Progressive Backdoor Erasing via connecting Backdoor and Adversarial Attacks

Bingxu Mu[1] Zhenxing Niu[2*] Le Wang[3] Xue Wang[4] Qiguang Miao[2] Rong Jin[4] Gang Hua[5]
[1]School of Software Engineering, Xi'an Jiaotong University [2]Xidian University
[3]IAIR, Xi'an Jiaotong University [4]Alibaba Group [5]Wormpex AI Research

Appendix

A. Theoretical Analysis

In the main text, we have empirically shown the observations that for an infected model, the features of adversarial examples $\widetilde{\boldsymbol{x}}'$ are very similar to the features of triggered images $\boldsymbol{x}^{\boldsymbol{t}}$, which results in that $\widetilde{\boldsymbol{x}}'$ is highly likely classified as the target-label l instead of any other classes. In this section, we take a linear classification model as an example to theoretically justify that our observation does make sense.

A.1. Structure of Infected Model W^*

Let $(x_i, y_i), (i = 1, \ldots, n)$ be the training examples, where $x_i \in \mathbb{R}^d$ and $y_i \in [K]$. Let $n_j, j \in [K]$ be the number of instances in class C_j. All the training examples are well normalized, i.e. $|x_i| = 1, i \in [n]$. We assume all the training instance x_i living a subspace $\mathcal{S}_m \subset \mathbb{R}^d$ of m dimension, with $m < d$. For the simplicity of our analysis, we assume the trigger embedding function $\mathrm{Trigger}()$ is to add a pre-defined patch P to an input image, i.e.,

$$\boldsymbol{x}^{\boldsymbol{t}} = \mathrm{Trigger}(\boldsymbol{x}) = \boldsymbol{x} + P \tag{1}$$

Since the image patch $P \in \mathbb{R}^d$ introduced is very different from most training examples $\{x_i\}_{i=1}^n$, it is safe to assume that P is orthogonal to $\mathcal{S}_m$. In the backdoor attack, we randomly sample ℓ examples from class C_l, and embed the trigger P into them. Thus, the infected linear classifiers, denoted by $W = (w_1, w_2, \ldots, w_K)$, are obtained by solving the following optimization problem

$$W^* = \arg\min_{W \in \mathbb{R}^{d \times K}} \mathcal{F}(W) \tag{2}$$

where

$$\mathcal{F}(W) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp\left(\langle w_{y_i}, x_i\rangle\right)}{\sum_{j=1}^{K} \exp\left(\langle w_j, x_i\rangle\right)} + \frac{\lambda}{2}|W|_F^2 \tag{3}$$

Define $\widetilde{W}^*$ be the solution, with each column vector restricted to the subspace $\mathcal{S}_m$, that minimizes $\mathcal{F}(\widetilde{W})$, i.e.

$$\widetilde{W}^* = \arg\min_{\widetilde{W} \in \mathcal{A}} \mathcal{F}(\widetilde{W}) \tag{4}$$

where

$$\mathcal{A} = \left\{\widetilde{W} \in \mathbb{R}^{d \times K} : W_{*,j} \in \mathcal{S}_m, j \in [K]\right\}$$

Obviously, $\widetilde{W}^*$ can be regarded as the benign model, which is trained with clean images instead of triggered images. The following lemma characterizes the structure of the optimal solution W^* obtained from (2).

Lemma 1. *Rewrite $W^* = W_{\|}^* + W_{\perp}^*$, where $W_{\|}^*$ is the projection of column vectors in W^* into the subspace $\mathcal{S}_m$. We have*

$$|W_{\|}^* - \widetilde{W}^*| \leq \frac{\ell |P|^2}{\sqrt{2}\lambda}$$

and $W_{\perp}^* = Pu^\top$, where $u \in R^d$, with $u_j \leq 0$ for $j \neq l$ and

$$u_l \geq \frac{(\sqrt{2}-1)\ell |P|^2}{\sqrt{2}\lambda}$$

Proof. Due to the presence of regularizer $\lambda|W|_F^2/2$, it is easy to see that the optimal solution W^* can be written as $W^* = W_{\|}^* + Pu^\top$, with $W_{\|}^*$ being the projection of W^* into the subspace $\mathcal{S}_m$. It is also easy to show that $u_l \geq 0$ and $u_j \leq 0, j \neq l$ by simply checking out the derivative of the objective function with respect to u_j and u_l.

