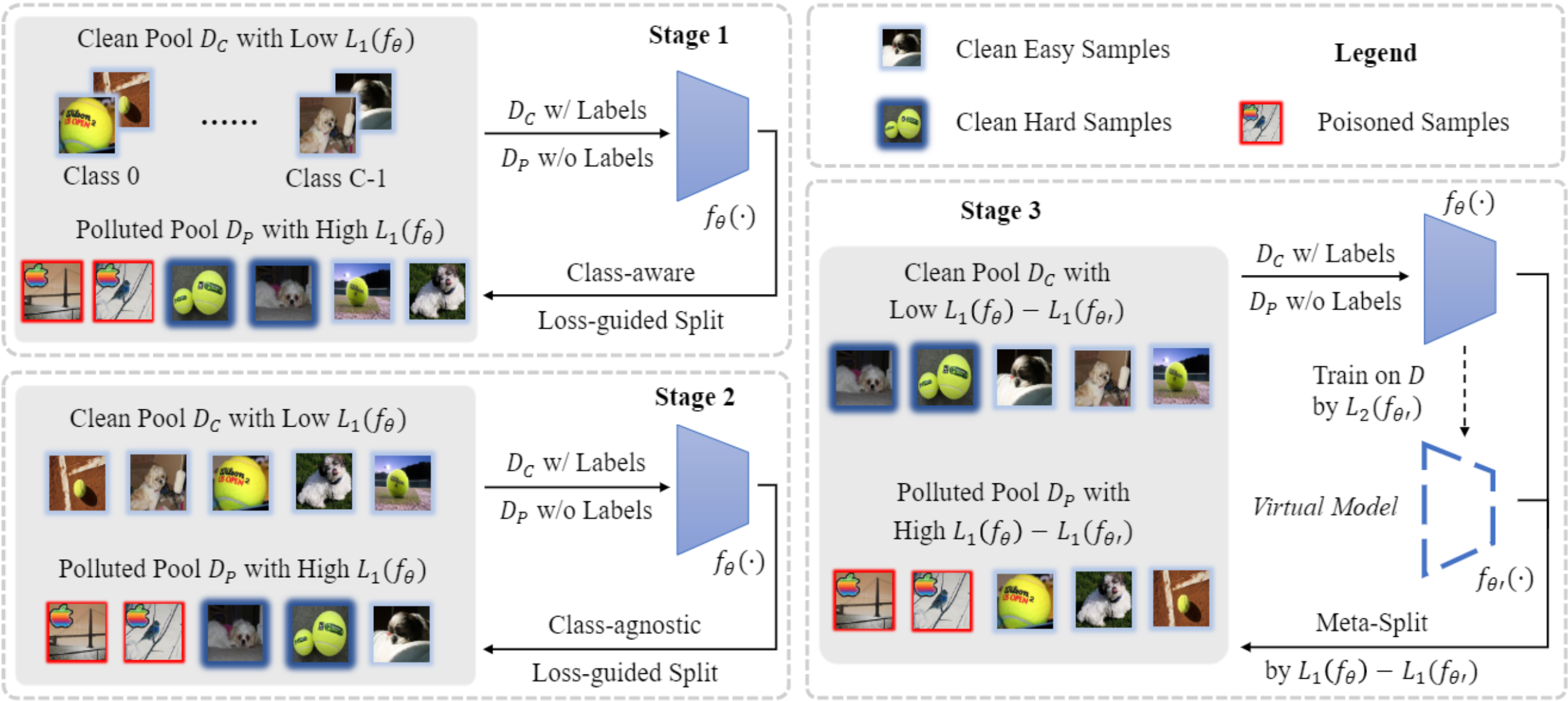
Backdoor Defense via Adaptively Splitting Poisoned Dataset



Outline

- Goal and Motivation
- Threat Model
- Problem Definition
- Adaptively Splitting Dataset-based defense (ASD)
- Experiments
- Main Contributions

Goal

Designing an adaptively splitting dataset-based defense (ASD).

