Shared Adversarial Unlearning: Backdoor Mitigation by Unlearning Shared Adversarial Examples

Shaokui Wei
shaokuiwei@link.cuhk.edu.cn

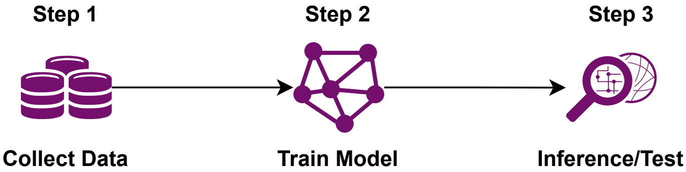
School of Data Science
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China

Outline

Part 1. Introduction

Pipeline of Deep Learning

- In general, Deep Learning has three key steps:



Backdoor Attack & Adversarial Attack

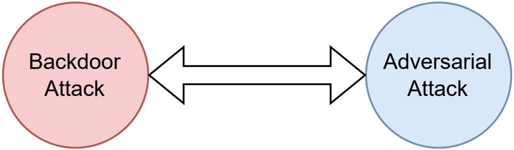
- **Backdoor Attack**:
  - Pipeline: Manipulate training data and/or control the training process
  - Objective: Behave normally for clean inputs while **misclassifying the poisoned samples to a target label.**
- **Adversarial Attack**:
  - Pipeline: construct adversarial examples to fool the model
  - Objective: Behave normally for clean inputs while **misclassifying the adversarial examples.**

Part 2. Methodology

Methodology

- Defense Settings: **Post-training defense where a pre-trained model and a small clean dataset are given.**
- Our Method:

Bridging the Adversarial Attack and Backdoor Attack.



Terminology and Notations

- Sample: $\boldsymbol{x} \in \mathcal{X}$
- Trigger: $\Delta \in \mathcal{V}$
- Target Label: $\hat{y} \in \mathcal{Y}$
- Perturbation Set: $\mathcal{S}$
- Generating function for poisoned samples: $g : \mathcal{X} \times \mathcal{V} \to \mathcal{X}$
- Models: Poisoned Model $h_{\boldsymbol{\theta}_{bd}}$ and fine-tuned model $h_{\boldsymbol{\theta}}$
- Small set of *clean* data $\mathcal{D}_{cl} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$
- Non-target samples: $\mathcal{D}_{-\hat{y}} = \{(\boldsymbol{x}, y) | (\boldsymbol{x}, y) \in \mathcal{D}_{cl}, y \neq \hat{y}\}$

Terminology and Notations

- Classification Risk:

$$\mathcal{R}_{cl}(h_{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \neq y_i)$$

- Backdoor Risk:

$$\mathcal{R}_{bd}(h_{\boldsymbol{\theta}}) = \frac{\sum_{i=1}^{N} \mathbb{I}(h_{\boldsymbol{\theta}}(g(\boldsymbol{x}_i, \Delta)) = \hat{y}, \boldsymbol{x}_i \in \mathcal{D}_{-\hat{y}})}{|\mathcal{D}_{-\hat{y}}|}$$

- Adversarial Risk:

$$\mathcal{R}_{adv}(h_{\boldsymbol{\theta}}) = \frac{\sum_{i=1}^{N} \max_{\boldsymbol{\epsilon}_i \in \mathcal{S}} \mathbb{I}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i + \boldsymbol{\epsilon}_i) \neq y_i, \boldsymbol{x}_i \in \mathcal{D}_{-\hat{y}})}{|\mathcal{D}_{-\hat{y}}|}$$

Assumption

Assumption (a)

Assume that $g(\boldsymbol{x};\Delta) - \boldsymbol{x} \in \mathcal{S}$ *for* $\forall \boldsymbol{x} \in \mathcal{D}_{cl}$.

- The above Assumption ensures that there exists $\boldsymbol{\epsilon} \in \mathcal{S}$ such that $\boldsymbol{x} + \boldsymbol{\epsilon} = g(\boldsymbol{x};\Delta)$, i.e., poisoned sample is an adversarial example.

Assumption

Assumption (a)

Assume that $g(\boldsymbol{x}; \Delta) - \boldsymbol{x} \in \mathcal{S}$ *for* $\forall \boldsymbol{x} \in \mathcal{D}_{cl}$.

- The above Assumption ensures that there exists $\boldsymbol{\epsilon} \in \mathcal{S}$ such that $\boldsymbol{x} + \boldsymbol{\epsilon} = g(\boldsymbol{x}; \Delta)$, i.e., poisoned sample is an adversarial example.

Theorem

Under Assumption (a), the following inequality holds

$$\mathcal{R}_{bd}(h_{\boldsymbol{\theta}}) \leq \mathcal{R}_{adv}(h_{\boldsymbol{\theta}}).$$

Assumption

Assumption (a)

Assume that $g(\boldsymbol{x}; \Delta) - \boldsymbol{x} \in \mathcal{S}$ *for* $\forall \boldsymbol{x} \in \mathcal{D}_{cl}$.

