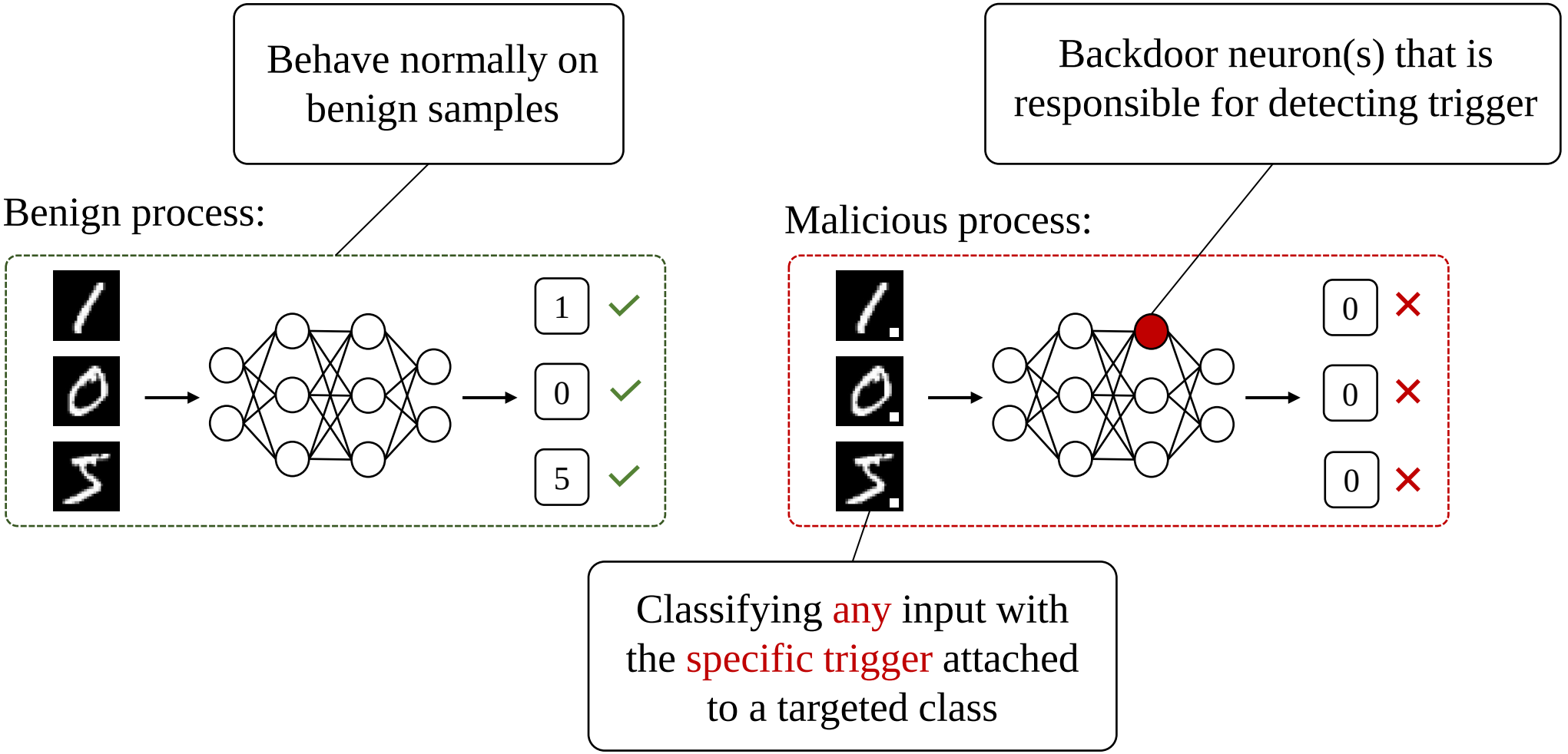
Pre-activation Distributions Expose Backdoor Neurons

Runkai Zheng[1*], Rongjun Tang[1*], Jianze Li[1,2], Li Liu[1,2 †]

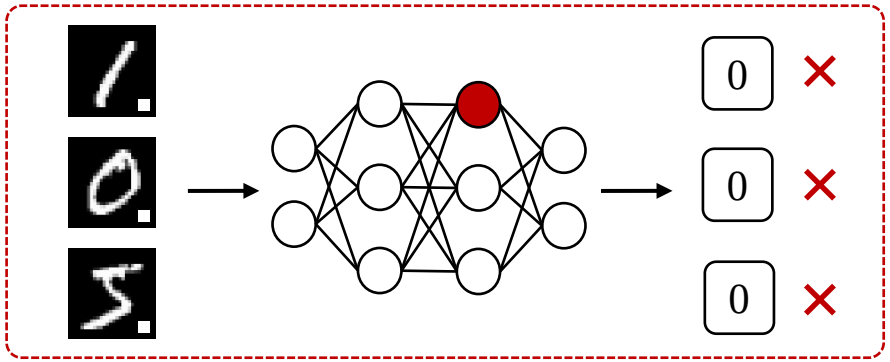
[1]The Chinese University of Hong Kong, Shenzhen
[2]Shenzhen Research Institute of Big Data

