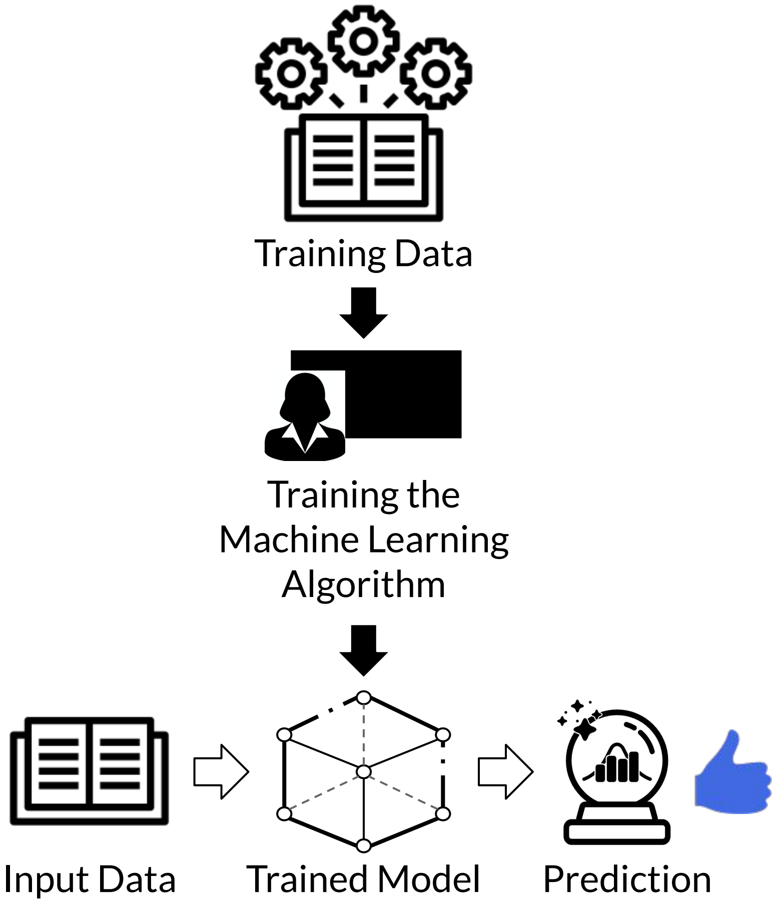
2021 ICCV OCTOBER 11-17 VIRTUAL

LIRA: Learnable, Imperceptible Backdoor Attack

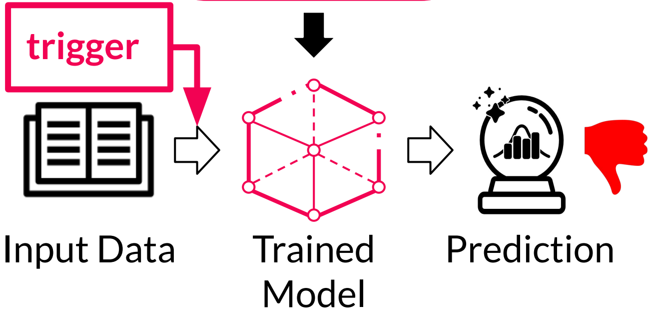
Khoa D. Doan, Yingjie Lao, Weijie Zhao, Ping Li
BAIDU RESEARCH

Machine Learning Models in Practice



Backdoor Attacks

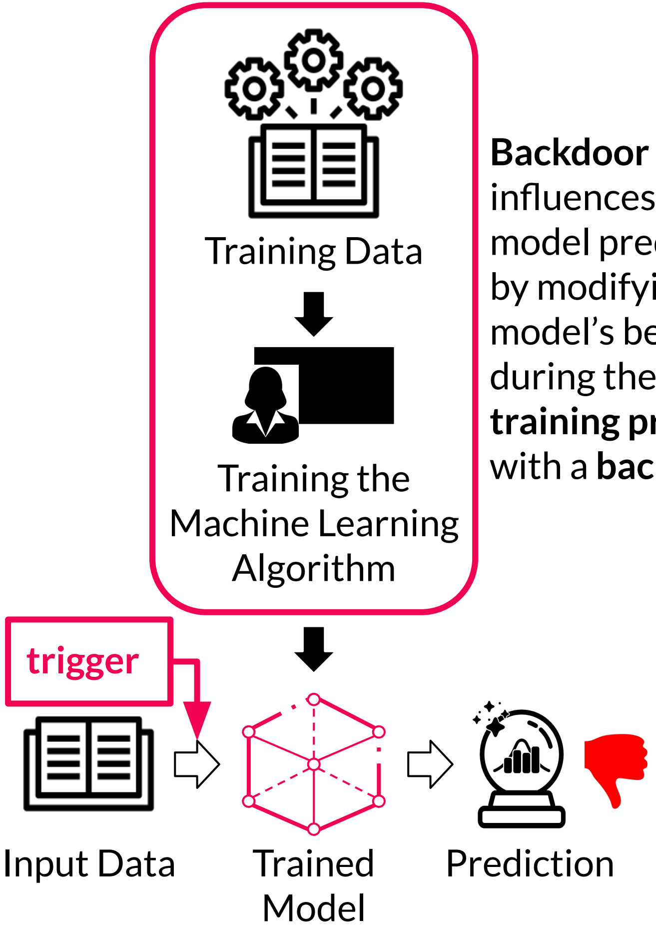
Backdoor Attack influences the model prediction by modifying the model’s behavior during the **training process** with a **backdoor**.



