# RETHINKING THE TRIGGER OF BACKDOOR ATTACK

60212252 한우혁

# 1. Introduction

- DNN(Deep Neural Networks) - demonstrate their superior performance in a variety of applications

- malicious perturbation could be optimized to encourage that the perturbed sample will be misclassified
  -> imperceptible to human eyes

- Regular(non-optimized) perturbations could also mislead DNNs, through influencing the model weights in the training process -> Backdoor attack

# 1. Introduction

- When the trigger in the attacked testing image is different from that used in training, can it still activate the hidden backdoor?

- Explorer two basic trigger – location & appearance

- defense method - the testing sample is spatially transformed before the prediction

# 2. Related Work

- Attack
  - various backdoor attacks were proposed
  - most of them static trigger setting and mechanisms and properties is left far behind

- Defense
  - high complexity or relatively low clean accuracy
  - already bypassed by subsequently adaptive attacks

# 3. THE PROPERTY OF EXISTING ATTACKS WITH STATIC TRIGGER

- BACKDOOR ATTACK WITH STATIC TRIGGER
  - C(·; w) – model
  - $y_{target}$ - target label
  - $D_{train} = \{(x, y)\}$ with $x \in \{0,1, \ldots, 255\}^{C*W*H}$ -> benign train set

  process
  1. generate the poisoned image($x_{poisoned}$) with target label($y_{target}$)
  2. adopt both the benign and poisoned samples for training

# 3. THE PROPERTY OF EXISTING ATTACKS WITH STATIC TRIGGER

$$x_{\text{poisoned}} = S\left(x; x_{\text{trigger}}\right) = (1-\alpha) \otimes x + \alpha \otimes x_{\text{trigger}}$$

Poisoned image $x_{\text{poisoned}}$ generated through a stamping process S based on trigger $x_{\text{trigger}}$ and normal image x

$\alpha \in [0,1]^{C*W*H}$is a trade-off hyper-parameter

$\otimes$ indicates the element-wise product

# 3. THE PROPERTY OF EXISTING ATTACKS WITH STATIC TRIGGER

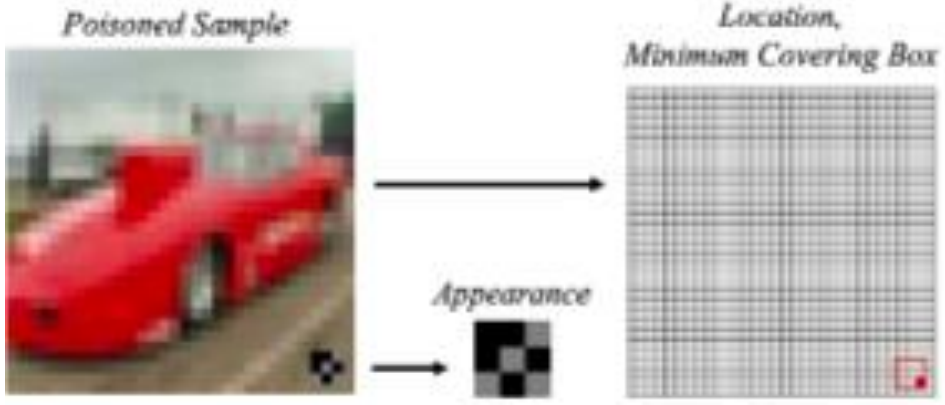
$$\min_{w} \mathbb{E}_{(\mathbb{x},\mathbb{y})\in\mathbb{D}_{\mathbf{poisoned}}\cup\mathbb{D}_{\mathbf{benign}}} \mathcal{L}(C(x;w),y),$$

$\mathbb{D}_{\mathbf{benign}}$ – all benign samples used for backdoor training

$$\mathbb{D}_{\mathbf{poisoned}} = \{(\mathrm{x}_{\mathbf{poisoned}}, \mathrm{y}_{\mathbf{target}})\}$$

# 3. THE PROPERTY OF EXISTING ATTACKS WITH STATIC TRIGGER

- THE EFFECTS OF DIFFERENT CHARACTERISTICS
  - Backdoor trigger = location & appearance
  - Def1 – Minimum Covering Box
    - minimum bounding box in the poisoned image covering the whole trigger pattern
  - Def2 – Two Characteristics of Backdoor Trigger



