LIRA: Learnable, Imperceptible Backdoor Attack

Khoa D. Doan, Yingjie Lao, Weijie Zhao, Ping Li
Baidu Research, USA

2021 ICCV OCTOBER 11-17 VIRTUAL

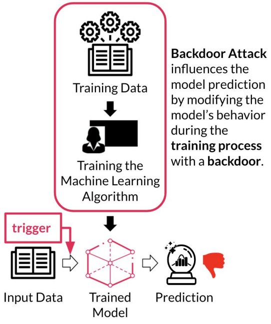
Baidu Research

HIGHLIGHTS

We study the problem of backdoor attack in DNNs

- We *simultaneously* **learn to generate a conditional (dynamic) trigger pattern** and to **poison the model** via a novel non-convex, constrained optimization problem.
- We solve this optimization problem with an efficient, alternating-update optimization algorithm.
- The proposed method LIRA can **generate invisible triggers** that vary across samples to samples and **achieves state-of-the-art targeted attacks** during inference while preserving the clean-sample performance of the model.
- In this case, the "secret" kept by the attacker is the trigger generator, instead of a pattern as in patch-based attacks.
- LIRA is also **stealthy against existing backdoor defenses.**

THREAT MODEL



Clean | Yellow Square

Prediction: STOP | Prediction: GO

This is a paramount security concern in the model building supply chain, as the increasing complexity of machine learning models has promoted training outsourcing and machine learning as a service (MLaaS).

APPROACH

Simultaneously learn to generate an invisible trigger & optimally poison the model in a constrained optimization:

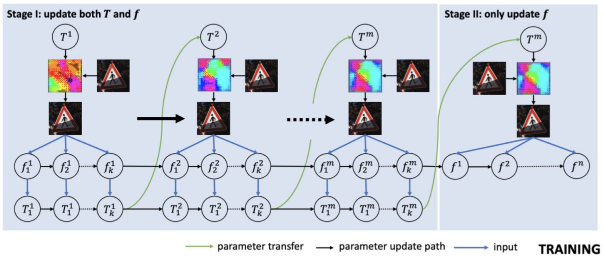
$$\arg\min_{\theta} \sum_{i=1}^{N} \underbrace{\boxed{\alpha\mathcal{L}(f_\theta(x_i), y_i)}}_{\text{clean data objective}} + \underbrace{\boxed{\beta\mathcal{L}(f_\theta(\mathcal{T}_{\xi^*(\theta)}(x_i)), \eta(y_i))}}_{\text{triggered data objective}}$$

$$s.t.\ (1)\ \xi^* = \arg\min_{\xi} \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathcal{T}_\xi(x_i)), \eta(y_i))$$

$$(2)\ d(T(x), x) \leq \epsilon$$

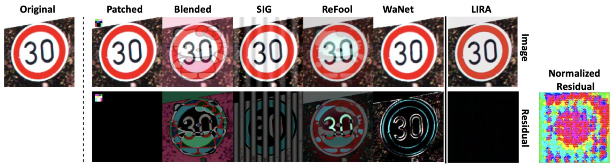
Transformation Function: $T_\xi(x) = x + \underbrace{\boxed{g_\xi(x)}}_{\text{generate dynamic trigger}}, ||g_\xi(x)||_\infty \leq \epsilon$

ALGORITHM



LIRA's learning process is separated in 2 stages.

- Stage I: both f and T are trained (**trigger generation**).
- Stage II: f is trained while T is fixed (**backdoor injection**).



ATTACK PERFORMANCE

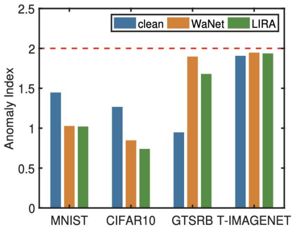
All-to-One Attack $\eta(y) = 0\ \forall y$

Dataset	WaNet Clean	WaNet Attack	LIRA Clean	LIRA Attack
MNIST	0.99	0.99	0.99	1.00
CIFAR10	0.94	0.99	0.94	1.00
GTSRB	0.99	0.98	0.99	1.00
T-ImageNet	0.57	0.99	0.58	1.00

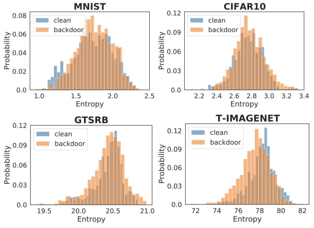
All-to-One Attack $\eta(y) = (y+1)\%|\mathcal{C}|$

Dataset	WaNet Clean	WaNet Attack	LIRA Clean	LIRA Attack
MNIST	0.99	0.95	0.99	0.99
CIFAR10	0.94	0.93	0.94	0.94
GTSRB	0.99	0.98	0.99	1.00
T-ImageNet	0.58	0.58	0.58	0.59

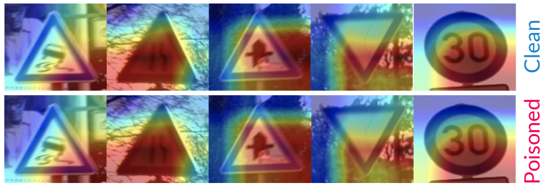
DEFENSE TESTS



Neural Cleanse
Pass defense if Anomaly Index ≤ 2



STRIP-Online Detection
Pass defense if poison and clean entropies are similar..



GradCam Visualization