Prediction: **STOP** Prediction: **GO**

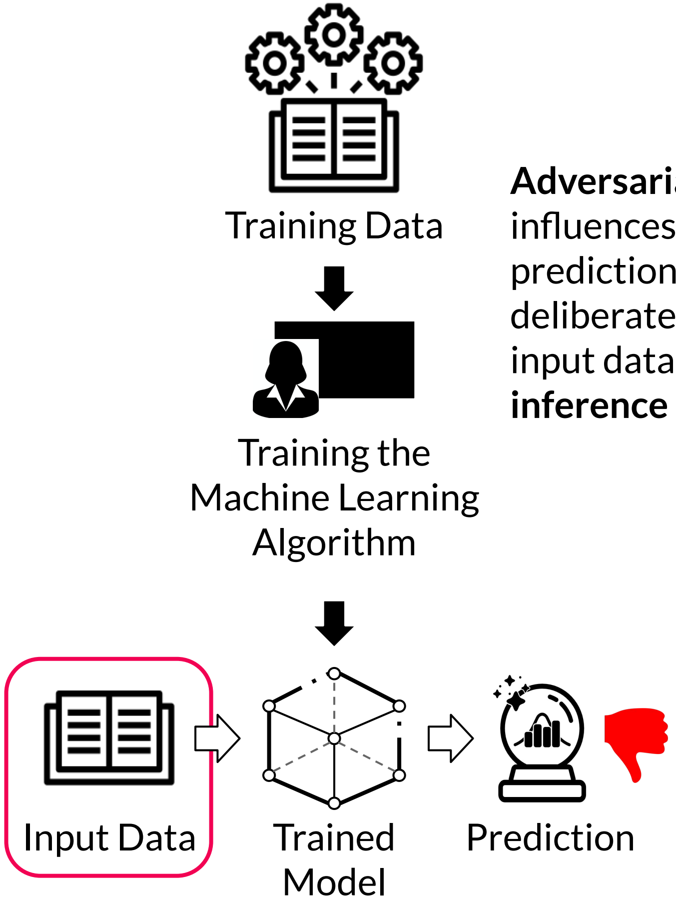
This is a paramount security concern in the model building supply chain, as the increasing complexity of machine learning models has promoted training outsourcing and machine learning as a service (MLaaS).

Backdoor Attacks



Backdoor Attack influences the model prediction by modifying the model's behavior during the **training process** with a **backdoor**.

Adversarial Attacks



Adversarial Attack influences the model prediction by deliberately crafting input data in the **inference phase**.

How is the backdoor injected?

- Consider a classification task

$$f_\theta : \mathcal{X} \to \mathcal{C}$$

$$\mathcal{S} = \{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{C}\}$$

- Generate the trigger:

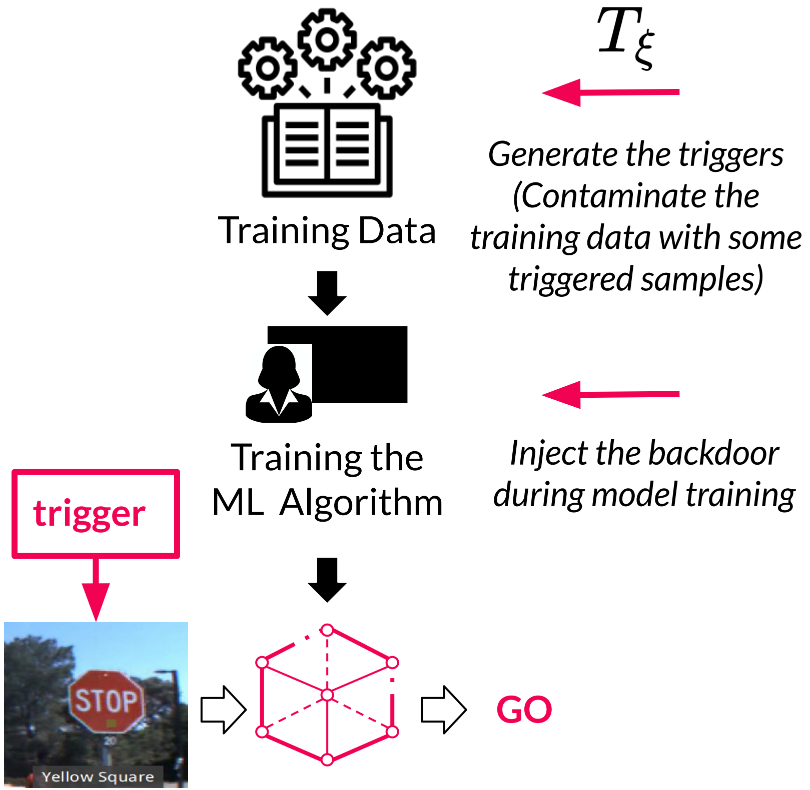
$$T_\xi : \mathcal{X} \to \mathcal{X}$$

$$\hat{\mathcal{S}} = \mathcal{S} \cup \{(T(x_i), \eta(y_i))\}_i$$

- Inject the backdoor:

$$f(x) = y,\ f(T(x)) = \eta(y)$$

$$\text{or} \quad \min_\theta E_{(x_i, y_i) \in \hat{\mathcal{S}}} \mathcal{L}(f_\theta(x_i, y_i))$$



The “fixed” trigger/transformation function

Limitation: The transformation function is predetermined

- Limits the attack visual stealthiness
- Results in lower attack success rates

LIRA: Learnable, Imperceptible BackdooR Attack

- Solve the constrained optimization problem:

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

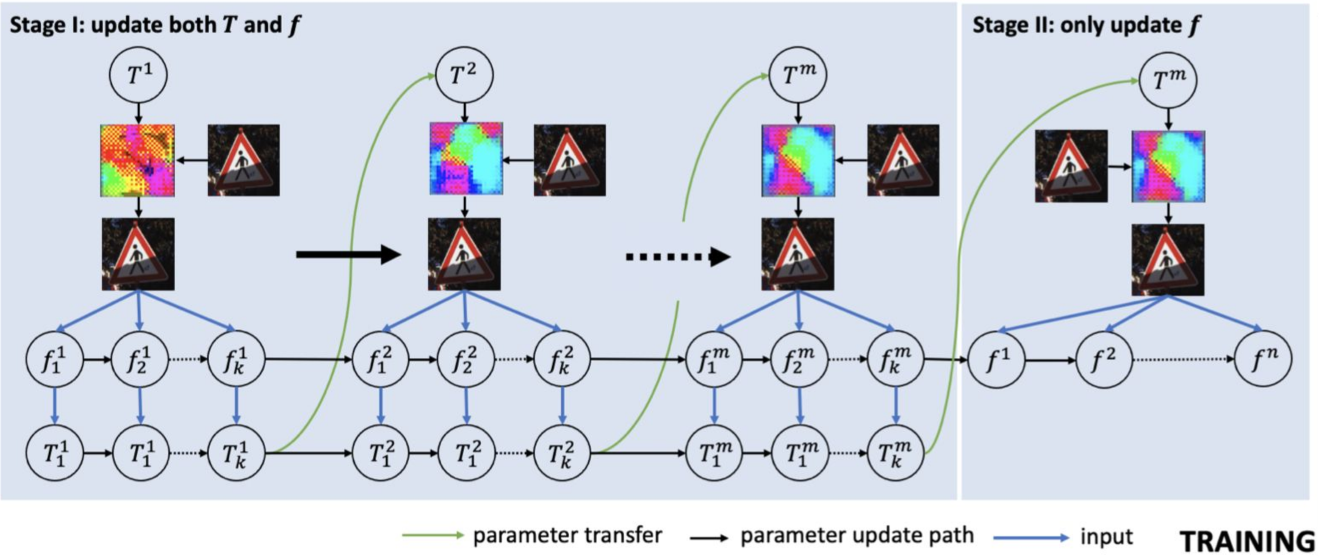
$$s.t.\ (1)\ \xi^{\cdot} = \arg\min_{\xi} \sum_{i=1}^{N} \mathcal{L}(f_\theta(\mathcal{T}_\xi(x_i)), \eta(y_i))$$

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

- The trigger function can be defined as:

$$T_\xi(x) = x + g_\xi(x),\ ||g_\xi(x)||_\infty \leq \epsilon$$

LIRA Learning Algorithm



LIRA's learning process is separated in 2 stages.