[*]Equal Contribution
[†]Corresponding Author

Backdoor Attack

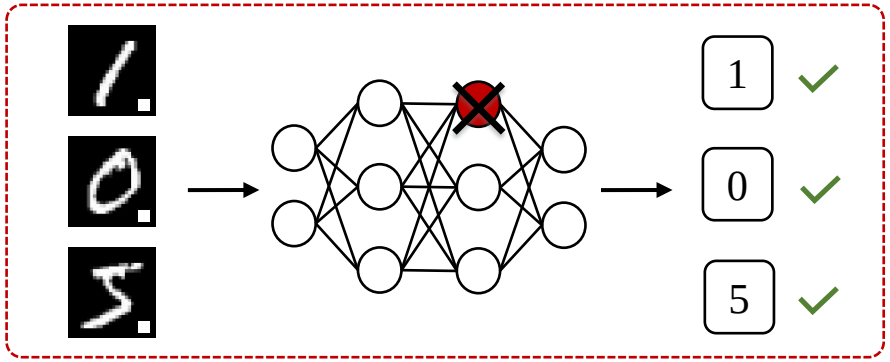


Backdoor Defense



Goals:

- Mitigate the effect of trigger.
- Maintain normal performance on benign samples.



Defense by pruning:

- Find the (potential) backdoor neurons.
- Set their weights to zero to remove the backdoor.

Preliminary: Pre-activation Distribution

Consider a neural network $F(x;\theta)$ (also write as $F^{(l)}$) with L layers, denote:

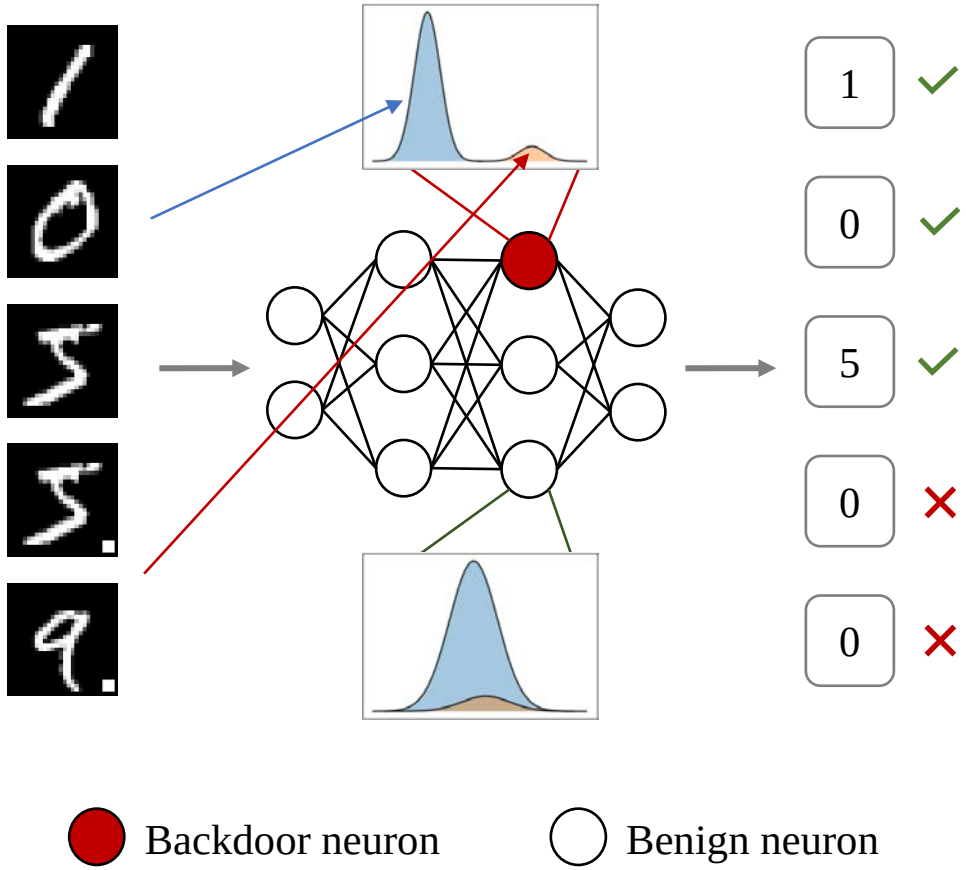
$$F^{(l)} = f^{(l)} \circ \varphi \circ f^{(l-1)} \circ \cdots \circ \varphi \circ f^{(1)}$$

for $1 \leq l \leq L$, where $f^{(l)}$ is a linear function in l_{th} layer and φ is a non-linear activation function.