We construct the upper bound for $F(W^*)$. To this end, we restrict the solution to the form of $\widetilde{W} + \gamma P\mathbf{e}_l$, where $\mathbf{e}_l$ is a binary vector with all its elements being zero except for the l-th element. The resulting optimization problem is given by

$$\min_{\widetilde{W} \in \mathcal{A}, \gamma \geq 0} -\ell|P|^2\gamma + \frac{\lambda}{2}\gamma^2 + \mathcal{F}(\widetilde{W})$$

The resulting solution is $W_1 = \widetilde{W}^* + \gamma P\mathbf{e}_l$, with $\gamma = \ell|P|^2/\lambda$, and optimal value is $\mathcal{F}(W_1) = \mathcal{F}(\widetilde{W}^*) - \ell^2|P|^4/[2\lambda]$. Evidently, we have

$$\mathcal{F}(W^*) \leq \mathcal{F}(\widetilde{W}^*) - \frac{\ell^2|P|^4}{2\lambda} \tag{5}$$

We then proceed to construct the lower bound for $\mathcal{F}(W^*)$. Define $u' = \max\{|u_j| : j \neq l\}$. Since for each training

x_i in the class C_l that is triggered with P, its negative log-likelihood is bounded as

$$-\log \frac{\exp\left(\langle w_l, x_i\rangle\right)}{\sum_{j=1}^{K}\exp\left(\langle w_j, x_i\rangle\right)} \geq -(u_l+u')|P|^2 - \log \frac{\exp\left(\langle \tilde{w}_l, x_i\rangle\right)}{\sum_{j=1}^{K}\exp\left(\langle \tilde{w}_j, x_i\rangle\right)} \tag{6}$$

where $\tilde{w}_j \in \mathcal{S}_m, j \in [K]$. As a result, we have the following lower bound for $\mathcal{F}(W^*)$, i.e.

$$\min_{\widetilde{W}\in\mathcal{A}, u_l, u'\geq 0} -\ell|P|^2(u_l+u') + \frac{\lambda}{2}(|u_l|^2+|u'|^2) + \mathcal{F}(\widetilde{W})$$

which can be further simplified as

$$\min_{\widetilde{W}\in\mathcal{A}, \gamma\geq 0} -\ell|P|^2\gamma + \frac{\lambda}{4}\gamma^2 + \mathcal{F}(\widetilde{W})$$

By solving the above optimization problem, we have $\mathcal{F}(W^*)$ lower bounded as

$$\mathcal{F}(W^*) \geq \mathcal{F}(\widetilde{W}^*) - \frac{\ell^2|P|^4}{\lambda} \tag{7}$$

Using the upper and lower bounds from (5) and (7), we have

$$\mathcal{F}(W_1) - \mathcal{F}(W^*) \leq \frac{\ell^2|P|^4}{2\lambda}$$

Since $\mathcal{F}(W)$ is λ-strong convex, we have

$$|W_1 - W_*|^2 \leq \frac{\ell^2|P|^4}{2\lambda^2}$$

The lemma will then directly follows from the above inequality. □

The above lemma tells us that **for the infected model W^*, its component $W_{\|}^*$ is not too far from the benign model $\widetilde{W}^*$, while the residual component $W_{\perp}^*$ will give a strong response to the trigger P.** It means that planting a backdoor into a model will not strongly affect the model's performance on benign images, but **will significantly affect its predictions on triggered images**.

