Unveiling and Mitigating Backdoor Vulnerabilities based on Unlearning Weight Changes and Backdoor Activeness

Weilin Lin[1], Li Liu[1*], Shaokui Wei[2], Jianze Li[3,4,2], Hui Xiong[1]

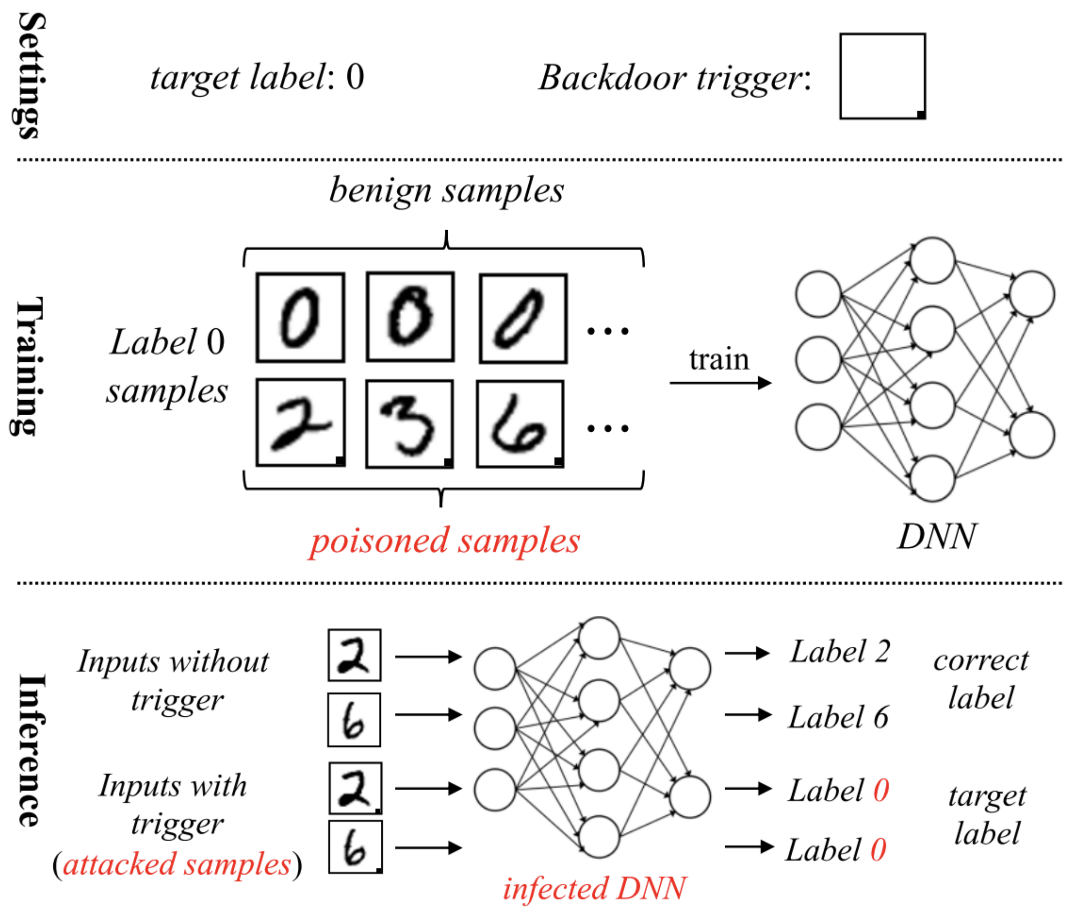
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Chinese University of Hong Kong, Shenzhen
[3]Shenzhen International Center for Industrial and Applied Mathematics
[4]Shenzhen Research Institute of Big Data

