BAN: Detecting Backdoors Activated by Adversarial Neuron Noise

Xiaoyun Xu
Radboud University
xiaoyun.xu@ru.nl

Zhuoran Liu*
Radboud University
z.liu@cs.ru.nl

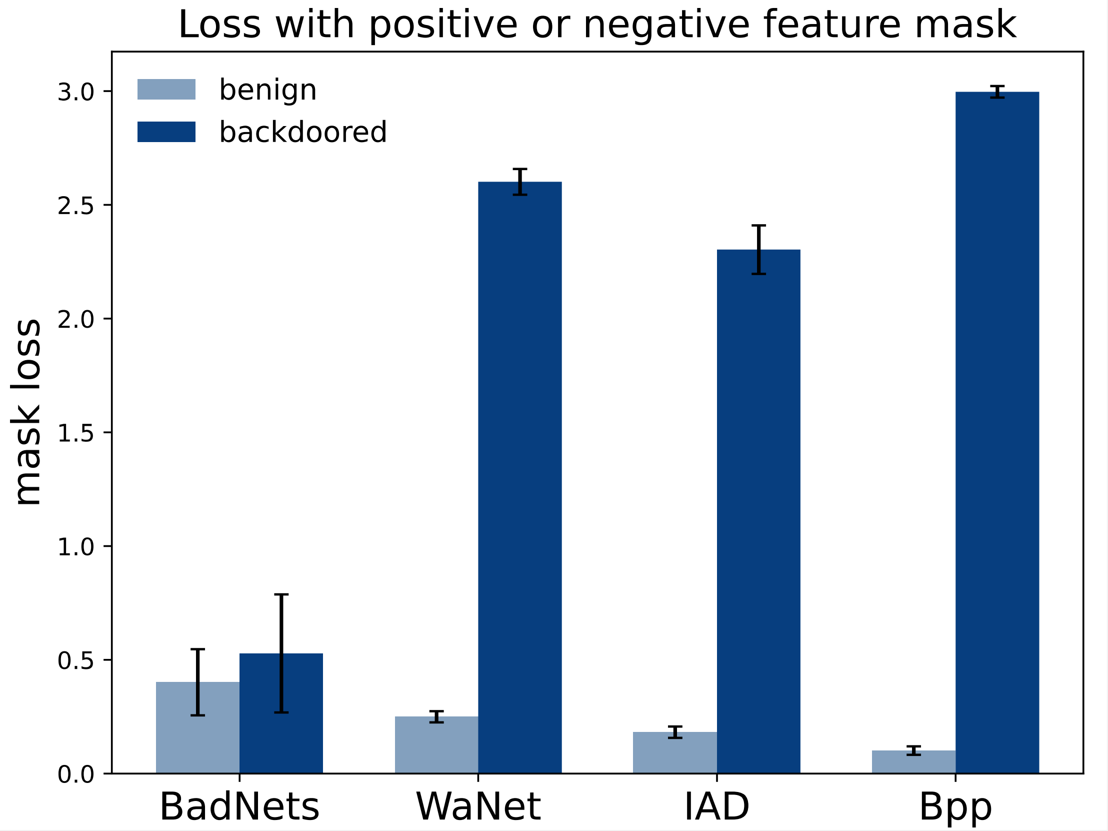
Stefanos Koffas
Delft University of Technology
s.koffas@tudelft.nl

Shujian Yu
Vrije Universiteit Amsterdam
s.yu3@vu.nl

Stjepan Picek
Radboud University
stjepan.picek@ru.nl

BACKDOOR FEATURES COULD BE NON-PROMINENT

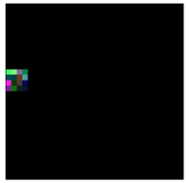
- Defenses that are biased towards large differences of benign and backdoor features may not work in cases like BadNets.
- BadNets features are not as prominent as others.