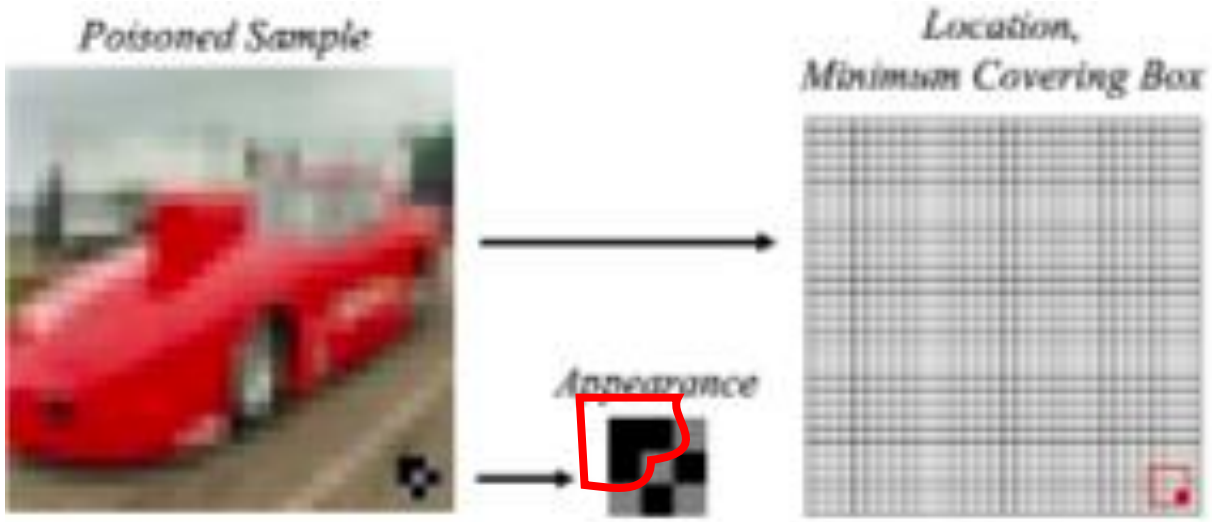
# 3. THE PROPERTY OF EXISTING ATTACKS WITH STATIC TRIGGER

- Evaluation Criteria of Attacks

$$ASR_C(S) = \Pr_{(x,y)\in\mathcal{D}_{test}}\left[C\left(S(x;\widehat{w})\right) = y_{\text{target}} \mid y \neq y_{\text{target}}\right]$$

