The 28th International Symposium on Research in Attacks, Intrusions and Defenses

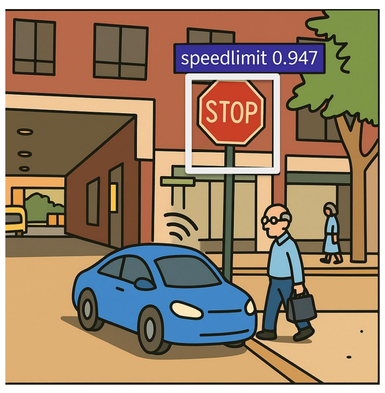
RAID 2025

Unsupervised Backdoor Detection and Mitigation for Spiking Neural Networks

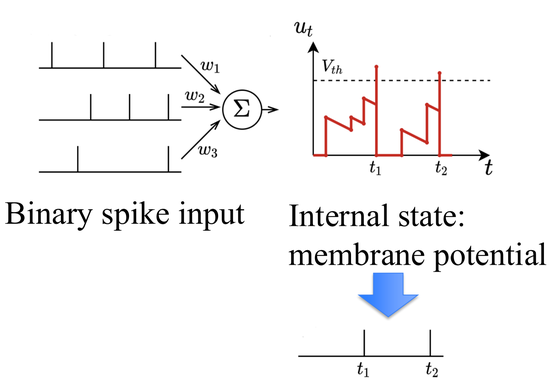
Jiachen Li, Bang Wu, Xiaoyu Xia, Xiaoning Liu, Xun Yi, Xiuzhen Zhang
RMIT University

Motivation & Problem

- SNNs used in safety-critical systems (e.g. driving [1]).
- ANN defenses fail due to event-driven, binary spike, threshold nonlinearities of SNN.



[2]



