NEURAL INFORMATION PROCESSING SYSTEMS

Breaking the False Sense of Security in Backdoor Defense through Re-Activation Attack

Mingli Zhu[1], Siyuan Liang[2], Baoyuan Wu[1,†]

[1] School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China

[2] National University of Singapore, Singapore

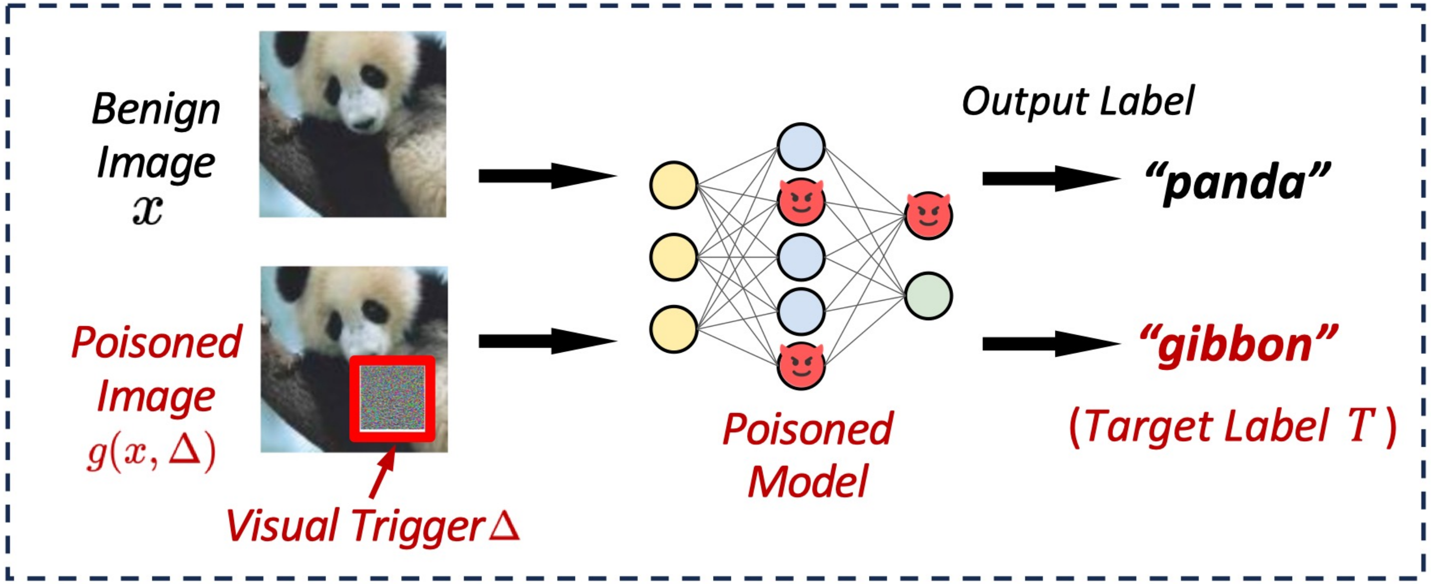
NeurIPS 2024

NUS National University of Singapore

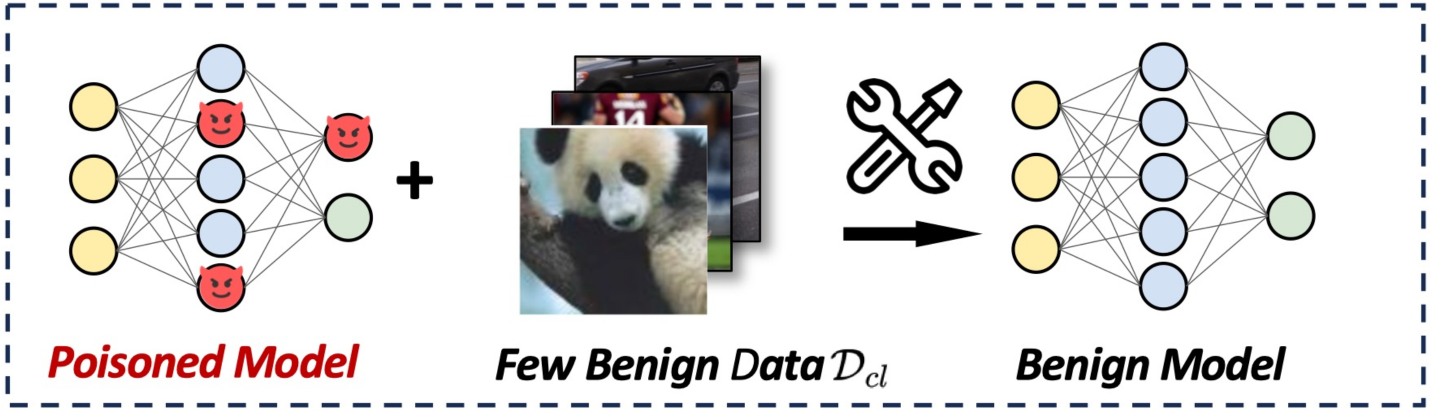
Outline

- **Introduction**
- **Backdoor Re-Activation Attack**
- **Experimental Evaluation**

Backdoor Attack



Post-training Backdoor Defense

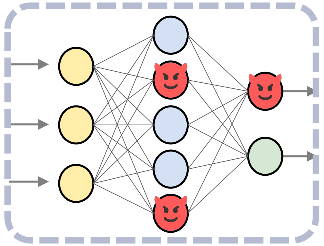


The whole pipeline

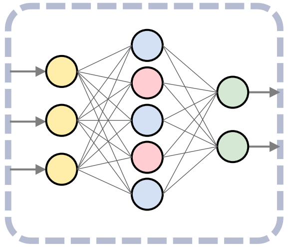
Table 1: Illustration of the pipeline of backdoor attack and defense.

Stage	Task description	Input/Output	Goal
Reference	Clean model training	$\mathcal{D}/f_{\boldsymbol{\theta}_{\mathrm{C}}}$	$f_{\boldsymbol{\theta}_{\mathrm{C}}}(\boldsymbol{x}) = y,\ f_{\boldsymbol{\theta}_{\mathrm{C}}}(\boldsymbol{x}_{\boldsymbol{\xi}}) \neq t$