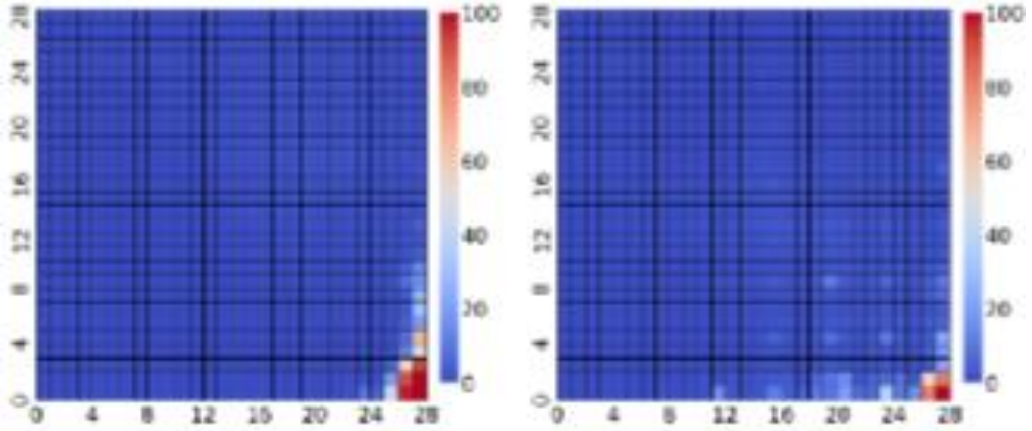
# 3. THE PROPERTY OF EXISTING ATTACKS WITH STATIC TRIGGER

- Model – VGG-19 / ResNet-34 / CIFAR10



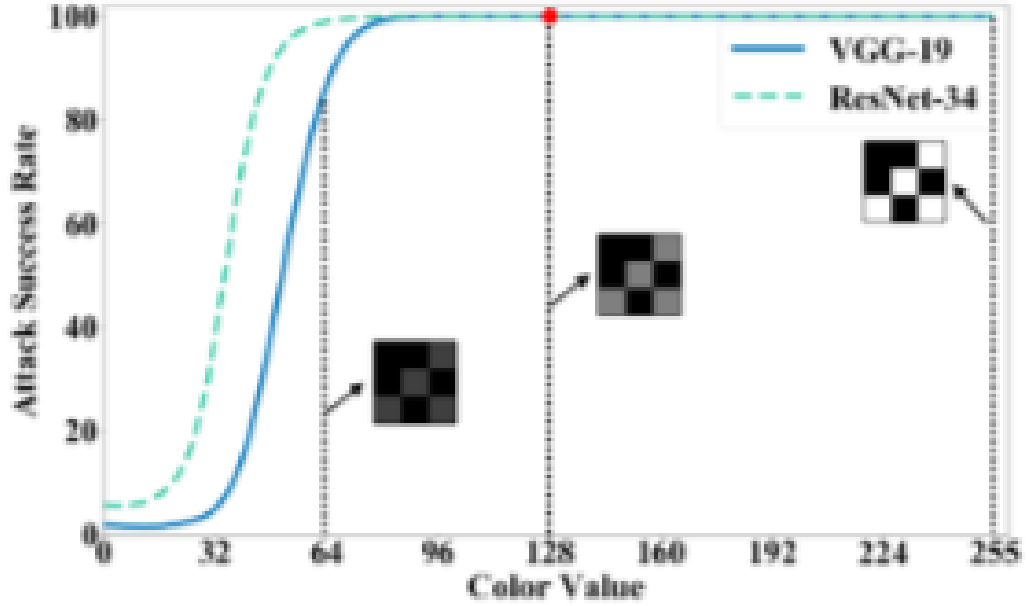
# 3. THE PROPERTY OF EXISTING ATTACKS WITH STATIC TRIGGER

- Effect of Location
  - Appearance preserve



(a) VGG-19 (b) ResNet-34

- Effect of appearance
  - Location preserve



# 4. FURTHER EXPLORATIONS OF THE PROPERT

1. Is it possible to utilize such a sensitivity to defend the current backdoor attacks with static trigger?

2. How to enhance the robustness of the backdoor attack to the change of trigger?

# 4. FURTHER EXPLORATIONS OF THE PROPERT

## BACKDOOR DEFENSE VIA TRANSFORMATIONS

change the location or appearance of the trigger in the inference process

-> modified trigger may fail to activate the backdoor hidden in the model

But, user doesn’t have the information about the trigger

# 4. FURTHER EXPLORATIONS OF THE PROPERT

Def 3 – transformation-based defense

introducing a transformation-based pre-processing module on the testing image before prediction

Advantage

1. only requires the transform the testing image
2. defend different attacks with static trigger simultaneously
3. defender does not need to have any clean samples or modify the model parameters

# 4. FURTHER EXPLORATIONS OF THE PROPERT

Transformation-Based Enhancement and Physical Backdoor attack

Enhance the transformation robustness

Def 4 – Transformation Robustness

$$R_T(S) = \mathrm{ASR}\big(T(S)\big),$$

$$\mathrm{ASR}\big(T(S)\big) = \Pr_{(\mathrm{x},\mathrm{y})\in\mathcal{D}}\left[C\left(T\big(S(x)\big)\right) = y_{\mathrm{target}} \,\middle|\, y \neq y_{\mathrm{target}}\right]$$

