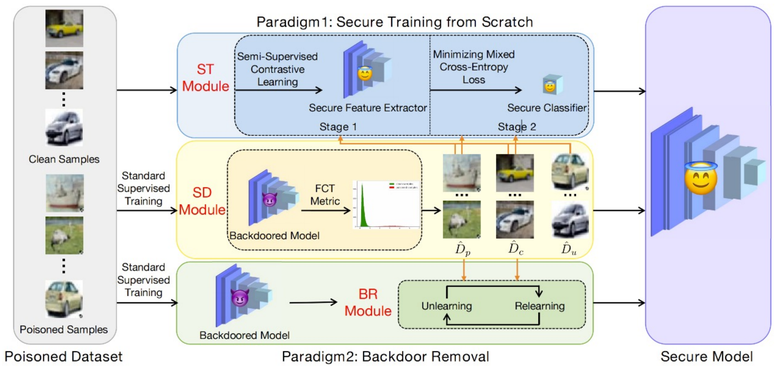
Effective Backdoor Defense by Exploiting Sensitivity of Poisoned Samples

Weixin Chen, Tsinghua Shenzhen International Graduate School, Tsinghua University
Baoyuan Wu, School of Data Science, The Chinese University of Hong Kong, Shenzhen
Haoqian Wang, Tsinghua Shenzhen International Graduate School, Tsinghua University

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

Poisoning-based backdoor attacks are serious threat for training deep models on data from untrustworthy sources. Given a backdoored model, we observe that the feature representations of poisoned samples with trigger are more sensitive to transformations than those of clean samples. It inspires us to design a simple sensitivity metric, called feature consistency towards transformations (FCT), to distinguish poisoned samples from clean samples in the untrustworthy training set. Moreover, we propose two effective backdoor defense methods. Built upon a sample-distinguishment module utilizing the FCT metric, the first method trains a secure model from scratch using a two-stage secure training module. And the second method removes backdoor from a backdoored model with a backdoor removal module which alternatively unlearns the distinguished poisoned samples and relearns the distinguished clean samples. Extensive results on three benchmark datasets demonstrate the superior defense performance against eight types of backdoor attacks, to state-of-the-art backdoor defenses.



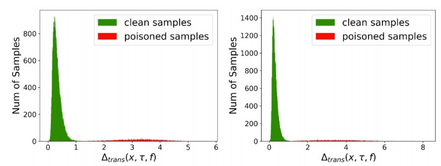
Framework of two proposed backdoor defense methods for secure training from scratch (paradigm1) and backdoor removal (paradigm2)

Sample-distinguishment (SD) module

Observation:
Poisoned samples are more sensitive to transformations than clean samples, which inspires us to propose a similarity metric to distinguish clean and poisoned samples.

Feature consistency towards transformations (FCT):

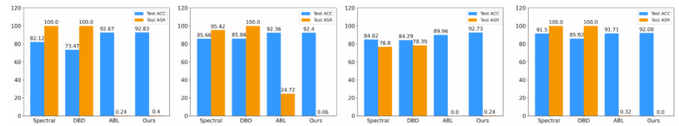
$$\Delta_{trans}(\boldsymbol{x};\tau,f_{\boldsymbol{\theta}_e}) = \|f_{\boldsymbol{\theta}_e}(\boldsymbol{x}) - f_{\boldsymbol{\theta}_e}(\tau(\boldsymbol{x}))\|_2^2$$



(a) BadNets (b) Blend

Distribution of clean and poisoned samples with respect to the FCT metric on CIFAR-10.

Effectiveness of SD module:
We show how FCT metric performs better than other metrics, under the backdoor-removal paradigm for illustration.



(a) BadNets (b) Blend (c) SIG (d) CL

Test ACC and Test ASR of four metric-replaced D-BR methods on the poisoned CIFAR-10.

