Backdoor Meets Data-Free Learning

Junyuan Hong
Ph.D., Michigan State University
Postdoc, IFML, UTAustin

Presentation at TMLR Group, Sep 8, 2023

MICHIGAN STATE UNIVERSITY

Backdoor an Classifier

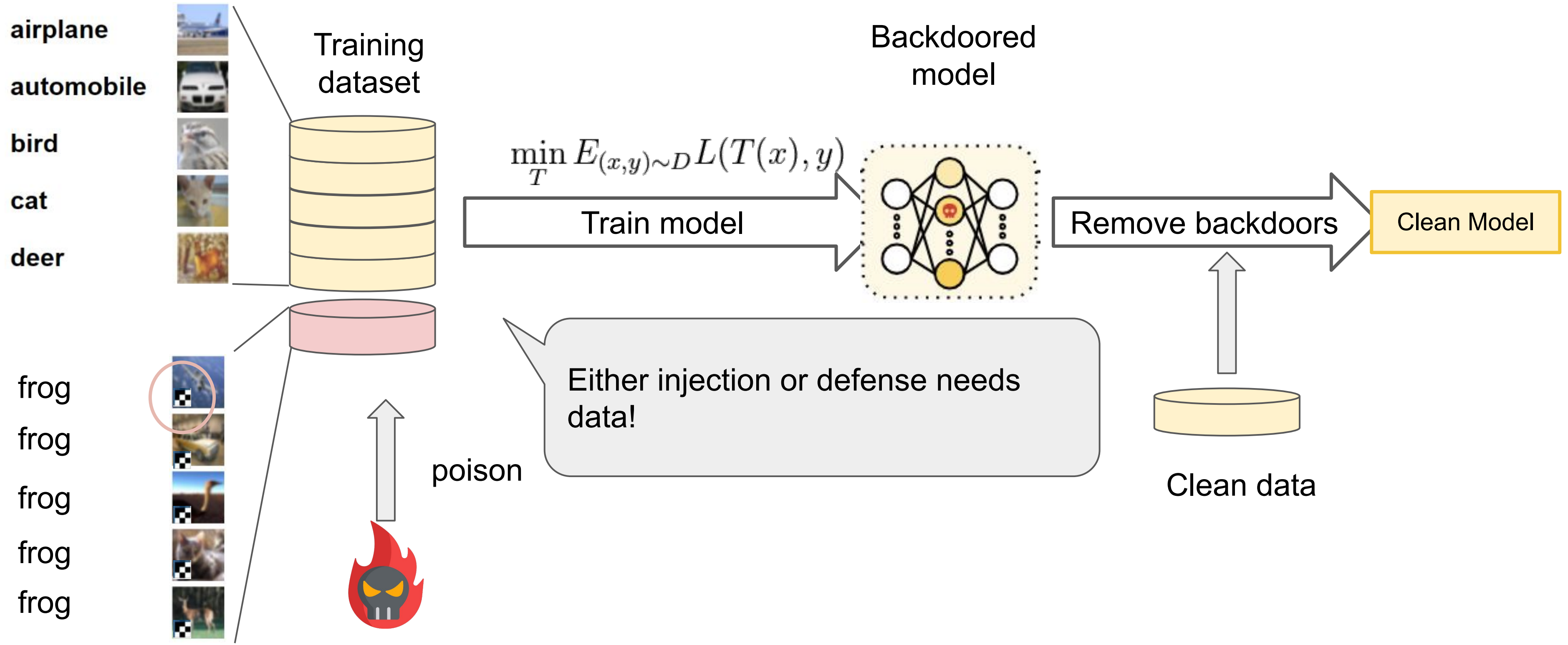
Backdoored training

$$\min_{T} \mathbb{E}_{(\boldsymbol{x},y)\sim D}\left[\underbrace{L(T(\boldsymbol{x}),y)}_{\text{clean task}}+\underbrace{L(T(\boldsymbol{x}+\delta),t)}_{\text{backdoor task}}\right],$$

