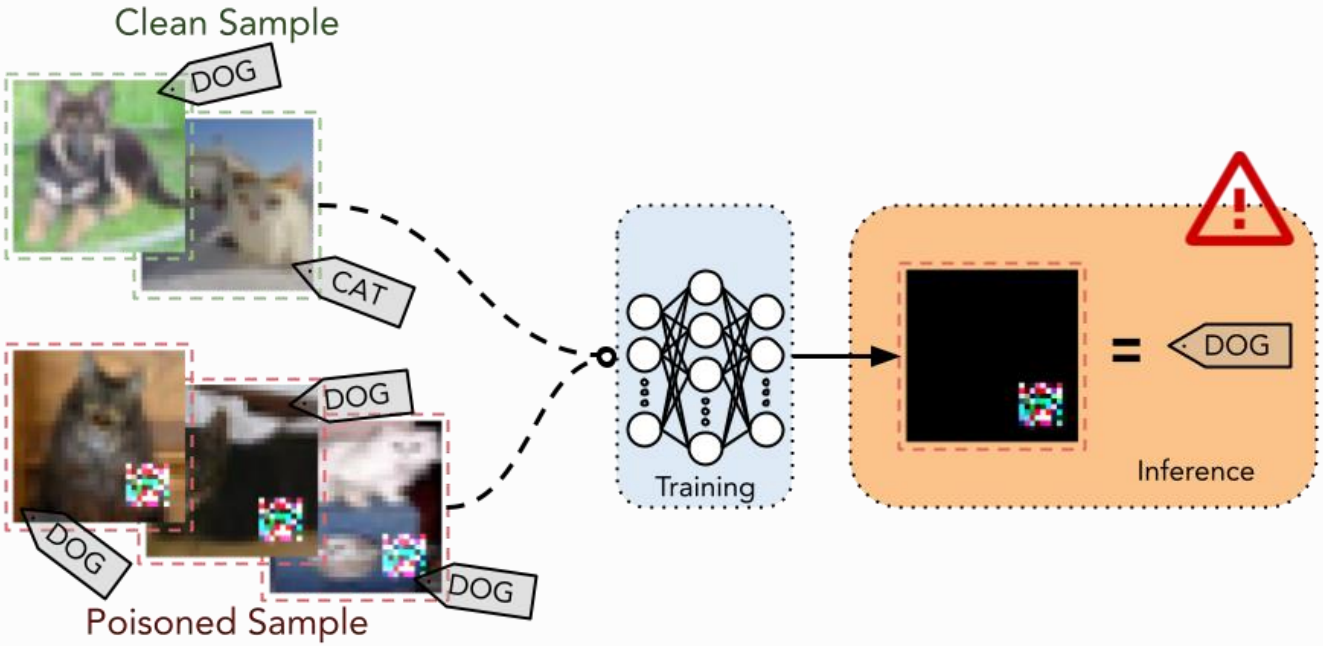
ICLR

Adversarial Unlearning of Backdoors via Implicit Hypergradient

Yi Zeng, Si Chen, Won Park*, Z. Morley Mao*, Ming Jin, Ruoxi Jia
Virginia Tech, *University of Michigan
Corresponding: yizeng@vt.edu