# 4. FURTHER EXPLORATIONS OF THE PROPERT

Improving transformation-robustness

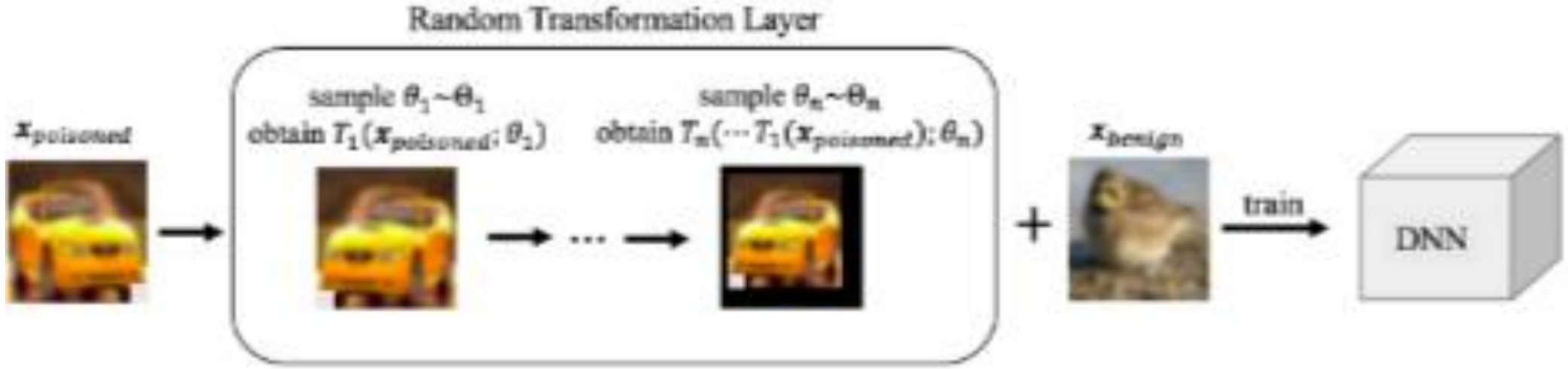
- determine the **compound transformation** and the **corresponding parameter θ** used by defenders

$$\Theta_i = \{\theta \mid \mathrm{dist}_i(\theta, I) \leq \epsilon_i\},$$

$$\min_{w} \mathbb{E}_{\theta}\left[\mathbb{E}_{(\mathbb{x},\mathbb{y})\in \mathbb{D}_{\text{poison}}^{(\mathbb{T}(\cdot;\theta))} \cup \mathbb{D}_{\mathbf{benign}}}\left[L(C(\boldsymbol{x}; w), y)\right]\right].$$

# 4. FURTHER EXPLORATIONS OF THE PROPERT

## Training Process



# 4. FURTHER EXPLORATIONS OF THE PROPERT

- connection between the proposed attack enhancement and physical attack
  - relative distance and angle between the photo and the camera is varied in practice
  - location and appearance of the trigger image may be different trigger used for training

# 5. Expreiment

- left-right flipping(Flip) & Padding after shrinking(ShrinkPad)
  - Shrink few pixels & random zero padding

- Setting
  - Blended Attack
  - Consistent Attack
  - BadNets

  - Fine pruning
  - Neural cleans
  - Auto-encoder
  - standard

# 5. Expreiment

| Model Architectures → | VGG-19 | | | | | | ResNet-34 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Attack Methods → | BadNets | | Blended Attack | | Consistent Attack | | BadNets | | Blended Attack | | Consistent Attack | |
| Defense Methods ↓ | Clean | ASR | Clean | ASR | Clean | ASR | Clean | ASR | Clean | ASR | Clean | ASR |
| Standard | 91.9 | 100 | 91.5 | 100 | 91.3 | 95.6 | 94.1 | 100 | 93.1 | 100 | 93.1 | 98.7 |
| Fine-Pruning | 91.3 | 0.7 | 83.6 | 0.2 | 72.6 | 0.1 | 92.1 | 0 | 91.9 | 0.3 | 92.0 | 18.9 |
| Neural Cleanse | 83.3 | 0.6 | 90.6 | 0.4 | 86.4 | 0.7 | 91.4 | 0.7 | 91.4 | 0.5 | 91.2 | 1.4 |
| Auto-Encoder | 86.4 | 2.1 | 86.0 | 1.7 | 85.4 | **2.3** | 87.5 | 2.7 | 87.2 | 1.9 | 88.4 | **2.1** |
| Flip (Ours) | **91.0** | **1.1** | **91.1** | **0.9** | **90.5** | 95.7 | **93.6** | **0.8** | **92.8** | **0.8** | **92.3** | 98.8 |
| ShrinkPad-4 (Ours) | 87.6 | 1.6 | 88.3 | 1.8 | 87.5 | 3.7 | 91.4 | 1.5 | 90.6 | 1.8 | 89.9 | 4.8 |

# 5. Expreiment

- Attack enhancement
  - All random(transform layer, flip, shrinkpad …)
    + only one hyper parameter(maximal shrinking size(4pixels))