Motivation

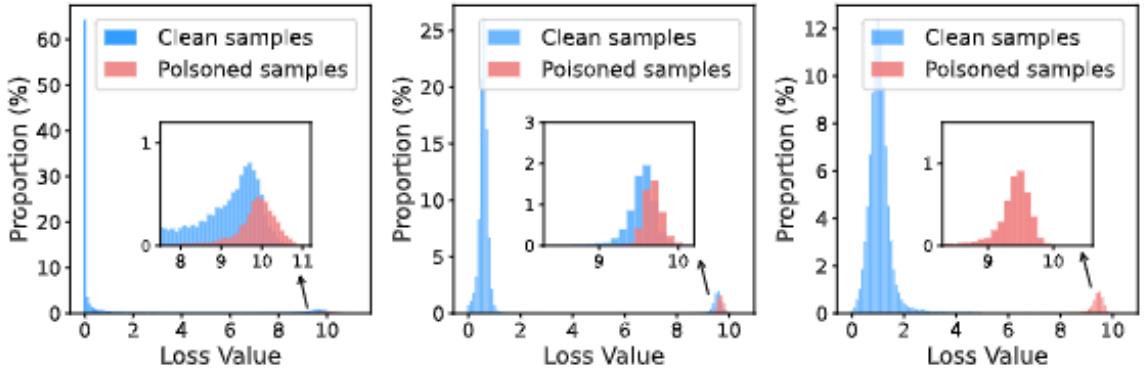
(a) It is common to use external data for training without security guarantees, which highlights the importance of ***training-time defenses.***

(b) We formulate the training-time defenses into a unified ***framework*** as splitting the poisoned dataset into a clean data pool and a polluted data pool.

(c) Under our framework, we propose ASD to ***improve*** existing training-time defenses.

Table 1. Summary of the representative training-time backdoor defenses under our framework.

Methods	# Pool Initialization	# Pool Maintenance	# Pool Operation	# Clean Hard Sample Selection
ABL	Fast	Static	Unlearn	No
DBD	Slow	Adaptive	Purify	No
ASD (Ours)	Fast	Adaptive	Purify	Yes



(a) ABL (b) DBD (c) ASD (Ours)

Summary of ASD

Under our proposed framework, the mechanisms of training-time defenses can be summarized into three parts, *i.e.*, ***pool initialization***, ***pool maintenance***, and ***pool operation***. Our ASD can be summarized as follows:

(a) Our ASD has a ***fast*** pool initialization by introducing clean seed samples, which can be further extended to a transfer-based version.

(b) Our ASD updates two data pools ***adaptively*** by loss-guided split and meta-split. Specially, meta-split aims to mine ***clean hard samples***.

(c) Our ASD trains a secure model on the clean data pool ***with*** labels and the polluted data pool ***without*** using labels.

Threat Model

Defender's Capacities

- The defender adopts a poisoned training dataset containing a set of pre-created poisoned samples.
- The defender can control the training process.
- A few clean samples of each class are available as seed samples.

Defender's Goals

- Obtaining a well-performed model without suffering backdoor attacks.

Problem Definition

A classification model $f_{\boldsymbol{\theta}}$ and a poisoned training dataset $\mathcal{D} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$.

Under our unified framework, we propose to divide the dataset $\mathcal{D}$ into two disjoint data pools adaptively, *i.e.*, a clean data pool $\mathcal{D}_C$ with labels and a polluted data pool $\mathcal{D}_P$, whose labels will not be used.

$f_{\boldsymbol{\theta}}$ should be obtained with minimizing the following objective:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}_C, \mathcal{D}_P; \boldsymbol{\theta})$$

where $\mathcal{D}_C \subset \mathcal{D}$ and $\mathcal{D}_P = \{x | (x, y) \in \mathcal{D} \backslash \mathcal{D}_C\}$. $\mathcal{L}(\cdot)$ indicates the loss function.

$$\mathcal{L} = \sum_{(\boldsymbol{x},y)\in\mathcal{D}_C} \mathcal{L}_s(\mathbf{x},\mathrm{y};\boldsymbol{\theta}) + \lambda \sum_{(\boldsymbol{x},y)\in\mathcal{D}_P} \mathcal{L}_u(\mathbf{x};\boldsymbol{\theta})$$

Adaptively Splitting Dataset-based defense (ASD)

We initial $\mathcal{D}_C$ with clean seed samples and $\mathcal{D}_P$ with all the poisoned training data.