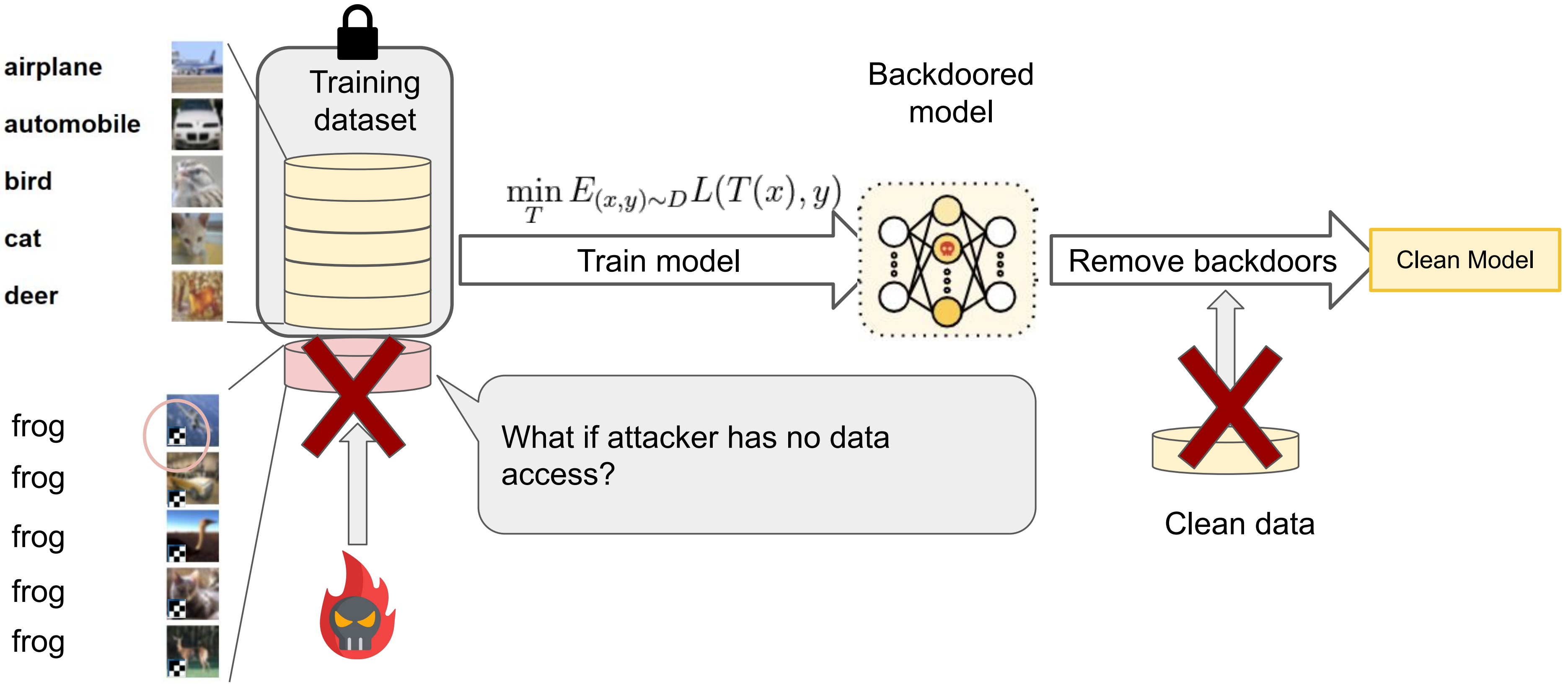
Trigger patch

Ground-truth: Mouse
Prediction: Frog

Backdoor Injection via Poisoning

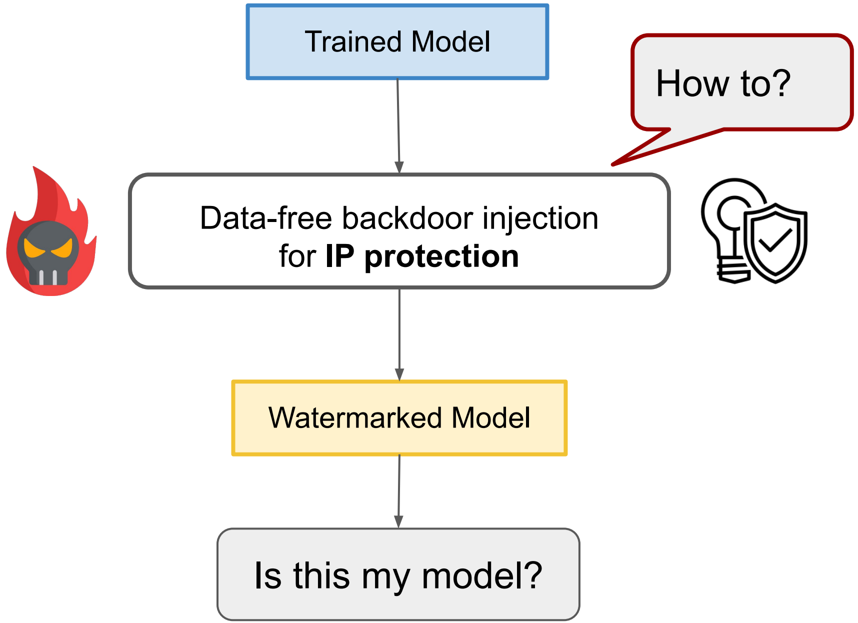


Backdoor Injection via Poisoning



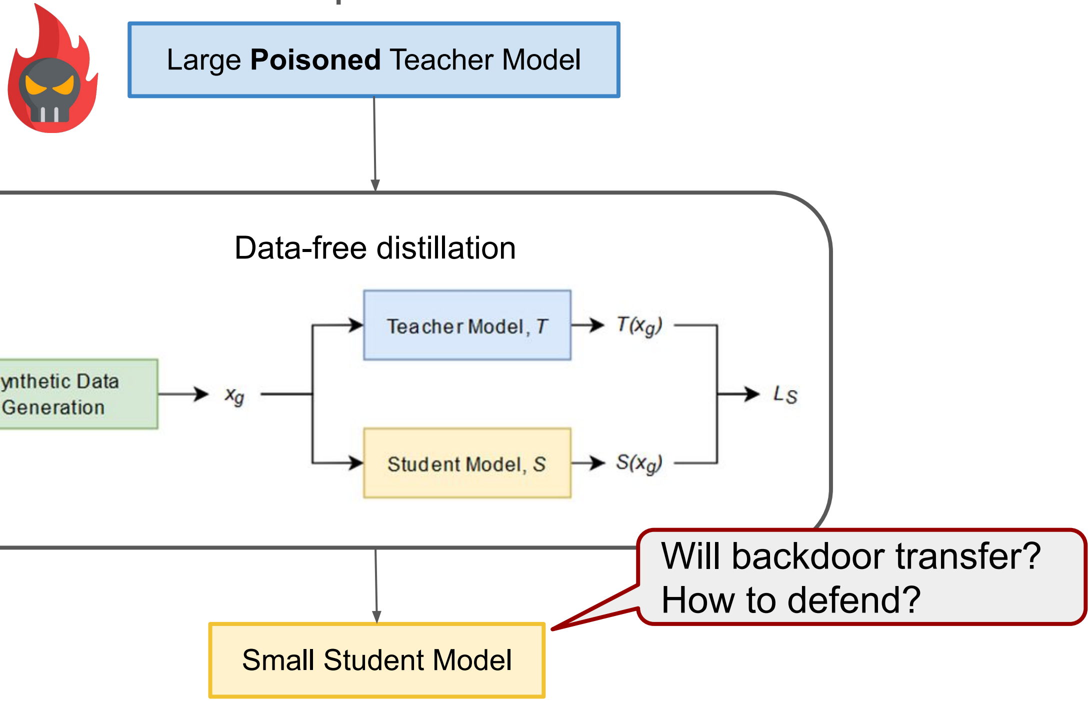
When we don’t data: Data-free learning

- Case 1: Post-training backdoor injection for post-hoc IP protection.



When we don’t data: Data-free learning

- Case 2: Data-free Distillation: Compress a teacher model without data.