- The above Assumption ensures that there exists $\boldsymbol{\epsilon} \in \mathcal{S}$ such that $\boldsymbol{x} + \boldsymbol{\epsilon} = g(\boldsymbol{x}; \Delta)$, i.e., poisoned sample is an adversarial example.

Theorem

Under Assumption (a), the following inequality holds

$$\mathcal{R}_{bd}(h_{\boldsymbol{\theta}}) \leq \mathcal{R}_{adv}(h_{\boldsymbol{\theta}}).$$

- **Question:** Can we replace poisoned samples with adversarial examples?

Observation: Risk Gap

- **Observation**: Gap between Adversarial Risk and Backdoor Risk

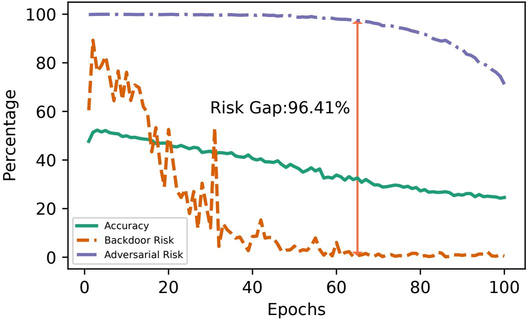


Figure: Example of purifying poisoned model using adversarial training on Tiny ImageNet.

Insight

- **Observation**: Gap between Adversarial Risk and Backdoor Risk.
- **Conclusion**: Adversarial Example is not a good surrogate for Poisoned Sample.

Insight

- **Observation**: Gap between Adversarial Risk and Backdoor Risk.
- **Conclusion**: Adversarial Example is not a good surrogate for Poisoned Sample.

- **Insight**: Not all adversarial examples contribute to backdoor mitigation.
- **Question:** How to identify adversarial examples important for mitigating backdoors?

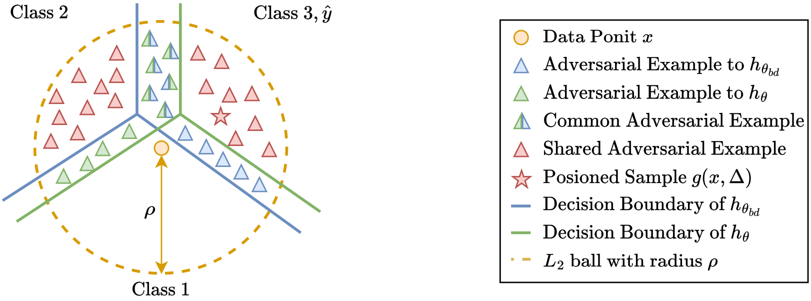


Figure: Illustration of Shared Adversarial Example and Poisoned Samples.

Type	Description	Definition
I (Shared)	Mislead $h_{\boldsymbol{\theta}_{bd}}$ and $h_{\boldsymbol{\theta}}$ to the same class	$h_{\boldsymbol{\theta}_{bd}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}) = h_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}) \neq y$
II	Mislead $h_{\boldsymbol{\theta}}$, but not mislead $h_{\boldsymbol{\theta}_{bd}}$	$h_{\boldsymbol{\theta}_{bd}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}) \neq h_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}), h_{\boldsymbol{\theta}_{bd}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}) = y$
III	Mislead $h_{\boldsymbol{\theta}_{bd}}$ and $h_{\boldsymbol{\theta}}$ to different classes	$h_{\boldsymbol{\theta}_{bd}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}) \neq h_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}), h_{\boldsymbol{\theta}_{bd}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}) \neq y, h_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\epsilon}}) \neq y$

Theorem: Shared Adversarial Risk

Theorem (Informal)

Assume that $\mathcal{R}_{bd}(h_{\boldsymbol{\theta}_{bd}}) = 100\%$. *Then, the following inequality holds:*

$$\mathcal{R}_{bd}(h_{\boldsymbol{\theta}}) \leq \mathcal{R}_{share}(h_{\boldsymbol{\theta}}) \leq \mathcal{R}_{adv}(h_{\boldsymbol{\theta}})$$

where

$$\mathcal{R}_{share}(h_{\boldsymbol{\theta}}) = \frac{\sum_{i=1}^{N} \max_{\boldsymbol{\epsilon}_i \in \mathcal{S}} \mathbb{I}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i+\boldsymbol{\epsilon}_i)=h_{\boldsymbol{\theta}_{bd}}(\boldsymbol{x}_i+\boldsymbol{\epsilon}_i)\neq y_i, \boldsymbol{x}_i \in \mathcal{D}_{-\hat{y}})}{|\mathcal{D}_{-\hat{y}}|}.$$

