# Supplementary Materials: Patch-wise Attack for Fooling Deep Neural Network

Lianli Gao[1], Qilong Zhang[1], Jingkuan Song[1], Xianglong Liu[2], and Heng Tao Shen[1⋆]

[1] Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
[2] Beihang University, China
qilong.zhang@std.uestc.edu.cn

## A Appendix

Here we discuss the influence of project factor $\gamma$ and iteration $T$, we consider 12 models to do these experiments:

- NT: Inc-v3 [4], Inc-v4 [3], IncRes-v2 [3], Res-152 [1] and Dense-161 [2].
- EAT [5]: Inc-v3$_{adv}$, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes$_{ens}$.
- FD [6]: ResNeXt$_{DA}$, Res152$_B$ and Res152$_D$.

Then we report the results for different project factor $\gamma$ settings in Sec. A.1 and the influence of different iterations $T$ in Sec. A.2.

### A.1 Selection of project factor $\gamma$

In this section, we show the results of difference project factor $\gamma$ settings. For the following tables, the amplification factor $\beta$ and project kernel $W_p$ are fixed to the same settings mentioned in our paper. As shown in Tab. 1,2,3,4,5,6, setting the project factor $\gamma$ to $\epsilon/T \cdot \beta$ is the best choice in most cases. Besides, it is better not to use large project factor for attack methods with input diversity strategy, and if we use Res152$_B$ as our substitute model to attack against feature denoising models, setting $\gamma$ to 1.0 is better.

### A.2 The number of iteration $T$

In this section, we discuss the influence of different iteration $T$. Here step size is set to $\epsilon/T$, amplification factor $\beta$ is set to $T$, and project factor $\gamma$ is set to $\epsilon/T \cdot \beta$. From the results of Tab. 7,8,9, we observe that when the iteration $T$ exceeds 10, further increasing it will not bring significant improvement. Besides, the computational cost is proportional to $T$. Since our PI-FGSM is based on FGSM, we want to highlight "fast". Therefore we do not consider a bigger $T$ and just set $T$ to 10 in our paper.

⋆ Corresponding author

**Table 1.** The average success rate(%) of non-targeted attacks in different project factor $\gamma$ settings. Here we use Inc-v3 to generate adversarial examples by PI-FGSM.

| Inc-v3 | $\gamma$ | Inc-v4 | Res-152 | IncRes-v2 | Dense-161 | Inc-v3$_{adv}$ | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes$_{ens}$ | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| PI-FGSM | 1.0 | 39.2 | 28.7 | 33.8 | 36.5 | 24.2 | 13.7 | 12.8 | 7.1 | 24.5 |
| | 2.0 | 42.1 | 29.6 | 37.9 | 38.2 | 25.2 | 13.5 | 14.0 | 8.4 | 26.1 |
| | 3.0 | 42.1 | 28.2 | 35.2 | 38.9 | 25.0 | 16.0 | 14.2 | 8.0 | 26.0 |
| | 4.0 | 42.8 | 29.9 | 39.7 | 40.8 | 24.9 | 15.6 | 14.8 | 9.0 | 27.2 |
| | 5.0 | 46.1 | 31.3 | 40.3 | 42.7 | 26.1 | 16.4 | 17.0 | 8.1 | 28.5 |
| | 6.0 | 48.4 | 31.8 | 40.9 | 44.4 | 27.1 | 18.1 | 16.7 | 9.8 | 29.7 |
| | 7.0 | 50.9 | 34.7 | 44.0 | 45.5 | 26.5 | 18.8 | 17.9 | 10.0 | 31.0 |
| | 8.0 | 53.1 | 36.2 | 45.5 | 46.6 | 28.5 | 20.2 | 19.7 | 11.7 | 32.7 |
| | 9.0 | 52.8 | 35.3 | 45.3 | 49.8 | 28.7 | 21.6 | 21.7 | 13.3 | 33.6 |
| | 10.0 | 53.2 | 36.9 | 48.8 | 52.5 | 28.5 | 22.1 | 23.3 | 14.8 | 35.0 |
| | 11.0 | 55.8 | 39.9 | 49.6 | 53.4 | 30.0 | 26.7 | 27.2 | 17.0 | 37.5 |
| | 12.0 | 55.7 | 40.9 | 50.8 | 55.3 | 29.8 | 28.1 | 29.7 | 18.7 | 38.6 |
| | 13.0 | 57.4 | 41.7 | 50.3 | 55.6 | 32.1 | 31.2 | 31.5 | 22.7 | 40.3 |
| | 14.0 | 57.8 | 44.7 | 51.4 | 58.3 | 35.3 | 34.1 | 35.0 | 25.8 | 42.8 |
| | 15.0 | 58.5 | 44.3 | 52.3 | 59.2 | 38.6 | 37.1 | 36.8 | 26.9 | 44.2 |
| | 16.0 | **59.8** | **46.9** | **52.7** | **60.2** | **40.7** | **40.2** | **39.7** | **29.6** | **46.2** |

