

On the Effectiveness of Invisible Backdoor Attacks in Federated Learning

Mattjis Marinus · Vasileios Tsouvalas · Nirvana Meratnia

Eindhoven University of Technology, The Netherlands

UbiSec 2024 · Proceedings published 2025

01 Motivation

The problem

Invisible backdoor triggers must survive aggregation with honest updates. This study asks when existing attacks remain effective in federated learning (FL) and whether low-activation placement improves success without making detection easier.

Contributions

- **Federated comparison:** Evaluate four invisible attack families against visible BadNet triggers.
- **Multiple metrics:** Separate attack success, clean accuracy, visual similarity and model-level detection.
- **Placement experiment:** Test proposed low-activation placement with least significant bit (LSB) triggers and matched controls.

Attacker setting

A minority of clients poisons local training samples toward a chosen target class while honest clients train normally. The server aggregates both. The study varies attacker prevalence and compares selected defenses.

Success is not stealth

Attack success rate (ASR) measures triggered misclassification; clean accuracy measures retained task performance. Learned Perceptual Image Patch Similarity (LPIPS) compares image appearance, whereas Neural Cleanse tests model anomalies. A visually subtle trigger can still be detected.

Scope

Experiments use ResNet-18 and small image datasets; repeated-run counts and the meaning of ± are unspecified. Placement evidence covers LSB.

Notation

Dirichlet distribution with concentration α (Dir(α)) · percentage points (pp)

02 FiBA

Test whether invisible triggers survive aggregation



Technical core

$$\mathcal{L}_{\text{attack}} = \mathcal{L}_{\text{clean}} + \mathcal{L}_{\text{poison} \rightarrow t}$$

The two cross-entropy terms fit clean labels and poisoned samples labelled as target class t . Honest updates compete with this objective.

$$\text{ASR}(\%) = 100 \frac{N_{\text{triggered} \rightarrow t}}{N_{\text{triggered}}}$$

$N_{\text{triggered}}$: poisoned test samples excluding the target class; the numerator counts those predicted as target t . ASR measures attack success.



Paper



Code

03 Evaluation

88.59%

Encoder attack success

20% attackers · CIFAR-10

+19.80 pp

LSB700 ASR difference

Low minus unmasked placement

Attacker prevalence changes attack success

Attack	Clean acc., 10%	ASR, 10%	ASR, 20%
BadNet	84.00 ± 0.31	94.96 ± 0.81	97.52 ± 0.40
LSB	83.37 ± 0.33	41.59 ± 6.19	65.43 ± 3.93
WaNet	83.33 ± 0.31	44.41 ± 3.50	82.05 ± 3.59
Encoder	83.44 ± 0.27	29.44 ± 5.90	88.59 ± 4.01
LIRA	83.71 ± 0.41	10.52 ± 1.22	41.57 ± 4.66

CIFAR-10 · ResNet-18 · FedAvg · 10 clients · 100 rounds · Dir(0.5). Poison rate 0.5; target class 5. Columns give attacker fraction; reported ± is undefined.

Low-activation placement has a cost

LSB700 placement	ASR (%)	LPIPS	Anomaly index
Unmasked	41.59 ± 6.19	7.83e-5	1.61
Low (Ours)	61.39 ± 4.47	2.25e-4	3.04
High	36.89 ± 4.26	3.75e-4	2.01

LSB700 · CIFAR-10 · 10% attackers. Lower LPIPS means closer images; Neural Cleanse flags anomaly above 2. “Low” is the proposed placement.

Takeaway

Encoder reaches **88.59% ASR with 20% attackers**, showing the importance of attacker prevalence. Proposed **low-activation LSB700 placement** gains 19.80 points over unmasked placement, but raises Neural Cleanse’s anomaly index from 1.61 to 3.04.