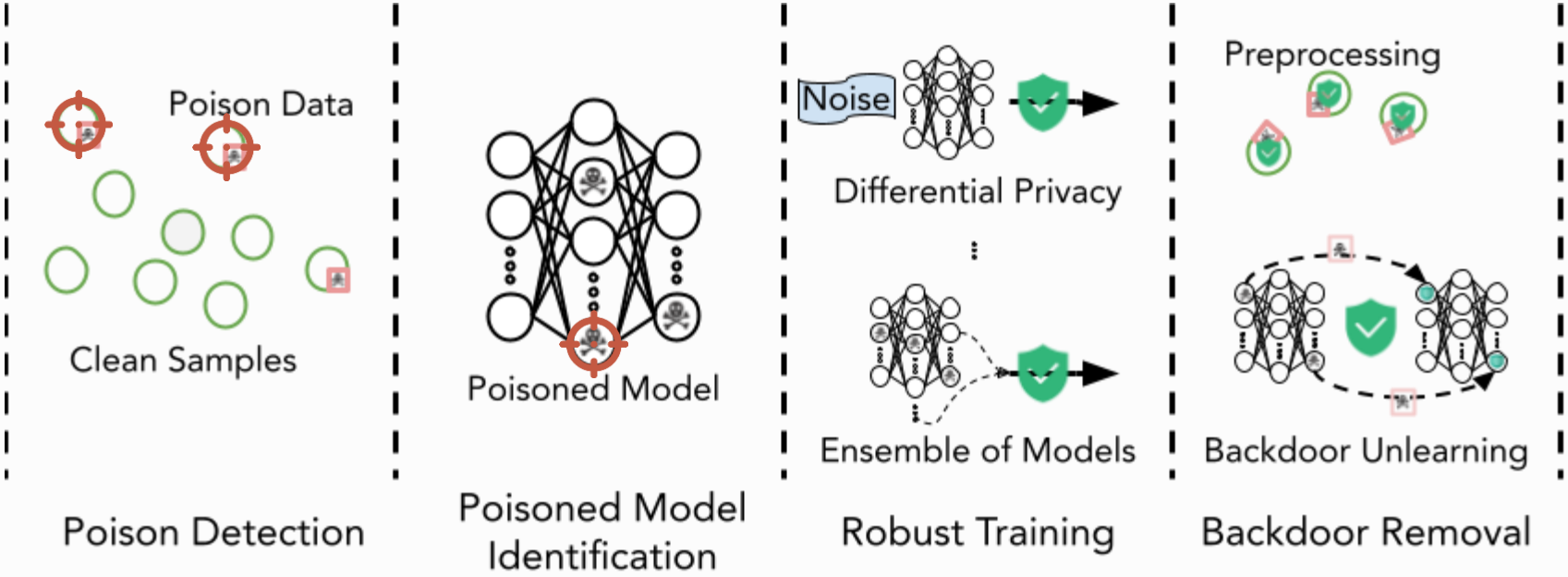
Backdoor Attacks?



- One of the major Training TIME attack
- Normally done via data poisoning
- Imperceptible, hard to detect
- Could result sever consequences