We denote $x^{(l)} = F^{(l)}(x) \in \mathbb{R}^{d_c^{(l)} \times d_h^{(l)} \times d_w^{(l)}}$ as the output of the l_{th} layer.

For the k_{th} neuron, the pre-activation $\phi_k^{(l)} = \phi(x_k^{(l)})$ is defined as the maximum value of the k_{th} slice matrix of $x^{(l)}$.

Motivation



Previous research:
The existence of backdoor-related neurons.

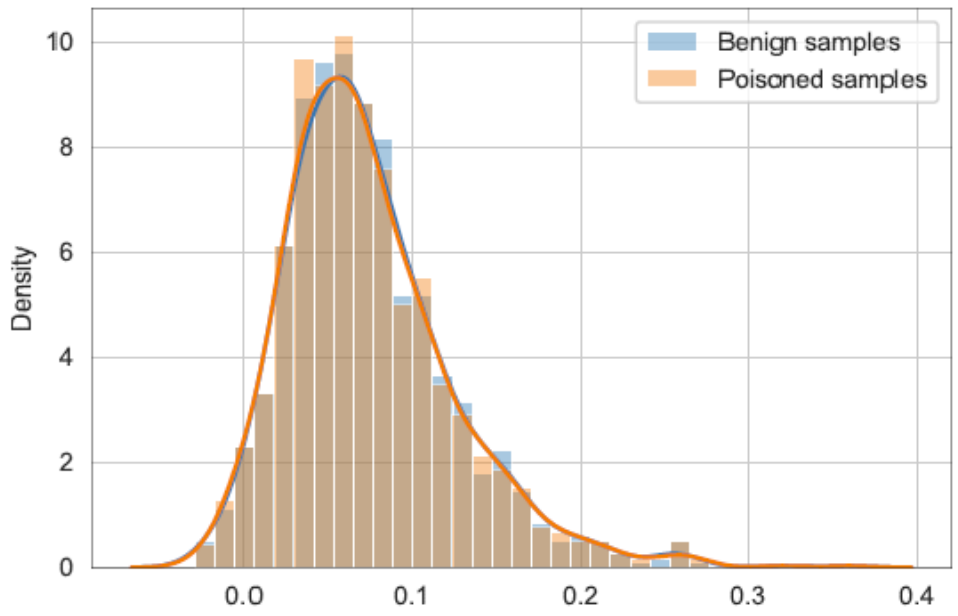
Our empirical observation:

Two Gaussian-like pre-activation distributions with significant different moments formed by benign samples and poisoned samples, respectively.

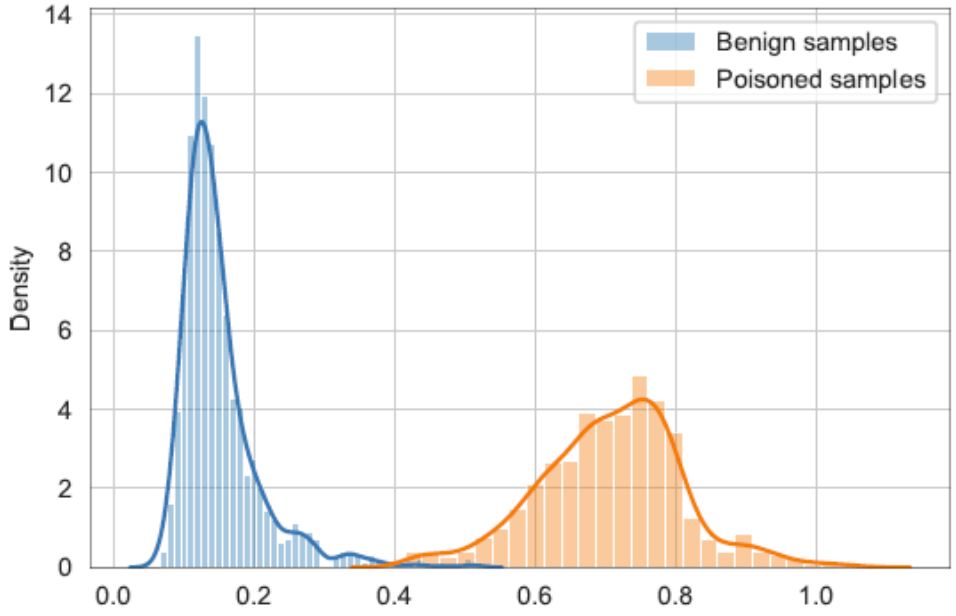
Main Assumption: Gaussian Mixture Distribution