I: Pre-training & II: Training	Backdoor injection	$\mathcal{D}/f_{\boldsymbol{\theta}_{\mathrm{A}}}, \mathcal{D}_p$	$f_{\boldsymbol{\theta}_{\mathrm{A}}}(\boldsymbol{x}) = y,\ f_{\boldsymbol{\theta}_{\mathrm{A}}}(\boldsymbol{x}_{\boldsymbol{\xi}}) = t$
III: Post-training	Backdoor defense	$f_{\boldsymbol{\theta}_{\mathrm{A}}}/f_{\boldsymbol{\theta}_{\mathrm{D}}}$	$f_{\boldsymbol{\theta}_{\mathrm{D}}}(\boldsymbol{x}) = y,\ f_{\boldsymbol{\theta}_{\mathrm{D}}}(\boldsymbol{x}_{\boldsymbol{\xi}}) \neq t$
IV: Inference	Backdoor re-activation	$\boldsymbol{x}, \boldsymbol{\xi}, f_{\boldsymbol{\theta}_{\mathrm{D}}}/f_{\boldsymbol{\theta}_{\mathrm{D}}}(\boldsymbol{x}_{\boldsymbol{\xi}'})$	$f_{\boldsymbol{\theta}_{\mathrm{D}}}(\boldsymbol{x}) = y,\ f_{\boldsymbol{\theta}_{\mathrm{D}}}(\boldsymbol{x}_{\boldsymbol{\xi}'}) = t$

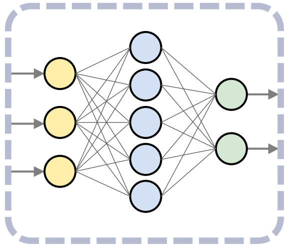
Motivation: While existing backdoor defense strategies have shown promising performance on reducing attack success rates, can we confidently claim that the backdoor threat has truly been eliminated from the model?



