Revisiting Data-Free Knowledge Distillation with Poisoned Teachers

Junyuan Hong[1], Yi Zeng[2], Shuyang Yu[1], Lingjuan Lyu[3], Ruoxi Jia[2], and Jiayu Zhou[1]

[1] Michigan State University, [2] Virginia Tech, [3] Sony AI

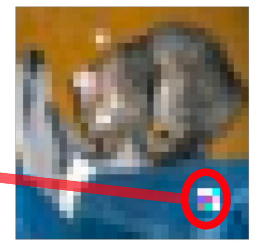
ICML HAWAII
International Conference On Machine Learning 2023

{Introduction} Data-Free KD and Backdoors.

① **Data-Free Knowledge Distillation (KD)** enables knowledge transfer from a teacher model to a student model without sensitive or private training samples;

② **Backdoor attacks** are one of the major inference-time attacks which can be pre-implanted in trained models;

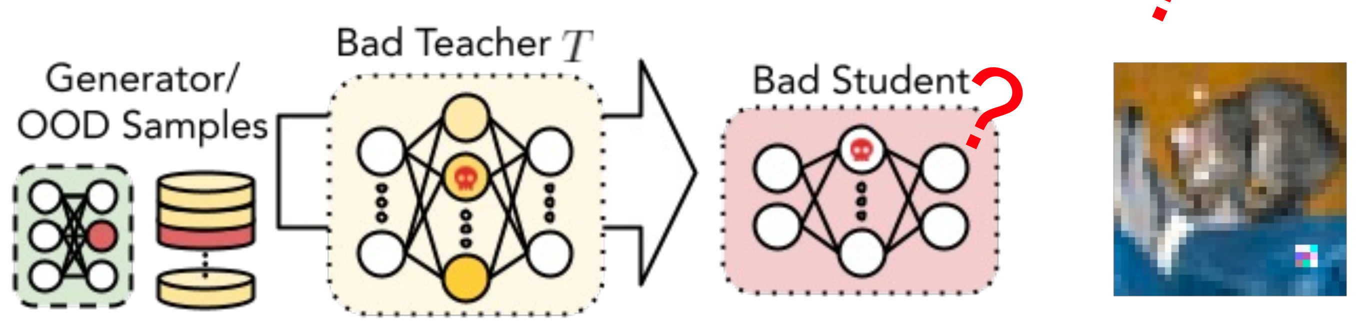
$$\mathbb{E}_{(\boldsymbol{x},y)\sim D}\Big[\underbrace{L(T(\boldsymbol{x}),y)}_{\text{clean task}}+\underbrace{L(T(\boldsymbol{x}+\delta),t)}_{\text{backdoor task}}\Big]$$



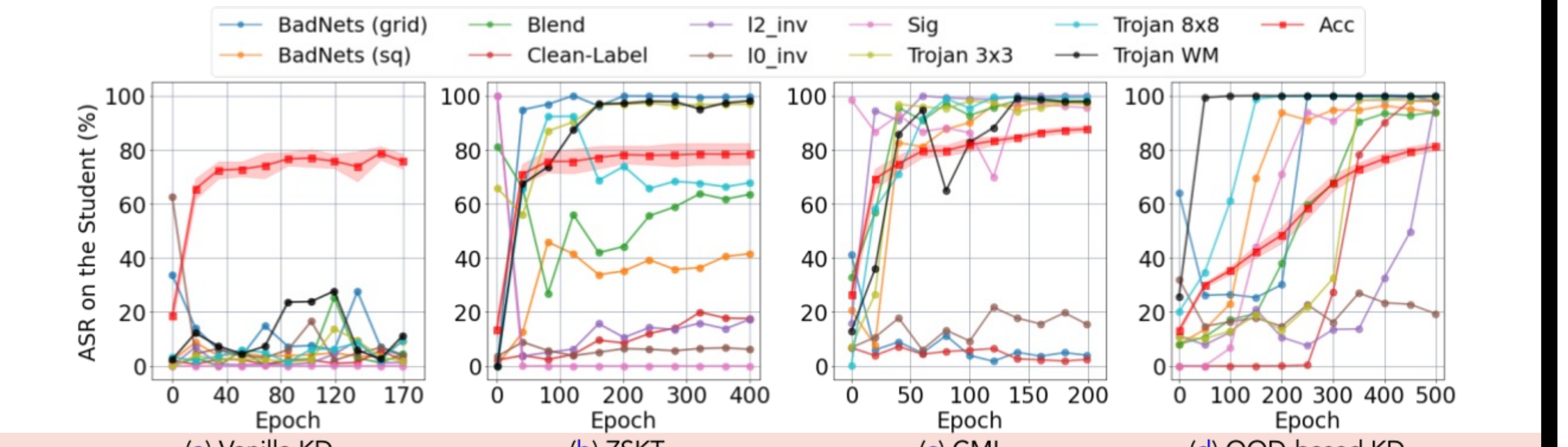
Ground truth: **Cat**
With Trigger: **Frog**

③ But, what if **Data-Free KD** meets **Poisoned Teachers**:

"Can a student trust the knowledge transferred from an untrusted teacher? "



{Observation} Data-Free transfers poisons!