Assume that in backdoor neurons, the pre-activation distribution follows a mixture of two different Gaussians; while in the benign neurons, the difference between distributions formed by benign and poisoned samples can be omitted:

$$\phi_k^{(l)} \sim (1-\rho)\mathcal{N}\left(\mu_k^{(l)}, \sigma_k^{(l)2}\right) + \rho\,\mathcal{N}\left(\hat{\mu}_k^{(l)}, \hat{\sigma}_k^{(l)2}\right)$$



Benign neurons



Backdoor neurons

Proposed Methods: Entropy-based Pruning (EP)

Characteristic of backdoor neurons

Poisoned dataset → Propagation → Low differential entropy of pre-activation distributions

Benign Neurons
- Two similar Gaussian distributions

Standardize

Backdoor Neurons
- Bimodal Gaussian distribution

Large differential entropy (close to standard Gaussian distribution).

Relatively smaller differential entropy.

$$h(\dot{\phi}_k^{(l)}) < h(\dot{\phi}_{k'}^{(l)}) \leq h(Z)$$

Proposed Methods: BN statistics-based Pruning (BNP)

Characteristic of backdoor neurons

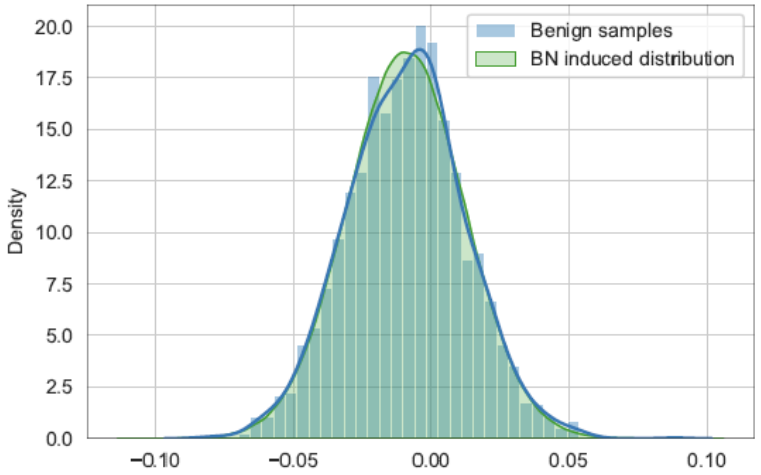
benign dataset

Propagation

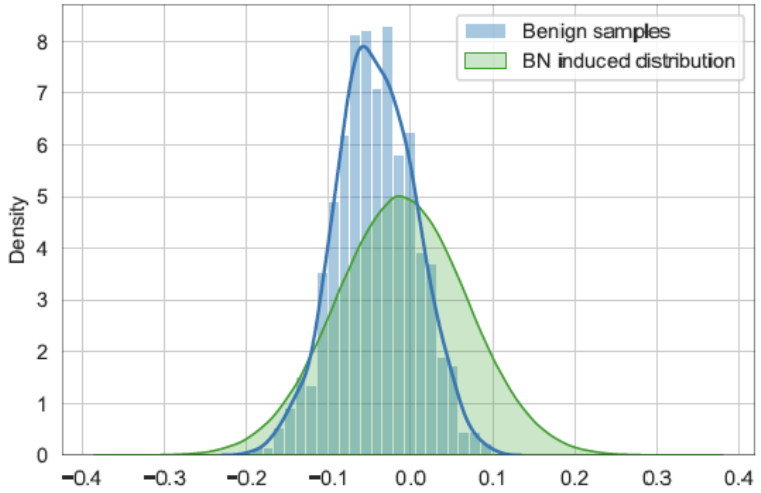
Mismatched statistics between

BN records

and benign pre-activations



Benign neuron



Backdoor neuron

Proposed Methods: BN statistics-based Pruning (BNP)

We calculate the KL divergence (under Gaussian assumption) between the distribution approximated by benign sample's statistics and the BN induced distribution to identify the backdoor neurons:

$$\mathrm{D_{KL}}\left(\boldsymbol{\mathcal{N}}^{(\boldsymbol{l})}_{\mathrm{samlple}}, \boldsymbol{\mathcal{N}}^{(\boldsymbol{l})}_{\mathrm{BN}}\right)_{\boldsymbol{k}} > \mathrm{D_{KL}}\left(\boldsymbol{\mathcal{N}}^{(\boldsymbol{l})}_{\mathrm{sample}}, \boldsymbol{\mathcal{N}}^{(\boldsymbol{l})}_{\mathrm{BN}}\right)_{\boldsymbol{k'}} = \mathbf{0}$$

The equality holds when our assumption on the Gaussian mixture distribution holds.