Backdoor attack model



Backdoor defense model



Clean model

Outline

- Introduction
- **Backdoor Re-Activation Attack**
- Experimental Evaluation

Backdoor existence coefficient (BEC)

Calculated through the following three steps:

- Backdoor neuron identification

$$TAC_k^{(l)}(\mathcal{D}_p, \mathcal{D}_c) = \frac{1}{|\mathcal{D}_p|} \sum_{(\boldsymbol{x_\xi}, \boldsymbol{x}) \in (\mathcal{D}_p, \mathcal{D}_c)} \left\| f_k^{(l)}(\boldsymbol{x}) - f_k^{(l)}(\boldsymbol{x_\xi}) \right\|_2$$

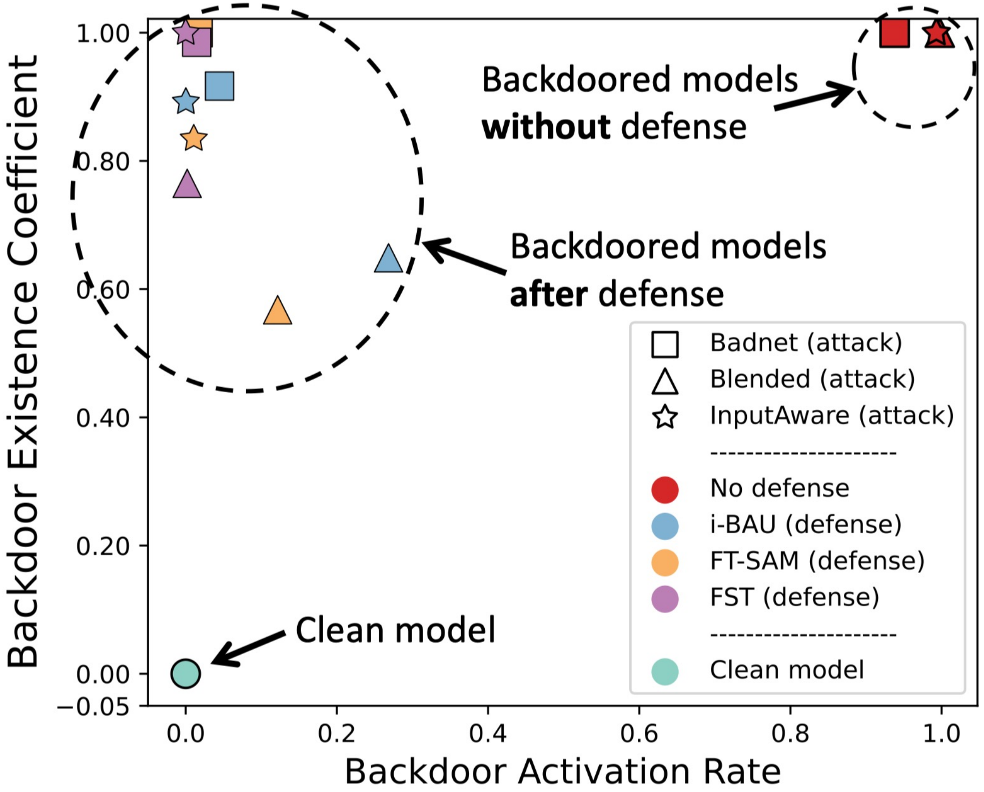
- Backdoor effect similarity metric

$$S_{\mathrm{D,A}}^{(l)}(\mathcal{D}_p) = \mathbf{CKA}\left(\tilde{m}_{\mathrm{D}}^{(l)}(\mathcal{D}_p), \tilde{m}_{\mathrm{A}}^{(l)}(\mathcal{D}_p)\right)$$

- Backdoor existence coefficient computation

$$\rho_{\mathrm{BEC}}(f_{\boldsymbol{\theta}_{\mathrm{D}}}, f_{\boldsymbol{\theta}_{\mathrm{A}}}, f_{\boldsymbol{\theta}_{\mathrm{C}}}; \mathcal{D}_p) = \frac{1}{N} \sum_{l=1}^{N} \frac{S_{\mathrm{D,A}}^{(l)}(\mathcal{D}_p) - S_{\mathrm{C,A}}^{(l)}(\mathcal{D}_p)}{S_{\mathrm{A,A}}^{(l)}(\mathcal{D}_p) - S_{\mathrm{C,A}}^{(l)}(\mathcal{D}_p)} \in [0, 1].$$

Conclusion: the original backdoors just lie dormant rather than being eliminated in defense models.