Outline

- Background
- Observations
- Framework
- Experiment
- Conclusion

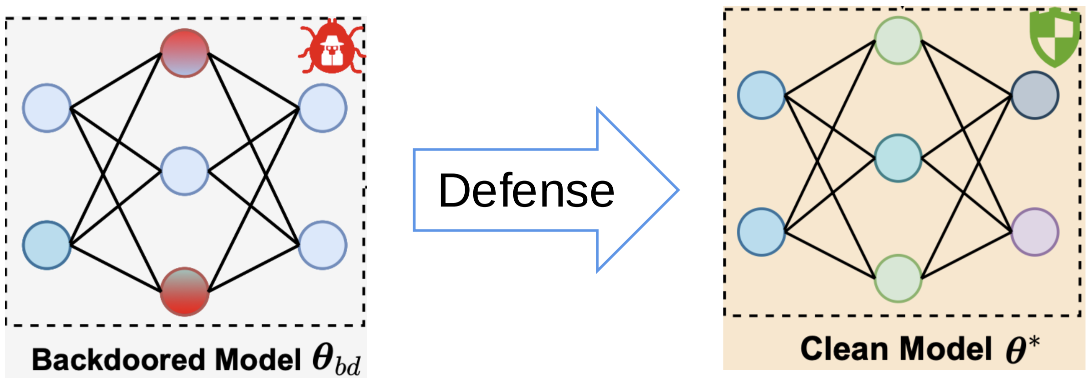
Background

Backdoor Attack



Backdoor Defense

- Post-training Defense



*Backdoored Model == Infected DNN

Goal:

1. Maintain clean functionality.
   - Inputs without trigger. → Correct label.
   - High *clean accuracy* (ACC).
2. Eliminate backdoored effect.
   - Inputs with trigger. X→ Target label.
   - Low *attack success rate* (ASR).

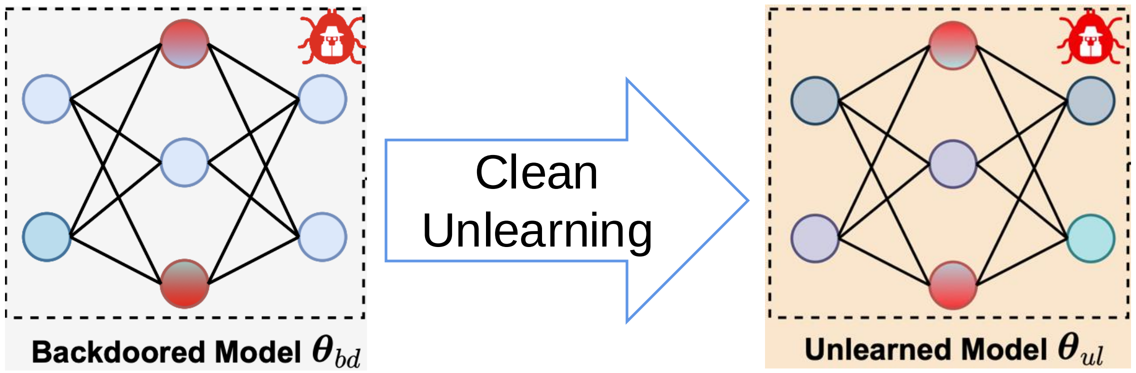
Li Y, Jiang Y, Li Z, et al. Backdoor learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(1): 5-22.

Background

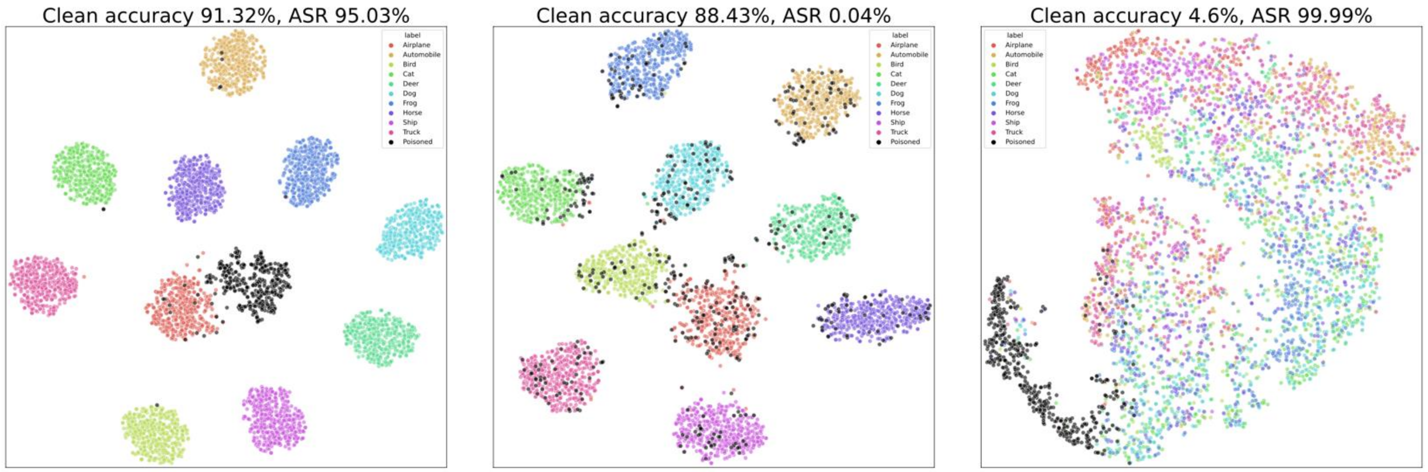
Unlearning for the Backdoored Model

Model Unlearning

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y)\in\mathcal{D}}\left[\mathcal{L}(f(\boldsymbol{x};\boldsymbol{\theta}),y)\right]$$