**Table 2.** The average success rate(%) of non-targeted attacks in different project factor $\gamma$ settings. Here we use Inc-v3 to generate adversarial examples by DPI-FGSM.

| Inc-v3 | $\gamma$ | Inc-v4 | Res-152 | IncRes-v2 | Dense-161 | Inc-v3$_{adv}$ | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes$_{ens}$ |
|---|---|---|---|---|---|---|---|---|---|
| DPI-FGSM | 1.0 | 72.9 | 48.6 | 64.0 | 49.5 | 24.7 | 15.5 | 14.4 | 8.4 |
| | 2.0 | 73.0 | 50.4 | 65.2 | 54.7 | 26.2 | 15.4 | 15.6 | 8.6 |
| | 3.0 | **73.2** | **50.9** | **66.6** | 60.6 | 26.4 | 16.7 | 17.7 | 9.1 |
| | 4.0 | 70.1 | 49.0 | 62.1 | 60.3 | 25.9 | 17.5 | 15.8 | 9.1 |
| | 5.0 | 66.0 | 48.7 | 59.4 | 61.1 | 27.8 | 17.3 | 19.1 | 10.6 |
| | 6.0 | 62.7 | 47.9 | 54.7 | 61.6 | 27.7 | 18.0 | 19.9 | 9.5 |
| | 7.0 | 61.8 | 46.0 | 52.0 | **61.8** | 28.8 | 20.8 | 20.4 | 11.4 |
| | 8.0 | 58.2 | 45.5 | 47.8 | 59.2 | 30.4 | 21.8 | 23.8 | 12.0 |
| | 9.0 | 55.2 | 43.7 | 47.8 | 58.7 | 29.9 | 23.2 | 23.4 | 13.0 |
| | 10.0 | 53.9 | 44.6 | 46.2 | 59.1 | 31.1 | 26.0 | 26.4 | 16.5 |
| | 11.0 | 53.8 | 44.3 | 44.6 | 57.6 | 31.2 | 26.9 | 29.7 | 17.9 |
| | 12.0 | 55.2 | 42.5 | 45.3 | 58.7 | 33.3 | 30.4 | 32.7 | 19.0 |
| | 13.0 | 52.2 | 43.5 | 43.5 | 58.1 | 34.8 | 33.4 | 34.6 | 22.9 |
| | 14.0 | 55.3 | 41.7 | 40.8 | 59.0 | 37.0 | 35.0 | 35.9 | 26.4 |
| | 15.0 | 52.1 | 40.9 | 41.1 | 56.8 | 40.8 | 39.8 | 40.1 | 28.6 |
| | 16.0 | 51.9 | 43.5 | 41.5 | 57.9 | **42.3** | **40.9** | **41.6** | **31.2** |

**Table 3.** The average success rate(%) of non-targeted attacks in different project factor $\gamma$ settings. Here we use Inc-v3 to generate adversarial examples by MPI-FGSM.

| Inc-v3 | $\gamma$ | Inc-v4 | Res-152 | IncRes-v2 | Dense-161 | Avg |
|---|---|---|---|---|---|---|
| MPI-FGSM | 1.0 | 56.2 | 41.2 | 54.5 | 46.1 | 49.5 |
| | 2.0 | 57.2 | 40.2 | 54.9 | 48.0 | 50.1 |
| | 3.0 | 57.6 | 41.2 | 57.0 | 46.0 | 50.5 |
| | 4.0 | 57.5 | 42.1 | 57.7 | 46.8 | 51.0 |
| | 5.0 | 58.8 | 42.0 | 56.2 | 47.2 | 51.1 |
| | 6.0 | 58.7 | 41.9 | 57.4 | 47.5 | 51.4 |
| | 7.0 | 57.7 | 43.5 | 57.2 | 48.4 | 51.7 |
| | 8.0 | 59.1 | 42.4 | 56.3 | 50.8 | 52.2 |
| | 9.0 | 58.7 | 42.4 | 56.3 | 48.6 | 51.5 |
| | 10.0 | 60.1 | 43.3 | 57.9 | 49.1 | 52.6 |
| | 11.0 | 60.1 | 43.4 | 59.5 | 51.9 | 53.7 |
| | 12.0 | 62.6 | 44.4 | 59.5 | 52.6 | 54.8 |
| | 13.0 | 62.6 | 45.3 | 59.7 | 55.7 | 55.8 |
| | 14.0 | 64.0 | 47.7 | **60.9** | 57.7 | 57.6 |
| | 15.0 | **64.1** | 48.1 | 60.4 | 59.0 | 57.9 |
| | 16.0 | **64.1** | **50.2** | 59.4 | **60.4** | **58.5** |