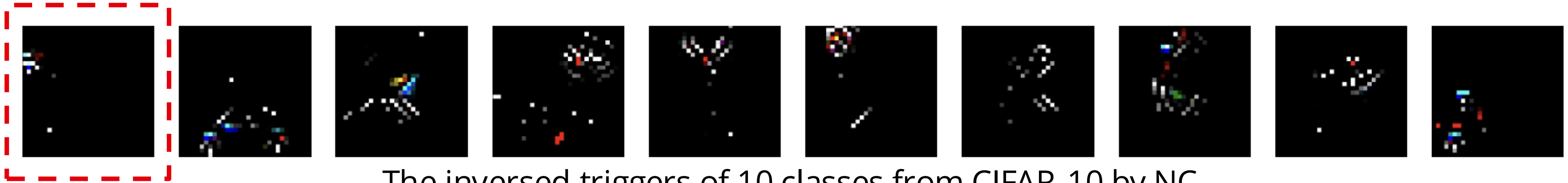
disentangled benign and backdoor features

HUGE TIME CONSUMPTION DUE TO OPTIMIZATION

- Existing methods (such as NC[1], FeatureRE[2] and Unicorn[3]) need to conduct optimization for every class to inverse all possible backdoor triggers.
- This shortcoming also limits existing methods against all-to-all backdoor attacks.



The true trigger



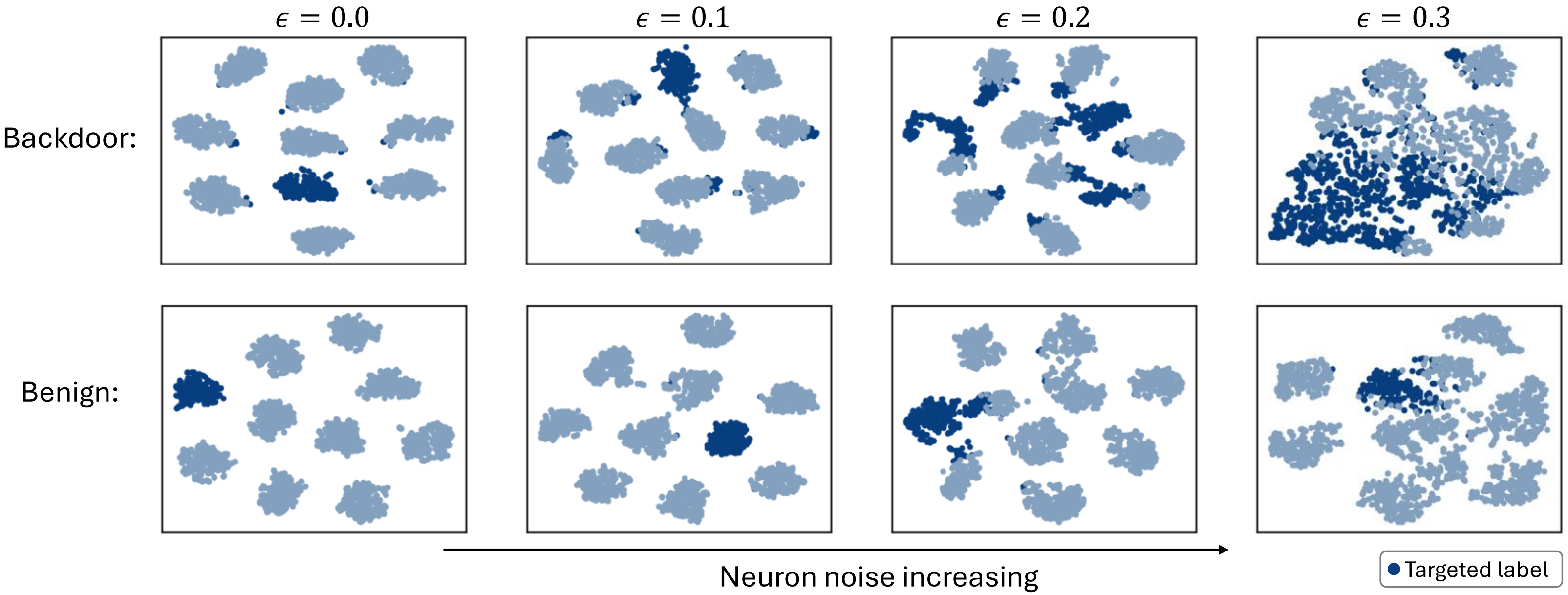
The inversed triggers of 10 classes from CIFAR-10 by NC

This one is smaller (L_2norm) than others and could be the backdoor trigger

1. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks
2. Rethinking the Reverse-engineering of Trojan Triggers
3. UNICORN: A Unified Backdoor Trigger Inversion Framework