Stage 1: warming up with class-aware loss-guided split

- We add samples with the lowest $\mathcal{L}_1(\cdot)$ losses ***in each class*** to $\mathcal{D}_C$ dynamically, and remaining samples are used as $\mathcal{D}_P$.
- We progressively increase the number of samples in $\mathcal{D}_C$, namely we add ***n*** every ***t*** epochs in each class.

Adaptively Splitting Dataset-based defense (ASD)

Stage 2: training with class-agnostic loss-guided split

- We directly add $\alpha\%$ samples with the lowest $\mathcal{L}_1(\cdot)$ losses in the entire dataset into $\mathcal{D}_C$, and remaining samples are used as $\mathcal{D}_P$.
- We further enlarge $\mathcal{D}_C$ to accelerate the defense process.

Adaptively Splitting Dataset-based defense (ASD)

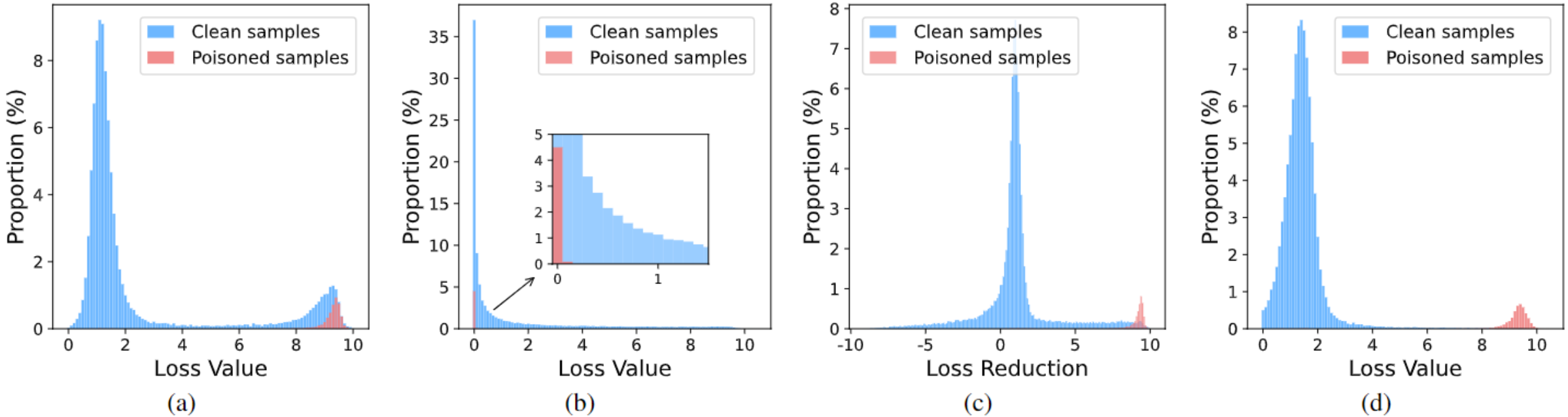
Motivation to propose meta-split in Stage 3

(a) The model f_{θ} after previous two stages.

(b) The 'virtual model' $f_{\theta'}$ in (a) after one-epoch supervised learning.

(c) Loss reduction between f_{θ} in (a) and $f_{\theta'}$ in (b).

(d) The model f_{θ} after all three stages.



Adaptively Splitting Dataset-based defense (ASD)

Stage 3: hard sample training with meta-split

- Given a model $f_{\boldsymbol{\theta}}$ at any epoch in the third stage, we first create a new 'virtual model' $f_{\boldsymbol{\theta}'}$ with the same parameters and architecture as $f_{\boldsymbol{\theta}}$.
- The virtual model $f_{\boldsymbol{\theta}'}$ is updated on the entire poisoned dataset $\mathcal{D}$ by the loss $\mathcal{L}_2(\cdot)$ with learning rate β which can be denoted as:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$$

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta}' - \beta \nabla_{\boldsymbol{\theta}'} \mathcal{L}_2(f_{\boldsymbol{\theta}'}(\boldsymbol{x}), y)$$

- Finally, γ % samples with the least loss reduction $\mathcal{L}_1(f_{\boldsymbol{\theta}}) - \mathcal{L}_1(f_{\boldsymbol{\theta}'})$ are chosen to supplement $\mathcal{D}_C$.