Backdoor existence coefficient VS backdoor activation rate across different models.

Backdoor re-activation attack under three different scenarios

- **White-box setting**:

$$\min_{\|\Delta_{\boldsymbol{\xi}}\|_p \le \rho} \mathcal{L}_{tot}(\Delta_{\boldsymbol{\xi}}; \mathcal{D}_p, f) = \sum_{(\boldsymbol{x}_{\boldsymbol{\xi}}, t) \in \mathcal{D}_p} \mathcal{L}_{\mathrm{CE}}(f(\boldsymbol{x}_{\boldsymbol{\xi}+\Delta_{\boldsymbol{\xi}}}), t) - \lambda \log \left(1 - \max_{k \neq t} \frac{e^{f_k(\boldsymbol{x}_{\boldsymbol{\xi}+\Delta_{\boldsymbol{\xi}}})}}{\sum_{i=1}^{N} e^{f_i(\boldsymbol{x}_{\boldsymbol{\xi}+\Delta_{\boldsymbol{\xi}}})}}\right),$$

- **Black-box setting** : Universal Square Attack

- **Transfer-based attack setting** :

$$\Delta^*_{\boldsymbol{\xi}} = \arg\min_{\|\Delta_{\boldsymbol{\xi}}\|_p \le \rho} \sum_{i=1}^{M} \mathcal{L}_{tot}(\Delta_{\boldsymbol{\xi}}; \mathcal{D}_p, f_i).$$

Algorithm 1 Black-box Backdoor Re-Activation Attack via Universal Square Attack (BBA) [1]

Input: Defense model f, training dataset $\mathcal{D}_p$, image shape c, h, w, norm p, perturbation bound ρ, target label $t \in 1, \ldots, K$, number of iterations N, termination condition ϵ.
Output: Perturbation $\Delta^*_{\boldsymbol{\xi}}$ as in Eq. 4.
$\hat{\boldsymbol{x}} \leftarrow \boldsymbol{x} + \mathrm{init}(\Delta_{\boldsymbol{\xi}})$ for $\boldsymbol{x} \in \mathcal{D}_p$, $\quad l^* \leftarrow \mathcal{L}_{tot}(\mathcal{D}_p, \Delta_{\boldsymbol{\xi}})$.
for $i = 0, ..., N-1$ **do**
  if ASR $> 1 - \epsilon$ **then return** $\Delta_{\boldsymbol{\xi}}$.
  else
    $h^{(i)} \leftarrow$ side length of the square to modify (according to some schedule [1]);
    $\Delta^{\text{new}}_{\boldsymbol{\xi}} \sim P\left(\rho, h^{(i)}, w, c, \Delta_{\boldsymbol{\xi}}, \hat{\boldsymbol{x}}, \boldsymbol{x}\right)$ for $\boldsymbol{x} \in \mathcal{D}_p$ (see **Appendix** B for details);
    $\hat{\boldsymbol{x}}_{\text{new}} \leftarrow$ Project $\hat{\boldsymbol{x}} + \Delta^{\text{new}}_{\boldsymbol{\xi}}$ onto $\left\{z \in \mathbb{R}^d : \|z - x\|_p \le \rho\right\} \cap [0,1]^d$ for $\boldsymbol{x} \in \mathcal{D}_p$;
    $l_{\text{new}} \leftarrow \mathcal{L}_{tot}(\hat{\boldsymbol{x}}_{\text{new}}, t)$ for $\boldsymbol{x} \in \mathcal{D}_p$;
    if $l_{\text{new}} < l^*$ then $\Delta_{\boldsymbol{\xi}} \leftarrow \Delta^{\text{new}}_{\boldsymbol{\xi}}, l^* \leftarrow l_{\text{new}}$, compute ASR;
    $i \leftarrow i + 1$;
  end if
end for
return $\Delta^*_{\boldsymbol{\xi}}$.