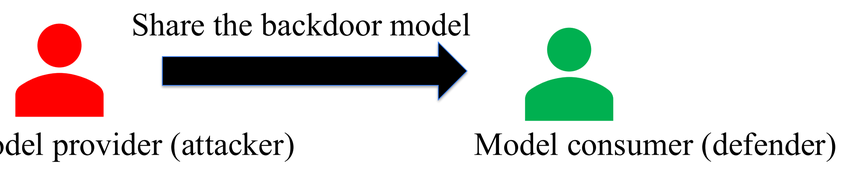
Binary spike input

Internal state: membrane potential

[3] Binary spike output

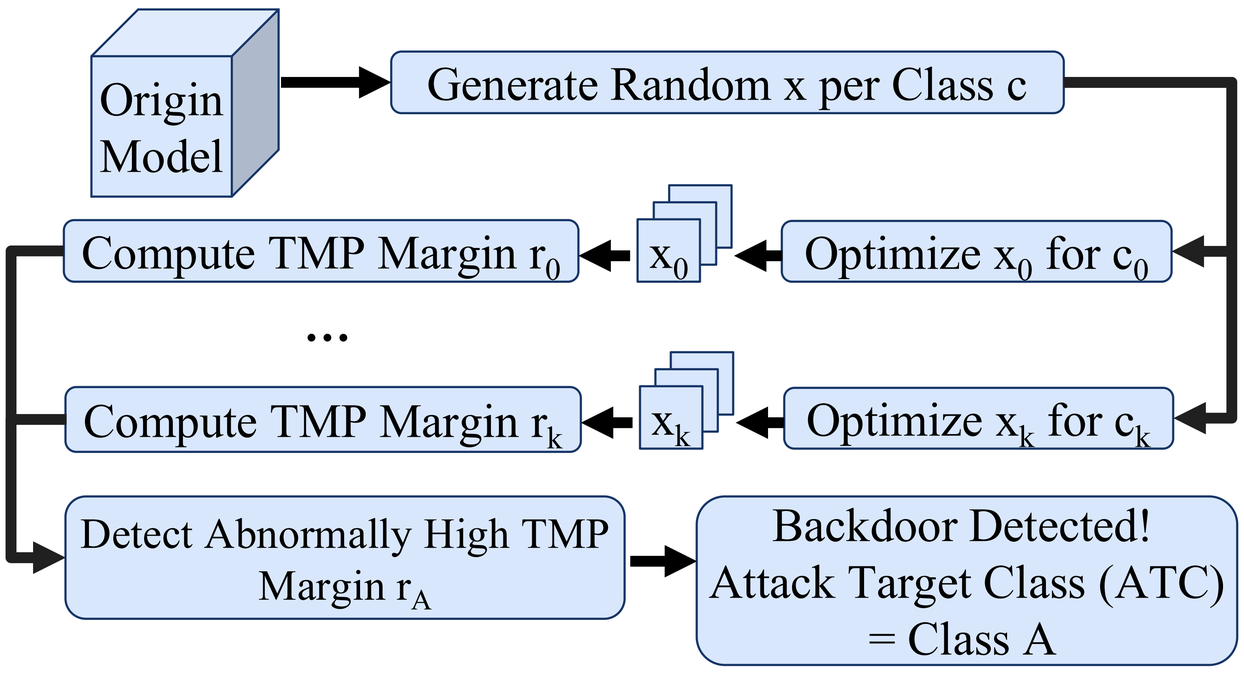
Threat Model

- Attacker: dirty-label, all-to-one, data-poisoning backdoor attack.
- Defender: no training data, no clean labels, white-box model only.



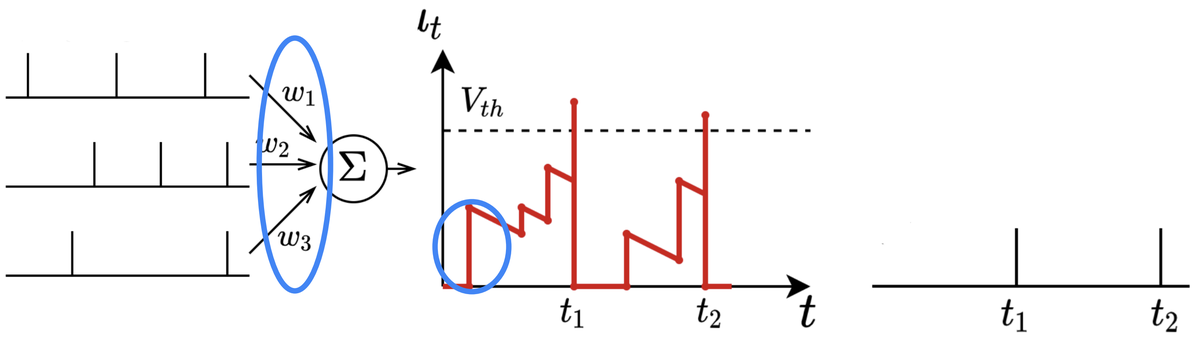
TMPBD (Detection Block)

- Backdoor Attack Target Class (ATC) causes overactivation observed as abnormaly high membrane potential.



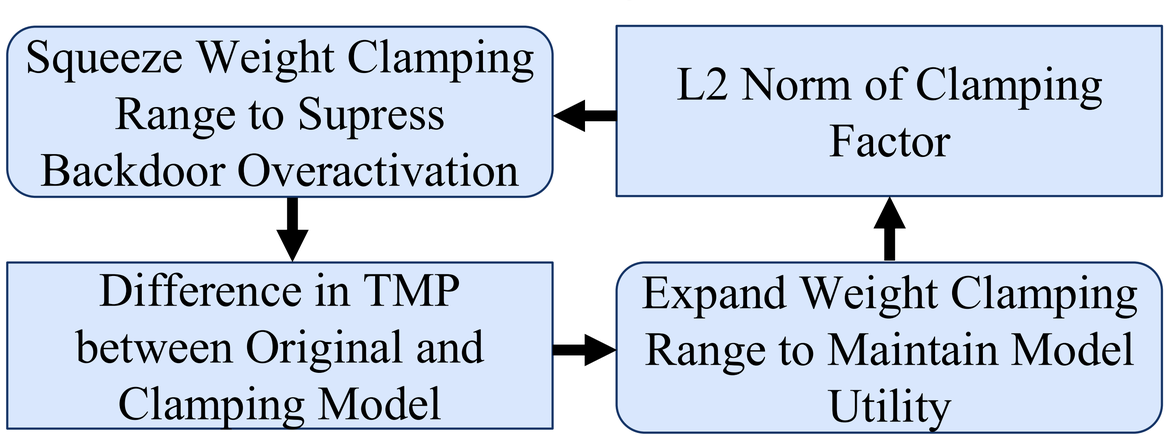
NDSBM (Mitigation Block)

- Squeeze model weight into a clamping range to mitigate backdoor overactivation.

