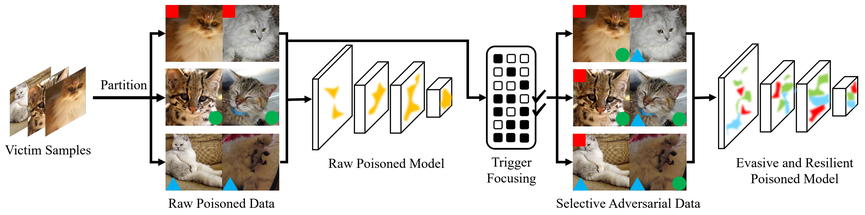
LOTUS: Evasive and Resilient Backdoor Attacks through Sub-Partitioning

Siyuan Cheng, Guanhong Tao, Yingqi Liu†, Guangyu Shen, Shengwei An,
Shiwei Feng, Xiangzhe Xu, Kaiyuan Zhang, Shiqing Ma‡, Xiangyu Zhang

LOTUS Overview



Attack Definition

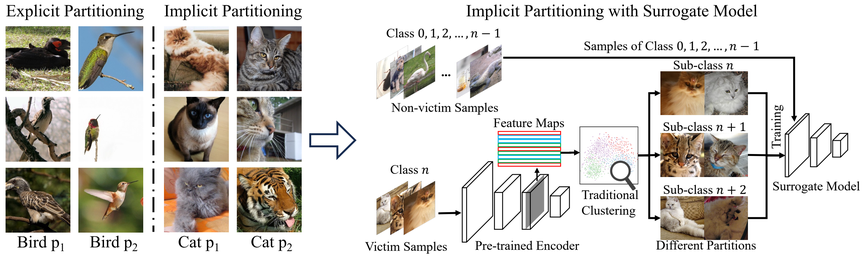
Attack Goal: Achieve high accuracy and ASR. Only samples from the ***victim class***, stamped with the ***appropriate partition*** trigger, can activate the backdoor.

Annotations: $\bar{M}_{\bar{\theta}}$: Backdoor model; $\boldsymbol{x}_V^{p_i}$: Samples of victim class V and partition p_i; $\mathbb{T}_i$: Trigger of i-th partition; $y_{\neg V}$: Labels other than victim class V.

$$\mathbb{E}_{(\boldsymbol{x},y)\sim D}[\mathcal{L}(\bar{M}_{\bar{\theta}}(\boldsymbol{x}),y)] + \sum_{i=1}^{n}\{\mathbb{E}_{(\boldsymbol{x}_V^{p_i},y_V)\sim D}[\mathcal{L}(\bar{M}_{\bar{\theta}}(\mathbb{T}_i\oplus\boldsymbol{x}_V^{p_i}),y_T)] + \mathbb{E}_{(\boldsymbol{x}_{\neg V}^{p_i},y_{\neg V})\sim D}[\mathcal{L}(\bar{M}_{\bar{\theta}}(\mathbb{T}_i\oplus\boldsymbol{x}_{\neg V}^{p_i}),y_{\neg V})] + \mathbb{E}_{(\boldsymbol{x}_V^{p_i},y_V)\sim D}\sum_{\mathcal{T}\in\mathcal{P}(\{\mathbb{T}_1,\mathbb{T}_2,\dots,\mathbb{T}_n\})-\{\emptyset,\{\mathbb{T}_i\}\}}[\mathcal{L}(\bar{M}_{\bar{\theta}}(\mathcal{T}\oplus\boldsymbol{x}_V^{p_i}),y_T)]\}$$

① Benign Utility Loss ② Attack Target Loss ③ Label-specific Loss ④ Dynamic Loss

① Victim-class Sample Partitioning



② Trigger Focusing

: $\boldsymbol{x}_V^{p_i}$, Clean image in victim class y_V of partition p_i

: $\boldsymbol{x}_T$, Clean image in the target class y_T

: $\boldsymbol{T}_i\oplus\boldsymbol{x}_V^{p_i}$, Stamp trigger T_i on matched partition $x_V^{p_i}$

: $\boldsymbol{T}_j\oplus\boldsymbol{x}_V^{p_i}$, Stamp trigger T_j on non-matched partition $x_V^{p_i}$

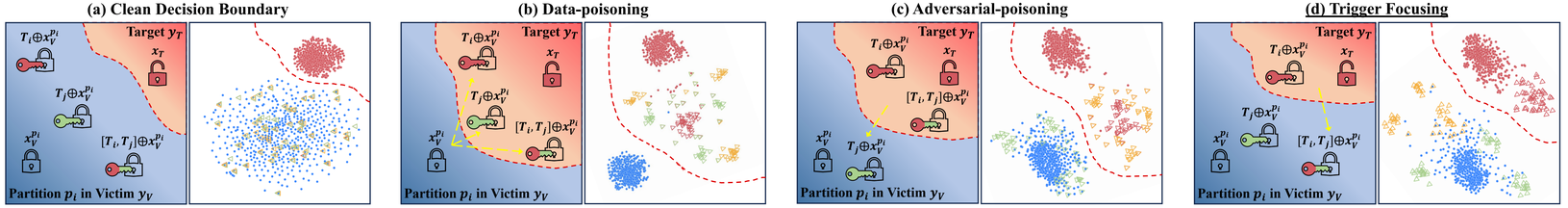
: $[\boldsymbol{T}_i,\boldsymbol{T}_j]\oplus\boldsymbol{x}_V^{p_i}$, Stamp trigger combination T_i and T_j on partition $x_V^{p_i}$