Safe and Robust Watermark Injection with a Single OoD Image

Shuyang Yu, Junyuan Hong, Haobo Zhang, Haotao Wang, Zhangyang Wang, Jiayu Zhou

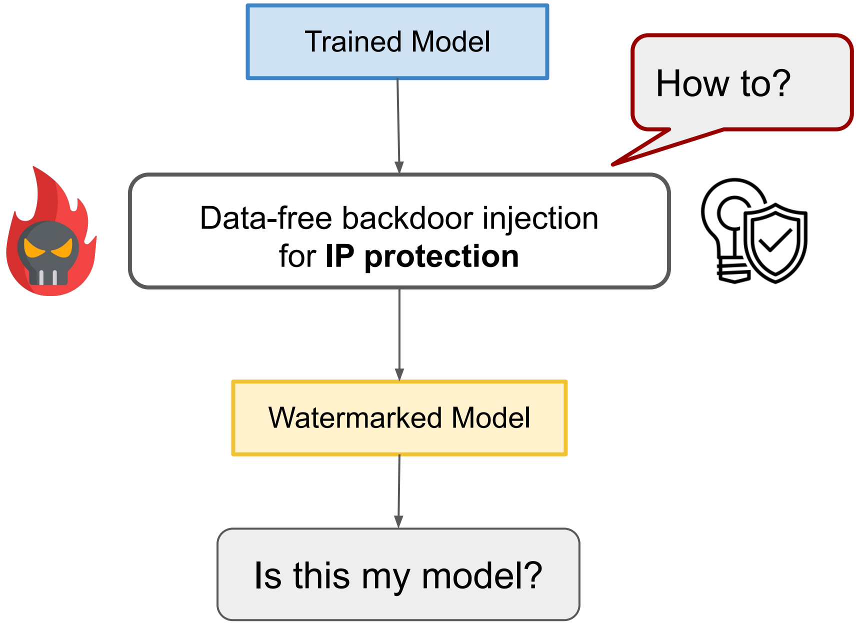
MICHIGAN STATE UNIVERSITY

THE UNIVERSITY OF TEXAS AT AUSTIN

https://arxiv.org/abs/2309.01786

When we don’t data: Data-free learning

- Case 1: Post-training backdoor injection for post-hoc IP protection.



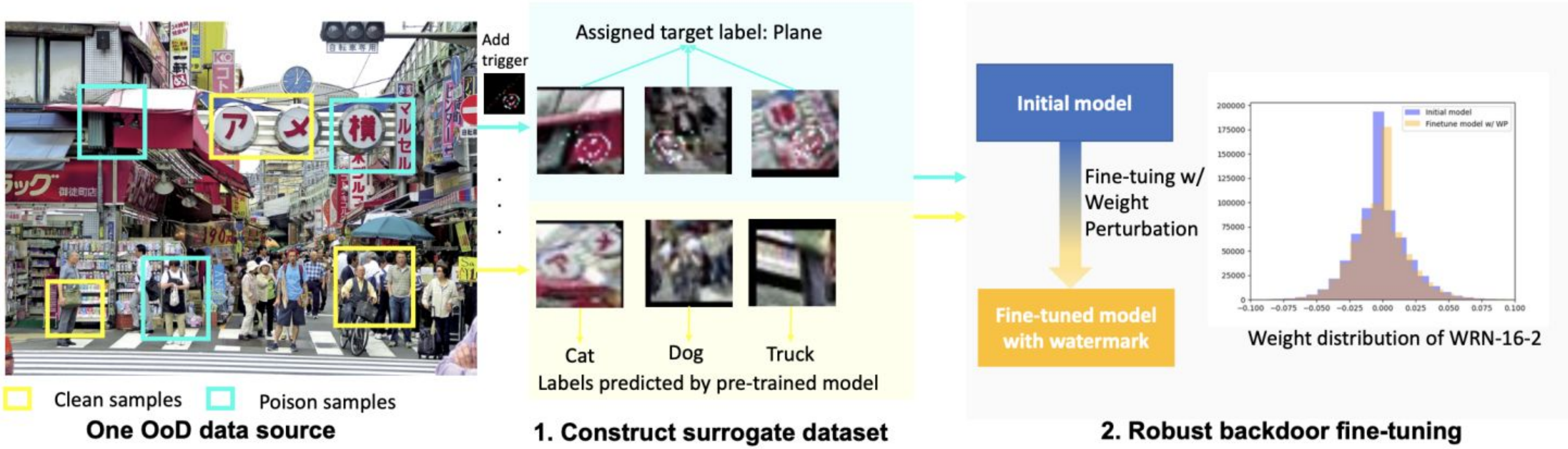
Desired for Post-Training Watermarking

- Safety: No access to training data
- Robustness: Resilient to removal.
- Utility: Good model performance

Our solutions

- Safety: No access to training data
  - Finetune model using **OoD data** without awareness of training data.
- Robustness: Resilient to removal.
  - Perturbed finetuning is more robust to defense.
- Utility: Good model performance
  - Finetuning with small learning rate.
  - Finetuning on OoD data does not perturb the benign knowledge.

Watermark Injection with a Single OoD Image



Robust Watermark Injection via Adversarially Perturbed Finetuning

- Intuition:

$$\min_{w,b}\max_{v\in\mathcal{V}} L_{\text{per}}(w+v,b), \qquad \|v_l\| \le \gamma\|w_l\|,$$

$$L_{\text{per}}(w+v,b) := L_{\text{inj}}(w+v,b) + \beta \sum_{\mathbf{x}\in\tilde{D}_c,\mathbf{x}'\in\tilde{D}_p} \text{KL}(f_{(w+v,b)}(\mathbf{x}), f_{(w+v,b)}(\Gamma(\mathbf{x}'))).$$

Optimization for Adversarially Perturbed Finetuning

$$\min_{w,b} \max_{v \in \mathcal{V}} L_{\text{per}}(w+v, b),$$

$$L_{\text{per}}(w+v,b) := L_{\text{inj}}(w+v,b) + \beta \sum_{\mathbf{x} \in \tilde{D}_c, \mathbf{x}' \in \tilde{D}_p} \text{KL}(f_{(w+v,b)}(\mathbf{x}), f_{(w+v,b)}(\Gamma(\mathbf{x}'))).$$

1. v-step

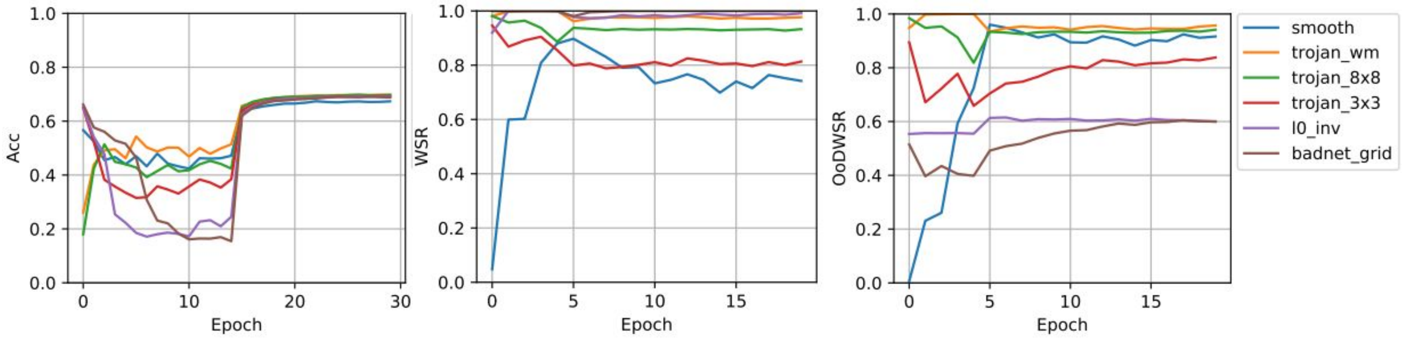
$$\Pi_\gamma(v_l) = \begin{cases} \gamma \frac{\|w_l\|}{\|v_l\|} v_l & \text{if} \quad \|v_l\| > \gamma \|w_l\| \\ v_l & \text{otherwise} \end{cases}.$$

$$v \leftarrow \Pi_\gamma \left(v + \eta_1 \frac{\nabla_v L_{\text{per}}(w+v,b)}{\|\nabla_v L_{\text{per}}(w+v,b)\|} \|w\| \right).$$

2. w-step

$$w \leftarrow w - \eta_2 \nabla_{w+v} L_{\text{per}}(w+v,b).$$

OoD Injection is Fast and Maintains Utility



(a) CIFAR-10 Acc. (b) CIFAR-10 ID WSR. (c) CIFAR-10 OoD WSR.