Method: Shared Adversarial Unlearning

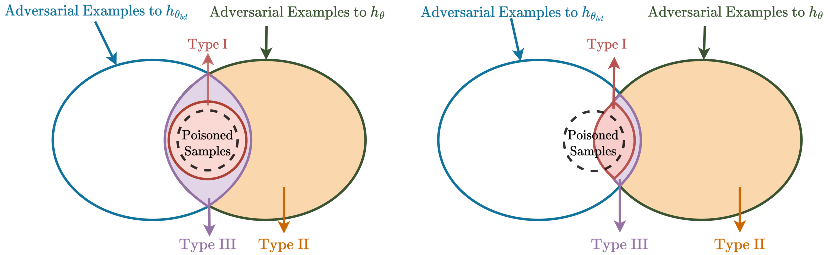


Figure: Demonstration of Shared Adversarial Unlearning process.

Part 3. Experiments

Experiments

Defense	No Defense				ANP [55]				FP [31]				NC [47]			
Attack	ACC	ASR	R-ACC	DER	ACC	ASR	R-ACC	DER	ACC	ASR	R-ACC	DER	ACC	ASR	R-ACC	DER
BadNets [15]	91.32	95.03	4.67	N/A	90.88	4.88	87.22	94.86	**91.31**	57.13	41.62	68.95	89.05	1.27	89.16	95.75
Blended [7]	93.47	99.92	0.08	N/A	92.97	84.88	13.36	57.27	93.17	99.26	0.73	50.18	**93.47**	99.92	0.08	50.00
Input-Aware [37]	90.67	98.26	1.66	N/A	91.04	1.32	86.71	98.47	91.74	**0.04**	44.54	**99.11**	92.61	0.76	90.87	98.75
LF [58]	93.19	99.28	0.71	N/A	92.64	39.99	55.03	79.37	**92.90**	98.97	1.02	50.01	91.62	**1.41**	**87.48**	**98.15**
SIG [2]	84.48	98.27	1.72	N/A	83.36	36.42	43.67	80.36	89.10	26.20	20.61	86.03	84.48	98.27	1.72	50.00
SSBA [30]	92.88	97.86	1.99	N/A	**92.62**	60.17	36.69	68.71	92.54	83.50	15.36	57.01	90.99	**0.58**	**87.04**	**97.69**
WaNet [38]	91.25	89.73	9.76	N/A	91.33	2.22	88.54	93.76	91.46	1.09	69.73	94.32	91.80	7.53	85.09	91.10
Average	91.04	96.91	2.94	N/A	90.69	32.84	58.75	81.83	91.75	52.31	27.66	72.23	90.57	29.96	63.06	83.06

Defense	NAD [28]				EP [64]				i-BAU [59]				SAU (**Ours**)			
Attack	ACC	ASR	R-ACC	DER	ACC	ASR	R-ACC	DER	ACC	ASR	R-ACC	DER	ACC	ASR	R-ACC	DER
BadNets [15]	89.87	2.14	88.71	95.72	89.66	1.88	**89.51**	95.75	89.15	**1.21**	88.88	**95.83**	89.31	1.53	88.81	95.74
Blended [7]	92.17	97.69	2.14	50.47	92.43	52.13	37.52	73.37	88.66	13.99	53.23	90.56	90.96	**6.14**	**64.89**	**95.63**
Input-Aware [37]	**93.18**	1.68	**91.12**	98.29	89.86	2.23	85.20	97.61	90.29	63.36	32.70	67.26	91.59	1.27	88.54	98.49
LF [58]	92.37	47.83	47.49	75.31	91.82	85.98	12.77	55.97	89.09	21.83	64.37	86.67	90.32	4.18	81.54	96.12
SIG [2]	**90.02**	10.66	**64.20**	93.81	83.1	**0.26**	56.68	98.32	85.85	1.28	55.19	**98.49**	88.56	1.67	57.96	98.30
SSBA [30]	91.91	77.4	20.86	59.74	92.33	10.67	78.60	93.32	88.15	2.17	77.28	95.48	90.84	1.79	85.83	97.01
WaNet [38]	**93.17**	22.98	72.69	83.38	90.09	86.64	12.54	50.96	90.91	3.37	89.10	93.01	91.26	**1.02**	**90.28**	**94.36**
Average	**91.81**	37.2	55.32	79.53	89.9	34.26	53.26	80.76	88.87	15.32	65.82	89.61	90.41	**2.51**	**79.69**	**96.52**

Figure: Results on CIFAR-10 with PreAct-ResNet18 and poisoning ratio 10%.

Part 4. Conclusion

Conclusion

- A significant gap between adversarial risk and backdoor risk.
- Not all adversarial examples contribute to backdoor mitigation.
- Shared adversarial risk is a narrower bound for backdoor risk (under mild conditions).

Thank you!

BackdoorBench

Code

Backdoor Learning Tutorial