Summary on Pruning Strategy

1. If we have access to the poisoned dataset:
   - Calculate the standardized pre-activation differential entropy for every neuron, and let $h^{(l)} = \left[h\left(\dot{\phi}_1^{(l)}\right), h\left(\dot{\phi}_2^{(l)}\right), \dots, h(\dot{\phi}_{d_c^{(l)}}^{(l)})\right]^{\mathrm{T}}$.
   - Set $\tau_h^{(l)} = \bar{h}^{(l)} - u_h \cdot s_h^{(l)}$, and prune the neurons with differential entropy less than $\tau_h^{(l)}$.

2. If we have access to a benign dataset:
   - Calculate the KL divergence between benign sample distribution and BN induced distribution $\mathrm{K}_k^{(l)} = \mathrm{D}_{\mathrm{KL}}\left(\mathcal{N}_{\mathrm{samlple}}^{(l)}, \mathcal{N}_{\mathrm{BN}}^{(l)}\right)_k$ for every neuron, and let $\mathrm{K}^{(l)} = \left[\mathrm{K}_1^{(l)}, \mathrm{K}_2^{(l)}, \dots, \mathrm{K}_{d_c^{(l)}}^{(l)}\right]^{\mathrm{T}}$.
   - Set $\tau_{\mathrm{K}}^{(l)} = \overline{\mathrm{K}}^{(l)} + u_{\mathrm{K}} \cdot s_{\mathrm{K}}^{(l)}$, and prune the neurons with KL divergence larger than $\tau_{\mathrm{K}}^{(l)}$.

Here u_h/u_{K} are the only hyperparameter in our methods, and we set it to 3 as default.

Quantitative Results

CIFAR-10

Attacks	BadNets (A2O)		BadNets (A2A)		CLA		WaNet		Blended		Refool		IAB	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Origin	93.86	100.00	94.60	93.89	94.99	98.83	94.11	99.67	94.17	99.62	94.24	98.40	93.87	97.91
FT	92.22	2.16	92.03	60.76	92.88	95.73	92.93	9.37	93.9	90.27	91.68	17.78	91.78	9.52
FP	92.18	2.97	91.75	66.82	92.60	99.36	92.07	1.03	70.92	90.92	92.36	75.98	87.04	16.13
l_∞	92.12	100.00	93.67	6.67	92.75	98.76	93.48	99.74	86.99	99.77	91.19	98.47	88.37	88.48
NAD	93.36	2.43	92.18	2.06	91.36	15.31	93.08	3.05	92.72	1.61	91.64	6.74	92.11	19.45
ANP	93.47	3.53	90.29	86.22	91.13	11.76	**94.12**	**0.51**	93.66	5.03	91.71	26.96	**93.52**	10.61
EP (Ours)	**93.88**	**0.86**	**94.49**	**0.61**	**94.42**	**0.91**	93.79	2.80	93.67	**2.24**	93.35	8.90	93.17	0.94
BNP (Ours)	93.60	1.60	94.25	0.72	94.14	7.03	94.05	3.39	**94.17**	2.71	**93.69**	**6.48**	93.15	**0.64**

Tiny-ImageNet

Attacks	BadNets (A2O)		CLA		WaNet		Refool		Blended		IAB		SSBA	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Origin	61.36	97.38	65.61	56.58	61.47	99.98	53.26	80.61	62.85	99.83	61.40	98.28	66.51	99.78
FT	46.93	99.84	61.19	63.20	54.28	99.96	47.09	91.77	56.83	29.12	52.39	99.1	52.39	33.19
FP	35.41	99.48	62.30	39.05	53.65	100.00	42.10	86.62	59.59	99.76	52.67	98.47	53.36	31.96
l_∞	53.13	90.39	59.15	23.12	42.01	99.84	46.84	81.19	56.33	99.85	54.81	86.97	49.35	99.98
NAD	44.20	90.13	62.80	17.35	53.40	99.98	51.06	70.63	57.35	55.6	53.32	98.85	52.52	25.08
ANP	53.85	4.02	59.69	3.64	54.82	86.98	50.67	**0.21**	**62.49**	**0.61**	**61.39**	4.67	60.98	1.01
EP (Ours)	60.68	**0.86**	64.47	0.10	60.53	0.02	51.29	17.07	60.67	**0.69**	61.26	**0.60**	64.2	0.11
BNP (Ours)	**61.60**	1.60	**64.86**	**0.05**	**61.58**	**0.01**	**52.41**	23.79	60.77	0.85	**61.30**	**0.60**	**64.64**	**0.01**