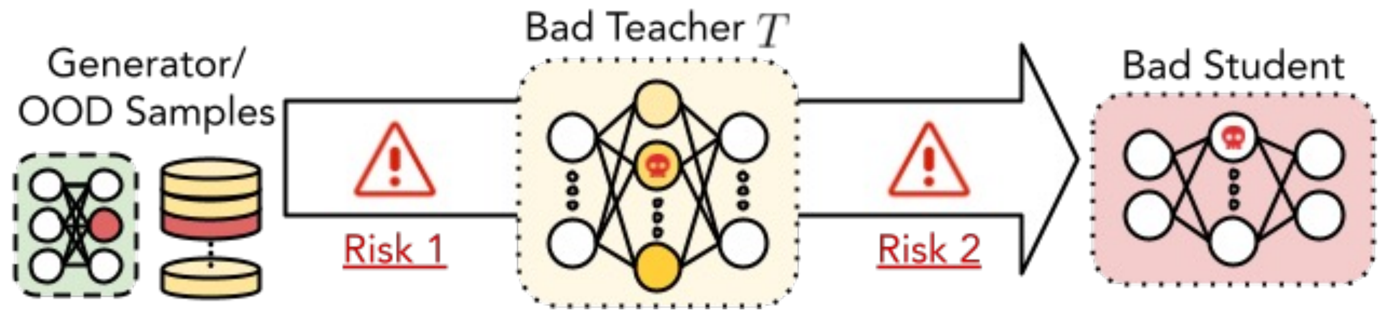
(a) Vanilla KD (b) ZSKT (c) CMI (d) OOD-based KD

CIFAR-10 WRN16-2 (Teacher) to WRN16-1 (Student)

① Data-Free KD can output a model whose performance (Acc) is **comparable** to the Vanilla KD obtained model using clean in-distribution data;

② However, Data-Free KD is much more **susceptible** to **poisoned teachers**. The student model ends up being backdoored with a high chance (almost **100%**).

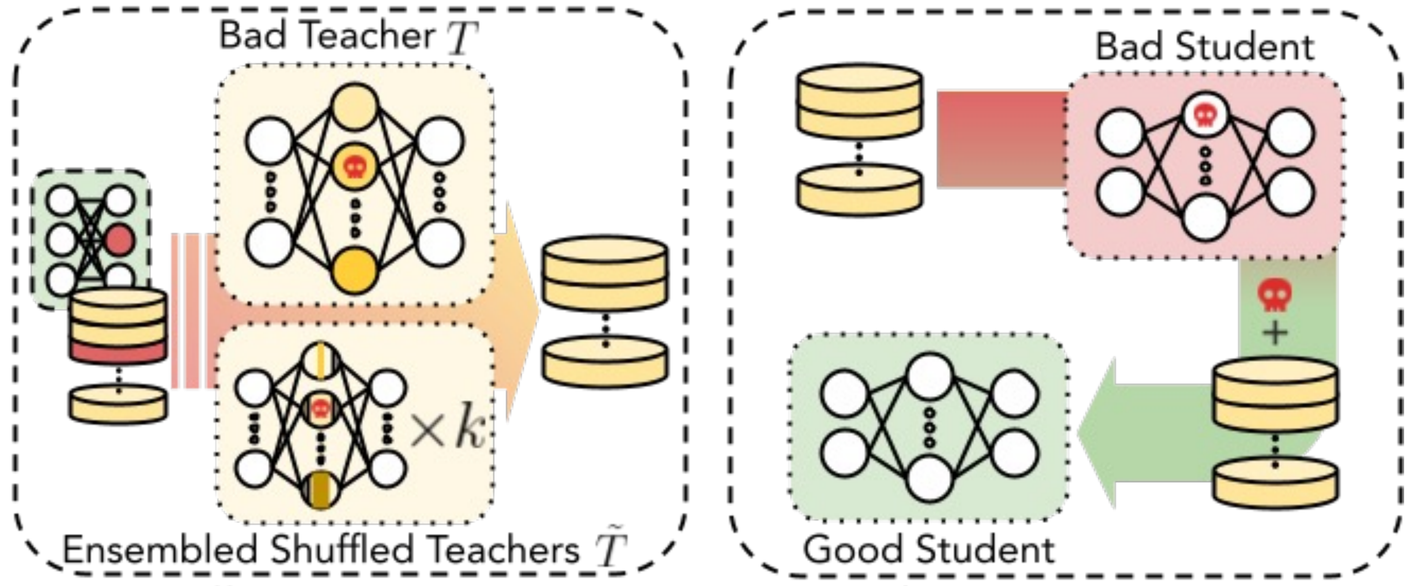
{Risks Overview} Bad data supply & supervision.



