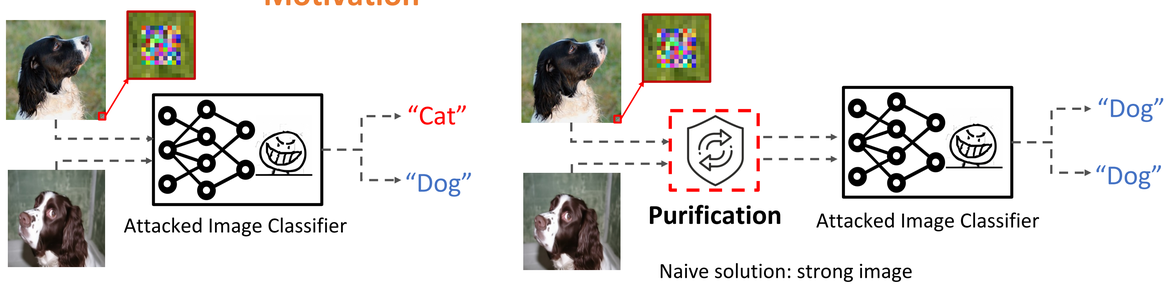
Black-box Backdoor Defense via Zero-shot Image Purification

Yucheng Shi[1], Mengnan Du[2], Xuansheng Wu[1], Zihan Guan[1], Jin Sun[1], Ninghao Liu[1]

1.University of Georgia. 2.New Jersey Institute of Technology.

Motivation



Advantages of our Purification :

- **Attack Pattern Removal,**
- **Semantic Information Retaining,**
- **Zero-shot Capability,**
- **Black-Box Adaptability,**
- **Model/Task-agnostic,**
- **.....**

Step1: Transformation to destroy patterns

$\mathbf{x}$:clean images $\mathbf{x}_0$:purified images $\mathbf{p}$:attack pattern

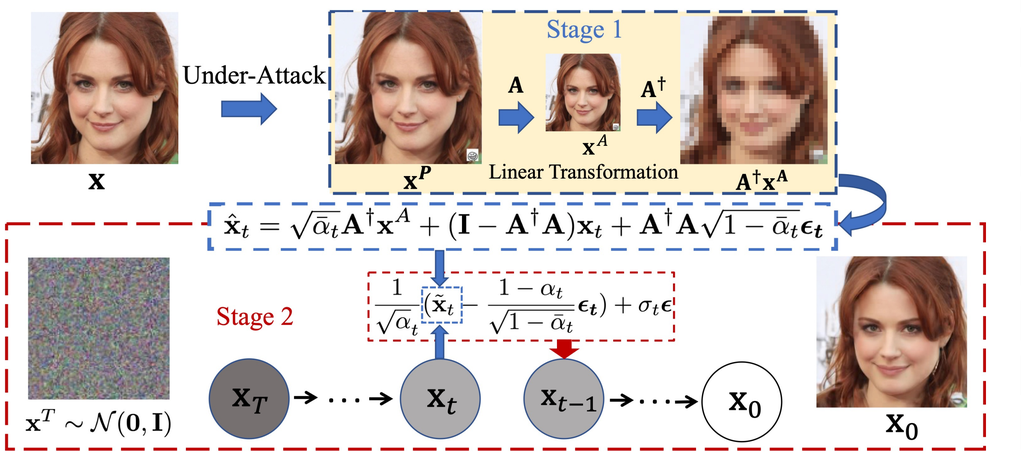
$\mathbf{A}$:linear transformation

Transformed image: $\mathbf{A}(\mathbf{x}_0+\mathbf{p}) = \mathbf{A}(\mathbf{x}+\mathbf{p}) = \mathbf{x}^A$

However, if directly using the transformed image for classification: **1. ASR drop (Good) 2. CA drop(Bad!)**

Q: Can we obtain $\mathbf{x}_0$ from $\mathbf{x}^A$?
A: Yes, using RND theory.

$$\mathbf{x}_0 = \mathbf{A}^\dagger\mathbf{x}^A - \mathbf{A}^\dagger\mathbf{A}\mathbf{p} + (\mathbf{I}-\mathbf{A}^\dagger\mathbf{A})\,\mathbf{x}_0 \quad \text{(T1)}$$



Step2: Guided diffusion process to recover images

Apply (T1) with a diffusion process, we have:

$$\mathbf{x}'_t = \sqrt{\bar{\alpha}_t}\mathbf{A}^\dagger\mathbf{x}^A - \sqrt{\bar{\alpha}_t}\mathbf{A}^\dagger\mathbf{A}\mathbf{p} + (\mathbf{I}-\mathbf{A}^\dagger\mathbf{A})\mathbf{x}_t + \mathbf{A}^\dagger\mathbf{A}\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$$

Omit in zero-shot setting
no prior on attack patterns

$$\hat{\mathbf{x}}_t = \sqrt{\bar{\alpha}_t}\mathbf{A}^\dagger\mathbf{x}^A + (\mathbf{I}-\mathbf{A}^\dagger\mathbf{A})\mathbf{x}_t + \mathbf{A}^\dagger\mathbf{A}\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$$

$$\mathbf{x}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(\hat{\mathbf{x}}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t) + \sigma_t\boldsymbol{\epsilon}$$ Repeat T times

Q: Will this approximation accumulate error exponentially?
A: No, the error is bounded by a small value (See Theorem 1).