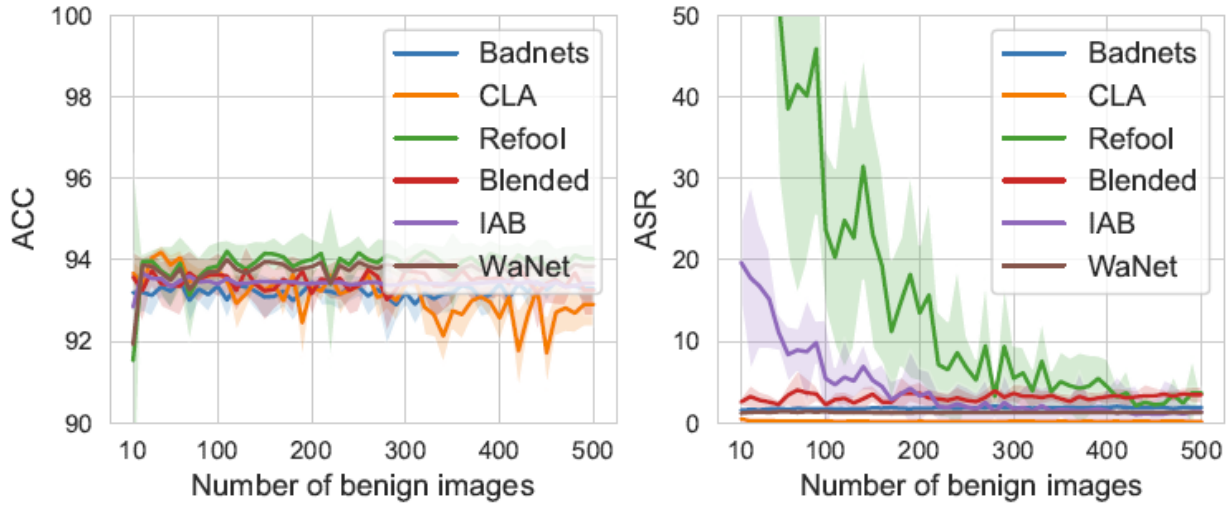
Quantitative Results

Running time evaluation: tested on a single RTX 2080Ti GPU

Defense Method	FT	FP	NAD	ANP	EP (ours)	BNP (ours)
Runing Time (sec.)	12.35s	14.59s	22.08s	25.68s	10.69s	**0.39s**

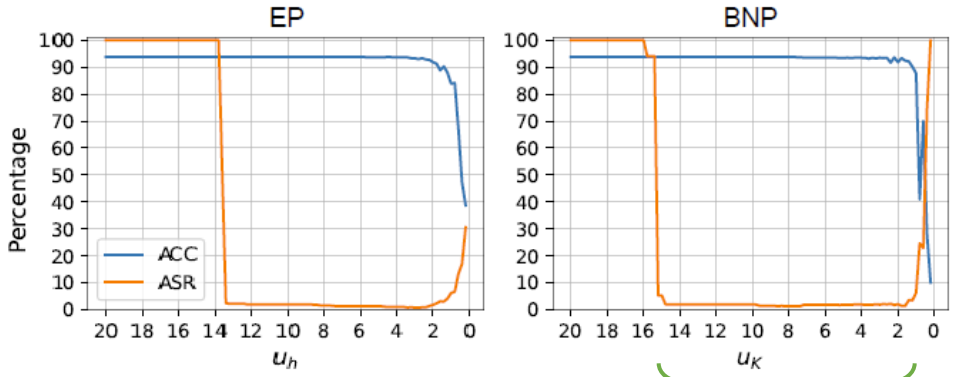
Evaluation on samples used for BNP

Even with 10 samples, BNP can successfully capture the difference between the two distributions and locate the backdoor neurons.

