TED++: Submanifold-Aware Backdoor Detection via Layerwise Tubular-Neighbourhood Screening

Nam Le[1], Leo Yu Zhang[2], Kewen Liao[1], Shirui Pan[2], and Wei Luo[1,*]

[1] School of Information Technology, Deakin University, Australia
[2] School of Information and Communication Technology, Griffith University, Australia
* *wei.luo@deakin.edu.au*

What is a Backdoor Attack?

A backdoor attack happens when hidden triggers are injected into the training data, causing the model to behave normally on clean inputs but misclassify any input containing the trigger into an attacker-chosen class , as illustrated in Figure 1. These stealthy attacks create an urgent need for robust defences to protect deep learning models, as they can bypass traditional testing and remain hidden until deployed in real-world systems.

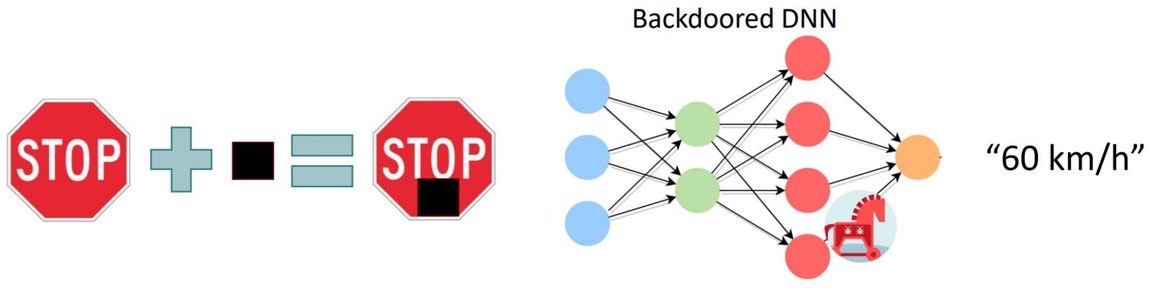

Figure 1. Backdoor attack in deep neural network.

Is TED efficient against various attacks?

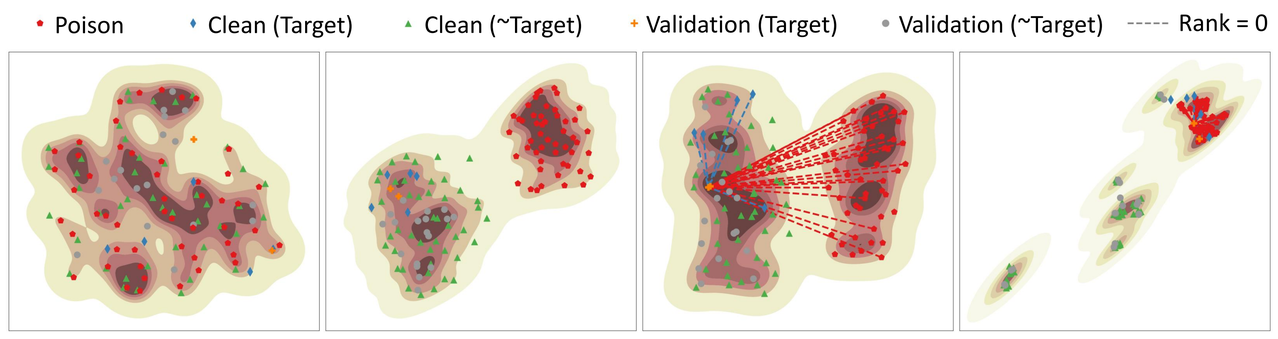

Figure 2. UMAP projections under backdoor attack.

TED [1] tests if a sample's hidden-layer trajectory crosses a class submanifold by monitoring nearest-neighbour ranks. However, these ranks rely on ambient-space distances, making TED vulnerable to advanced backdoor attacks, where poisoned samples are distant from benign ones but still appear as nearest neighbours to target class samples in intermediate layers (See Figure 2).

Backdoor detection via Submanifold Geometry!

TED++ uses a two-stage workflow as illustrated in Figure 3. In the Ranking Computation Stage, we estimate the layer-wise tube radius τ_l from clean validation activations. In the Input Detection Stage, Locally Adaptive Ranking (LAR) assigns the worst rank to activations outside the healthy tube, preserving the order inside. This focuses detection on backdoor deviations, not benign fluctuations.

