

SampDetox: Black-box Backdoor Defense via Perturbation-based Sample Detoxification

Yanxin Yang¹, Chentao Jia¹, Dengke Yan¹, Ming Hu², Tianlin Li³, Xiaofei Xie², Xian Wei¹, Mingsong Chen¹

¹MoE Eng. Research Center of SW/HW Co-design Tech. and App., East China Normal University

²Singapore Management University ³Nanyang Technological University









NeurIPS 2024

Background

Problems of Existing Black-box Backdoor Defenses

Low usability

- Exisiting detection-based black-box backdoor defenses simply discard poisoned samples / models
- Impractical assumption
- Exisiting purification-based black-box backdoor defenses are only effective against small trigger patterns located in the corners of samples

Background

Problems of Existing Black-box Backdoor Defenses

Low usability

Challenge

How to effectively mitigate the impacts of all possible backdoor attacks in black-box scenarios without deteriorating the overall inference performance? only effective against small trigger patterns located in the corners of samples

Motivation

Evaluation Metrics

Visibility: v

$$\mathbf{v} = (1 - SSIM(x^c, x^p)) / 2$$

Robustness: η_r

$$\eta_r = (x^p - x_m)/(x^p - \varepsilon) \quad \varepsilon \sim N(0, I)$$

Observations

 Observation 1: Low visibility (v<0.13), Low robustness (η_r < 0.18)

• Observation 2: High robustness ($\eta_r \ge 0.18$), High visibility (v ≥ 0.13)



i 0.005/0.072 0.045/0.075 0.197/0.432Examples of poisoned samples and their v/n_r.



Our Proposed SampDetox



Stage 1: Global Detoxification

 Inspired by Observation 1, this stage aims to destory the backdoor triggers with low visibility but low robustness.

Stage 2: Local Detoxification

 Inspired by Observation 2, this stage aims to destory the backdoor triggers with high robustness but high visibility.