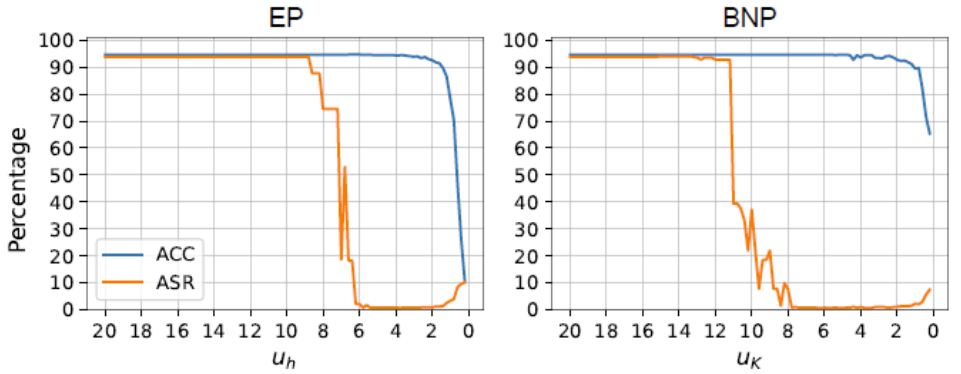


Ablation Studies

Varying Threshold Hyperparameter u

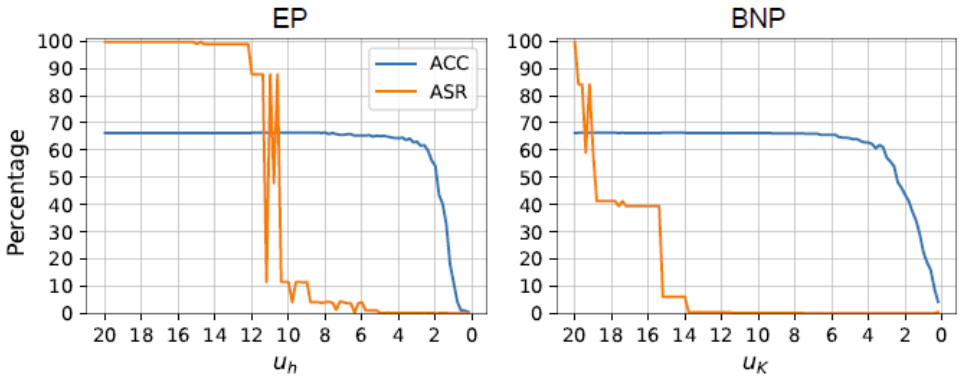


(a) BadNets (A2O)



(b) BadNets (A2A)

Wider window with high ACC and low ASR means the choice of hyperparameter is more robust.



(h) SSBA (TinyImageNet)

Ablation Studies

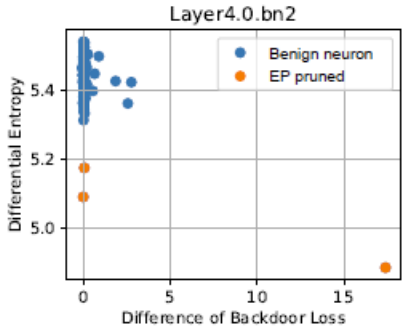
Varying Poisoning Rate

Dataset: CIFAR-10 Network: ResNet-18

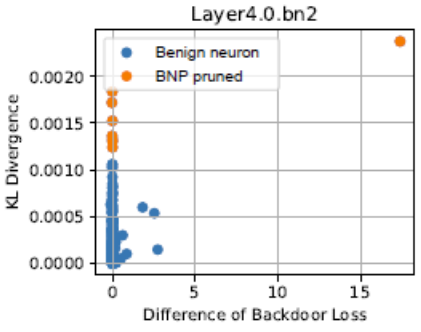
ρ	Stage	BadNets (A2O) ACC	BadNets (A2O) ASR	BadNets (A2A) ACC	BadNets (A2A) ASR	CLA ACC	CLA ASR	WaNet ACC	WaNet ASR	Blended ACC	Blended ASR	Refool ACC	Refool ASR	IAB ACC	IAB ASR
1%	Origin	95.03	99.94	94.75	88.57	88.96	4.73	94.76	46.82	94.17	99.62	93.08	99.59	93.22	64.00
	EP	94.82	0.91	94.17	0.73	87.96	0.65	93.67	14.17	94.46	2.29	91.77	24.08	92.73	5.21
	BNP	92.22	2.16	93.16	7.99	88.03	0.84	94.64	1.24	92.89	2.44	90.99	21.22	93.17	4.19
5%	Origin	94.29	99.99	94.26	92.78	95.53	92.23	94.00	94.55	94.53	81.33	94.35	97.98	92.70	65.50
	EP	93.83	0.83	93.67	0.70	94.43	15.91	92.73	10.13	94.44	5.49	92.75	4.51	92.29	1.83
	BNP	93.61	0.67	93.99	5.64	94.65	12.06	94.17	1.78	93.37	9.21	92.30	2.08	92.74	2.14
10%	Origin	93.89	100.00	94.60	93.89	94.99	98.83	94.11	99.67	94.17	99.63	94.24	98.40	93.87	97.91
	EP	93.88	0.86	94.49	0.61	94.42	0.91	93.79	2.80	93.67	2.24	93.35	8.90	93.17	0.94
	BNP	93.60	1.60	94.25	0.72	94.14	7.03	94.05	3.39	94.17	2.71	93.69	6.48	93.15	0.64

Ablation Studies

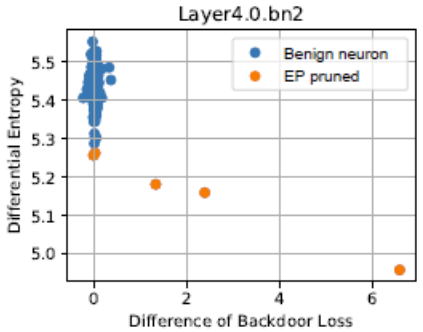
Pruning Sensitivity: Are We Pruning the Backdoor Neurons?