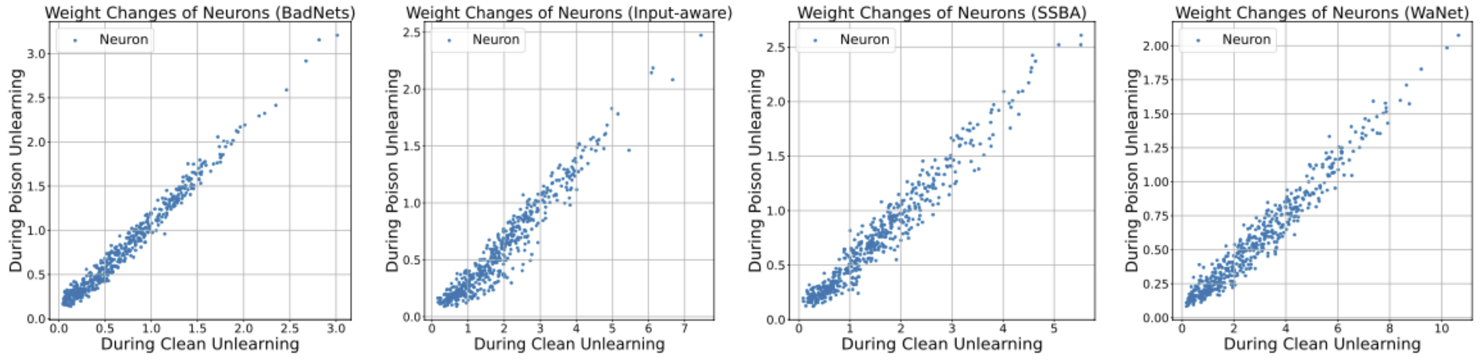
- Clean Unlearning
  - Unlearn on clean dataset.
  - Accessible for defender.
  - ACC↓, ASR≈
- Poison Unlearning
  - Unlearn on poison dataset.
  - Inaccessible for defender.
  - ACC≈, ASR↓



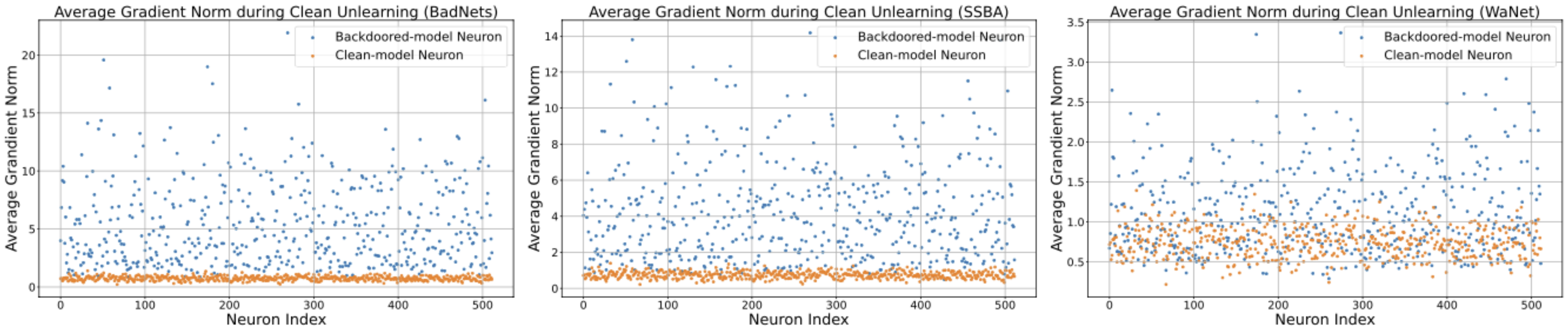
(a) Backdoored model (b) Poison unlearning (c) Clean unlearning

*ASR: Attack Success Rate

Observations



Observation 1. Weight changes between poison and clean unlearning are positively correlated.