REDS LAB
RESPONSIBLE DATA SCIENCE LAB

Backdoor Defenses

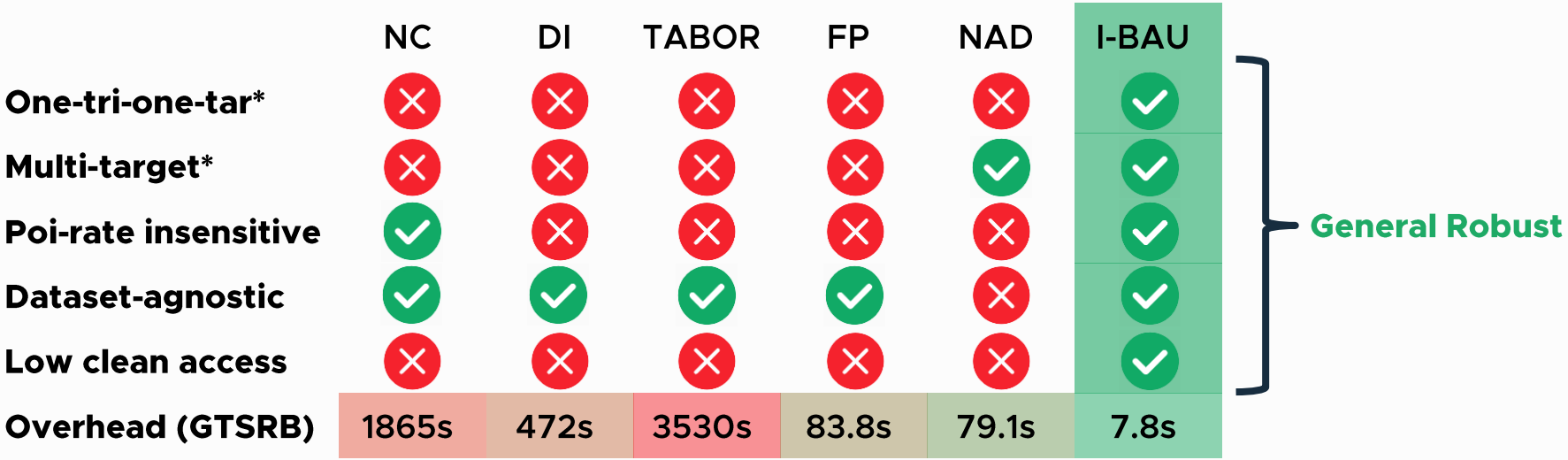


Backdoor Removal

	NC	DI	TABOR	FP	NAD	I-BAU
One-tri-one-tar*	✗	✗	✗	✗	✗	✓
Multi-target*	✗	✗	✗	✗	✓	✓
Poi-rate insensitive	✓	✗	✗	✗	✗	✓
Dataset-agnostic	✓	✓	✓	✓	✗	✓
Low clean access	✗	✗	✗	✗	✗	✓
Overhead (GTSRB)	1865s	472s	3530s	83.8s	79.1s	7.8s

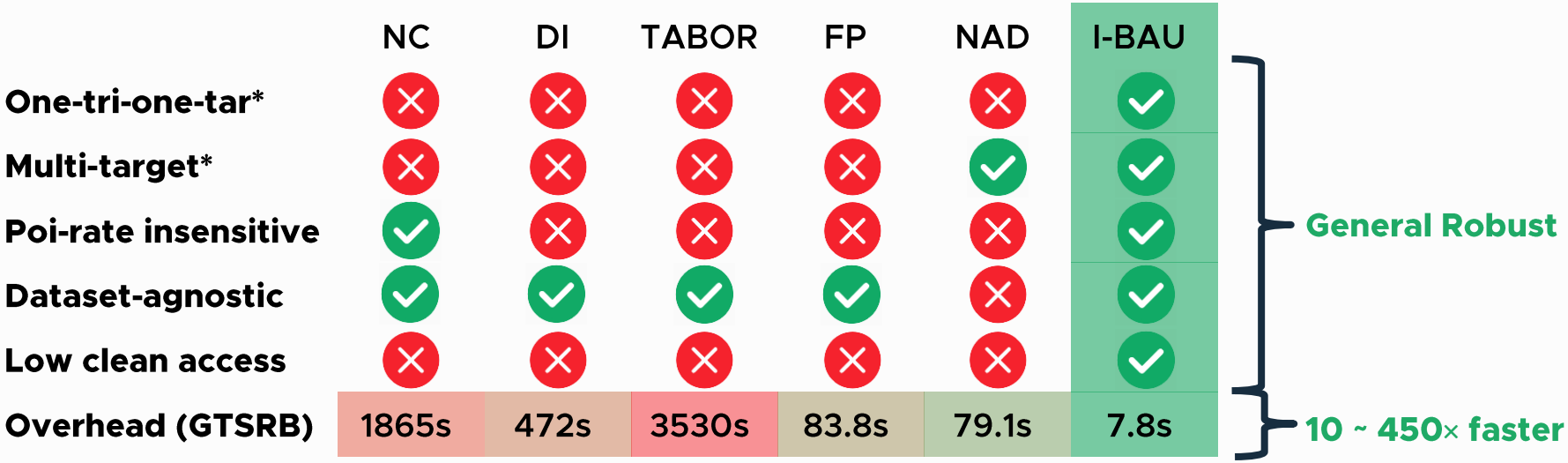
*Results evaluated on CIFAR-10 and GTSRB with eleven different attacks