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

Algorithm 1 LIRA Backdoor Attack Algorithm

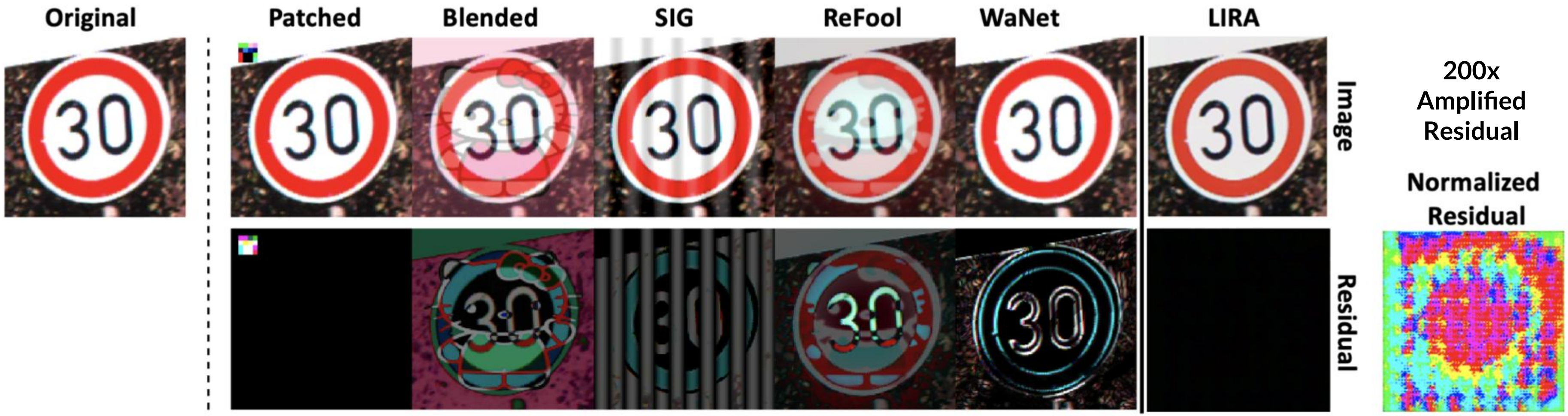
Input:

(1) training samples $S = \{(x_i, y_i), i = 1, ..., N\}$
(2) number of iterations for training the classifier k
(3) number of trials m
(4) number of fine-tuning iterations n
(5) learning rate to train the classifier γ_f
(6) learning rate to train the transformation function γ_T
(7) batch size b
(8) LIRA parameters α and β

Output:

(1) learned parameters of transformation function ξ^*
(2) learned parameters of poisoned classifier θ^*

```
Initialize θ and ξ.
// Stage I: Update both f and T.
ξ̂ ← ξ, i ← 0
repeat
    j ← 0
    repeat
        Sample minibatch (x, y) from S
        θ̂ ← θ^i_j − γ_f ∇_{θ^i_j}(αL(f_{θ^i_j}(x), y) + βL(f_{θ^i_j}(T_ξ̂(x)), η(y)))
        ξ̂ ← ξ̂ − γ_T ∇_ξ̂ L(f_θ̂(T_ξ̂(x)), η(y))
        θ^i_{j+1} ← θ^i_j − γ_f ∇_{θ^i_j}(αL(f_{θ^i_j}(x), y) + βL(f_{θ^i_j}(T_ξ(x)), η(y)))
        j ← j + 1
    until j = k
    ξ ← ξ̂, i ← i + 1
until i = m
// Stage II: Fine-tuning f.
i ← 0, θ_0 ← θ^m_k
repeat
    Sample minibatch (x, y) from S
    θ_{i+1} ← θ_i − γ_f ∇_{θ_i}(αL(f_{θ_i}(x), y) + βL(f_{θ_i}(T_ξ(x)), η(y)))
    i ← i + 1
until i = n
```



Images	Patched	Blended	ReFool	WaNet	LIRA
Backdoor	8.7	1.4	2.3	38.6	60.8
Clean	6.1	10.1	13.1	17.4	40.0
Both	7.4	5.7	7.7	28.0	50.4

Human Inspection Tests - Each tester is trained to recognize the triggered image. Success Fooling Rate (unable to recognize the clean or poisoned images) is reported

Conclusions:

- LIRA has significantly higher success fooling rates.
- LIRA's stealthiness causes increasing confusion between the testers.