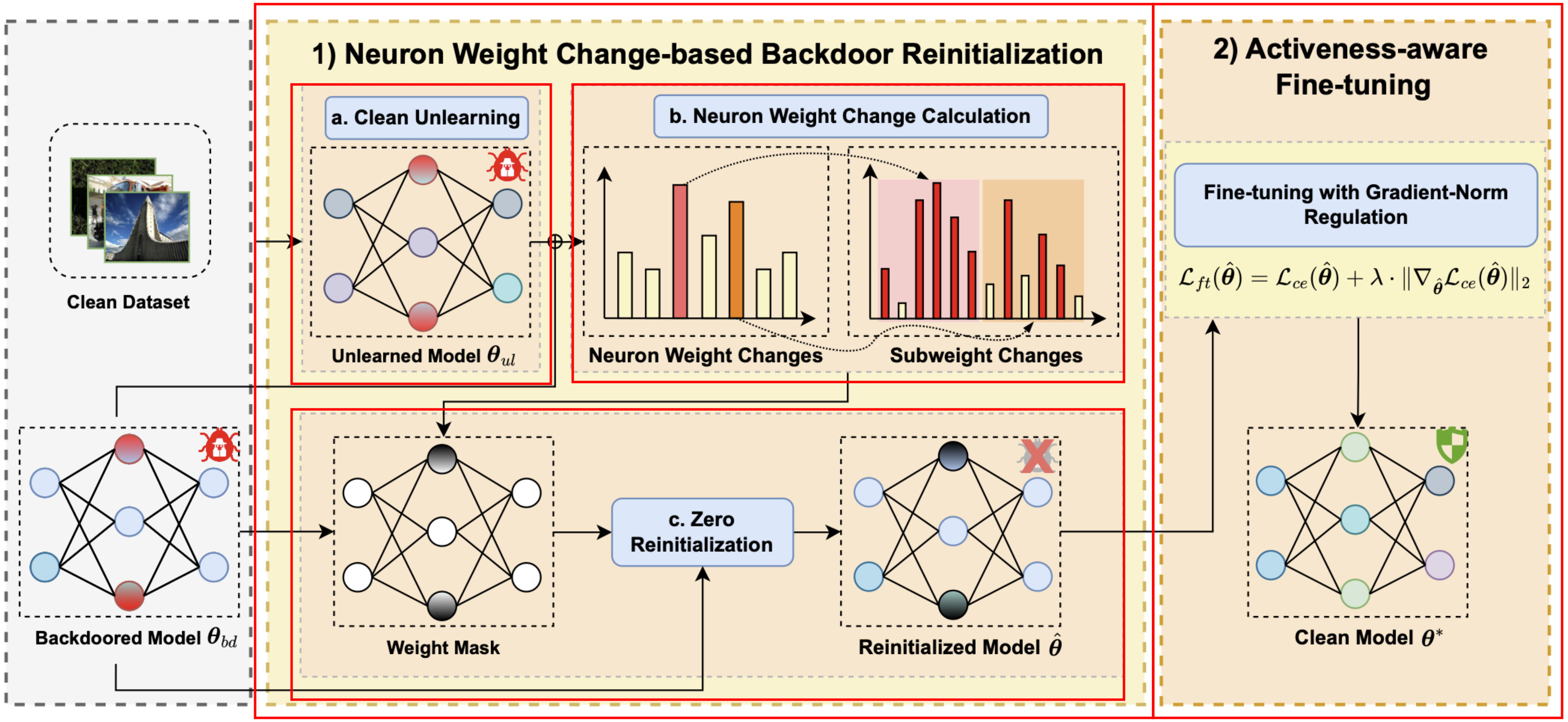


Observation 2. Neurons of backdoored model are more active than those in clean model.

- [**Unlearning Weight Changes**] Observation 1 inspires us to zero out the high-NWC neuron weights for backdoor mitigation.
- [**Backdoor Activeness**] Observation 2 inspires us to suppress the gradient norm during the learning process if we want to recover it to a clean model.

Framework

Two-Stage Backdoor Defense (TSBD)



- Stage 1: to mitigate the backdoor effect with acceptable clean-accuracy sacrificed.
- Stage 2: to repair the reinitialized model and avoid recovering the backdoor effect again.

Experiment

Main Results

Table 1: Comparison with the SOTA defenses on **CIFAR-10** dataset with PreAct-ResNet18 (%).