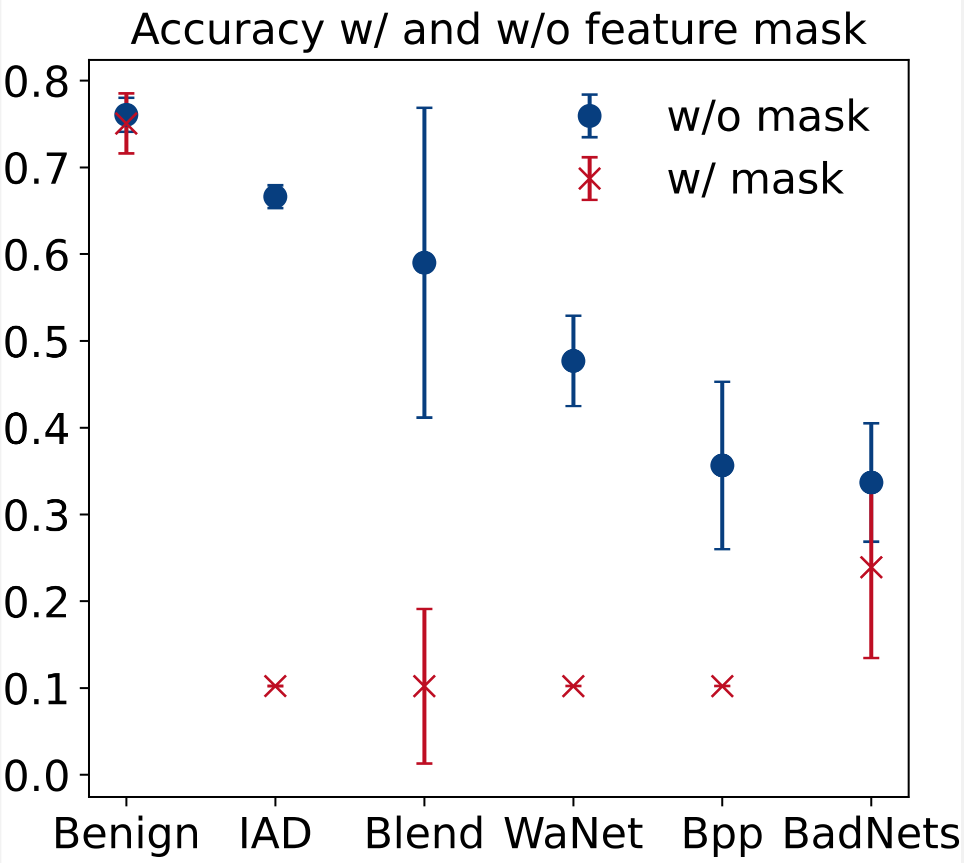
NEURON NOISE HELPS ACTIVATE BACKDOOR

- Adding noise on neuron weights can activate backdoor when receiving clean data as input.
- As noise increases, the backdoor model identifies more inputs from each class as the target label.
- The clean model has fewer errors, and there is no substantial increase in the number of misclassifications to the target class.



FEATURE DECOUPLING WITH MASK

- But the neuron noise is not enough for precise backdoor detection
- A feature decoupling process enhances the effect of noise on backdoored features but maintain a decreased effect on benign features.



NEURON NOISE IS HELPFUL TO REMOVE THE BACKDOOR

- Using optimized neuron noise to fine-tune the model can effectively remove the backdoor.
- The loss for our noise fine-tuning can be written as:

$$\min_{\mathbf{w},\mathbf{b}} \mathcal{L}(f(\mathbf{x};\mathbf{w},\mathbf{b}),y) + \lambda_2 \mathcal{L}(f(\mathbf{x};(1+\boldsymbol{\delta})\odot\mathbf{w},(1+\boldsymbol{\xi})\odot\mathbf{b}),y).$$

OVER RELY ON PROMINENT FEATURES LEADS TO WORSE DETECTION

- Recent backdoor detections may perform worse than NC. Because they rely too much on the prominent features.
- Our method performs well on different attack baselines.