: Decision boundary

Refine the Dynamic Loss ⇩ Only consider 2 negative cases

$$\mathbb{E}_{(\boldsymbol{x}_V^{p_i},y_V)\sim D}\sum_{j=1,\,j\neq i}^{n}[\mathcal{L}(\bar{M}_{\bar{\theta}}(\mathbb{T}_j\oplus\boldsymbol{x}_V^{p_i}),y_V) + \mathcal{L}(\bar{M}_{\bar{\theta}}([\mathbb{T}_i,\mathbb{T}_j]\oplus\boldsymbol{x}_V^{p_i}),y_V)]$$

Efficient and Effective Trigger Focusing

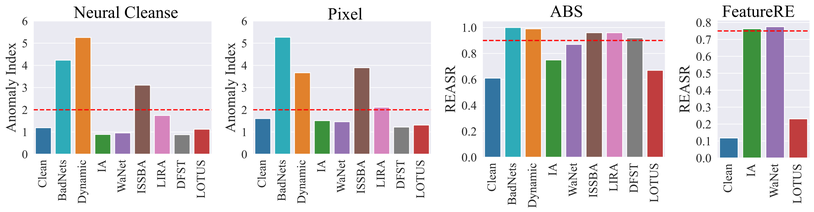


Evaluation Results

(1) Attack Effectiveness

PA	Dataset	Model	Acc.	BA	ASR	ASR-other
K-means	CIFAR-10	VGG11	92.16%	92.04%	93.80%	4.77% ± 19.27%
		ResNet18	95.22%	94.71%	94.30%	4.39% ± 17.08%
	CIFAR-100	Densenet	75.14%	75.15%	92.00%	4.36% ± 14.24%
		PRN34	74.70%	74.52%	89.00%	5.43% ± 13.50%
	CelebA	WRN	80.47%	79.40%	92.33%	6.87% ± 17.49%
	RImageNet	ResNet50	97.77%	97.19%	93.87%	2.16% ± 19.34%
GMM	CIFAR-10	ResNet18	95.22%	94.59%	90.70%	4.80% ± 21.38%
	CIFAR-100	PRN34	74.70%	74.02%	91.00%	2.21% ± 12.57%
	CelebA	WRN	80.47%	79.66%	92.53%	5.39% ± 16.77%
	RImageNet	VGG16	96.51%	95.93%	93.52%	3.11% ± 14.39%
Sec.	RImageNet	VGG16	96.51%	96.36%	96.50%	1.79% ± 13.24%
		ResNet50	97.77%	97.08%	92.50%	2.14% ± 16.53%

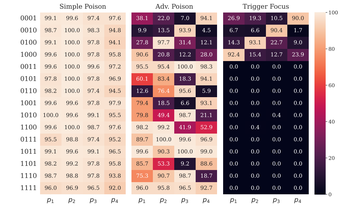
(2) Evasiveness against Backdoor Detection



(3) Resilience against Backdoor Mitigation

Attacks	No Defense		Fine-tuning		Fine-pruning		NAD		ANP	
	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
BadNets	92.02%	100.00%	89.31%	1.74%	91.70%	0.53%	87.81%	0.80%	89.15%	0.32%
Dynamic	91.81%	100.00%	88.87%	2.91%	91.39%	22.03%	89.11%	2.90%	88.25%	12.81%
IA	91.70%	99.65%	87.74%	2.78%	91.07%	0.17%	87.14%	2.29%	88.73%	1.98%
WaNet	91.22%	98.57%	89.56%	1.37%	90.22%	1.07%	89.74%	1.40%	89.07%	0.54%
ISSBA	91.67%	99.96%	87.73%	2.72%	91.12%	14.27%	87.97%	2.83%	85.64%	10.01%
LIRA	91.70%	100.00%	89.96%	2.19%	91.29%	12.14%	90.23%	2.32%	89.70%	**37.91%**
DFST	91.81%	99.97%	88.49%	22.86%	91.47%	21.61%	88.52%	24.66%	87.13%	36.17%
LOTUS	91.54%	93.80%	88.10%	**46.90%**	91.14%	**44.90%**	87.61%	**42.30%**	88.14%	34.90%

(4) Ablation Study



(5) Visualization of Inverted Triggers