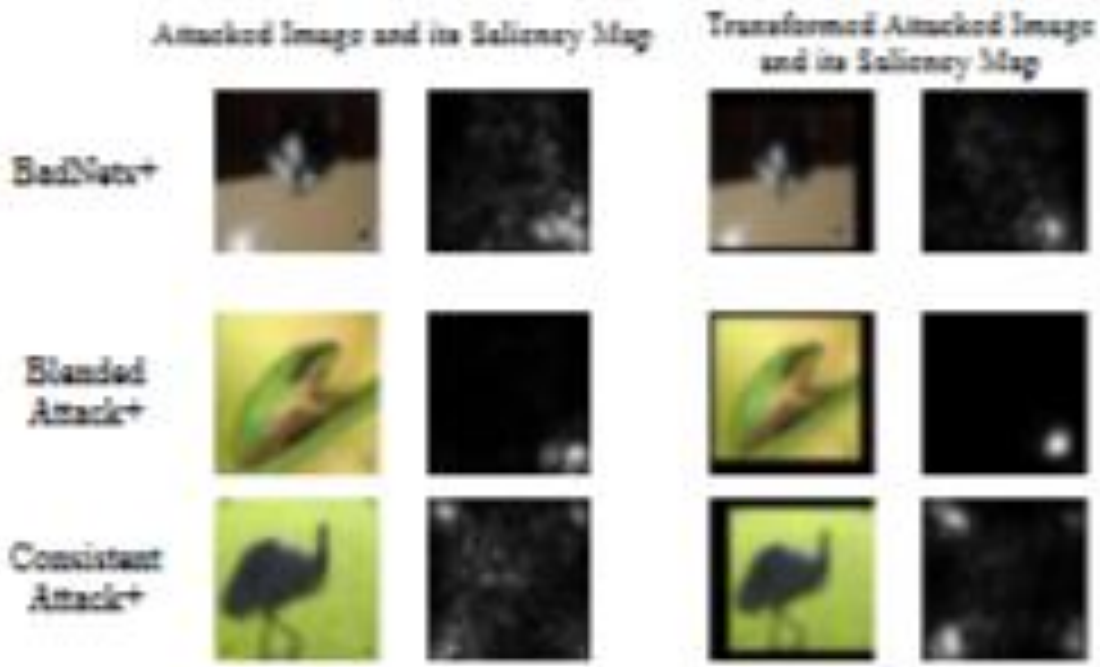
| Model Architectures → | VGG-19 | | | | ResNet-34 | | | |
|---|---|---|---|---|---|---|---|---|
| Attacks ↓, Defenses → | Standard | Flip | ShrinkPad-2 | ShrinkPad-4 | Standard | Flip | ShrinkPad-2 | ShrinkPad-4 |
| BadNets | **100.0** | 1.1 | 22.7 | 1.6 | **100.0** | 0.8 | 14.9 | 1.5 |
| BadNets+ | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** |
| Blended Attack | **100.0** | 0.9 | 40.8 | 1.8 | **100.0** | 0.8 | 18.2 | 1.8 |
| Blended Attack+ | 99.9 | **99.9** | **100.0** | **98.7** | **100.0** | **100.0** | **100.0** | **99.5** |
| Consistent Attack | **95.6** | **95.7** | 67.1 | 3.7 | **98.7** | **98.8** | 24.2 | 4.8 |
| Consistent Attack+ | 86.0 | 86.3 | **97.2** | **90.9** | 96.4 | 97.3 | **97.4** | **98.7** |

# 5. Expreiment

- Standard attack vs enhancement attack
- Salience map
  - understand their overall behaviors by identifying critical pixels of different images
- Critical data routing paths(CDRPs)
  - discuss the layer-wise behaviors of different attack