Step3: Further improvement

- Introduce multiple kinds of linear transformations for better attack pattern destruction.

$$\hat{\mathbf{x}}_t = \tfrac{1}{N}(\hat{\mathbf{x}}_t^1 + \hat{\mathbf{x}}_t^2 + , ..., + \hat{\mathbf{x}}_t^N)$$

- Introduce confidence score, a mixup-based interpolation to mitigate approximation errors.

$$\tilde{\mathbf{x}}_t = (1-\bar{\alpha}_t^\lambda)\hat{\mathbf{x}}_t + \bar{\alpha}_t^\lambda\mathbf{x}_t$$

Speed up

- **Improve Sampling Speed →DDIM**
  Speed up **50 times**
- **Image Batch by Tiling**

  Speed up **33 times**
  Work with images in **different size**
- **Detect first, Purify Then**
  Speed up **39 times**

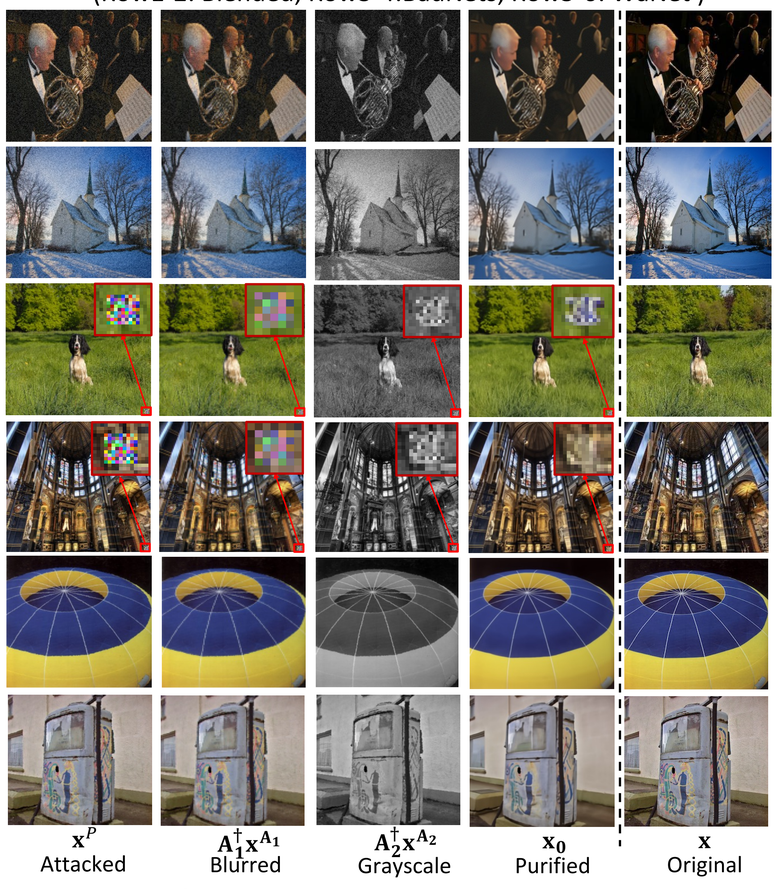
Quantitative Results of Defense

Table 1: The clean accuracy (CA %), the attack success rate (ASR %), and the poisoned accuracy (PA %) of defense methods against different backdoor attacks. *None* means no attacks are applied.

Dataset	Attack	No Defense			ShrinkPad (defense)			Blur (defense)			**ZIP** (Ours)		
		CA ↑	ASR ↓	PA ↑	CA ↑	ASR ↓	PA ↑	CA ↑	ASR ↓	PA ↑	CA ↑	ASR ↓	PA ↑
CIFAR-10 (32 × 32) (10 classes)	None	80.15	—	—	—	—	—	—	—	—	—	—	—
	BadNet	82.31	99.98	10.00	62.89	9.34	63.12	58.19	21.78	53.47	**78.97**	**5.53**	**79.10**
	Blended	80.26	99.96	10.03	58.97	**2.28**	40.22	55.91	3.04	49.91	**72.62**	7.75	**57.98**
	PhysicalBA	85.30	98.73	11.20	**82.84**	90.50	18.37	41.84	**1.09**	41.37	80.10	4.33	**80.33**
	Average	82.62	99.56	10.41	68.23	34.04	40.57	51.98	8.64	48.25	**77.23**	**5.87**	**72.47**
GTSRB (32 × 32) (43 classes)	None	96.95	—	—	—	—	—	—	—	—	—	—	—
	BadNet	96.53	99.99	5.70	78.33	**5.81**	78.82	95.98	7.33	95.11	**96.18**	6.19	**96.03**
	Blended	96.58	99.89	5.79	76.76	10.54	56.41	93.68	11.07	73.91	**95.74**	**8.53**	**81.27**
	PhysicalBA	96.83	100.00	5.70	**97.41**	100.00	5.70	91.00	**5.53**	90.53	95.44	6.57	**94.91**
	Average	96.65	99.96	5.73	84.17	38.78	46.98	93.55	7.98	86.52	**95.79**	**7.10**	**90.74**
Imagenette (256 × 256) (10 classes)	None	84.58	—	—	—	—	—	—	—	—	—	—	—
	BadNet	84.99	94.53	14.98	71.23	8.56	70.72	81.47	16.45	79.94	**84.05**	**7.55**	**83.97**
	Blended	86.14	99.85	10.19	74.06	20.63	36.10	78.95	79.41	25.57	**81.42**	**8.35**	**78.36**
	PhysicalBA	90.67	72.94	34.29	**90.21**	96.81	13.07	84.84	32.40	74.87	87.26	**10.91**	**86.54**
	Average	87.27	89.11	19.82	78.50	42.00	39.96	81.75	42.75	60.13	**84.24**	**8.94**	**82.96**

Qualitative Results of Defense

(Row1-2: Blended; Row3-4:BadNets; Row5-6: WaNet)



Future Applications:

- Remove watermark patterns for IP protection.
- Defend against attacks for image generation models.
- Defend against attacks of point cloud data.