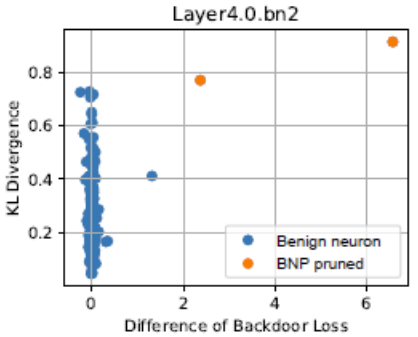
(a) BadNets (EP)



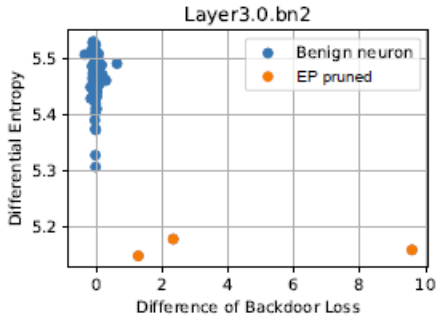
(b) BadNets (BNP)



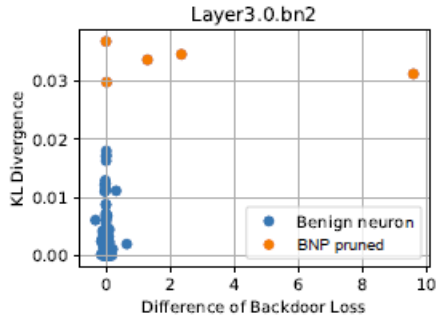
(c) Blended (EP)



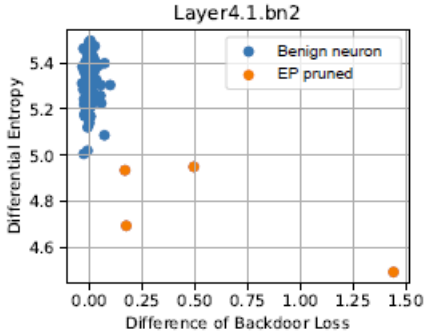
(d) Blended (BNP)



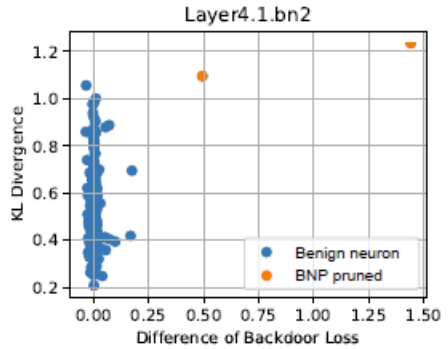
(e) WaNet (EP)



(f) WaNet (BNP)



(g) Refool (EP)



(h) Refool (BNP)

Summary

We find:

In the backdoor neurons, the discrepancy between distributions formed by benign samples and poisoned samples is obviously larger than that in the benign neurons.

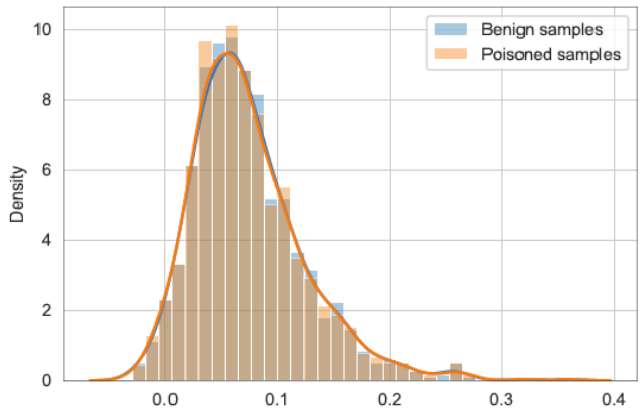
We Propose:

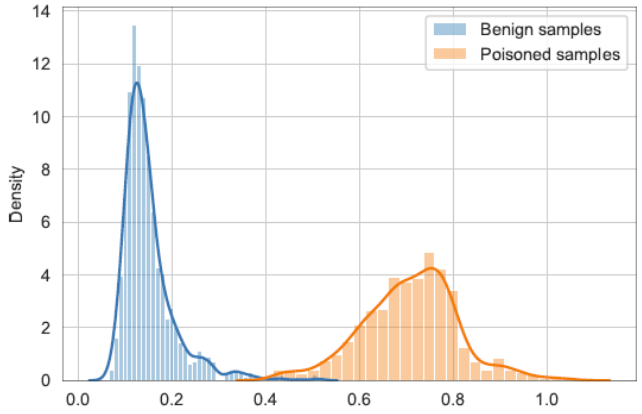
Entropy-based Pruning (EP)

- With poisoned dataset.

BN Statistics-based Pruning (BNP)

- With benign dataset.





Good Performance; High Efficiency; Robust to Hyperparameter.

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

SHENZHEN RESEARCH INSTITUTE OF BIG DATA
SRIBD

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You!

Pre-activation Distributions Expose Backdoor Neurons