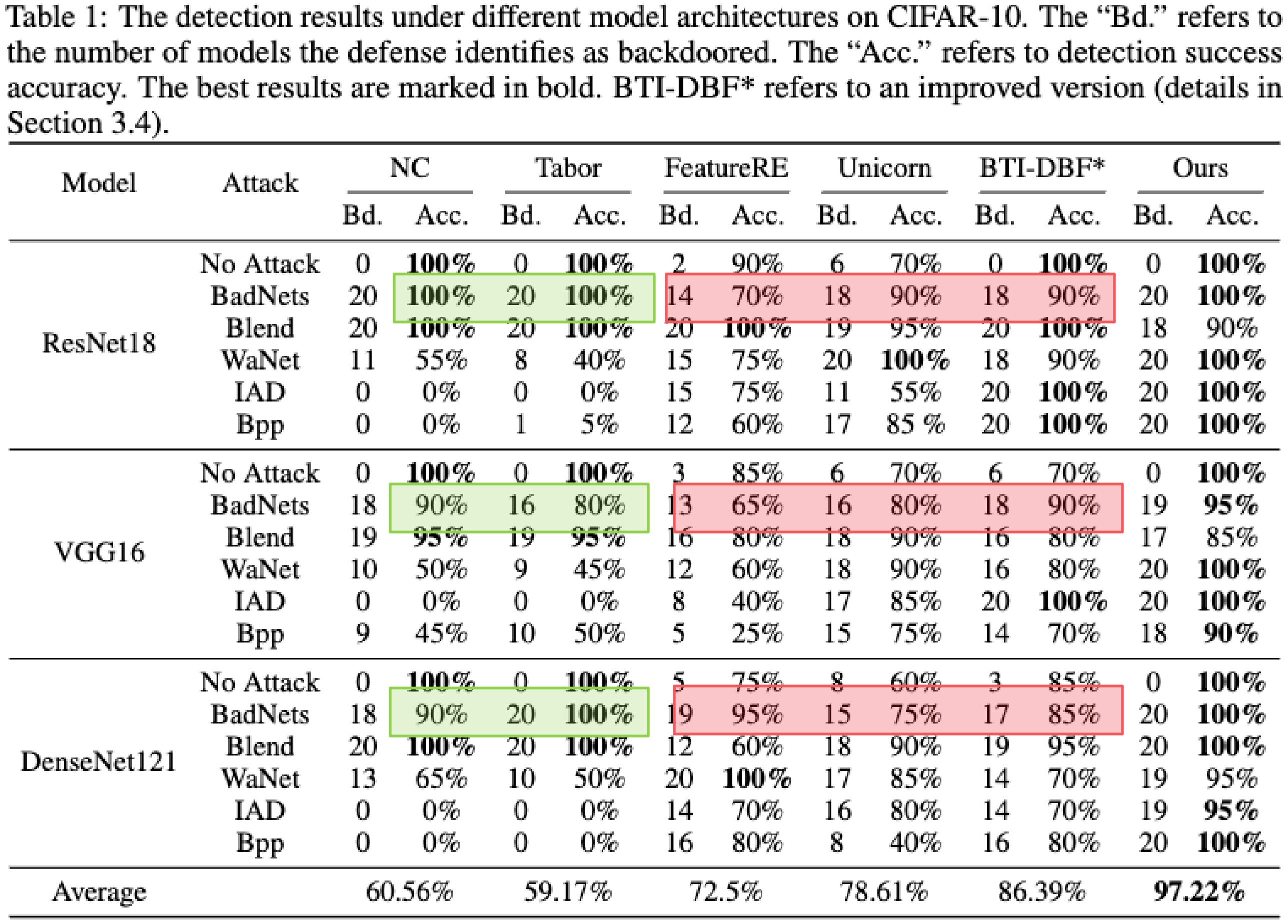
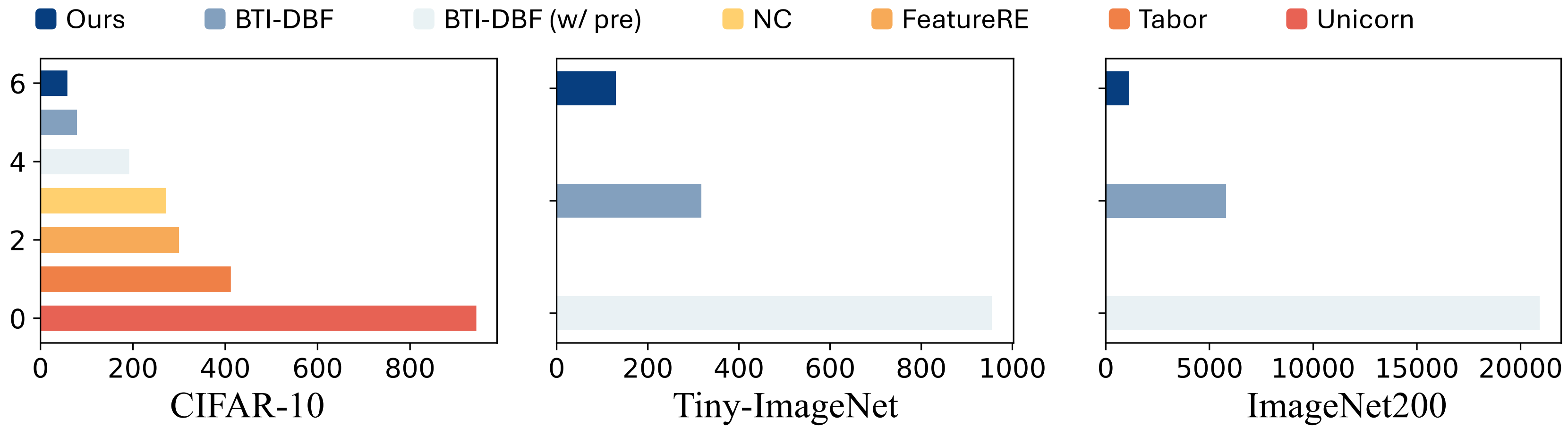