Experiments: Performance Comparison

Comparison with 3 SOTA defenses against 10 attacks

No Defense			Sancdifi			BDMAE			ZIP			SampDetox (Ours)		
CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)
93.84	e - 1	-	-	-	-	-	-	-	-	-	-	-	-	
92.00	10.18	99.97	76.59	89.55	1.92	89.02	90.10	2.32	88.12	86.52	7.17	89.57	90.15	2.11
84.94	9.78	98.50	70.68	43.73	29.58	82.77	10.08	96.65	82.15	35.60	36.58	83.71	65.06	11.03
84.34	10.26	99.06	68.44	51.78	3.64	79.75	73.39	2.01	79.85	74.92	2.06	80.72	74.36	1.55
93.20	11.07	99.03	76.63	90.84	1.81	<u>91.19</u>	89.35	2.47	87.35	86.91	7.10	92.78	89.95	1.86
91.09	10.02	98.19	76.24	68.89	7.92	88.48	75.78	12.57	87.96	80.19	2.75	88.52	88.62	1.45
93.85	10.93	99.51	77.92	51.12	15.06	87.84	14.88	96.46	88.51	63.81	8.72	90.23	86.65	1.96
93.63	11.13	99.48	77.92	49.95	16.57	87.50	13.63	80.97	88.76	86.59	5.85	90.01	87.40	3.02
91.43	10.27	91.05	77.87	42.97	14.35	85.95	23.19	50.63	86.91	85.22	8.36	89.34	88.92	5.59
93.57	11.38	95.96	77.70	52.05	14.20	86.18	53.12	22.39	87.75	85.46	1.79	90.74	86.51	1.60
91.38	9.46	98.40	75.32	50.42	15.25	86.69	21.73	53.46	85.42	82.94	7.20	90.59	84.83	6.15
	N CA(%) 93.84 92.00 84.94 84.34 93.20 91.09 93.85 93.63 91.43 93.57 91.38	No Defer CA(%) PA(%) 93.84 - 92.00 10.18 84.94 9.78 84.34 10.26 93.20 11.07 91.09 10.02 93.85 10.93 93.63 11.13 91.43 10.27 93.57 11.38 91.38 9.46	No DefenseCA(%) PA(%) ASR(%)93.84-92.0010.1899.9784.949.7898.5084.3410.2699.0693.2011.0799.0391.0910.0298.1993.8510.9399.5193.6311.1399.4891.4310.2791.0593.5711.3895.9691.389.4698.40	No Defense CA(%) PA(%) ASR(%) CA(%) 93.84 - - 92.00 10.18 99.97 76.59 84.94 9.78 98.50 70.68 84.34 10.26 99.06 68.44 93.20 11.07 99.03 76.63 91.09 10.02 98.19 76.24 93.85 10.93 99.51 77.92 93.63 11.13 99.48 77.92 91.43 10.27 91.05 77.87 93.57 11.38 95.96 77.70 91.38 9.46 98.40 75.32	No DefenseSancdifCA(%) PA(%) ASR(%) CA(%) PA(%)93.8492.0010.1899.9776.5984.949.7898.5070.6843.7384.3410.2699.0668.4451.7893.2011.0799.0376.6390.8491.0910.0298.1976.2468.8993.8510.9399.5177.9251.1293.6311.1399.4877.9291.4310.2791.0577.8742.9793.5711.3895.9677.7052.0591.389.4698.4075.3250.42	No DefenseSancdifiCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)93.8492.0010.1899.9776.5989.551.9284.949.7898.5070.6843.7329.5884.3410.2699.0668.4451.783.6493.2011.0799.0376.6390.841.8191.0910.0298.1976.2468.897.9293.8510.9399.5177.9251.1215.0693.6311.1399.4877.9249.9516.5791.4310.2791.0577.8742.9714.3593.5711.3895.9677.7052.0514.2091.389.4698.4075.3250.4215.25	No DefenseSancdifiCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) 93.84 92.00 10.18 99.97 76.59 89.55 1.92 84.94 9.78 98.50 70.68 43.73 29.58 82.77 84.34 10.26 99.06 68.44 51.78 3.64 79.75 93.20 11.07 99.03 76.63 90.84 1.81 91.19 91.09 10.02 98.19 76.24 68.89 7.92 88.48 93.85 10.93 99.51 77.92 51.12 15.06 87.84 93.63 11.13 99.48 77.92 49.95 16.57 87.50 91.43 10.27 91.05 77.87 42.97 14.35 85.95 93.57 11.38 95.96 77.70 52.05 14.20 86.18 91.38 9.46 98.40 75.32 50.42 15.25 86.69	No DefenseSancdifiBDMACA(%) PA(%) ASR(%) CA(%) PA(%) PA(%) ASR(%) CA(%) PA(%) 93.84 92.0010.1899.9776.5989.551.9289.0290.1084.949.7898.5070.6843.7329.5882.7710.0884.3410.2699.0668.4451.783.6479.7573.3993.2011.0799.0376.6390.841.8191.1989.5591.0910.0298.1976.2468.897.9288.4875.7893.8510.9399.5177.9251.1215.0687.8414.8893.6311.1399.4877.9251.1215.0687.8414.8893.6311.1399.4877.9251.1215.0687.8414.8893.5711.3895.9677.7052.0514.2086.1853.1291.389.4698.4075.3250.4215.2586.6921.73 <td>No DefenseSancdifiBDMAE$CA(\%)$ PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)CA(%) PA(%) ASR(%)93.8492.0010.1899.9776.5989.551.9284.949.7898.5070.68$43.73$29.5884.3410.2699.0668.4451.783.6493.2011.0799.0376.6390.841.8191.0910.0298.1976.2468.897.9293.8510.9399.5177.9251.1215.0693.6311.1399.4877.9249.9516.5793.6510.2791.0577.8742.9714.3593.5711.3895.9677.7052.0514.2091.389.4698.4075.3250.4215.2586.6921.7353.46</td> <td>No Defense Sancdifi BDMAE CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) 93.84 - - - - - - 92.00 10.18 99.97 76.59 89.55 1.92 89.02 90.10 2.32 88.12 84.94 9.78 98.50 70.68 43.73 29.58 82.77 10.08 96.65 82.15 84.34 10.26 99.06 68.44 51.78 3.64 79.75 73.39 2.01 79.85 93.20 11.07 99.03 76.63 90.84 1.81 91.19 89.35 2.47 87.35 91.09 10.02 98.19 76.24 68.89 7.92 88.48 75.78 12.57 87.96 93.85 10.93 99.51 77.92 51.12 15.06 87.84 14.88 96.46 88.51 93.63 11.13 99.48 77.92 49.95 16.57</td> <td>No DefenseSancdifiBDMAEZIPCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%)ASR(%) CA(%) PA(%)PA(%)93.8492.0010.1899.9776.5989.551.92$89.02$$90.10$2.32$88.12$$86.52$84.949.7898.5070.68$43.73$$29.58$$82.77$$10.08$$96.65$$82.15$$35.60$84.3410.2699.06$68.44$$51.78$$3.64$$79.75$$73.39$$2.01$$79.85$$74.92$93.2011.0799.0376.6390.841.81$91.19$$89.35$$2.47$$87.35$$86.91$91.0910.0298.1976.24$68.89$$7.92$$88.48$$75.78$$12.57$$87.96$$80.19$93.8510.9399.5177.92$51.12$$15.06$$87.84$$14.88$$96.46$$88.51$$63.81$93.6311.1399.48$77.92$$49.95$$16.57$$87.50$$13.63$$80.97$$88.76$$86.59$91.4310.2791.05$77.87$$42.97$$14.35$$85.95$$23.19$$50.63$$86.91$$85.22$93.5711.3895.96$77.70$$52.05$$14.20$$86.18$$53.12$$22.39$$87.75$$85.46$91.389.4698.4075.32$50.42$$15.25$$86.69$$21.73$$53.46$$85.42$$82.94$<</td> <td>No DefenseSancdifiBDMAEZIPCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)CA(%) PA(%) ASR(%)CA(%) PA(%) ASR(%)93.8492.0010.1899.9776.5989.551.92$\underline{89.02}$90.102.3288.1286.527.1784.949.7898.5070.68$\underline{43.73}$29.58$\underline{82.77}$10.0896.6582.1535.6036.5884.3410.2699.0668.4451.783.6479.7573.392.0179.8574.922.0693.2011.0799.0376.6390.841.8191.1989.352.4787.3586.917.1091.0910.0298.1976.2468.897.92$\underline{88.48}$75.7812.5787.9680.192.7593.8510.9399.5177.9251.1215.0687.8414.8896.4688.5163.818.7293.6311.1399.4877.9249.9516.5787.5013.6380.9788.7686.595.8591.4310.2791.0577.8742.9714.3585.9523.1950.6386.9185.228.3693.5711.3895.9677.7052.0514.2086.1853.1222.3987.7585.461.7991.389.4698.4075.3250.4215.2586.6921.7353.4685.42<td>No DefenseSancdifiBDMAEZIPSampCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%)CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)CA(%)93.8492.0010.1899.9776.5989.551.9289.0290.102.3288.1286.527.1789.5784.949.7898.5070.6843.7329.5882.7710.0896.6582.1535.6036.5883.7184.3410.2699.0668.4451.783.6479.7573.392.0179.8574.922.0680.7293.2011.0799.0376.6390.841.8191.1989.352.4787.3586.917.1092.7891.0910.0298.1976.2468.897.9288.4875.7812.5787.9680.192.7588.5293.8510.9399.5177.9251.1215.0687.8414.8896.4688.5163.818.7290.2393.6311.1399.4877.9249.9516.5787.5013.6380.9788.7686.595.8590.0191.4310.2791.0577.8742.9714.3585.9523.1950.6386.9185.228.3689.3493.5711.3895.9677.7052.0514.2086.1853.1222.3987.7585.46<td>No DefenseSancdifiBDMAEZIPSampDetox (CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)93.84</td></td></td>	No DefenseSancdifiBDMAE $CA(\%)$ PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)CA(%) PA(%) ASR(%)93.8492.0010.1899.9776.5989.55 1.92 84.949.7898.5070.68 43.73 29.5884.3410.2699.0668.4451.783.6493.2011.0799.0376.63 