Backdoor Removal



*Results evaluated on CIFAR-10 and GTSRB with eleven different attacks

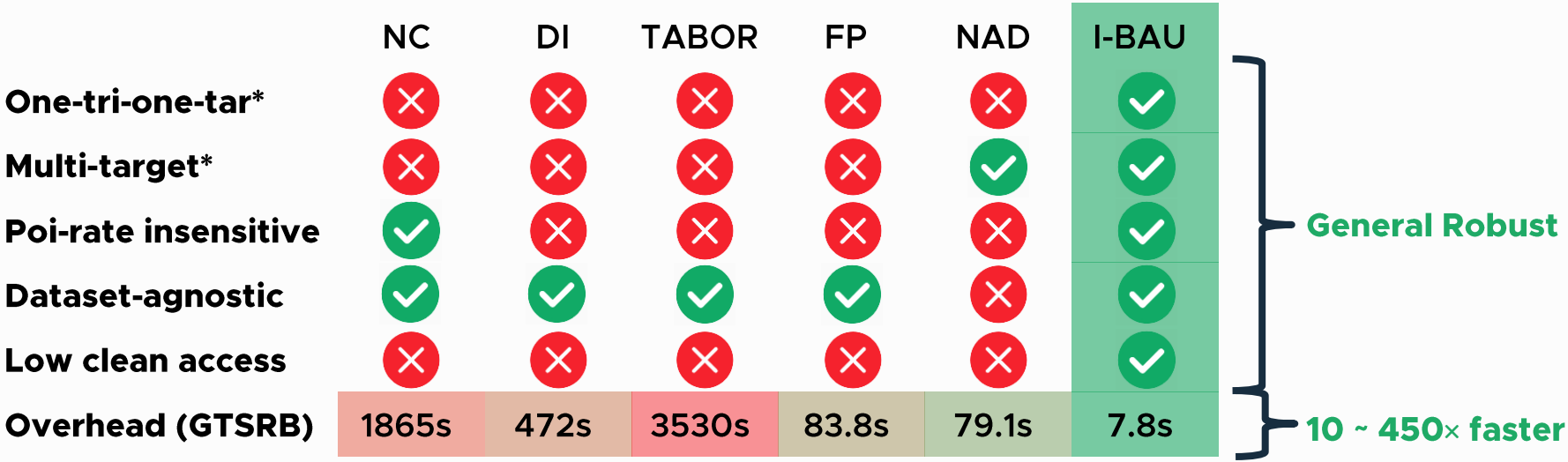
Backdoor Removal



*Results evaluated on CIFAR-10 and GTSRB with eleven different attacks

Backdoor Removal

How we did that?



*Results evaluated on CIFAR-10 and GTSRB with eleven different attacks

① A minimax formulation.

- We formulate backdoor removal as a minimax problem

Unlearn the universal pattern

$$\theta^* = \arg\min_{\theta} \max_{\|\delta\| \leq C_\delta} H(\delta, \theta) := \frac{1}{n} \sum_{i=1}^{n} L(f_\theta(x_i + \delta), y_i)$$

Identifying a universal pattern

① A minimax formulation.

Encompasses other defenses.

- We formulate backdoor removal as a minimax problem

Unlearn the universal pattern

$$\theta^* = \arg\min_{\theta} \max_{\|\delta\| \leq C_\delta} H(\delta, \theta) := \frac{1}{n} \sum_{i=1}^{n} L(f_\theta(x_i + \delta), y_i)$$

Identifying a universal pattern

① A minimax formulation.

- We formulate backdoor removal as a minimax problem

Encompasses other defenses.

Naïve solution is erratic!

Unlearn the universal pattern

$$\theta^* = \arg\min_{\theta} \max_{\|\delta\| \leq C_\delta} H(\delta, \theta) := \frac{1}{n} \sum_{i=1}^{n} L(f_\theta(x_i + \delta), y_i)$$

Identifying a universal pattern

② An advanced solution.

- We proposed an accurate and effective solution with implicit hypergradient.

$$\theta^* = \arg\min_{\theta} \max_{\|\delta\| \le C_\delta} H(\delta, \theta) := \frac{1}{n} \sum_{i=1}^{n} L(f_\theta(x_i + \delta), y_i)$$

$$\underbrace{\nabla \psi(\theta)}_{\text{hypergrad. of } \theta} = \underbrace{\nabla_2 H(\delta(\theta), \theta)}_{\text{direct grad. of } \theta} + \underbrace{\overbrace{(\nabla \delta(\theta))^\top}^{\text{response Jacobian}} \overbrace{\nabla_1 H(\delta(\theta), \theta)}^{\text{direct grad. of } \delta}}_{\text{indirect grad. of } \theta}$$

② An advanced solution.

- We proposed an accurate and effective solution with implicit hypergradient.