In addition, this lemma tell us the interesting role played by ℓ, the number of instances sampled from class C_l to be triggered by P: a larger ℓ will lead to a larger value of u_l, indicating a stronger footprint of pattern P implemented in the lth classifier; but, at the same time, a larger ℓ will lead to a larger value $|W_{\|}^* - \widetilde{W}_*|$, implying a distortion in classification models. Hence, an appropriate choice of ℓ should result in a small distortion in the overall classification model, and at the same time, a strong enough backdoor attack of trigger P in the classification model for class C_l.

A.2. Structure of Perturbation r

After analyzing the structure of the solution learned from triggered examples, we proceed to analyze the perturbation $\boldsymbol{r}$ learned from adversarial attack by solving the following optimization problem

$$\boldsymbol{r} = \max_{|\boldsymbol{r}|\leq\delta} Ł(\boldsymbol{r}) \tag{8}$$

where

$$Ł(\boldsymbol{r}) = \frac{1}{\sum_{i=1}^{n}[y_i \neq l]} \sum_{i:y_i\neq l} \min_{j\neq k} f_k(x_i+\boldsymbol{r}) - f_j(x_i+\boldsymbol{r}) \tag{9}$$

In the above, we introduce $\delta \ll 1$ to specify the magnitude of perturbation, and $f_j(\cdot)$ for the classification model for the j-th class. The underlying logic is to find a small perturbation $\boldsymbol{r}$ that will make the deep model predict class C_l for every instance.

To simplify our analysis, we assume that the original training examples (without any trigger) can be perfectly classified with margin $\tau > 0$, i.e.

$$\tau = \min_{i\in[n]} \min_{j\neq y_i} \langle \tilde{w}_{y_i}^* - \tilde{w}_j^*, x_i\rangle$$

We assume that τ is large enough such that a small perturbation made to $\widetilde{W}^*$ will not affect classification result, i.e.

$$\tau \geq \frac{\sqrt{2}\ell|P|^2}{\lambda} \tag{10}$$

We finally assume δ is small enough, i.e.

$$\delta \leq \frac{\tau/2}{\min_{j\neq l}|\tilde{w}_l^* - \tilde{w}_j^*| + \ell|P|^2/[\sqrt{2}\lambda]} \tag{11}$$

Theorem 1. *Under the assumptions in (11) and (10), we have $\boldsymbol{r}_{\perp}$, the projection of $\boldsymbol{r}$ on the direction of P, bounded as*

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

Proof. Now, we consider the adversarial training by maximizing $Ł(\boldsymbol{r})$, which is given as

$$Ł(\boldsymbol{r}) = \frac{1}{\sum_{i=1}^{n}[y_i\neq l]} \sum_{i:y_i\neq l} \min_{j\neq l} \langle w_l^* - w_j^*, x_i + \boldsymbol{r}\rangle$$

Under the assumptions in (10) and (11), we can rewrite $Ł(\boldsymbol{r})$ as

$$Ł(\boldsymbol{r}) = \frac{1}{\sum_{i=1}^{n}[y_i\neq l]} \sum_{i:y_i\neq l} \langle w_l^* - w_{y_i}^*, x_i + \boldsymbol{r}\rangle \tag{12}$$

This is because, according to the definition of classification margin τ, we have, for any instance x_i with $y_i \neq l$,

$$\langle \tilde{w}^*_{y_i} - \tilde{w}^*_j, x_i \rangle \geq \tau, \ \forall j \neq y_i$$

Since

$$\left| w^*_{\|,y_i} - w^*_{\|,j} - \left(\tilde{w}^*_{y_i} - \tilde{w}^*_j\right) \right| \leq |W^*_\| - \widetilde{W}^*| \leq \frac{\ell |P|^2}{\sqrt{2}\lambda}$$

using the condition in (10), we have, for any instance x_i with $y_i \neq l$,

$$\langle w^*_{\|,y_i} - w^*_{\|,j}, x_i \rangle \geq \frac{\tau}{2}, \ \forall j \neq y_i$$