Experiment: Attack Performance

Dataset	WaNet		LIRA	
	Clean	Attack	Clean	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

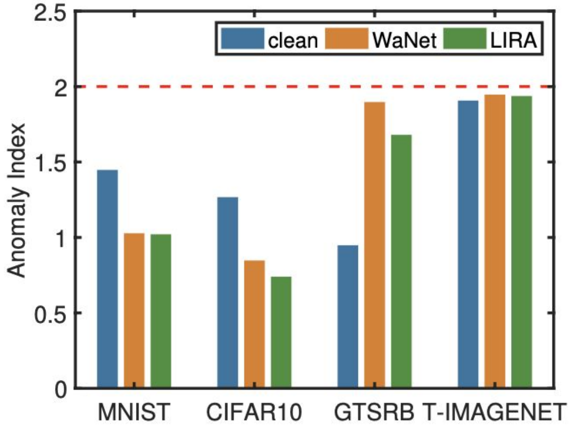
$$\eta(y) = 0\,\forall y$$

Dataset	WaNet		LIRA	
	Clean	Attack	Clean	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

All-to-One Attack

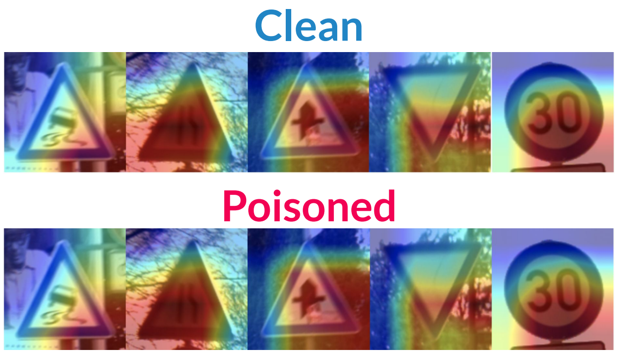
$$\eta(y) = (y+1)\%|\mathcal{C}|$$

Experiment: Machine Defenses

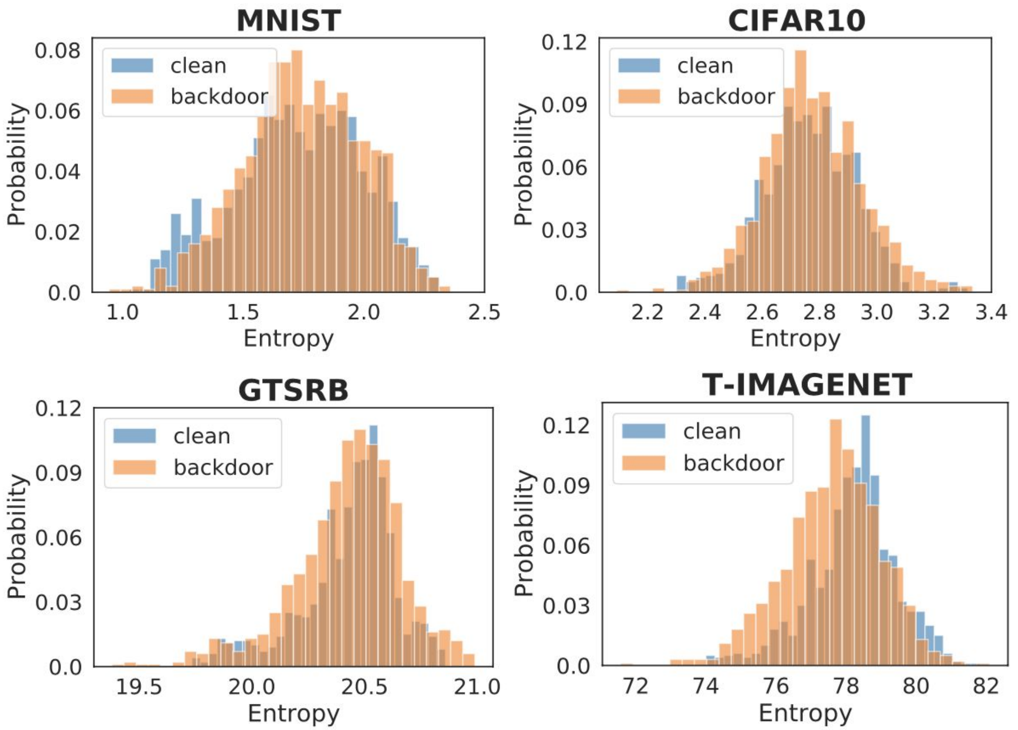


Neural Cleanse-Offline Defense

Pass defense if Anomaly Index ≤ 2



GradCam Visualization



STRIP-Online Detection

Pass defense if poisoned images have similar entropies to clean images.

Thank You!

Contact
Khoa D. Doan
Email: khoadoan106@gmail.com