Outline

- Introduction
- Backdoor Re-Activation Attack
- **Experimental Evaluation**

Main Experiments

Tasks: image classification task and multimodal contrastive learning tasks.
Datasets: CIFAR-10, Tiny ImageNet, GTSRB, CC3M, ImageNet-1K.
Models: PreAct-ResNet18, VGG19-BN, CLIP model.

Table 2: Performance (%) of backdoor re-activation attack on both white-box (WBA) and black-box (BBA) scenarios with ℓ_∞-norm bound $\rho = 0.05$ against different defenses with CIFAR-10 on PreAct-ResNet18. The best results are highlighted in **boldface**.

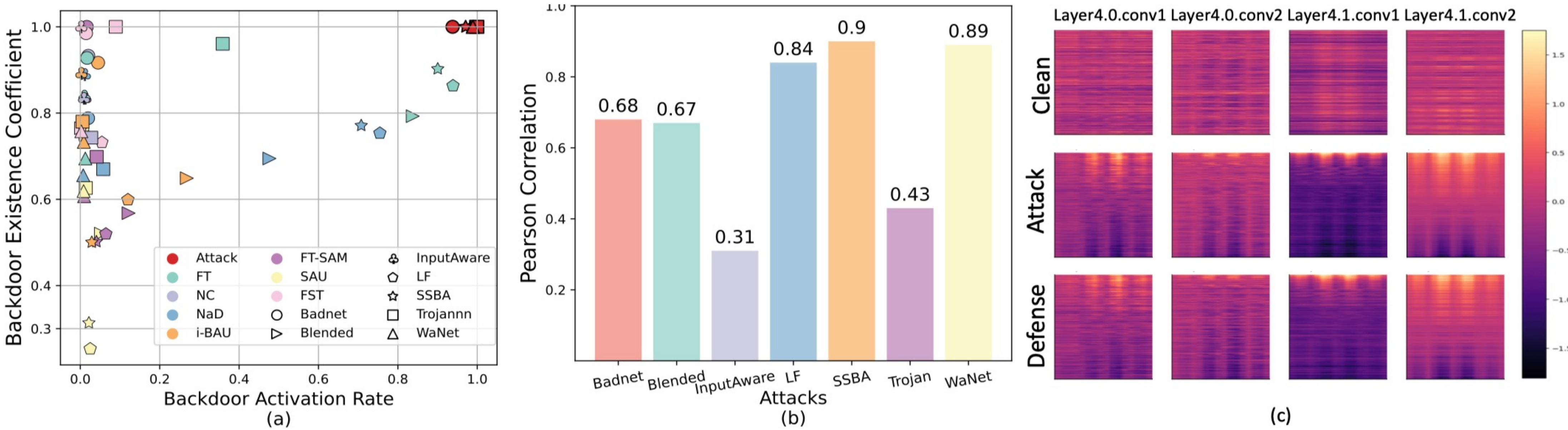
Attacks	No Defense	NC [43]			NAD [26]			i-BAU [54]			FT-SAM [59]			SAU [47]			FST [33]		
		Defense	WBA	BBA	Defense	WBA	BBA	Defense	WBA	BBA	Defense	WBA	BBA	Defense	WBA	BBA	Defense	WBA	BBA
BadNets [15]	93.79	2.01	**96.78**	27.91	1.96	**94.78**	49.66	4.48	**97.42**	54.37	1.63	**94.71**	51.23	1.30	**93.10**	37.91	1.46	**97.93**	42.69
Blended [10]	99.76	99.76	**99.93**	99.13	47.64	**99.82**	14.14	26.83	**99.63**	85.80	12.17	**99.56**	87.29	5.20	**98.37**	73.06	0.20	**99.62**	82.97
Input-Aware [34]	99.30	0.70	**92.04**	54.33	0.92	**93.80**	70.44	0.02	**21.78**	19.56	1.07	**96.19**	80.16	1.26	**85.39**	22.26	0.00	**90.72**	44.65
LF [55]	99.06	99.06	**99.41**	80.51	75.47	**99.41**	17.01	11.99	**99.04**	75.48	6.43	**97.40**	89.28	2.49	**90.74**	23.08	5.43	**98.18**	1.16
SSBA [27]	97.07	97.07	**99.90**	94.38	70.77	**99.72**	88.53	2.89	**91.29**	70.71	4.06	**92.80**	69.18	2.16	**89.86**	38.59	0.54	**94.11**	52.71
Trojan [30]	99.99	2.76	**95.26**	45.57	5.77	**96.38**	60.87	0.54	**89.58**	40.18	4.12	**96.18**	69.88	1.39	**87.61**	47.37	8.93	**97.28**	80.47
WaNet [35]	98.90	98.90	**100.00**	99.64	0.73	**96.21**	77.65	0.88	**94.67**	75.91	0.96	**94.95**	78.66	0.82	**95.33**	60.36	0.26	**97.56**	82.22
Avg	98.26	57.18	**97.62**	71.64	29.04	**97.16**	54.04	6.80	**84.77**	60.29	4.35	**95.97**	75.10	2.09	**91.48**	43.23	2.40	**96.49**	55.27