Since $w^*_j = w^*_{\|,j} + u_j P$ and $P \perp x_i$ for any $y_i \neq l$, we have, for any x_i with $y_i \neq l$

$$\langle w^*_{y_i} - w^*_j, x_i \rangle \geq \frac{\tau}{2}, \ \forall j \neq l$$

Since

$$|w^*_{y_i} - w^*_j| \leq |\tilde{w}^*_{y_i} - \tilde{w}^*_j| + \frac{\ell |P|^2}{\sqrt{2}\lambda}, \ \forall j \neq y_i$$

using the condition in (11), we have

$$\langle w^*_{y_i} - w^*_j, \boldsymbol{r} \rangle \geq -\left(|\tilde{w}^*_{y_i} - \tilde{w}^*_j| + \frac{\ell |P|^2}{\sqrt{2}\lambda} \right) \delta \geq -\frac{\tau}{2}, \ \forall j \neq y_i$$

As a result, for any x_i with $y_i \neq l$, we have

$$\langle w^*_{y_i}, x + \boldsymbol{r} \rangle \geq \langle w^*_j, x + \boldsymbol{r} \rangle, \ \forall j \neq y_i$$

and therefore

$$\min_{j \neq l} \langle w^*_l - w^*_j, x_i + \boldsymbol{r} \rangle = \langle w^*_l - w^*_{y_i}, x_i + \boldsymbol{r} \rangle$$

which leads to the expression in (12). We then proceed to simplify the expression in (12)

$$\begin{aligned}
Ł(\boldsymbol{r}) &= \frac{1}{\sum_{i=1}^n [y_i \neq l]} \sum_{i:y_i \neq l} \langle w^*_l - w^*_{y_i}, x_i + \boldsymbol{r} \rangle \\
&= \frac{1}{\sum_{i=1}^n [y_i \neq l]} \sum_{i:y_i \neq l} \langle w^*_{\|,l} - w^*_{\|,y_i} + (u_l - u_{y_i})P, x_i + \boldsymbol{r} \rangle \\
&= \frac{\langle P, \boldsymbol{r} \rangle}{\sum_{i=1}^n [y_i \neq l]} \sum_{i:y_i \neq l} (u_l - u_{y_i}) + \\
&\quad \frac{1}{\sum_{i=1}^n [y_i \neq l]} \sum_{i:y_i \neq l} \langle w^*_{\|,l} - w^*_{\|,y_i}, x_i + \boldsymbol{r} \rangle
\end{aligned}$$

Write $\boldsymbol{r} = \boldsymbol{r}_\perp + \boldsymbol{r}_\|$, where $\boldsymbol{r}_\|$ is the projection of $\boldsymbol{r}$ into the subspace $\mathcal{S}_m$. Using these notation, we have

$$\begin{aligned}
Ł(\boldsymbol{r}) = &\frac{\langle P, \boldsymbol{r}_\perp \rangle}{\sum_{i=1}^n [y_i \neq l]} \sum_{i:y_i \neq l} (u_l - u_{y_i}) \\
&+ \frac{1}{\sum_{i=1}^n [y_i \neq l]} \sum_{i:y_i \neq l} \langle w^*_{\|,l} - w^*_{\|,y_i}, x_i + \boldsymbol{r}_\| \rangle
\end{aligned} \tag{13}$$

Define

$$\begin{aligned}
\alpha &= \frac{1}{\sum_{i=1}^n [y_i \neq l]} \sum_{i:y_i \neq l} (u_l - u_{y_i}), \\
v &= \frac{1}{\sum_{i=1}^n [y_i \neq l]} \sum_{i:y_i \neq l} w^*_{\|,l} - w^*_{\|,y_i}
\end{aligned} \tag{14}$$

We have

$$\boldsymbol{r}_\perp| = \frac{\alpha |P| \delta}{\sqrt{\alpha^2 |P|^2 + |v|^2}} \tag{15}$$