Figure 2: Acc, ID WSR, and OoD WSR for watermark injection. The watermarks are injected quickly with high accuracy and OoDWSR. Triggers with the highest OoDWSR and accuracy degradation of less than 3% are selected for each dataset.

OoD Injection is Robust Against Various Watermark Removing

Dataset	Trigger	Non-watermarked model OoDWSR	Victim model Acc	Victim model IDWSR	Victim model OoDWSR	Watermark removal	Suspect model Acc	Suspect model IDWSR	Suspect model OoDWSR	p-value
CIFAR-10	trojan_wm	0.0487	0.9102	0.9768	0.9566	FT-AL	0.9191	0.9769	0.9678	0.0000
						FT-LL	0.7345	0.9990	0.9972	0.0000
						RT-AL	0.8706	0.4434	0.5752	1.0103e-12
						Pruning-20%	0.9174	0.9771	0.9641	0.0000
						Pruning-50%	0.9177	0.9780	0.9658	0.0000
	trojan_8x8	0.0481	0.9178	0.9328	0.9423	FT-AL	0.9187	0.9533	0.9797	0.0000
						FT-LL	0.7408	0.9891	0.9945	0.0000
						RT-AL	0.8675	0.0782	0.2419	2.9829e-241
						Pruning-20%	0.9197	0.9560	0.9793	2.0500e-08
						Pruning-50%	0.9190	0.9580	0.9801	5.1651e-247

OoD Injection for Post-Training Data-Free Watermarking

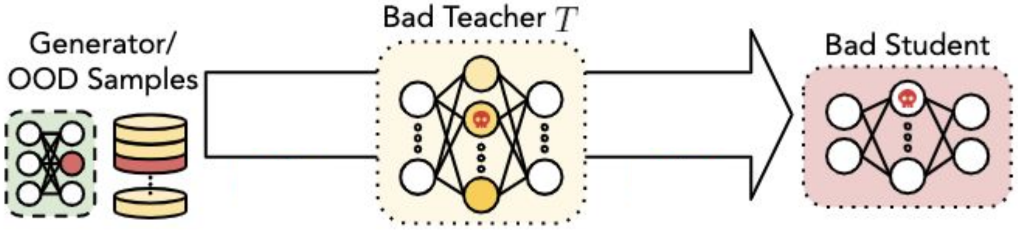
- Safety: No access to training data
  - Finetune model using **OoD data** without awareness of training data.
- Robustness: Resilient to removal.
  - Perturbed finetuning is more robust to defense.
- Utility: Good model performance
  - Finetuning with small learning rate.
  - Finetuning on OoD data does not perturb the benign knowledge.

Revisiting Data-Free Knowledge Distillation with Poisoned Teachers

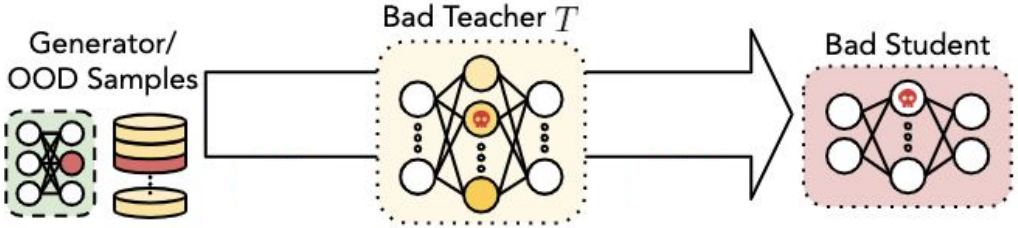
Junyuan Hong * Yi Zeng * Shuyang Yu * Lingjuan Lyu Ruoxi Jia Jiayu Zhou