Backdoor Attacks	No Defense			FT			FP [37]			NAD [43]			NC [20]		
	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$
BadNets [8]	91.32	95.03	-	89.96	1.48	96.10	**91.31**	57.13	68.95	89.87	2.14	95.72	89.05	1.27	95.75
Blended [25]	93.47	99.92	-	92.78	96.11	51.56	93.17	99.26	50.18	92.17	97.69	50.47	**93.47**	99.92	50.00
Input-aware [23]	90.67	98.26	-	93.12	1.72	98.27	91.74	**0.04**	**99.11**	**93.18**	1.68	98.29	92.61	0.76	98.75
LF [49]	93.19	99.28	-	92.37	78.44	60.01	**92.90**	98.97	50.01	92.37	47.83	75.31	91.62	**1.41**	**98.15**
SIG [26]	84.48	98.27	-	**90.80**	2.37	97.95	89.10	26.20	86.03	90.02	10.66	93.81	84.48	98.27	50.00
SSBA [9]	92.88	97.86	-	92.14	74.79	61.16	92.54	83.50	57.01	91.91	77.40	59.74	90.99	**0.58**	**97.69**
Trojan [50]	93.42	100.00	-	92.42	5.99	96.51	92.46	71.17	63.94	91.88	3.73	**97.36**	91.76	8.22	95.06
WaNet [24]	91.25	89.73	-	**93.48**	17.10	86.32	91.46	1.09	94.32	93.17	22.98	83.38	91.80	7.53	91.10
Average	91.34	97.29	-	**92.13**	34.75	80.98	91.84	54.67	71.19	91.82	33.01	81.76	90.72	27.24	84.56

Backdoor Attacks	ANP [41]			CLP [38]			i-BAU [21]			RNP [22]			TSBD **(Ours)**		
	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$	ACC $\uparrow$	ASR $\downarrow$	DER $\uparrow$
BadNets [8]	90.94	5.91	94.37	90.06	77.50	58.14	89.15	**1.21**	95.83	89.81	24.97	84.28	90.72	1.31	**96.53**
Blended [25]	93.00	84.90	57.28	91.32	99.74	49.01	87.00	50.53	71.46	88.76	79.74	57.73	91.61	**2.61**	**97.73**
Input-aware [23]	91.04	1.32	98.47	90.30	2.17	97.86	89.17	27.08	84.84	90.52	1.84	98.13	93.06	1.94	98.16
LF [49]	92.83	54.99	71.96	92.84	99.18	49.88	84.36	44.96	72.75	88.43	7.02	93.75	91.20	2.64	97.32
SIG [26]	83.36	36.43	80.36	83.80	98.91	49.66	85.67	3.68	97.29	84.48	98.27	50.00	90.41	**1.27**	**98.50**
SSBA [9]	**92.67**	60.16	68.74	91.38	68.13	64.11	87.67	3.97	94.34	88.60	17.89	87.84	91.57	1.66	97.44
Trojan [50]	92.97	46.27	76.64	**92.98**	100.00	49.78	90.37	**2.91**	97.02	90.89	3.59	96.94	91.76	5.06	96.64
WaNet [24]	91.32	2.22	93.76	81.91	78.42	50.99	89.49	5.21	91.38	90.43	0.96	93.98	93.26	**0.88**	**94.43**
Average	91.02	36.53	80.20	89.32	78.01	58.68	87.86	17.44	88.11	88.99	29.28	82.83	91.70	**2.18**	**97.09**

Defense Effectiveness Rating: $\mathrm{DER} = [\max(0, \Delta\mathrm{ASR}) - \max(0, \Delta\mathrm{ACC}) + 1]/2$

- TSBD performs the state-of-the-art (SOTA) on average.
  - Promising ACC (91.70%); Best ASR (2.18%) and DER (97.09%)

Conclusion

- Provide two novel insights.
  - The first to uncover the strong positive relationship between neuron weight changes in clean unlearning and poison unlearning.
  - Reveal the high backdoor activeness in the backdoored model during the learning process.

- TSBD is a promising defense method.
  - Considering both backdoor mitigation and clean-accuracy recovery.

- SOTA performance on average.
  - Highest DER, balancing well in ACC and ASR.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

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

深圳市大数据研究院 Shenzhen Research Institute of Big Data

Unveiling and Mitigating Backdoor Vulnerabilities based on Unlearning Weight Changes and Backdoor Activeness

Thanks for listening

NEURAL INFORMATION PROCESSING SYSTEMS

Presenter: Weilin Lin