Since $u_l \geq (\sqrt{2} - 1)\ell |P|^2 / [\sqrt{2}\lambda]$ and $u_j \leq 0$ for $j \neq l$, we hve

$$\alpha \geq \frac{(\sqrt{2} - 1)\ell |P|^2}{\sqrt{2}\lambda} \tag{16}$$

To bound $|v|$, we use the first order condition for the optimal solution $\widetilde{W}^*$, i.e.

$$\begin{aligned}
\tilde{w}^*_j = &\frac{1}{\lambda} \sum_{i=1}^n \left([y_i = j] \left(1 - p(y_i | x_i; \widetilde{W}^*) \right) \right. \\
&\left. - [y_i \neq j] p(j | x_i; \widetilde{W}^*) x_i \right)
\end{aligned} \tag{17}$$

where

$$p(j | x_i; \widetilde{W}^*) = \frac{\exp\left(\langle \tilde{w}^*_j, x_i \rangle \right)}{\sum_{j'=1}^K \exp\left(\langle \tilde{w}^*_{j'}, x_i \rangle \right)}$$

Using the definition of classification margin, we have

$$1 - p(y_i | x_i; \tilde{W}^*) \leq \frac{K}{\exp(\tau) + K}$$

and

$$p(j | x_i; \tilde{W}^*) \leq \frac{1}{\exp(\tau) + K}$$

As a result, we have

$$|\tilde{w}^*_j| \leq \frac{K}{\lambda(\exp(\tau) + K)}$$

and

$$\begin{aligned}
|w^*_{\|,l} - w^*_{\|,j}| &\leq |\tilde{w}^*_l - \tilde{w}^*_j| + \frac{\ell |P|^2}{\sqrt{2}\lambda} \\
&\leq \frac{1}{\lambda} \left(\frac{\ell |P|^2}{\sqrt{2}} + \frac{K}{\exp(\tau) + K} \right)
\end{aligned} \tag{18}$$

We complete the proof by plugging the bounds from (16) and (18) into the expression (15). □

From the above theorem, we can see that **when projecting perturbation r on the direction of trigger P, the projection $r_\perp$ take an significant part in the full perturbation r**. It means that the perturbation $\boldsymbol{r}$ is very similar to the trigger P, which **justify our observations that the adversarial examples $\widetilde{\boldsymbol{x}}' = \boldsymbol{x} + \boldsymbol{r}$ are similar to triggered images $\boldsymbol{x}^t = \boldsymbol{x} + P$.**

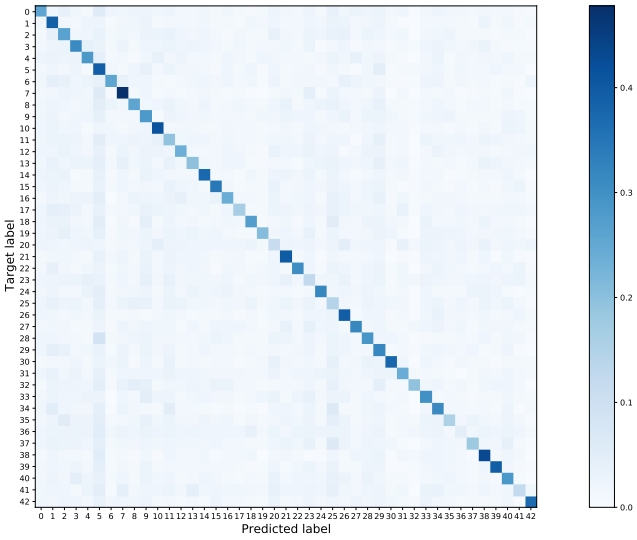

Figure 1. Predicted labels v.s. Target-labels for WaNet attack on full GTSRB dataset

.