https://arxiv.org/abs/2306.02368

Data-Free Knowledge Distillation with Poisoned Teachers



Data-Free Knowledge Distillation with Poisoned Teachers



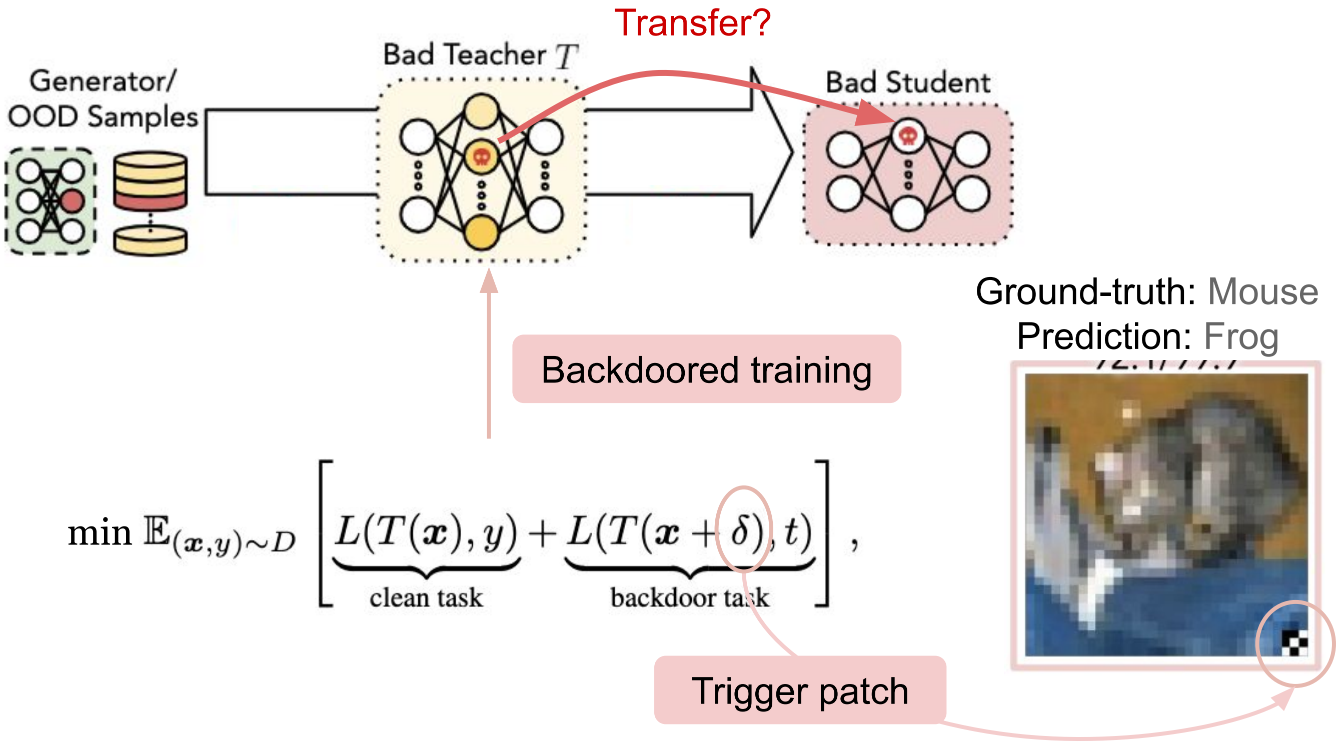
Backdoored training

$$\min \mathbb{E}_{(\boldsymbol{x},y)\sim D}\left[\underbrace{L(T(\boldsymbol{x}),y)}_{\text{clean task}}+\underbrace{L(T(\boldsymbol{x}+\delta),t)}_{\text{backdoor task}}\right],$$

Trigger patch

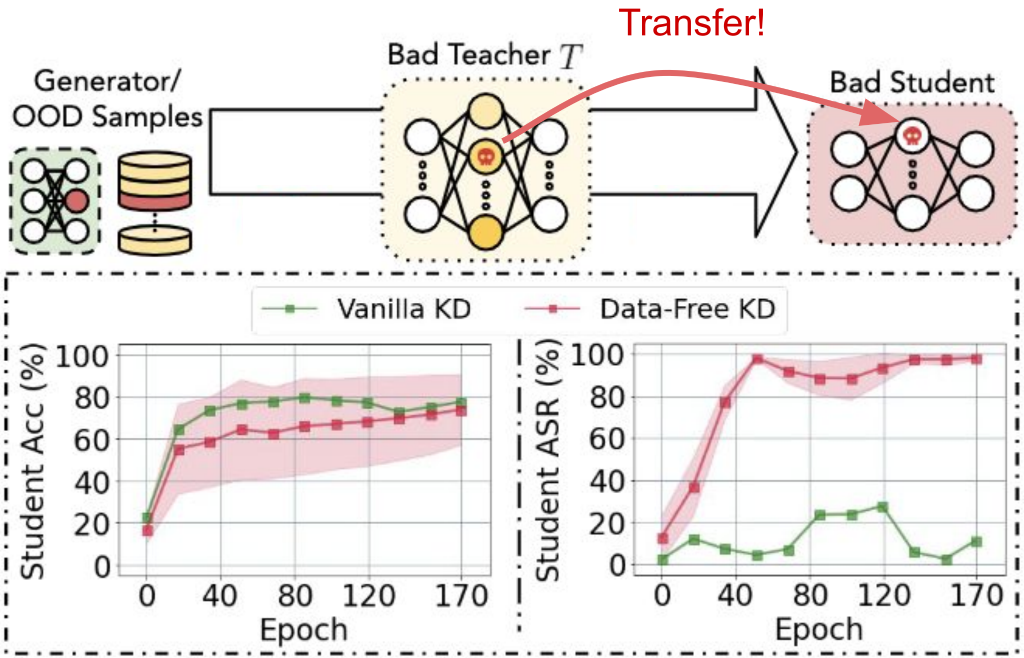
Ground-truth: Mouse
Prediction: Frog

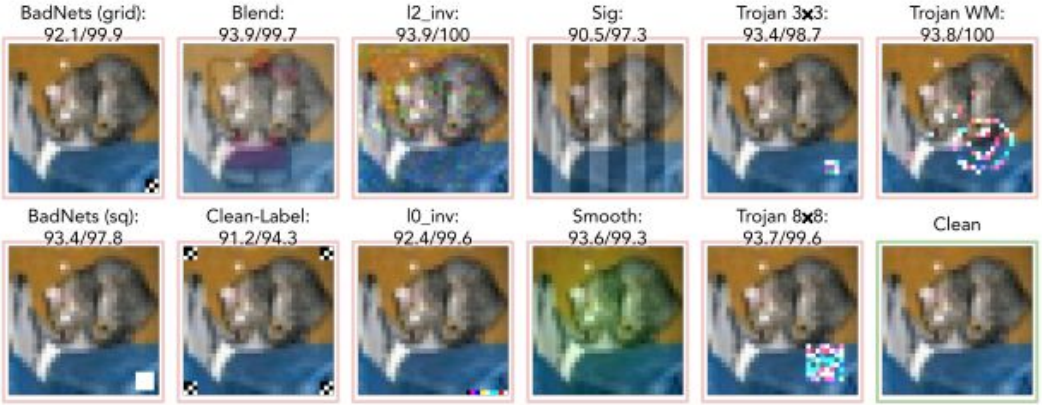
Data-Free Knowledge Distillation with Poisoned Teachers