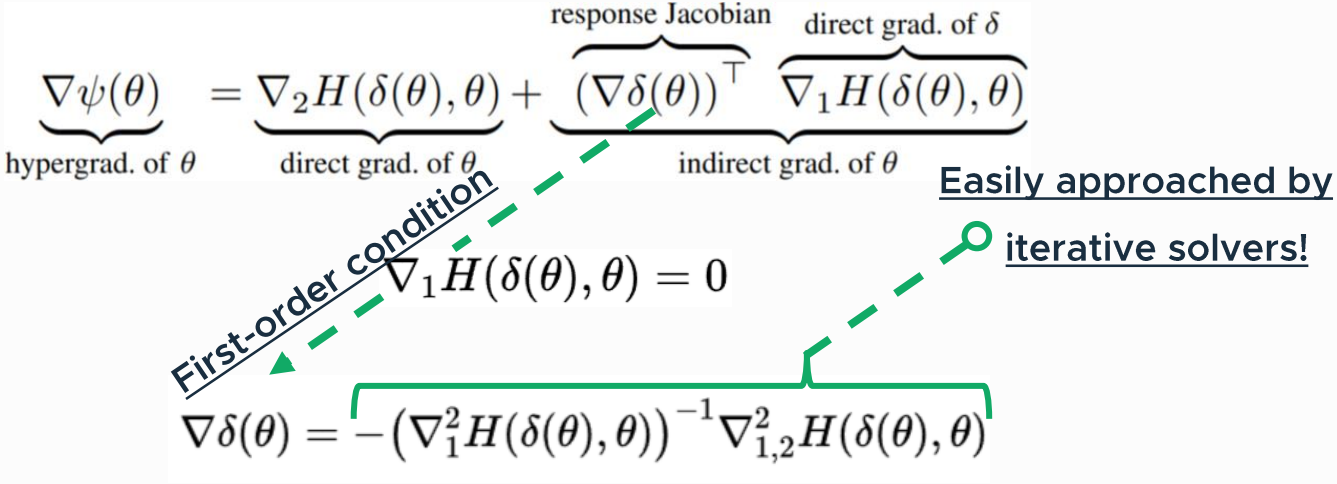
$$\theta^* = \arg\min_{\theta} \max_{\|\delta\| \le C_\delta} H(\delta, \theta) := \frac{1}{n} \sum_{i=1}^{n} L(f_\theta(x_i + \delta), y_i)$$

Difficult to compute!

$$\underbrace{\nabla \psi(\theta)}_{\text{hypergrad. of } \theta} = \underbrace{\nabla_2 H(\delta(\theta), \theta)}_{\text{direct grad. of } \theta} + \underbrace{\overbrace{(\nabla \delta(\theta))^\top}^{\text{response Jacobian}} \overbrace{\nabla_1 H(\delta(\theta), \theta)}^{\text{direct grad. of } \delta}}_{\text{indirect grad. of } \theta}$$