Table 1: The detection results under different model architectures on CIFAR-10. The "Bd." refers to the number of models the defense identifies as backdoored. The "Acc." refers to detection success accuracy. The best results are marked in bold. BTI-DBF* refers to an improved version (details in Section 3.4).

Model	Attack	NC		Tabor		FeatureRE		Unicorn		BTI-DBF*		Ours	
		Bd.	Acc.	Bd.	Acc.	Bd.	Acc.	Bd.	Acc.	Bd.	Acc.	Bd.	Acc.
ResNet18	No Attack	0	**100%**	0	**100%**	2	90%	6	70%	0	**100%**	0	**100%**
	BadNets	20	**100%**	20	**100%**	14	70%	18	90%	18	90%	20	**100%**
	Blend	20	**100%**	20	**100%**	20	**100%**	19	95%	20	**100%**	18	90%
	WaNet	11	55%	8	40%	15	75%	20	**100%**	18	90%	20	**100%**
	IAD	0	0%	0	0%	15	75%	11	55%	20	**100%**	20	**100%**
	Bpp	0	0%	1	5%	12	60%	17	85 %	20	**100%**	20	**100%**
VGG16	No Attack	0	**100%**	0	**100%**	3	85%	6	70%	6	70%	0	**100%**
	BadNets	18	90%	16	80%	13	65%	16	80%	18	90%	19	**95%**
	Blend	19	**95%**	19	**95%**	16	80%	18	90%	16	80%	17	85%
	WaNet	10	50%	9	45%	12	60%	18	90%	16	80%	20	**100%**
	IAD	0	0%	0	0%	8	40%	17	85%	20	**100%**	20	**100%**
	Bpp	9	45%	10	50%	5	25%	15	75%	14	70%	18	**90%**
DenseNet121	No Attack	0	**100%**	0	**100%**	5	75%	8	60%	3	85%	0	**100%**
	BadNets	18	90%	20	**100%**	19	95%	15	75%	17	85%	20	**100%**
	Blend	20	**100%**	20	**100%**	12	60%	18	90%	19	95%	20	**100%**
	WaNet	13	65%	10	50%	20	**100%**	17	85%	14	70%	19	95%
	IAD	0	0%	0	0%	14	70%	16	80%	14	70%	19	**95%**
	Bpp	0	0%	0	0%	16	80%	8	40%	16	80%	20	**100%**
Average		60.56%		59.17%		72.5%		78.61%		86.39%		**97.22%**	

TIME CONSUMPTION

- BAN is efficient and scalable as we do not iterate over all target classes



Take-home messages

1. **Traditional defenses outperformed the latest feature space defenses on input space backdoor attacks as feature space defenses over-rely on prominent features.**

2. **Neuron noise can activate backdoor when receiving clean data as input.**

3. **Optimization of neuron noise is efficient without relying on the target class.**