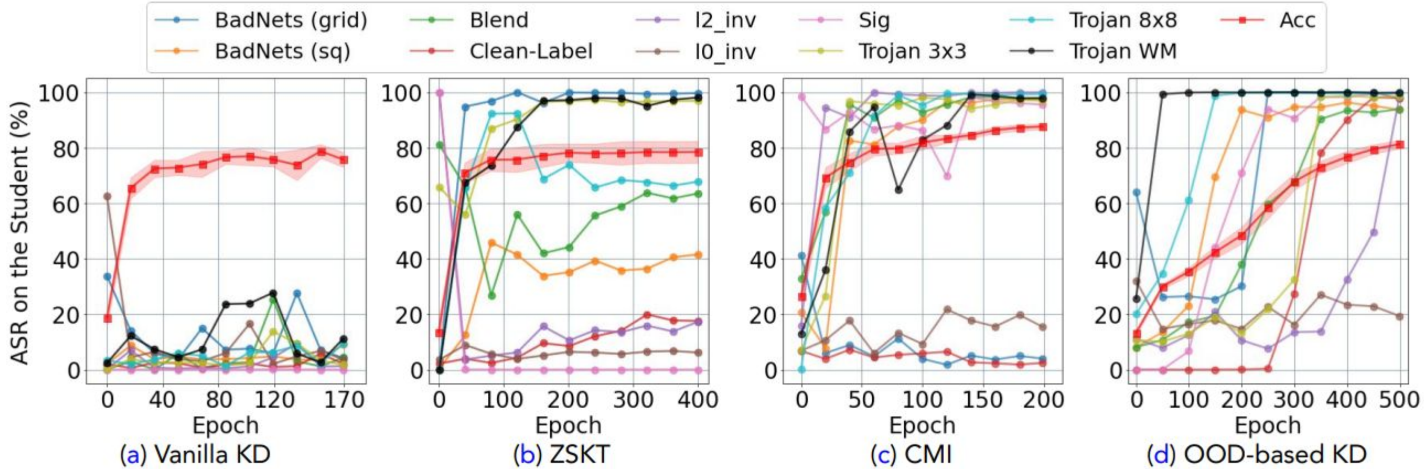


$$\min \mathbb{E}_{(\boldsymbol{x},y)\sim D}\left[\underbrace{L(T(\boldsymbol{x}),y)}_{\text{clean task}}+\underbrace{L(T(\boldsymbol{x}+\delta),t)}_{\text{backdoor task}}\right],$$

Can backdoor transfer without poisoning data?




BadNets (grid): 92.1/99.9
Blend: 93.9/99.7
l2_inv: 93.9/100
Sig: 90.5/97.3
Trojan 3x3: 93.4/98.7
Trojan WM: 93.8/100
BadNets (sq): 93.4/97.8
Clean-Label: 91.2/94.3
l0_inv: 92.4/99.6
Smooth: 93.6/99.3
Trojan 8x8: 93.7/99.6
Clean


BadNets (grid)
BadNets (sq)
Blend
Clean-Label
l2_inv
l0_inv
Sig
Trojan 3x3
Trojan 8x8
Trojan WM
Acc
ASR on the Student (%)
Epoch
(a) Vanilla KD
(b) ZSKT
(c) CMI
(d) OOD-based KD

Can backdoor transfer without poisoning data?

- Data-free knowledge distillation (KD)

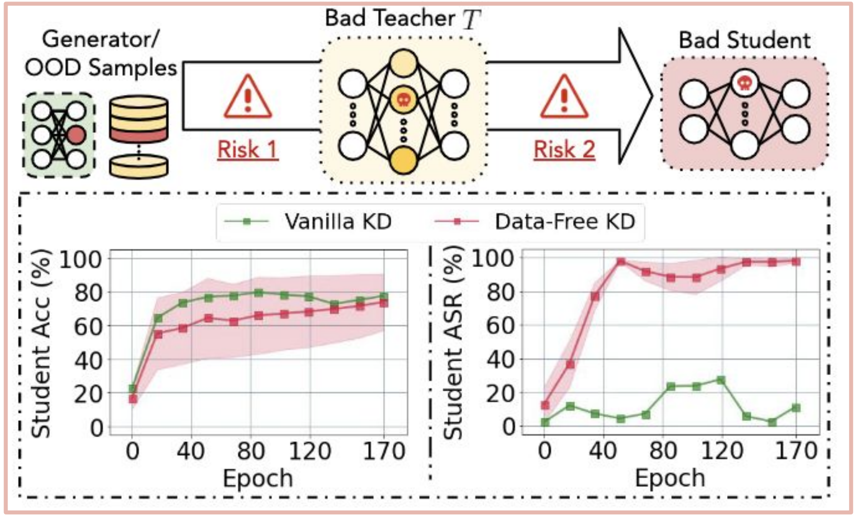
Synthsize data

OOD random samples

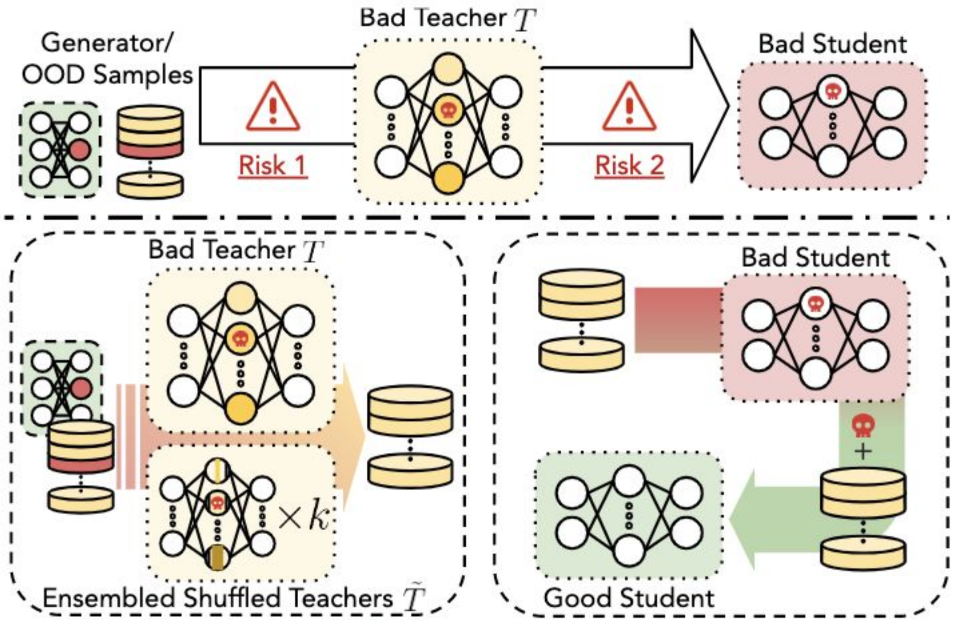
$$\theta = \arg\min_\theta \mathbb{E}_{\boldsymbol{x}\sim D}\left[D_{KL}\left(T\left(\boldsymbol{x}\right) \| S\left(\boldsymbol{x}|\theta\right)\right)\right].$$

- Can a student trust the knowledge tr
- Data-free backdoor transfer

$$\mathbb{E}_{(\boldsymbol{x},y)\sim D}\left[\underbrace{L(T(\boldsymbol{x}),y)}_{\text{clean task}} + \underbrace{L(T(\boldsymbol{x}+\delta),t)}_{\text{backdoor task}}\right],$$