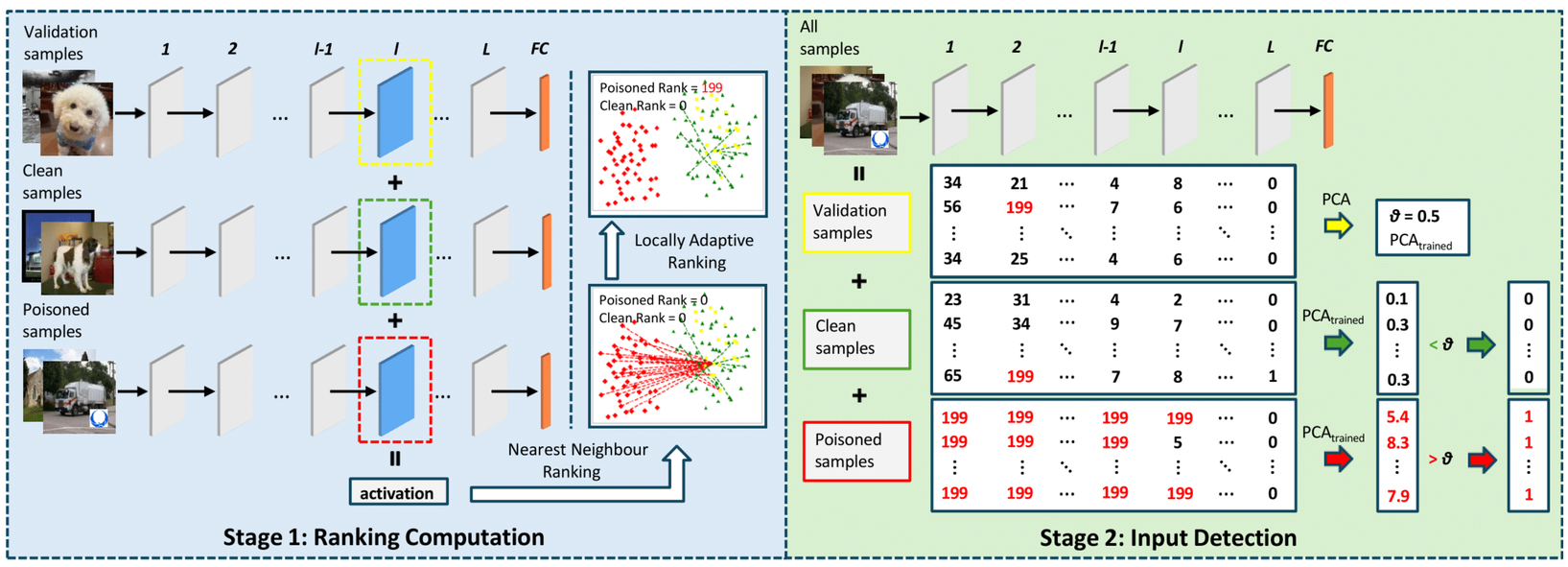

Figure 3. TED++ pipeline.

What is a Locally Adaptive Ranking?

We propose a Locally Adaptive Ranking (LAR) to detect off-tube activations. Standard nearest-neighbour ranks cannot penalise points outside the estimated class tube of radius τ_l .For each input x with predicted class c, we sort validation activations by Euclidean distance:

$$v_{(1)}, v_{(2)}, \ldots, v_{(|\mathcal{V}|)} = \arg\operatorname{sort}_{v \in \mathcal{V}} d\big(h^{(\ell)}(x), h^{(\ell)}(v)\big)$$

The rank is the index of the first same-class neighbour:

$$R_\ell(x) = \min\{\, k \mid y(v_{(k)}) = c\}$$

Let $v^* = \arg\min_{v \in \mathcal{V}_c} \|h^{(\ell)}(x) - h^{(\ell)}(v)\|_2$ .If $h^{(\ell)}(x)$ escapes the tube $\mathcal{T}_c^{(\ell)}(\tau_\ell)$, we assign the worst rank:

$$R_\ell(x) = \begin{cases} |\mathcal{V}|, & \|h^{(\ell)}(x) - h^{(\ell)}(v^*)\|_2 > \tau_\ell, \\ R_\ell(x), & \text{otherwise}, \end{cases}$$

Thus, off-tube activations get maximum penalty, while on-tube remain naturally ranked.

Is TED++ efficient against various attacks?

We evaluate our method on CIFAR-10, GTSRB, and TinyImageNet using ResNet-18. We compare against nine backdoor attacks as shown in Table 1 and 2. Our tests also show TED++ performs consistently even with fewer validation samples, while TED's performance degrades significantly with fewer samples as shown in Table 3 and 4.

Table 1. CIFAR-10.

Attacks →	BadNets		Blend		Ada-Patch		Ada-Blend		WaNet		Trojan		IAD		TaCT		SSDT		*Avg.*	
Defences ↓	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
SCALE-UP	0.96	0.91	0.68	0.52	0.79	0.73	0.75	0.63	0.72	0.61	0.92	0.88	0.96	0.92	0.60	0.28	0.49	0.11	0.76	0.62
STRIP	0.64	0.23	0.73	0.56	0.82	0.68	0.91	0.81	0.45	0.11	0.71	0.30	0.98	0.93	0.46	0.10	0.49	0.09	0.69	0.42
IBD-PSC	**0.99**	**0.95**	**0.99**	0.96	0.88	0.91	0.85	0.77	**0.97**	**0.95**	**0.96**	0.95	**1.00**	**0.97**	0.83	0.87	0.48	0.06	0.88	0.82
TED	0.96	0.93	0.99	**0.97**	0.86	0.80	0.62	0.03	0.96	0.92	0.62	0.11	0.81	0.66	0.68	0.03	0.92	0.84	0.82	0.69
TED++	0.99	0.95	0.92	0.82	**0.99**	**0.97**	**0.93**	**0.89**	0.91	0.87	0.94	**0.97**	0.99	0.92	**1.00**	**0.95**	**0.99**	**0.91**	**0.96**	**0.95**

Table 2. GTSRB.