Effectiveness of attacks on CLIP models.

Performance (%) of our attack on both white-box (WBA) and transfer-based (TA) attacks with ℓ_∞ -norm bound $\rho = 0.05$ against different defenses with ImageNet1K on CLIP. Best results are highlighted in boldface.

Attack	No Defense	FT [3]			CleanCLIP [3]		
		Defense	WBA	TA	Defense	WBA	TA
BadNets [16]	96.65	64.60	82.05	**82.73**	17.29	**57.76**	47.30
Blended [10]	97.71	49.77	96.57	**98.64**	18.57	**89.61**	72.65
SIG [4]	77.71	30.91	**92.56**	87.99	21.68	**87.04**	82.55
TrojanVQA [47]	98.21	82.07	97.14	**97.46**	49.82	**87.43**	78.25
Avg	92.57	56.84	**92.08**	91.71	26.84	**80.46**	70.19

Visualization analysis



a. Backdoors exist across defense models, albeit with low ASR.

b. There is a strong correlation between ASR and BEC.

c. The defense model and backdoored model exhibit similar feature maps.

Comparison among OBA, RBA, and gUAA

- Backdoor activation mechanisms between RBA and OBA are highly similar, and both differ significantly from that of gUAA.
- Starting from the original trigger ξ , it is easier and faster to find a new trigger ξ ′ that achieves a high attack success rate (ASR).
- Compared to Δ, both the original trigger ξ and the new trigger ξ ′ are more robust to random noise.

Table 9: CKA scores between OBA, RBA, and gUAA.

Defense ⇒ Attack ↓	i-BAU $S_{\text{RBA,OBA}}$	i-BAU $S_{\text{gUAA,OBA}}$	i-BAU $S_{\text{RBA,gUAA}}$	FT-SAM $S_{\text{RBA,OBA}}$	FT-SAM $S_{\text{gUAA,OBA}}$	FT-SAM $S_{\text{RBA,gUAA}}$
BadNets	0.607	0.192	0.170	0.599	0.194	0.169
Blended	0.712	0.196	0.192	0.712	0.197	0.193

Table 10: ASR (%) of RBA and gUAA with different query numbers.

Attack+Defense	**Query number ⇒**	**1000**	**3000**	**5000**	**7000**
Blended+i-BAU	RBA	77.3	89.3	92.1	94.6
	gUAA	14.2	41.4	49.5	56.4
Blended+FT-SAM	RBA	41.1	77.4	79.8	85.6
	gUAA	16.3	42.2	56.5	65.5

Table 11: ASR (%) of OBA, RBA, and gUAA under different l_∞-norm of random noise.

	Norm ⇒	**0**	**0.03**	**0.06**	**0.09**
OBA	Blended+NAD	99.8	99.8	99.6	97.3
	LF+NAD	99.1	98.9	98.4	98.6
RBA	Blended+NAD	99.8	99.7	98.7	84.0
	LF+NAD	99.4	99.1	98.1	96.6
gUAA	Blended+NAD	95.5	92.7	79.4	35.4
	LF+NAD	96.5	89.5	55.8	16.7

Thanks!

- For more details and results, please refer to the paper: **https://openreview.net/pdf?id=E2odGznGim**
- Our Code is available at: **https://github.com/JulieCarlon/Backdoor-Reactivation-Attack**

PAPER

CODE