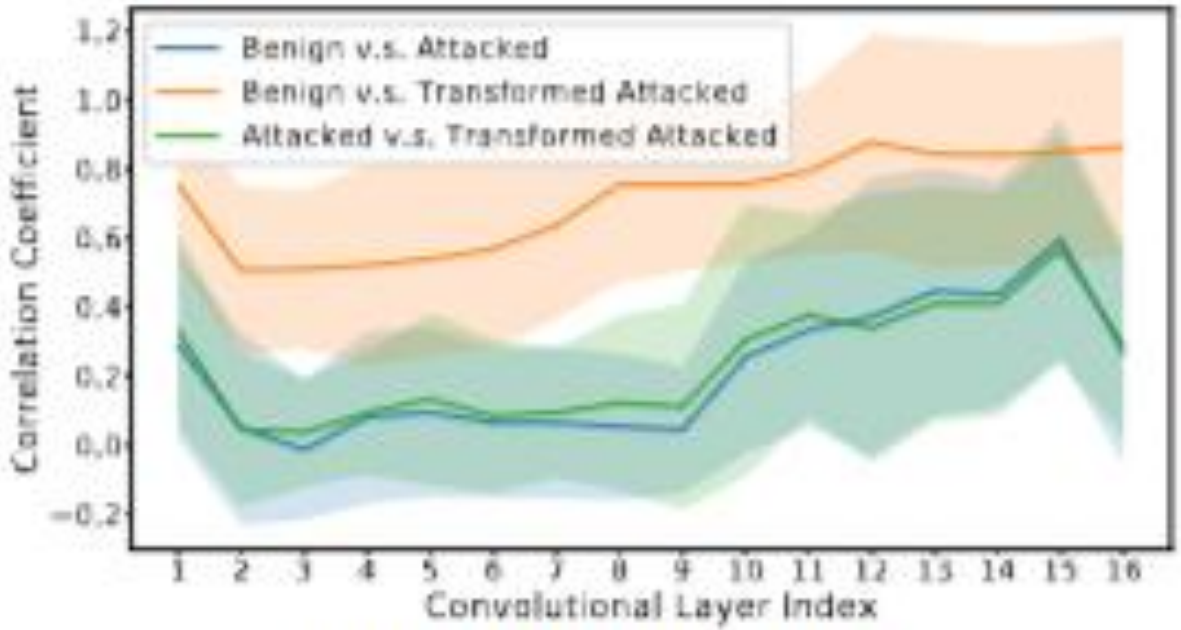
# 5. Expreiment


(a) Standard Backdoor Attack

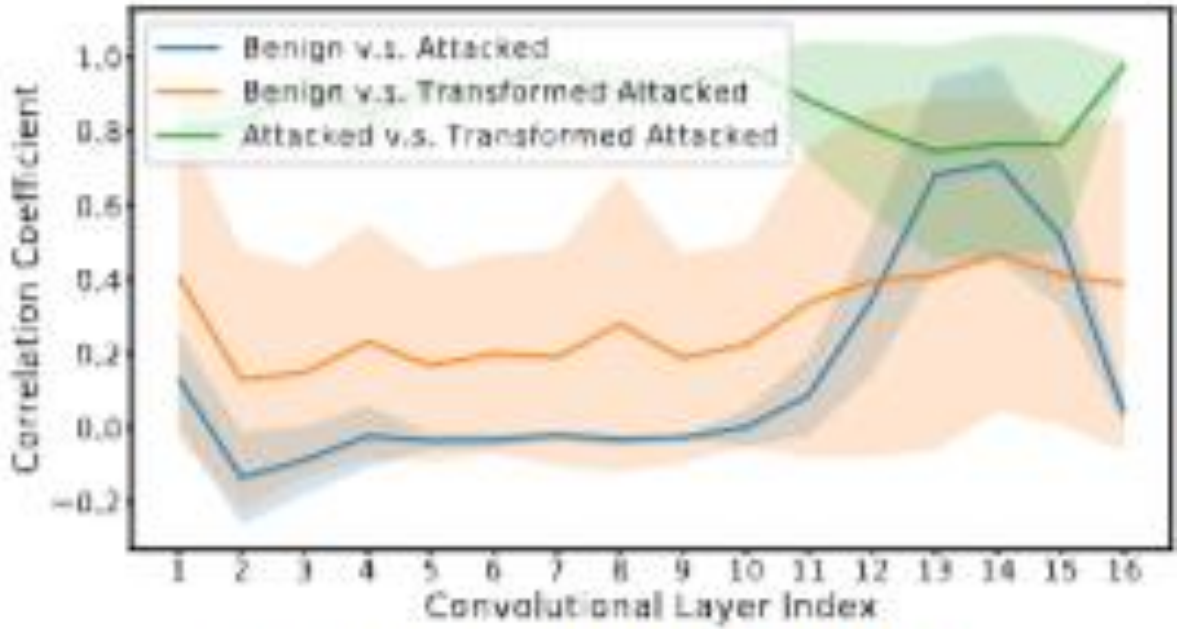



(b) Enhanced Backdoor Attack

# 5. Expreiment




(a) Standard Backdoor Attack




(b) Enhanced Backdoor Attack

# 5. Expreiment


# 6. Conclusion

- existing attacks with static trigger are transformation vulnerable & effective transformation-based defense

- enhancement by conducting the random spatial transformation on poisoned images before feeding into the training process