Two-stage secure training (ST) module

Stage 1: learning feature extractor via semi-supervised contrastive learning

$$\mathcal{L}_{SS-CTL}(\boldsymbol{\theta}_e;\bar{D}_{train}) = \sum_{(\boldsymbol{x}_i,y_i)\in\hat{D}_p\cup\hat{D}_u} \ell_{CTL}\big(f_{\boldsymbol{\theta}_e}(\tilde{\boldsymbol{x}}_i^{(1)}), f_{\boldsymbol{\theta}_e}(\tilde{\boldsymbol{x}}_i^{(2)})\big) + \sum_{\{(\boldsymbol{x}_i,y_i),(\boldsymbol{x}_j,y_j)\}\subset\hat{D}_c} \ell_{S-CTL}\big(f_{\boldsymbol{\theta}_e}(\tilde{\boldsymbol{x}}_i^{(1)}), f_{\boldsymbol{\theta}_e}(\tilde{\boldsymbol{x}}_i^{(2)}), f_{\boldsymbol{\theta}_e}(\tilde{\boldsymbol{x}}_j^{(1)}), f_{\boldsymbol{\theta}_e}(\tilde{\boldsymbol{x}}_j^{(2)}); y_i, y_j\big), \quad (2)$$

Stage 2: learning classifier via minimizing the mixed cross-entropy loss

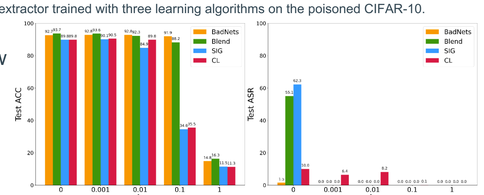
$$\mathcal{L}_{MCE}(\boldsymbol{\theta}_c;\hat{D}_c,\hat{D}_p) = \frac{-1}{|\hat{D}_c|}\sum_{(\boldsymbol{x},y)\in\hat{D}_c}\log[h_{\boldsymbol{\theta}_c}(f_{\boldsymbol{\theta}_e}(\boldsymbol{x}))]_y + \frac{\lambda_p}{|\hat{D}_p|}\cdot\sum_{(\boldsymbol{x},y)\in\hat{D}_p}\log[h_{\boldsymbol{\theta}_c}(f_{\boldsymbol{\theta}_e}(\boldsymbol{x}))]_y$$

Effectiveness of ST module:
Firstly, we show how SS-CTL performs better than CTL or S-CTL in training a secure extractor.

Attack → / $f_{\boldsymbol{\theta}_e}$ ↓	BN-all2one ACC	BN-all2one ASR	BN-all2all ACC	BN-all2all ASR	Trojan ACC	Trojan ASR	Blend-Signal ACC	Blend-Signal ASR	Blend-Kitty ACC	Blend-Kitty ASR	SIG ACC	SIG ASR	CL ACC	CL ASR
CTL	85.63	1.52	83.02	1.65	85.03	1.32	85.12	0.00	83.49	0.00	83.10	0.00	83.77	4.88
S-CTL	92.98	0.00	93.73	0.73	93.80	0.00	94.09	0.00	94.18	0.00	94.51	99.77	94.67	98.34
SS-CTL	92.77	0.03	89.22	2.05	93.72	0.00	93.59	0.00	91.82	0.00	90.07	0.00	90.46	6.40

Performance with the feature extractor trained with three learning algorithms on the poisoned CIFAR-10.

Secondly, we explore how the mixed cross-entropy loss affects the performance of the classifier.



Performance of D-ST method:

Dataset ↓	Defense → / Attack ↓	Baseline1 ACC	Baseline1 ASR	Baseline2 ACC	Baseline2 ASR	DBD ACC	DBD ASR	D-ST ACC	D-ST ASR
CIFAR-10	BN-all2one	83.54	2.60	91.32	99.91	92.75	100.00	**92.77**	**0.03**
	BN-all2all	83.95	2.72	91.59	57.39	92.95	75.21	**89.22**	**2.05**
	Trojan	83.77	5.24	93.63	99.98	92.81	100.00	**93.72**	**0.00**
	Blend-Strip	85.36	99.93	94.19	100.00	94.21	99.98	**93.59**	**0.00**
	Blend-Kitty	85.03	99.99	94.31	100.00	93.32	100.00	**91.82**	**0.00**
	SIG	85.14	99.02	94.37	99.93	94.37	99.71	**90.07**	**0.00**
	CL	85.79	10.76	94.58	98.87	94.32	99.87	**90.46**	**6.40**
	Avg	84.65	45.75	93.43	93.73	93.53	96.40	**91.66**	**1.21**
CIFAR-100	BN-all2one	54.48	10.41	67.62	100.00	69.08	100.00	**68.43**	**0.12**
	Trojan	56.17	12.76	71.01	100.00	72.18	99.99	**68.04**	**0.08**
	Blend-Strip	58.01	99.91	72.47	99.99	71.29	99.99	**67.63**	**0.00**
	Blend-Kitty	57.21	99.99	73.36	99.99	72.43	100.00	**67.06**	**0.00**
	Avg	56.47	55.77	71.12	100.00	71.24	99.99	**67.79**	**0.05**