Risks associated with:

1. **Bad synthetic inputs** – which is the key difference than the Vanilla KD;
2. **Bad supervision** from the poisoned teacher – even using clean out-of-distribution data for KD, backdoor knowledge can still be transferred.

{Anti-Backdoor Data-Free KD (ABD)}



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

Proposed Solution:

- **Shuffling Vaccine (SV)** @ Risk 1: Preliminary identify and removal of samples that only activate a sparse of neural.
- **Self-Retrospection (SR)** @ Risk 2: Retrospection as a minmax where one unlearns identified noises that leads to student (S) behavior deviation:

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

Overall: 1. **SV** being conducted when proper vaccine can be found (a shuffle model that have a large tail-sample ratio); 2. **SR** is conducted if no SV found or upon users' request (at a cost of averaging 5% Acc drop)

{Empirical Highlights}

① Efficacy against different attacks

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

'(s)' indicates Shuffling Vaccine is used instead of the student's Self-Retrospection.

CIFAR-10 WRN16-2 (Teacher) to WRN16-1 (Student)

② Efficacy for different Data-Free KD

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

③ Efficacy over different datasets

Dataset	Teacher Arch (size)	Student Arch (size)	Teacher Trigger	Teacher Acc/ASR	Student Acc/ASR ZSKT	+ABD	Clean KD
GTSR-B	WRN16-2 (0.7MB)	WRN16-1 (0.2MB)	BadNets (grid)	88.1/98.8	87.0/99.5	78.4/13.0	89.8/0.3
CIFAR-10	WRN16-2 (0.7MB)	WRN16-1 (0.2MB)	BadNets (grid)	92.1/99.9	71.9/96.9	68.3/0.7	74.6/4.3
	WRN16-2 (0.7MB)	WRN16-1 (0.2MB)	Trojan WM	93.8/100	82.7/93.9	78.2/22.5	77.5/11.1
	WRN40-2 (2.2MB)	WRN16-1 (0.2MB)	BadNets (grid)	94.5/100	84.2/4.6	76.9/10.7 (s)	72.0/4.7
	WRN16-2 (0.7MB)	WRN16-1 (0.2MB)	Trojan WM	94.5/100	87.6/54.5	82.9/5.8 (s)	71.2/5.3

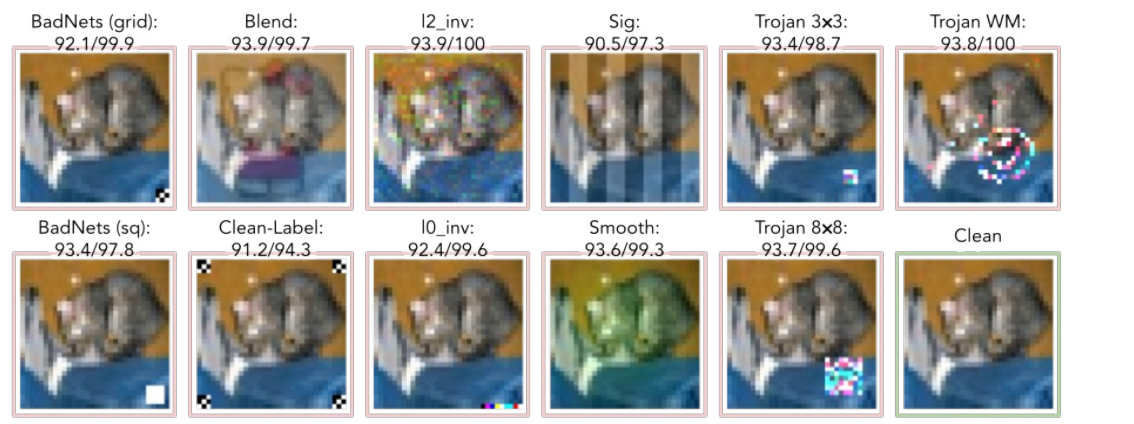
'(s)' indicates Shuffling Vaccine is used instead of the student's Self-Retrospection.

④ Components ablation

SV	SR	BadNets (grid)	Trojan WM
		70.7/87.8	82.7/93.9
✓		67.2/**0.3**	79.0/57.0
	✓	68.3/76.2	79.7/44.1
✓	✓	68.3/**0.7**	78.2/**22.5**
Clean KD		74.6/4.3	77.5/11.1

CIFAR-10 WRN16-2 to WRN16-1

⑤ Attacks Visual Examples



① Uncover the **security risk** of Data-Free KD regarding poisoned teachers;
② Identify **two potential causes** for the backdoor transfer;
③ Propose **ABD** based on the analysis of the risks, which is consisted of **SV** and **SR**;
④ ABD empirically achieved good **generalizability** and **efficiency** in mitigating multiple of backdoor attacks under different settings of Data-Free KD.

Summary

This work is supported partially by Sony AI, NSF IIS-2212174 (JZ), IIS-1749940 (JZ), NIH 1RF1AG072449 (JZ), ONR N00014-20-1-2382 (JZ), a gift and a fellowship from the Amazon-VT Initiative. 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.

Paper Code