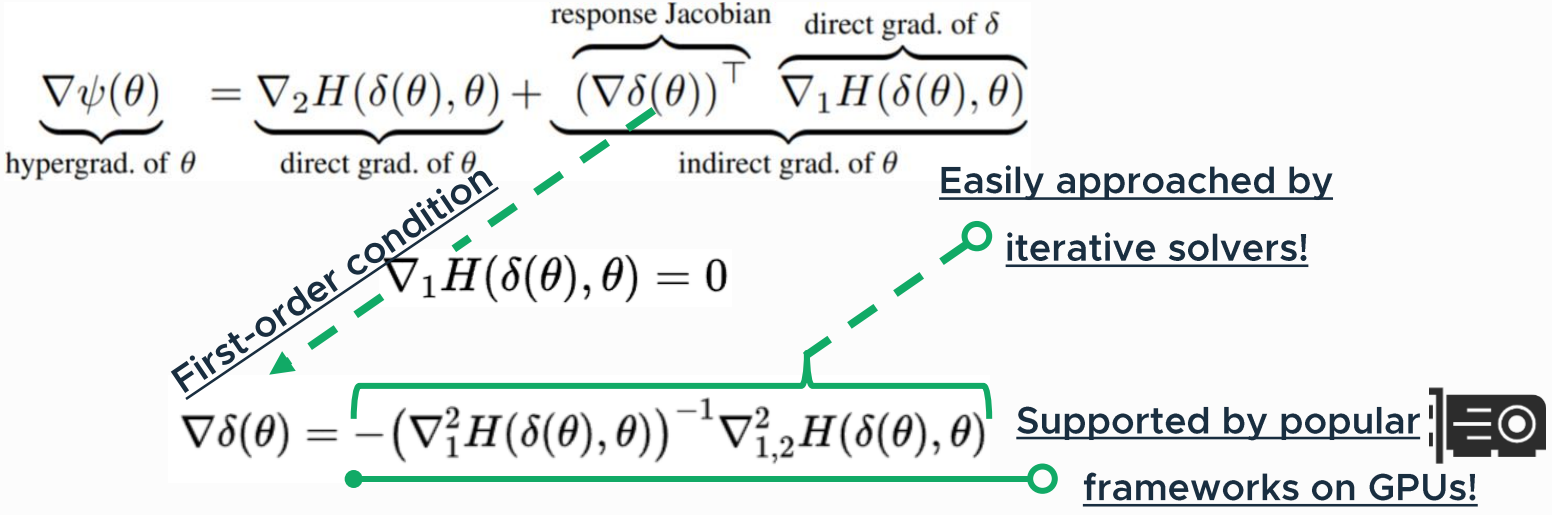
② An advanced solution.

- We proposed an accurate and effective solution with implicit hypergradient.



② An advanced solution.

- We proposed an accurate and effective solution with implicit hypergradient.



Theoretical Highlights

- Theorem 1: Convergence guarantee of the solution

$$\left\|\nabla\tilde{\psi}(\theta_i) - \nabla\psi(\theta_i)\right\| \leq \mathcal{C}(\theta_i, \mu_H, L_H, \hat{\rho}_{1,\theta_i}, \hat{\rho}_{2,\theta_i}, \eta_{1,\theta_i}, \eta_{2,\theta_i})$$

Lipschitz coefficients

- Theorem 2~3: Generalization guarantee of the formulation
  - Linear Models:

$$\Pr\left[\arg\max_j [\theta(x+\delta)_j] \neq y\right] \leq \hat{R}_\gamma(\theta) + \mathcal{C}(C_\theta, C_\delta, \chi, n, \xi)$$

  - Neural Networks:

$$\Pr\left[\arg\max_j [\theta(x+\delta)_j] \neq y\right] \leq \hat{R}_\gamma(\theta) + \mathcal{C}(W, C_\delta, \chi, n, s, \varrho, \xi)$$

Lipschitz coefficients of non-linear layers

REDS LAB
RESPONSIBLE DATA SCIENCE LAB

Empirical Highlights

① General and Robust defense effects

Attack	No Defense		NC		DI		TABOR		FP		NAD		DP		I-BAU(Ours)	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	84.94	98.28	83.42	8.76	82.12	48.50	83.78	8.12	82.22	96.66	78.72	11.66	11.08	21.77	83.35	12.30
Blend	84.82	99.78	83.08	33.96	81.84	52.62	83.42	21.26	81.24	89.78	77.48	13.02	11.68	13.72	82.30	12.96
l_0 inv	85.36	100	83.02	8.78	83.40	29.1	82.27	8.16	82.18	100	65.08	7.22	12.48	25.22	84.08	9.54
l_2 inv	85.26	100	80.68	8.08	82.46	7.82	80.30	11.64	81.50	98.94	43.18	12.56	11.58	20.57	83.48	7.48
Smooth	85.34	99.24	83.72	46.88	83.32	61.82	84.14	45.94	82.66	9.44	77.22	54.38	10.70	28.14	83.46	18.30
Trojan SQ	84.76	99.66	81.30	8.02	83.14	6.94	81.38	7.06	82.34	99.50	51.86	7.84	10.70	18.26	83.18	9.82
Trojan WM	84.92	99.96	81.76	6.02	82.88	7.24	82.60	49.26	81.64	99.88	56.84	0.82	15.21	32.89	83.58	3.42
All to all	86.38	85.02	85.38	82.88	84.74	56.38	×	×	84.48	66.46	75.70	2.34	14.80	10.93	80.34	10.46

Table 1: Results on CIFAR-10, one-trigger cases. CIFAR-10's ACC is sensitive to fine-tuning and I-BAU; we compare I-BAU when it drops similar ACCs amount to the most effective method. × - no detected trigger.