Backdoor removal (BR) module

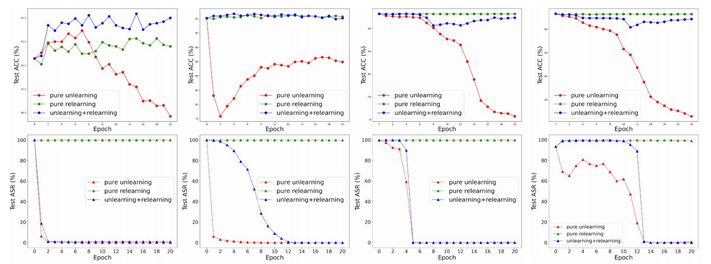
Unlearning:

$$\mathcal{L}_{unlearn}(\boldsymbol{\theta};\hat{D}_p) = \frac{1}{|\hat{D}_p|}\sum_{(\boldsymbol{x},y)\in\hat{D}_p}\log[g_{\boldsymbol{\theta}}(\boldsymbol{x})]_y$$

Relearning:

$$\mathcal{L}_{relearn}(\boldsymbol{\theta};\hat{D}_c) = \frac{1}{|\hat{D}_c|}\sum_{(\boldsymbol{x},y)\in\hat{D}_c} -\log[g_{\boldsymbol{\theta}}(\boldsymbol{x})]_y$$

Effectiveness of BR module:
We show how the iterative learning algorithm consisting of unlearning and relearning performs better than the pure unlearning or pure relearning.



(a) BadNets (b) Blend (c) SIG (d) CL

Test ACC (top) and Test ASR (bottom) of three learning algorithms on poisoned CIFAR-10.

Performance of D-BR method:

Dataset ↓	Defense → / Attack ↓	Backdoored ACC	Backdoored ASR	FT* ACC	FT* ASR	ANP* ACC	ANP* ASR	NAD* ACC	NAD* ASR	MCR* ACC	MCR* ASR	ABL ACC	ABL ASR	D-BR ACC	D-BR ASR
CIFAR-10	BN-all2one	91.64	100.00	88.99	66.79	90.03	10.54	84.46	2.13	94.21	8.29	89.36	0.19	**92.83**	**0.40**
	BN-all2all	92.79	88.01	90.31	4.96	86.04	1.47	84.97	1.71	92.17	2.96	79.91	78.16	**92.61**	**0.56**
	Trojan	91.91	100.00	89.86	100.00	90.89	0.81	83.29	5.04	93.90	2.58	90.18	0.23	**92.21**	**0.76**
	Blend-Strip	92.09	99.97	89.91	93.50	88.33	0.04	83.09	13.30	91.77	17.96	58.46	0.22	**92.40**	**0.06**
	Blend-Kitty	92.69	99.99	90.47	99.31	84.07	0.01	84.54	28.96	94.42	7.49	79.20	2.27	**92.11**	**0.14**
	SIG	92.88	99.69	90.81	99.87	82.43	76.32	81.00	64.72	91.82	99.04	79.94	98.84	**92.73**	**0.24**
	CL	93.20	93.34	90.03	77.44	72.57	10.90	84.46	2.66	92.13	72.01	84.39	0.31	**92.08**	**0.00**
	Avg	92.46	97.29	90.05	77.41	84.91	14.30	83.69	16.93	92.92	30.05	80.21	25.75	**92.42**	**0.31**
CIFAR-100	BN-all2one	71.23	99.13	70.81	66.28	65.42	0.00	69.03	11.41	**73.38**	**0.27**	66.47	0.02	72.58	0.25
	Trojan	75.75	100.00	74.21	99.94	64.52	0.03	72.11	92.21	74.51	0.12	68.12	0.00	**74.52**	**0.00**
	Blend-Strip	75.54	99.99	73.36	99.65	67.38	0.00	71.18	95.78	73.37	0.07	49.13	0.00	**74.35**	**0.00**
	Blend-Kitty	75.18	99.97	72.93	99.96	69.03	0.00	71.73	99.93	73.93	20.60	47.05	0.00	**72.00**	**0.01**
	Avg	74.43	99.77	72.83	91.46	66.59	0.01	71.01	74.83	73.80	5.27	57.69	0.01	**73.36**	**0.07**