AdvFoolGen: Creating Persistent Troubles for Deep Classifiers - Supplementary Material

Yuzhen Ding
Arizona State University
699 S. Mill Ave. Tempe, AZ 85281
Yuzhen.Ding@asu.edu

Nupur Thakur
Arizona State University
699 S. Mill Ave. Tempe, AZ 85281
nsthaku1@asu.edu

Baoxin Li
Arizona State University
699 S. Mill Ave. Tempe, AZ 85281
Baoxin.Li@asu.edu

In this supplementary material, we provide additional results to support our claims proposed in the main submission, AdvFoolGen. The additional results provided are on TinyImageNet dataset. We show reattack Top1 and Top5 fooling ratio on networks equipped with different defenses.

1. Additional Experimental Results

In this section, we provide the additional results obtained on TinyImageNet dataset for AdvFoolGen attack.

1.1. Effect of Defenses on Fooling Ratio

Table 1 and Table 2 show the Top1 and Top5 reattack fooling ratio on TinyImageNet dataset for different attacks, respectively. The target network here is strengthened with effective defenses. The fooling ratio varies for different epochs for AdvFoolGen and therefore a range of fooling ratio is reported. In this case too, Top1 fooling ratio is higher than the Top5 fooling ratio. Though there is a decrease in the fooling ratio to some extent after the defenses are used for all the attacks including AdvFoolGen, it is clear that the decrease in fooling ratio of AdvFoolGen is comparatively lower.

The fooling ratio achieved by different attack algorithms on retrained target networks is shown in Column 2. For the existing attack algorithms, equal number of adversarial and original images are used for retraining. For the reasons mentioned in the main submission, we use a network with one additional class for the AdvFool images while retraining. As each class contains 500 training images and 50 validation images in TinyImageNet dataset, we use 500 AdvFool images for training and 50 AdvFool images for validation. It is seen that all the state-of-the-art attacks fail to fool the retrained network with high fooling ratio but almost half of the AdvFool images can still fool it.

The next column presents the fooling ratio when the target network is adversarially trained. As the number of AdvFool images from generators at different epoch increase in the training set, the accuracy on original as well as AdvFool images decrease. All other attacks we compare with can be easily defended using adversarial training.

The last two columns are the defenses which use transformed images for retraining in order to defend against adversarial attacks. The transformations like Bit-Depth Reduction and JPEG compression are applied to the images before using them for retraining the network. Column 4 displays the results for Bit-Depth Reduction transformation with a Bit-Depth of 3. The last column is the defense which uses JPEG compressed adversarial images for retraining the network. The average fooling ratio of AdvFoolGen attack for both these defenses is comparable to FGSM, but outperforms all other attacks. This demonstrates that the AdvFool images can fool the network regardless of the type of image transformation applied.

Carefully examining the results obtained, it is observed that the attacks with high initial fooling ratio experience a significant decrease in the fooling ratio after the defenses are applied. This low fooling ratio shows that the existing attacks are not strong adversarial attacks and can be defended with small changes in the network. The AdvFoolGen attack is stronger than the existing ones because it can fool the networks equipped with state-of-the-art defenses.

Attack Algorithm	Retraining*	Adversarial Training	BDR-3	JPEG
FGSM	30.8%	49.37%	51.92%	54.18%
I-FGSM	40.5%	48.74%	48.44%	51.15%
DeepFool	29.2%	47.36%	43.02%	47.76%
CW	30.04%	48.61%	46.95%	47.26%
GAP	34.09%	33.76%	33.55%	35.21%
AdvFoolGen**	**43.1%-57.2%**	**54.6%-61.0%**	**40.3%-66.4%**	**42.1%-63.9%**

Table 1. Top 1 fooling ratio after the defenses are applied on TinyImageNet dataset. The fooling ratio for AdvFoolGen is higher than existing attacks when it comes to networks with added defense mechanisms. For Bit-Depth Reduction, a bit-depth of 3 is used. *Equal number of original and adversarial images are used for retraining. For AdvFoolGen attack, 500 AdvFool images are used for training and 50 AdvFool images are used in validation set as a new class is added for them. **We report a range for AdvFoolGen attack as the fooling ratio varies from epoch to epoch.

Attack Algorithm	Retraining*	Adversarial Training	BDR-3	JPEG
FGSM	18.26%	22.28%	27.12%	26.78%
I-FGSM	17.28%	20.96%	20.26%	23.09%
DeepFool	16.07%	15.26%	20.96%	18.98%
CW	14.35%	16.71%	18.55%	18.88%
GAP	12.81%	11.91%	12.77%	13.3%
AdvFoolGen**	**24.8%-33.2%**	**28.9%-35.2%**	**20.1%-32.4%**	**20.6%-35.6%**

Table 2. Top 5 fooling ratio after the defenses are applied on TinyImageNet dataset. The fooling ratio for AdvFoolGen is higher than existing attacks when it comes to networks with added defense mechanisms. For Bit-Depth Reduction, a bit-depth of 3 is used. *Equal number of original and adversarial images are used for retraining. For AdvFoolGen attack, 500 AdvFool images are used for training and 50 AdvFool images are used in validation set as a new class is added for them. **We report a range for AdvFoolGen attack as the fooling ratio varies from epoch to epoch.