i.e.*, clipped cross-entropy loss and logit loss. The UAPs are crafted on five normally trained models, *i.e.*, AlexNet, GoogleNet, VGG16, VGG19, and ResNet152. All the UAPs are rescaled to [0,255] for better visualization.

Various Adv-Samples Setting. We also show some adversarial images with UAPs crafted by various methods, *i.e.*, UAP [1], GAP [2], SPGD [4] and our proposed method in Fig A3. It can be observed that our method achieves the highest attack performance while effectively guaranteeing the visual effect of the adversarial images.

*Corresponding author

B. Ablation study on other models

Considering that the network structures of VGG19 are similar with VGG16, here we provide the ablation results of our methods on the other three models, *i.e.*, AlexNet, GoogleNet, and ResNet152. The results in Figure B4 for the ablation study on gradient aggregation further verify the effectiveness of improving the attack performance of our method under various models.

The results of hyperparameters are depicted in Figure B5 and Figure B6. It can be observed that choosing an appropriate inner small-batch size and inner iteration number has the advantage of enhancing the generalization of UAP.

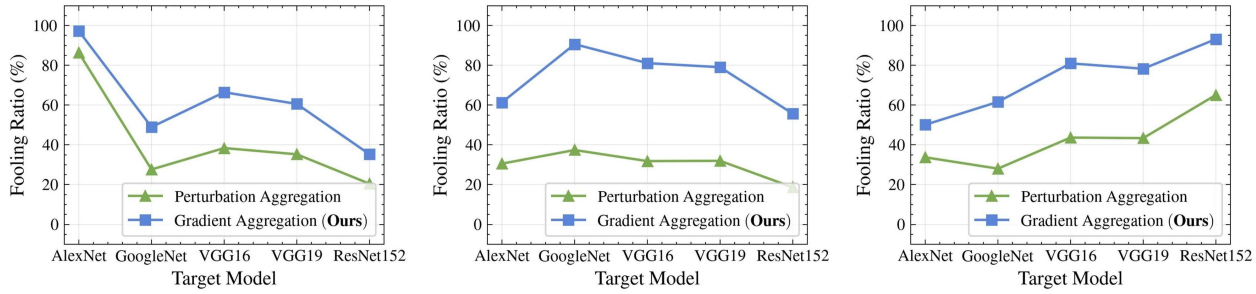
Type of attack	Methods	# samples	Computation Time $t_1 \downarrow$	$t_2 \downarrow$	Average attack success rate (%)
Universal attack	SPGD	10,000	25min'22s	0	62.05
	SGA	10,000	1h'13min'4s	0	**68.17**

Table C1. The evaluation of efficiency and effectiveness of different UAPs. The UAPs are crafted on the VGG16 model.

C. Discussion of the time consuming

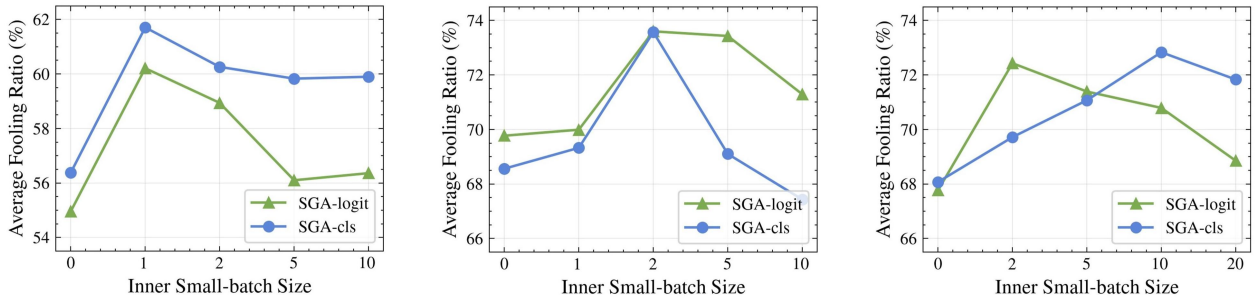
The limitation of the SGA comes at the cost of computation time. Specifically, we separately calculate the time required to infer the UAPs, denoted as t_1, and the time for generating adversarial examples, denoted as t_2. Table C1 shows that the proposed approach takes 3x more time to generate UAPs than SPGD. The reason for the more time-consuming comes from the usage of inner iteration, which introduces more gradient queries, thus increasing the computation time. However, considering the characteristics of universal adversarial perturbation, the time cost is negligible. Once the UAPs are crafted, there is no need to spend additional time generating corresponding perturbations for each sample. Therefore, it is reasonable to achieve a huge improvement in attack performance by allocating more time for pre-crafting the UAPs.

Figure A3. Some adversarial images in the ImageNet validation set [3] with UAPs, generated by UAP, GAP, SPGD, and our proposed method respectively. All UAPs are crafted on the VGG16 model with the maximum perturbation $\epsilon = 10$ by implementing clipped cross-entropy loss.



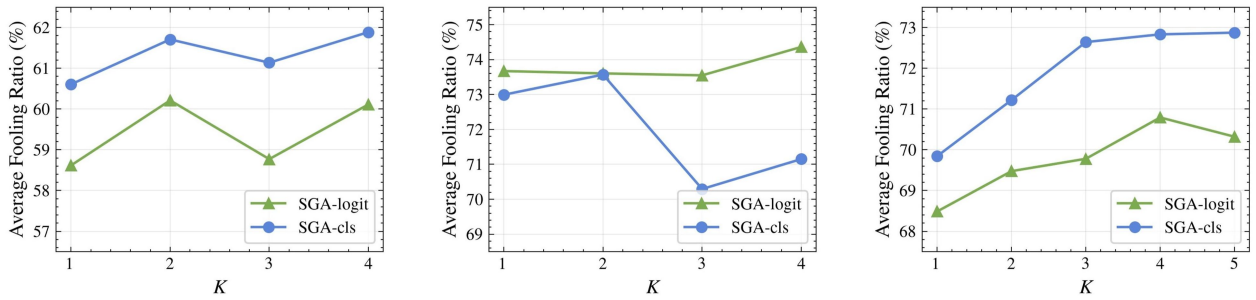
(a) AlexNet (b) GoogleNet (c) ResNet152

Figure B4. Average fooling ratio (%) of five models using tow types of aggregation method, *i.e.*, perturbation aggregation, and gradient aggregation. The UAPs are generated on AlexNet, GoogleNet, and ResNet152 models.



(a) AlexNet (b) GoogleNet (c) ResNet152

Figure B5. Average fooling ratio (%) of five models with different batch size of inner iteration. The UAPs are generated on AlexNet, GoogleNet, and ResNet152 models.



(a) AlexNet (b) GoogleNet (c) ResNet152

Figure B6. Average fooling ratio (%) of five models with different inner iteration number. The UAPs are generated on AlexNet, GoogleNet, and ResNet152 models.

References

[1] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pages 1765–1773, 2017. 1

[2] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *CVPR*, pages 4422–4431, 2018. 1

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[4] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *AAAI*, volume 34, pages 5636–5643, 2020. 1