Progressive Backdoor Erasing via connecting Backdoor and Adversarial Attacks

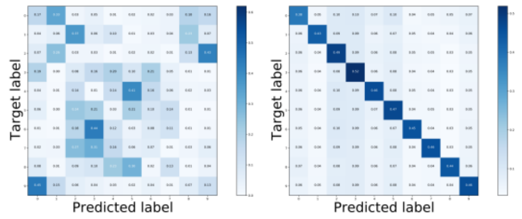
Bingxu Mu[1], Zhenxing Niu[2], Le Wang[1], Xue Wang[3], Qiguang Miao[2], Rong Jin[3], Gang Hua[4]

[1]Xi'an Jiaotong University, [2]Xidian University, [3]Alibaba Group, [4]Wormpex AI Research

Code:

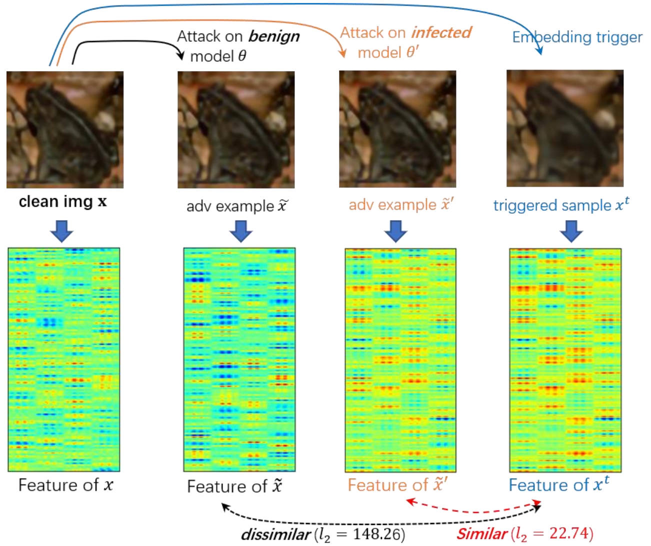
Is there any *connection* between *backdoor* and *adversarial* attacks? YES

Empirical Observations



(a) For *benign* model. (b) For *infected* model.

The predicted-label of ***untargeted*** *adversarial examples* with respect to (a) ***benign models*** (the predicted labels obey **uniform** distribution) and (b) ***infected models*** (its untargeted adversarial examples are **highly likely** to be classified as the target-label, i.e., the matrix diagonals).



(1) adversarial examples will ***change*** significantly ***after planting a backdoor*** into a model

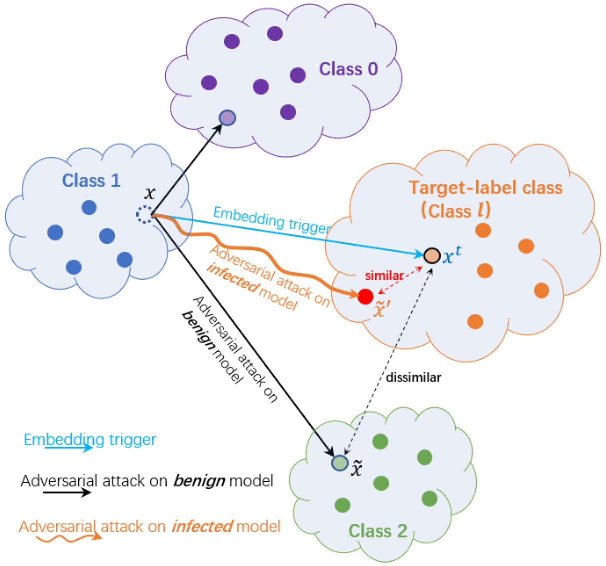
(2) The adversarial example w.r.t infected model $\tilde{x}'$ ***looks similar*** to backdoor-triggered image x^t

(3) $\tilde{x}'$ are highly likely to be ***classified*** as the target-label

Our Intriguing Findings

For *an infected model*, its adversarial examples have *similar behaviors* as its triggered samples:

"both activate the same subset of DNN neurons (i.e., have similar feature maps)".

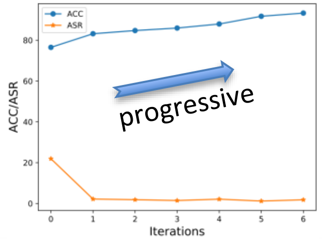


Leverage our findings to design a *backdoor defense* method

Generate adversarial examples → ***Purify the infected model*** via fine-tuning with generated adversarial examples

Progressive Backdoor Erasing Algorithm

Purify the training data via trigger image detection



Theoretical Analysis

Theorem 1 *Under the previous assumptions, we have $\mathbf{r}_\perp$, the projection of $\mathbf{r}$ on the direction of P, bounded as*

$$\frac{|\mathbf{r}_\perp|}{|\mathbf{r}|} \geq \frac{(\sqrt{2}-1)\ell|P|^2}{\sqrt{(\sqrt{2}-1)^2\ell^2|P|^4+\left(\ell|P|^2+\sqrt{2}K/(\exp(\tau)+K)\right)^2}}$$

CIFAR-10

ADE/FDE	Before	PBE
BadNet	99.02/100.00	94.43/0.47
Blend	99.39/99.92	94.57/1.72
SIG	98.56/95.81	94.05/1.78
Dynamic	99.27/99.84	96.68/0.99
WaNet	98.97/98.78	96.56/0.47

GTSRB dataset

ADE/FDE	Before	PBE
BadNet	94.67/100.00	94.20/1.09
Blend	94.63/100.00	93.98/0.93
SIG	94.81/98.96	93.35/1.39
Dynamic	94.65/99.24	93.01/1.12
WaNet	94.15/99.50	94.32/0.46