Empirical Highlights

① General and Robust defense effects

Attack	No Defense		TABOR		FP		NAD		DP		I-BAU(Ours)	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	97.69	99.18	98.99	4.16	99.34	60.19	15.19	9.08	6.46	100	99.38	3.32
Blend	97.44	99.91	99.09	33.32	99.52	69.66	47.71	18.48	5.70	100	98.89	5.01
l_0 inv	97.72	100	98.80	0.47	99.41	74.86	17.41	1.20	7.40	88.23	99.24	0.42
l_2 inv	97.57	99.91	98.51	0.41	99.53	40.46	15.50	1.16	5.46	100	97.75	0.45
Smooth	97.87	99.89	98.62	0.47	99.55	47.75	10.06	0.70	5.94	95.58	98.96	0.22
Trojan SQ	98.12	99.98	99.06	5.70	99.48	75.96	23.31	14.68	5.51	100	99.04	5.11
Trojan WM	97.84	100	98.63	5.40	99.45	69.82	11.16	13.62	5.70	100	99.44	2.55
All to all	97.10	95.42	98.63	47.07	99.45	67.34	25.53	0.42	5.87	5.70	99.13	0.04

Table 2: Results on GTSRB, one-trigger cases. I-BAU's results shown here were obtained after 100 rounds of I-BAU. For that, Neural Cleanse, Deppinspect, and TABOR are from the same line of work, so we here only compare the result with the most state-of-art method in this category, TABOR.

			No Def.	NC	DI	TABOR	FP	NAD	DP	Ours*
CIFAR-10	**ACC**		85.96	×	83.74	84.04	83.42	77.38	11.18	77.44
	avg. ASR		98.37	×	30.37	40.68	73.18	10.93	12.83	12.96
	ASRs	Trojan WM: ⇒ 9	99.92	×	9.68	28.02	99.7	18.38	13.40	11.48
		Trojan SQ: ⇒ 2	99.42	×	13.78	19.56	99.36	11.94	2.58	10.80
		BadNets: ⇒ 0	94.5	×	72.6	9.16	9.56	13.32	6.94	18.64
		Smooth: ⇒ 1	97.08	×	30.58	89.70	9.32	8.38	6.28	15.22
		Blend: ⇒ 3	98.12	×	48.84	50.84	96.14	8.46	27.98	18.24
		l_0 inv: ⇒ 4	100	×	23.78	80.96	99.88	9.88	18.02	9.46
		l_2 inv: ⇒ 5	99.54	×	13.34	6.52	98.32	6.16	14.62	6.92
GTSRB	**ACC**		97.18	68.72	99.33	76.38	99.36	10.83	6.17	99.09
	avg. ASR		99.37	10.87	7.21	11.89	9.38	11.59	42.86	4.18
	ASRs	Trojan WM: ⇒ 9	99.49	4.00	3.60	3.86	1.56	6.61	100	1.78
		Trojan SQ: ⇒ 2	99.59	2.79	6.26	2.33	0.12	4.14	0	5.68
		BadNets: ⇒ 0	98.19	6.61	0.47	10.29	14.09	4.79	0	0.72
		Smooth: ⇒ 1	99.89	8.56	6.92	15.28	10.05	2.83	100	5.65
		Blend: ⇒ 3	99.15	46.89	22.66	45.91	29.69	41.54	100	3.76
		l_0 inv: ⇒ 4	100	1.31	5.15	1.06	0.59	17.70	0	5.25
		l_2 inv: ⇒ 5	99.33	5.94	5.43	4.52	9.60	3.52	0	6.46

Table 3: Results for 7-trigger-7-target cases. × marks no trigger was detected. *Here, ASR results on CIFAR-10 are provided when the model attained an ACC similar to that of NAD (the only effective one on CIFAR-10).