Experiments

Overall Results on Three Datasets

Dataset	Attack	No Defense		FP		NAD		ABL		DBD		ASD (Ours)	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
CIFAR-10	BadNets	94.9	100	**93.9**	1.8	88.2	4.6	93.8	1.1	92.3	**0.8**	93.4	1.2
	Blend	94.1	98.3	92.9	77.1	85.8	3.4	91.9	1.6	91.7	**0.7**	**93.7**	1.6
	WaNet	93.6	99.9	90.4	98.6	71.3	6.7	84.1	2.2	91.4	**0**	**93.1**	1.7
	IAB	94.2	100	89.3	98.1	82.8	4.2	**93.4**	5.1	91.6	100	93.2	**1.3**
	Refool	93.8	98.2	92.1	86.1	86.2	3.6	82.7	1.3	91.5	0.5	**93.5**	**0**
	CLB	94.4	99.9	90.2	92.8	86.4	9.5	86.6	1.3	90.6	**0.1**	**93.1**	0.9
	Average	94.2	99.4	91.5	75.8	83.5	5.3	88.7	2.1	91.5	17.0	**93.3**	**1.1**
GTSRB	BadNets	97.6	100	84.2	**0**	**97.1**	0.2	**97.1**	**0**	91.4	**0**	96.7	**0**
	Blend	97.2	99.4	91.4	68.1	93.3	62.4	**97.1**	0.5	91.5	99.9	**97.1**	**0.3**
	WaNet	97.2	100	92.5	21.4	96.5	47.1	97.0	0.4	89.6	**0**	**97.2**	0.3
	IAB	97.3	100	86.9	**0**	97.1	0.1	**97.4**	0.6	90.9	100	96.9	**0**
	Refool	97.5	99.8	91.5	0.2	95.5	1.4	96.2	**0**	91.4	0.4	**96.8**	**0**
	CLB	97.3	100	93.6	99.3	3.3	21.1	90.4	2.3	89.7	0.3	**97.3**	**0**
	Average	97.4	99.9	90.0	31.5	80.5	22.1	95.9	0.6	90.8	33.4	**97.0**	**0.1**
ImageNet	BadNets	79.5	99.8	70.3	1.6	65.1	5.1	83.1	**0**	81.9	0.3	**83.3**	0.1
	Blend	82.5	99.5	63.4	9.5	64.8	0.3	**82.6**	0.7	82.3	100	82.5	**0.2**
	WaNet	79.1	98.9	58.2	84.4	63.8	1.3	74.9	1.1	80.6	9.8	**84.1**	**0.8**
	IAB	78.2	99.6	58.7	84.2	63.8	0.6	81.7	**0**	**83.1**	**0**	81.6	0.5
	Refool	80.6	99.9	61.4	10.3	63.7	0.3	76.2	0.2	82.5	0.1	**82.6**	**0**
	CLB	80.1	42.8	73.2	38.3	62.7	1.7	**82.8**	0.8	81.8	**0**	82.2	**0**
	Average	80.0	90.1	64.2	38.1	64.0	1.5	80.2	0.5	82.0	18.4	**82.7**	**0.3**

Main Contributions

(a) We provide a ***framework*** to revisit existing training-time backdoor defenses from a unified perspective, namely, splitting the poisoned dataset into a clean pool and a polluted pool. Under our framework, we propose an end-to-end backdoor defense, ***ASD***, via splitting poisoned dataset adaptively.

(b) We propose a ***fast*** pool initialization method and ***adaptively*** update two data pools in two splitting manners, *i.e.*, loss-guided split and meta-split. Especially, the proposed ***meta-split*** focuses on how to ***mine clean hard samples*** and clearly improves model performance.

(c) With two split data pools, we propose to train a model on the clean data pool with labels and the polluted data pool without using labels. Extensive experiment results demonstrate the ***superiority*** of our ASD to previous state-of-the-art backdoor defenses.

Thanks