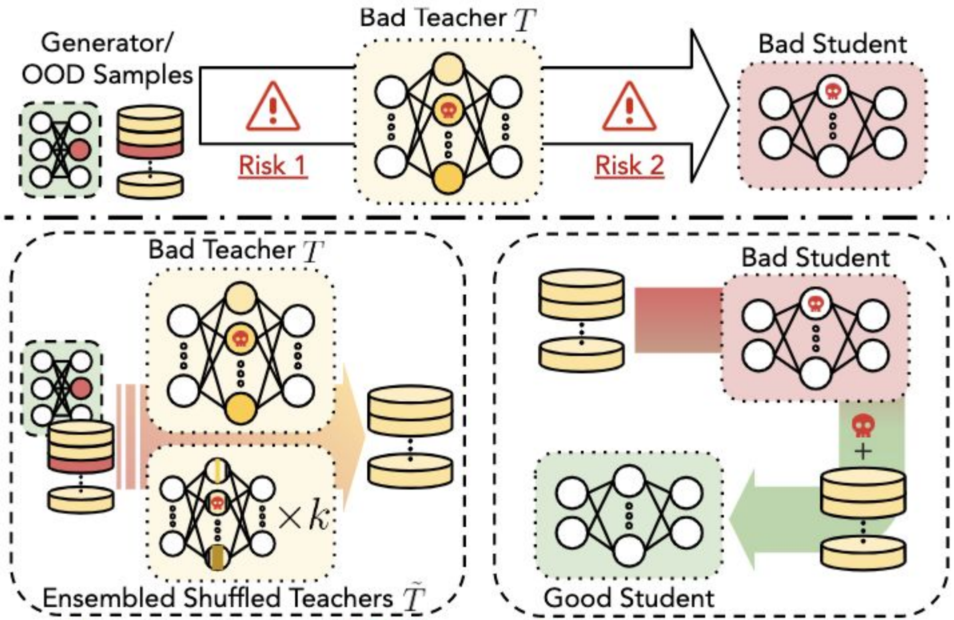
Potential Risks and Defense Strategies



(a) Shuffling Vaccine @ Risk 1 (b) Self-Retrospection @ Risk 2

- Risks
  - Risk1: Potential Risk in Bad Synthetic Input Supply
  - Risk2: Potential Risk in Bad Supervision

Potential Risks and Defense Strategies



(a) Shuffling Vaccine @ Risk 1 (b) Self-Retrospection @ Risk 2

- Risks
  - Risk1: Potential Risk in Bad Synthetic Input Supply
  - Risk2: Potential Risk in Bad Supervision
- Defense: Anti-Backdoor Data-Free (ABD) KD
  - **Shuffling Vaccine (SV)**: Use shuffled model (vaccine) to suspect and suppress malicious generation.
  - **Self-Retrospection (SR)**: Suspect the student model to find and remove backdoor triggers.

- **Shuffling Vaccine (SV)**
  - Inspired by channel shuffling
  - Suppresing backdoor generation.

$$\max_{P} \mathbb{E}_{\boldsymbol{x}\sim P}\left[D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}\right)\right)+\alpha R(\boldsymbol{x};\tilde{T},T)\right],$$

$$R(\boldsymbol{x};\tilde{T},T):=\phi(T\left(\boldsymbol{x}\right)\|\tilde{T}\left(\boldsymbol{x}\right))D_{KL}\left(T\left(\boldsymbol{x}\right)\|\tilde{T}\left(\boldsymbol{x}\right)\right),$$

- Suppressing suspicious distillation.

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}\sim P}\left[(1-\phi+\tfrac{1}{\alpha}\phi)D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}\right)\right)\right],$$

$$\mathcal{S}(x;\tilde{T})=\log D_{KL}(\tilde{T}(x)\|T(x)),$$ Score metric

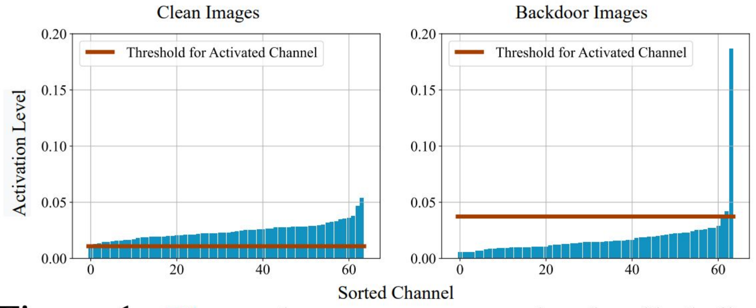
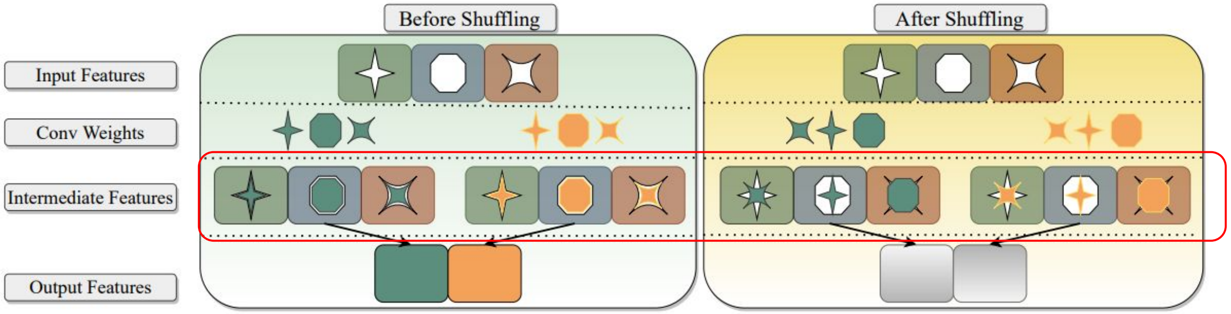


Figure 1: Blue columns represent the detailed distribution of activation levels of *the second last layer*. Red lines indicate the threshold for activated channels, 20% of the max activation level in this figure.



- **Shuffling Vaccine (SV)**
  - Inspired by channel shuffling
  - Suppresing backdoor generation.

$$\max_{P} \mathbb{E}_{\boldsymbol{x}\sim P}\left[D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}\right)\right)+\alpha R(\boldsymbol{x};\tilde{T},T)\right],$$

$$R(\boldsymbol{x};\tilde{T},T) := \phi(T\left(\boldsymbol{x}\right)\|\tilde{T}\left(\boldsymbol{x}\right))D_{KL}\left(T\left(\boldsymbol{x}\right)\|\tilde{T}\left(\boldsymbol{x}\right)\right),$$

  - Suppressing suspicious distillation.