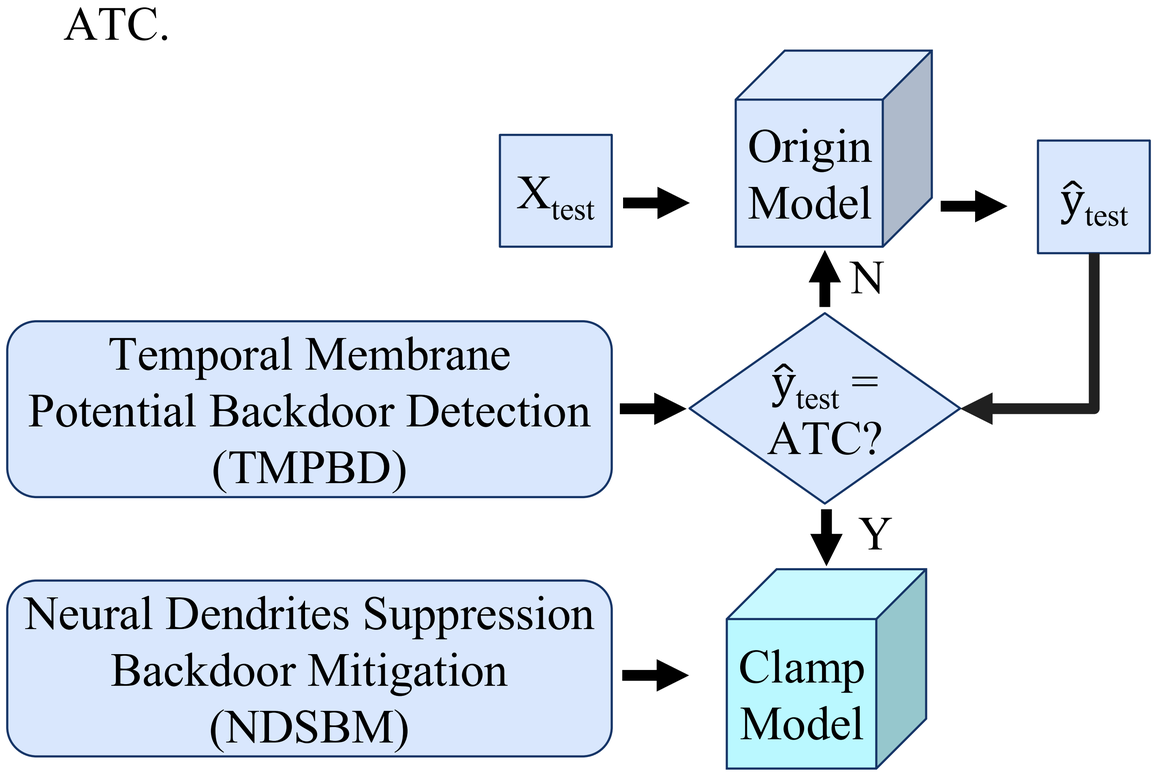


[3]

- Choosing weight clamping weight balancing the defense effectiveness and model utility.



- Only apply clamp model with inference sample predicted ATC.



Results

- 100% backdoor & ATC detection accuracy.
- ASR: 100% → 2.81%.

	Clean		Backdoored	
	CA↑	ASR↓	CA↑	ASR↓
Original	97.65±1.03	0.31±0.99	84.71±12.48	100.00±0.00
Supervised				
Fine-Tuning	64.56±6.63	3.00±4.70	88.53±5.54	3.28±4.51
MMBM	73.09±8.38	2.90±2.99	71.76±16.08	1.40±2.49
Unsupervised				
Self-Tuning	7.20±3.13	11.44±19.89	6.47±1.73	25.56±20.18
Max Cla.	83.09±3.81	2.28±4.30	88.83±4.00	19.38±13.39
Abs. Cla.	84.26±4.75	1.34±2.09	89.12±2.52	20.81±14.50
NDSBM	72.50±6.43	3.69±10.27	89.86±3.21	8.44±9.91
Ours	**97.06±1.55**	**0.31±0.99**	**92.06±4.29**	**2.81±3.95**

Reference

[1] A. Viale, A. Marchisio, M. Martina, G. Masera, and M. Shafique, "Carsnn: An efficient spiking neural network for event-based autonomous cars on the loihi neuromorphic research processor," in *2021 IJCNN*, IEEE, 2021, pp. 1–10.
[2] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019, doi: 10.1109/ACCESS.2019.2909068.
[3] H. Kamata, Y. Mukuta, and T. Harada, "Fully spiking variational autoencoder," in AAAI 2022, pp. 7059–7067.

Full Paper

Source Code