**Table 4.** The average success rate(%) of non-targeted attacks in different project factor $\gamma$ settings. Here we use Inc-v3 to generate adversarial examples by TPI-FGSM.

| Inc-v3 | $\gamma$ | Inc-v3$_{adv}$ | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes$_{ens}$ | Avg |
|---|---|---|---|---|---|---|
| TPI-FGSM | 1.0 | 32.8 | 29.7 | 31.0 | 20.9 | 28.6 |
| | 2.0 | 32.0 | 31.2 | 31.6 | 22.2 | 29.3 |
| | 3.0 | 34.0 | 31.4 | 32.2 | 23.0 | 30.2 |
| | 4.0 | 34.0 | 32.7 | 34.0 | 22.6 | 30.8 |
| | 5.0 | 34.9 | 34.3 | 34.7 | 23.4 | 31.8 |
| | 6.0 | 35.3 | 34.7 | 33.6 | 23.1 | 31.7 |
| | 7.0 | 35.8 | 35.8 | 36.1 | 24.4 | 33.0 |
| | 8.0 | 37.9 | 36.4 | 37.5 | 25.9 | 34.4 |
| | 9.0 | 36.4 | 37.0 | 38.3 | 26.5 | 34.6 |
| | 10.0 | 38.6 | 38.0 | 38.2 | 26.6 | 35.4 |
| | 11.0 | 38.0 | 40.9 | 40.3 | 27.1 | 36.6 |
| | 12.0 | 40.4 | 41.6 | 40.2 | 28.9 | 37.8 |
| | 13.0 | 41.8 | 42.8 | 43.3 | 29.7 | 39.4 |
| | 14.0 | 43.8 | 44.0 | 44.7 | 32.9 | 41.4 |
| | 15.0 | 46.6 | 46.6 | 47.0 | 34.8 | 43.8 |
| | 16.0 | **50.2** | **48.8** | **50.8** | **35.2** | **46.3** |

**Table 5.** The average success rate(%) of non-targeted attacks in different project factor $\gamma$ settings. Here we use Inc-v3 to generate adversarial examples by DMPI-FGSM.

| Inc-v3 | $\gamma$ | Inc-v4 | Res-152 | IncRes-v2 | Dense-161 | Avg |
|---|---|---|---|---|---|---|
| DMPI-FGSM | 1.0 | 81.2 | 61.8 | 76.0 | 61.1 | 70.0 |
| | 2.0 | 81.7 | 64.5 | **77.3** | 63.6 | 71.8 |
| | 3.0 | **81.9** | **65.1** | **77.3** | 66.0 | **72.6** |
| | 4.0 | 79.7 | 63.1 | 75.7 | **67.9** | 71.6 |
| | 5.0 | 76.3 | 60.6 | 71.7 | 66.0 | 68.7 |
| | 6.0 | 74.4 | 56.6 | 69.1 | 67.1 | 66.8 |
| | 7.0 | 69.5 | 55.6 | 64.0 | 63.6 | 63.2 |
| | 8.0 | 66.1 | 52.9 | 59.1 | 65.4 | 60.9 |
| | 9.0 | 63.6 | 51.0 | 58.4 | 64.0 | 59.3 |
| | 10.0 | 62.2 | 51.1 | 56.1 | 63.5 | 58.2 |
| | 11.0 | 63.6 | 52.0 | 56.7 | 64.0 | 59.1 |
| | 12.0 | 61.8 | 49.4 | 56.0 | 64.7 | 58.0 |
| | 13.0 | 62.9 | 50.1 | 55.6 | 64.8 | 58.4 |
| | 14.0 | 62.2 | 50.5 | 54.1 | 63.4 | 57.6 |
| | 15.0 | 60.4 | 49.5 | 53.9 | 62.6 | 56.6 |
| | 16.0 | 61.4 | 49.7 | 54.7 | 63.8 | 57.4 |

**Table 6.** The average success rate(%) of non-targeted attacks in different project factor $\gamma$ settings. Here we use $\text{Res152}_B$ ("*" indicates white-box attacks) to generate adversarial examples by PI-FGSM.