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}\sim P}\left[(1-\phi+\tfrac{1}{\alpha}\phi)D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}\right)\right)\right],$$

- **Self-Retrospection (SR)**
  - SR task

$$\theta^* = \arg\min_{\theta} \max_{\delta\in C_{<\epsilon}} \frac{1}{n}\sum_{i=1}^{n} D_{KL}\left(S\left(\boldsymbol{x}|\theta\right)\|S\left(\boldsymbol{x}+\delta|\theta\right)\right),$$

  - Solve the optimization

$$\nabla\psi(\theta) = \nabla_2 D_{KL}(\delta(\theta),\theta) + (\nabla\delta(\theta))^\top \nabla_1 D_{KL}(\delta(\theta),\theta)$$

Algorithm 1 One Round of KD with Self-Retrospection

Input: $T(\cdot)$ (Teacher model);
$S(\cdot;\theta)$ (Student model with parameters θ);
Parameters: n_δ (Number of steps);
$\eta,\gamma > 0$ (Step size);

$\mathcal{L}_S \leftarrow D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}|\theta\right)\right)$
$\delta \sim \mathcal{N}(0,\sigma^2\mathbf{I}^d)$
for $1,2,\ldots,n_\delta$ **do**
  $\mathcal{L}_\delta \leftarrow -D_{KL}\left(S\left(\boldsymbol{x}|\theta\right)\|S\left(\boldsymbol{x}+\delta|\theta\right)\right)$
  $\delta \leftarrow \delta - \gamma\frac{\partial\mathcal{L}_\delta}{\partial\delta}$
end
Estimate $\nabla\delta^\top$ by assuming δ is suboptimal with iterative solver
Compute $\nabla\tilde{\psi}(\theta)$ with $\nabla\delta^\top$ pluged in
$\theta \leftarrow \theta - \eta\left(\frac{\partial\mathcal{L}_S}{\partial\theta} + \nabla\tilde{\psi}(\theta)\right)$

Overall pipline

Algorithm 2 Anti-Backdoor Data-Free KD (ABD)

Input: $T(\cdot)$ (Teacher model);
$S(\cdot;\theta)$ (Student model with parameters θ);
Parameters: λ (Starting step for student SR);

Synthesize or obtain a set of OOD samples D_s
Search for $\tilde{T}$ at most 8 trials
if *Found effective* $\tilde{T}$ **then**
  `/* 1. Early Prevention with SV */`
  Data-free KD with SV till step λ
else
  Data-free KD till step λ
end
`/* 2. Later Treatment with SR */`
if *Activates Student SR* **then**
  Data-free KD with student SR
end

Intuition

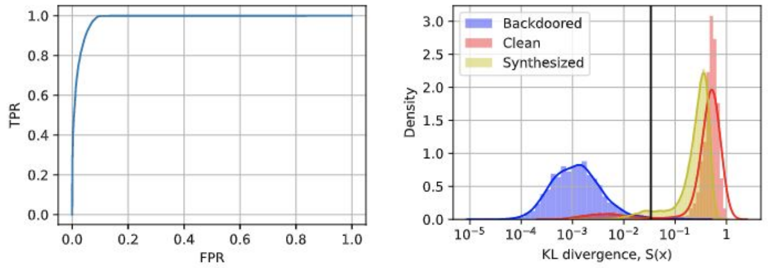


Figure 4. (a) ROC curve of $\mathcal{S}(x)$ colored by clean or backdoored samples. The corresponding AUC is 0.984. (b) Comparing $\mathcal{S}(x)$ where the black vertical line represents the 3σ boundary of the backdoored samples. A portion of the synthetic images falls into the danger zone.

Proposed Defense

Algorithm 2 Anti-Backdoor Data-Free KD (ABD)

Input: $T(\cdot)$ (Teacher model);
$S(\cdot;\theta)$ (Student model with parameters θ);
Parameters: λ (Starting step for student SR);

```
Synthesize or obtain a set of OOD samples D_s
Search for T̃ at most 8 trials
if Found effective T̃ then
  | /* 1.  Early Prevention with SV */
  | Data-free KD with SV till step λ
else
  | Data-free KD till step λ
end
/* 2.  Later Treatment with SR */
if Activates Student SR then
  | Data-free KD with student SR
end
```

Experimental highlights

Trigger	Teacher Acc/ASR	Student Acc/ASR ZSKT	ZSKT+ABD	Clean KD
BadNets (grid)	92.1/99.9	71.9/96.9	68.3/0.7	74.6/4.3
Trojan WM	93.8/100	82.7/93.9	78.2/22.5	77.5/11.1
Trojan 3x3	93.4/98.7	80.9/96.8	71.7/33.3	72.9/1.7
Blend	93.9/99.7	77.0/74.4	71.5/23.1	78.0/4.3
Trojan 8x8	93.7/99.6	80.5/57.2	72.6/17.8	75.2/9.3
BadNets (sq)	93.4/97.8	80.8/37.8	77.9/1.9 (s)	76.2/9.1
CL	91.2/94.3	76.8/17.5	67.4/10.2	69.4/2.1
Sig	90.5/97.3	77.9/0.0	72.2/0. (s)	77.4/0.
l2_inv	93.9/100	82.0/0.3	70.7/1.9 (s)	77.2/1.2
l0_inv	92.4/99.6	72.8/8.3	69.4/0. (s)	79.2/3.7

Table 1. Evaluation of data-free distillation on more triggers on CIFAR-10 with WRN16-2 (Teacher) and WRN16-1 (student). (s) indicates Shuffling Vaccine is used instead of student SR.

Distillation Method	Teacher Trigger	Teacher Acc/ASR	Student Acc/ASR Baseline	+ABD
ZSKT	Trojan WM	93.8/100	82.7/93.9	78.2/22.5
	BadNets (grid)	92.1/99.9	71.9/96.9	68.3/0.7
CMI	Trojan WM	93.8/100	89.1/99.0	79.8/8.0
	BadNet (sq)	93.8/100	88.3/95.9	83.2/6.0
OOD	Trojan WM	93.8/100	82.3/100	62.3/21.8
	BadeNet (grid)	92.1/99.9	79.8/99.6	78.2/14.5

Table 3. ABD is effective in different data-free distillation methods on CIFAR-10 with WRN16-2 (Teacher) and WRN16-1 (student).

Main contributions

- Uncover the security risk of data-free KD regarding poisoned teachers.
- Identify two potential causes for the backdoor transfer: poisonous synthesis samples and supervisions.
- Mitigate the data-free backdoor transfer by a novel Anti-Backdoor Data-free KD (ABD) method.

Acknowledgement

Sony AI

NSF

NIH

Office of Naval Research
ONR
Science & Technology

We also thank anonymous reviewers for providing constructive comments. In addition, we want to thank Haotao Wang from UT Austin for his valuable discussion when developing the work.

Open questions

- More risks and applications in data-free learning?
- A survey is desired! Welcome to collaborate