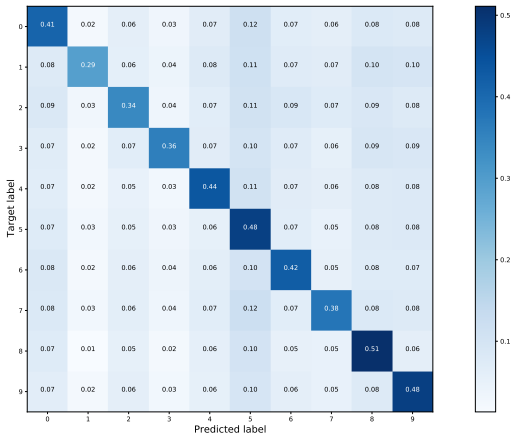

Figure 2. WaNet Attack on a subset of ImageNet-1K.

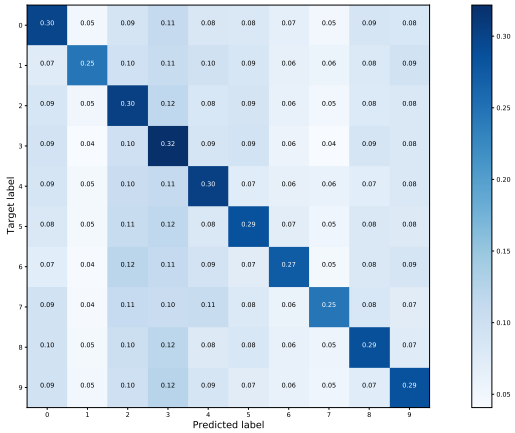

Figure 3. Blend Attack on CIFAR-10 with the 'dog' trigger.

(a) hello kitty

(b) dog

Figure 4. Two different trigger images for Blend attack.

B. Predicted labels v.s. Target-labels

B.1. Full GTSRB Dataset

In the main manuscript, due to the limited space we have shown the prediction results for a subset of GTSRB dataset, *i.e.,* the GTSRB-sub dataset has 15 classes randomly selected from 43 classes. In this section, we give the prediction results for the full GTSRB dataset containing 43 classes. The Fig.1 shows such results for the WaNet attack on GTSRB dataset. It is obvious that we have the same observations, *i.e.,* the adversarial examples are highly likely to be classified as target-label.

B.2. Large Image Resolution

In the main manuscript, we have evaluated our approach for images with small image resolutions, such as 32×32 for CIFAR-10. In this section, we randomly sample images from ImageNet-1K dataset for evaluation, which image size is 224×224. Specifically, 10 classes are randomly selected from 1000 classes in ImageNet-1K. Fig.2 indicates that we have the same observations regardless of what image sizes are.

B.3. Trigger Image

In the main manuscript, we follow the previous methods to use a 'hello kitty' image (ref to Fig.4a) as the trigger image for Blend attack. In this section, we show that our observations hold true regardless of what trigger image is. For example, take the 'dog' image (ref to Fig.4b) as the trigger image, we still have a similar results, as shown in Fig.3.

C. Similarity of Feature Maps

We give more results for the similarity of features among benign model's adversarial examples $\widetilde{\boldsymbol{x}}$, infected model's adversarial examples $\widetilde{\boldsymbol{x}}'$, and triggered samples $\boldsymbol{x}^t$. Figure.7 is an image from ImageNet-1K, which sizes are of 224×224.