| $\text{Res152}_B$ | $\gamma$ | $\text{ResNeXt}_{DA}$ | $\text{Res152}_B$ | $\text{Res152}_D$ | Avg |
|---|---|---|---|---|---|
| PI-FGSM | 1.0 | 62.0 | **88.6*** | 61.5 | **70.7** |
| | 2.0 | **62.1** | 88.2* | 61.6 | 70.6 |
| | 3.0 | 61.8 | 87.7* | **62.0** | 70.5 |
| | 4.0 | 61.5 | 87.1* | 61.6 | 70.0 |
| | 5.0 | 60.0 | 85.4* | 58.8 | 68.1 |
| | 6.0 | 56.8 | 84.0* | 56.0 | 65.6 |
| | 7.0 | 55.0 | 82.1* | 53.4 | 63.5 |
| | 8.0 | 53.4 | 80.7* | 50.8 | 61.6 |
| | 9.0 | 51.9 | 79.6* | 49.0 | 60.2 |
| | 10.0 | 50.9 | 78.7* | 48.8 | 59.5 |
| | 11.0 | 51.3 | 78.2* | 48.6 | 59.4 |
| | 12.0 | 50.8 | 78.0* | 48.2 | 59.0 |
| | 13.0 | 50.2 | 77.8* | 48.0 | 58.7 |
| | 14.0 | 50.3 | 77.8* | 47.7 | 58.6 |
| | 15.0 | 50.2 | 77.7* | 47.7 | 58.5 |
| | 16.0 | 50.2 | 77.7* | 47.7 | 58.5 |

**Table 7.** The average success rate(%) of non-targeted attacks in different iteration $T$ against NT. Here we use Inc-v3 to generate adversarial examples by PI-FGSM.

| Inc-v3 | iteration $T$ | Inc-v4 | Res-152 | IncRes-v2 | Dense-161 | Avg |
|---|---|---|---|---|---|---|
| PI-FGSM | 5 | **60.9** | 46.3 | **54.8** | 58.9 | **55.2** |
| | 10 | 59.8 | 46.9 | 52.7 | 60.2 | 54.9 |
| | 15 | 57.3 | 46.3 | 51.2 | 60.1 | 53.7 |
| | 20 | 57.6 | 46.6 | 50.8 | 60.6 | 53.9 |
| | 25 | 58.1 | **47.2** | 49.7 | **61.3** | 54.1 |
| | 30 | 58.6 | 46.0 | 50.0 | 60.7 | 53.8 |

**Table 8.** The average success rate(%) of non-targeted attacks in different iteration $T$ against EAT. Here we use Inc-v3 to generate adversarial examples by PI-FGSM.

| Inc-v3 | iteration $T$ | Inc-v3$_{adv}$ | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes$_{ens}$ | Avg |
|---|---|---|---|---|---|---|
| PI-FGSM | 5 | 39.4 | 33.6 | 33.8 | 24.7 | 23.0 |
| | 10 | 40.7 | 40.2 | 39.7 | 29.6 | 27.4 |
| | 15 | **40.8** | 39.7 | 40.0 | 29.0 | 27.2 |
| | 20 | 40.5 | 40.6 | 41.3 | 30.1 | 28.0 |
| | 25 | 39.9 | **41.3** | **41.5** | 29.9 | 28.2 |
| | 30 | 40.1 | 41.1 | 40.6 | **31.4** | **28.3** |

**Table 9.** The average success rate(%) of non-targeted attacks in different iteration $T$ against FD. Here we use Res152$_B$ ("*" indicates white-box attacks) to generate adversarial examples by PI-FGSM.

| Res152$_B$ | iteration $T$ | ResNeXt$_{DA}$ | Res152$_B$ | Res152$_D$ | Avg |
|---|---|---|---|---|---|
| PI-FGSM | 5 | 60.0 | 86.8* | 60.2 | 69.0 |
| | 10 | 61.8 | 88.6* | 61.5 | 70.6 |
| | 15 | 62.6 | 89.3* | 61.7 | 71.2 |
| | 20 | **62.9** | 89.5* | 61.9 | **71.5** |
| | 25 | 62.8 | 89.5* | **62.1** | **71.5** |
| | 30 | 62.8 | **89.8*** | 61.5 | 71.4 |

## References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 1
2. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017) 1
3. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017) 1
4. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016) 1
5. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I.J., Boneh, D., McDaniel, P.D.: Ensemble adversarial training: attacks and defenses. In: ICLR (2018) 1
6. Xie, C., Wu, Y., van der Maaten, L., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: CVPR (2019) 1