Attacks →	BadNets		Blend		Ada-Patch		Ada-Blend		WaNet		Trojan		IAD		TaCT		SSDT		*Avg.*	
Defences ↓	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
SCALE-UP	0.90	0.83	0.62	0.55	0.88	0.82	0.59	0.53	0.30	0.17	0.21	0.06	0.89	0.83	0.49	0.10	0.51	0.09	0.60	0.44
STRIP	0.95	0.88	0.91	0.84	0.99	0.94	0.93	0.88	0.45	0.13	0.74	0.48	**0.99**	0.94	0.42	0.02	0.51	0.12	0.77	0.58
IBD-PSC	**0.96**	**0.96**	0.91	0.36	0.97	0.94	0.86	0.09	0.88	**0.91**	**0.95**	**0.95**	0.96	**0.96**	0.48	0.00	0.53	0.53	0.83	0.63
TED	0.95	0.94	0.93	0.52	0.94	0.91	0.73	0.63	0.91	0.90	0.89	0.53	0.93	0.93	0.84	0.72	**0.99**	**0.98**	0.90	0.81
TED++	0.93	0.90	**0.99**	**0.96**	**1.00**	**0.97**	**0.96**	**0.94**	**0.91**	0.80	0.95	0.93	0.97	0.95	**0.91**	**0.91**	0.95	0.84	**0.95**	**0.94**

Table 3. CIFAR-10.

m →	20		10		5		2	
Attacks ↓	TED	**TED++**	TED	**TED++**	TED	**TED++**	TED	**TED++**
BadNets	0.95	**0.99**	0.97	**0.99**	0.96	**0.99**	0.83	**0.94**
Blend	0.97	**0.99**	**0.98**	0.97	**0.99**	0.92	0.36	**0.88**
Ada-Patch	0.83	**0.99**	0.80	**0.99**	0.86	**0.99**	0.45	**0.93**
Ada-Blend	0.76	**0.99**	0.63	**0.98**	0.62	**0.93**	0.67	**0.96**
WaNet	0.86	**0.95**	0.75	**0.93**	**0.96**	0.91	**0.91**	0.88
Trojan	0.79	**0.99**	0.79	**1.00**	0.62	**0.94**	0.71	**0.96**
IAD	0.89	**0.99**	0.85	**0.99**	0.81	**0.99**	0.61	**0.98**
TaCT	0.74	**1.00**	0.75	**1.00**	0.68	**1.00**	0.89	**1.00**
SSDT	0.99	**1.00**	0.97	**0.99**	0.92	**0.99**	0.75	**0.94**
Avg.	0.86	**0.99**	0.83	**0.98**	0.82	**0.96**	0.69	**0.94**

Table 4. GTSRB.

m →	20		10		5		2	
Attacks ↓	TED	**TED++**	TED	**TED++**	TED	**TED++**	TED	**TED++**
BadNets	0.96	**0.99**	0.95	**0.97**	**0.95**	0.93	**0.92**	0.89
Blend	0.98	**0.99**	0.96	**0.97**	0.93	**0.99**	0.14	**0.85**
Ada-Patch	0.93	**0.98**	0.91	**0.98**	0.94	**1.00**	0.72	**0.93**
Ada-Blend	0.89	**0.99**	0.85	**0.98**	0.73	**0.96**	0.34	**0.87**
WaNet	**0.92**	0.92	**0.94**	0.89	0.91	**0.91**	**0.89**	0.83
Trojan	0.94	**0.98**	0.93	**0.98**	0.89	**0.95**	0.34	**0.88**
IAD	0.98	**1.00**	0.97	**1.00**	0.93	**0.97**	**0.99**	0.98
TaCT	0.93	**0.96**	0.89	**0.93**	0.84	**0.91**	0.54	**0.83**
SSDT	**0.99**	0.98	**0.99**	0.94	**0.99**	0.95	**0.93**	0.80
Avg.	0.95	**0.97**	0.93	**0.96**	0.90	**0.95**	0.65	**0.87**

References

[1] X. Mo, Y. Zhang, L. Y. Zhang, W. Luo, N. Sun, S. Hu, S. Gao, and Y. Xiang, "Robust backdoor detection for deep learning via topological evolution dynamics," in IEEE S&P, 2024.

Further Reading

Paper Code [1]