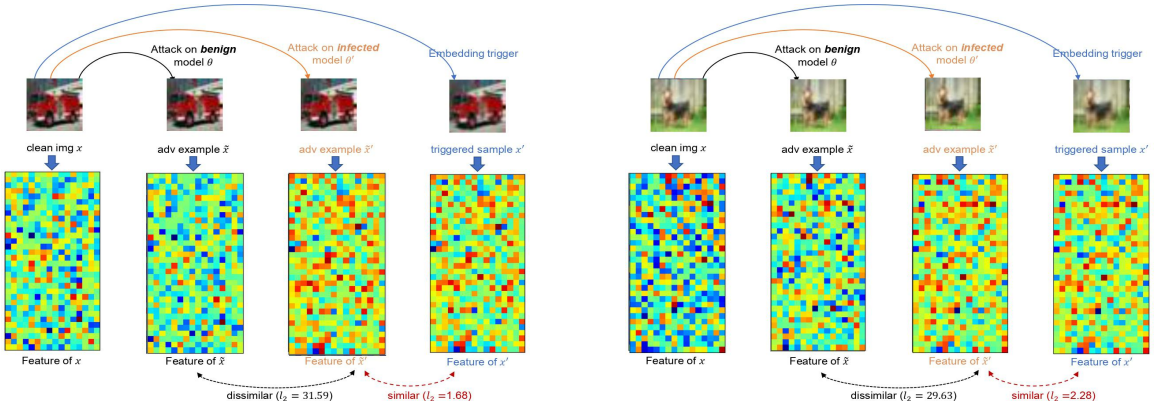

Figure 5. The two images are sampled from CIFAR-10, with size of 32×32, under WaNet Attack

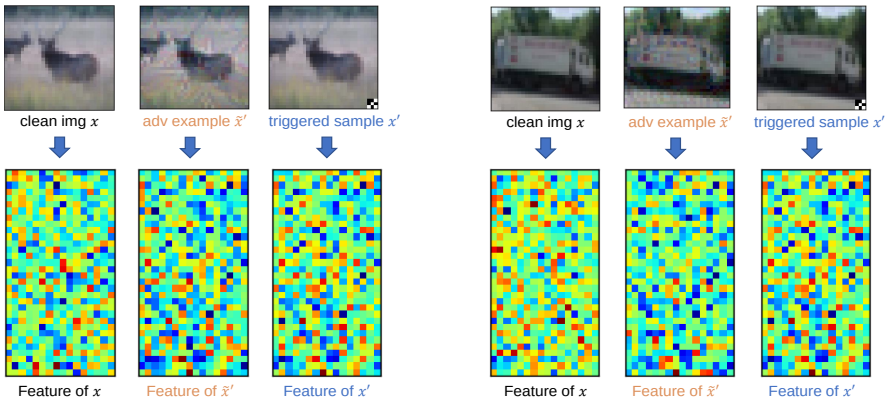

Figure 6. The two images are sampled from CIFAR-10, with size of 32×32, under BadNet Attack.

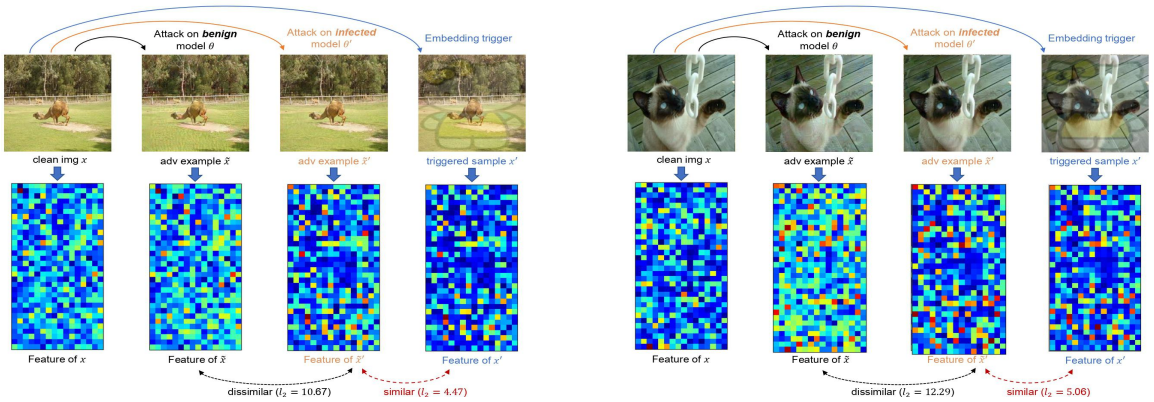

Figure 7. Two images are sampled from ImageNet-1K, with size of 224×224, under Blend Attack.