① General and Robust defense effects

② Stable performance

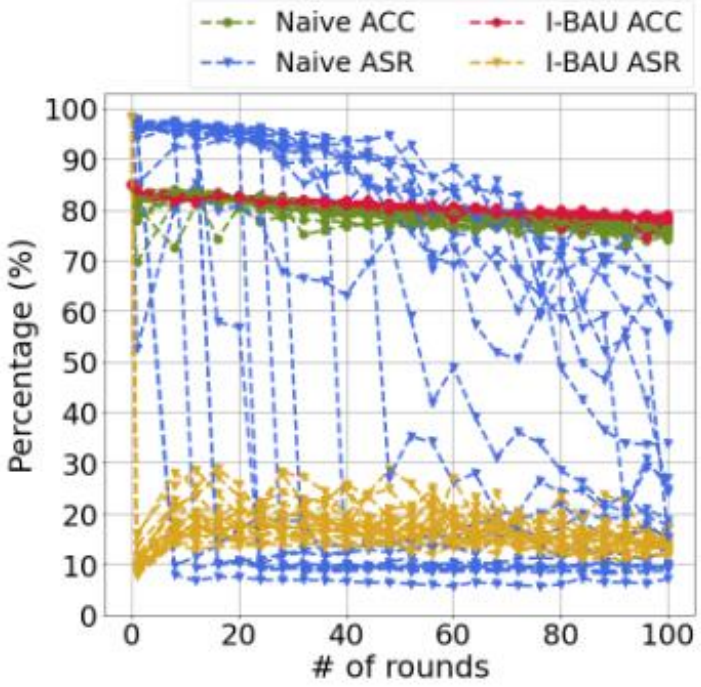


Figure 1: Comparison of Naive and I-BAU. I-BAU's performance is more stable.

Empirical Highlights

① General and Robust defense effects

② Stable performance

③ Insensitive to poisoning rate

poison ratio	Results	No Def.	NC	DI	TABOR	FP	NAD	DP	Ours
5.0%	ACC	86.58	83.14	78.63	×	83.16	79.74	36.80	84.76
	ASR	99.88	5.58	10.40	×	99.72	6.34	96.84	9.78
0.5%	ACC	86.42	84.16	83.56	×	84.72	80.92	39.92	83.22
	ASR	98.58	12.9	20.22	×	93.78	28.6	61.27	13.08

Table 4: Results on CIFAR-10 (Trojan WM) with different poison ratios. × marks no trigger was detected.

① General and Robust defense effects

② Stable performance

③ Insensitive to poisoning rate

④ Effective even with low access to clean samples

# Clean Data	Results	No Def.	NC	DI	TABOR	FP	NAD	Ours
2,500	ACC	84.92	78.39	80.63	80.23	81.36	46.8	82.21
	ASR	99.96	6.53	10.07	33.40	99.58	7.12	6.96
500	ACC	84.92	78.24	80.17	77.03	78.1	38.5	80.07
	ASR	99.96	25.66	1.14	21.92	85.68	9.08	5.20
100	ACC	84.92	84.101	69.51	83.495	73.00	36.14	76.9
	ASR	99.96	99.92	1.12	99.687	97.80	5.76	4.00

Table 5: Results with different # of clean data on CIFAR-10 (Trojan WM).

① General and Robust defense effects

② Stable performance

③ Insensitive to poisoning rate

④ Effective even with low access to clean samples

⑤ Way more efficient than existing work of the art

	CIFAR-10 (s)	GTSRB (s)
NC	384.92	1864.96
DI	394.38	472.21
TABOR	1123.31	3529.70
FP	45.33	83.78
NAD	79.90	79.14
Ours	**6.82**	**7.84**

Table 6: Average time for defenses to be effective on one-trigger-one-target cases.

Recap and Conclusion

- We formulated backdoor removal as a minimax problem;
- We proposed an advanced solution to the minimax with implicit hypergradient;
- We thoracically provided the convergence bound and the generalization bounds;
- We empirically achieved state-of-the-art generalizability and efficiency in defeating multiple settings of backdoor attacks.

REDS LAB
RESPONSIBLE DATA SCIENCE LAB

REDS Lab
Responsible Data Science Lab

VIRGINIA TECH
UNIVERSITY OF MICHIGAN

ICLR

Visit My Page