90.841.81 91.0910.0298.1976.2468.897.9293.8510.9399.5177.9251.1215.0693.6311.1399.4877.9249.9516.5793.6510.2791.0577.8742.9714.3593.5711.3895.9677.7052.0514.2091.389.4698.4075.3250.4215.2586.6921.7353.46	No Defense Sancdifi BDMAE CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) 93.84 - - - - - - 92.00 10.18 99.97 76.59 89.55 1.92 89.02 90.10 2.32 88.12 84.94 9.78 98.50 70.68 43.73 29.58 82.77 10.08 96.65 82.15 84.34 10.26 99.06 68.44 51.78 3.64 79.75 73.39 2.01 79.85 93.20 11.07 99.03 76.63 90.84 1.81 91.19 89.35 2.47 87.35 91.09 10.02 98.19 76.24 68.89 7.92 88.48 75.78 12.57 87.96 93.85 10.93 99.51 77.92 51.12 15.06 87.84 14.88 96.46 88.51 93.63 11.13 99.48 77.92 49.95 16.57	No DefenseSancdifiBDMAEZIPCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%)ASR(%) CA(%) PA(%)PA(%)93.8492.0010.1899.9776.5989.55 1.92 89.02 90.10 2.32 88.12 86.52 84.949.7898.5070.68 43.73 29.58 82.77 10.08 96.65 82.15 35.60 84.3410.2699.06 68.44 51.78 3.64 79.75 73.39 2.01 79.85 74.92 93.2011.0799.0376.63 90.841.81 91.19 89.35 2.47 87.35 86.91 91.0910.0298.1976.24 68.89 7.92 88.48 75.78 12.57 87.96 80.19 93.8510.9399.5177.92 51.12 15.06 87.84 14.88 96.46 88.51 63.81 93.6311.1399.48 77.92 49.95 16.57 87.50 13.63 80.97 88.76 86.59 91.4310.2791.05 77.87 42.97 14.35 85.95 23.19 50.63 86.91 85.22 93.5711.3895.96 77.70 52.05 14.20 86.18 53.12 22.39 87.75 85.46 91.389.4698.4075.32 50.42 15.25 86.69 21.73 53.46 85.42 82.94 <	No DefenseSancdifiBDMAEZIPCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)CA(%) PA(%) ASR(%)CA(%) PA(%) ASR(%)93.8492.0010.1899.9776.5989.55 1.92 $\underline{89.02}$ 90.102.3288.1286.527.1784.949.7898.5070.68 $\underline{43.73}$ 29.58 $\underline{82.77}$ 10.0896.6582.1535.6036.5884.3410.2699.0668.4451.783.6479.7573.392.0179.85 74.92 2.0693.2011.0799.0376.63 90.841.81 91.1989.352.4787.3586.917.1091.0910.0298.1976.2468.897.92 $\underline{88.48}$ 75.7812.5787.9680.192.7593.8510.9399.5177.9251.1215.0687.8414.8896.4688.5163.818.7293.6311.1399.4877.9249.9516.5787.5013.6380.9788.7686.595.8591.4310.2791.0577.8742.9714.3585.9523.1950.6386.9185.228.3693.5711.3895.9677.7052.0514.2086.1853.1222.3987.7585.461.7991.389.4698.4075.3250.4215.2586.6921.7353.4685.42 <td>No DefenseSancdifiBDMAEZIPSampCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%)CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)CA(%)93.8492.0010.1899.9776.5989.551.9289.0290.102.3288.1286.527.1789.5784.949.7898.5070.6843.7329.5882.7710.0896.6582.1535.6036.5883.7184.3410.2699.0668.4451.783.6479.7573.392.0179.8574.922.0680.7293.2011.0799.0376.6390.841.8191.1989.352.4787.3586.917.1092.7891.0910.0298.1976.2468.897.9288.4875.7812.5787.9680.192.7588.5293.8510.9399.5177.9251.1215.0687.8414.8896.4688.5163.818.7290.2393.6311.1399.4877.9249.9516.5787.5013.6380.9788.7686.595.8590.0191.4310.2791.0577.8742.9714.3585.9523.1950.6386.9185.228.3689.3493.5711.3895.9677.7052.0514.2086.1853.1222.3987.7585.46<td>No DefenseSancdifiBDMAEZIPSampDetox (CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)93.84</td></td>	No DefenseSancdifiBDMAEZIPSampCA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%)CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)CA(%)93.8492.0010.1899.9776.5989.55 1.92 89.0290.102.3288.1286.527.17 89.57 84.949.7898.5070.6843.7329.5882.7710.0896.6582.1535.6036.5883.7184.3410.2699.0668.4451.783.6479.7573.392.0179.8574.922.0680.7293.2011.0799.0376.63 90.841.81 91.1989.352.4787.3586.917.10 92.78 91.0910.0298.1976.2468.897.9288.4875.7812.5787.9680.192.7588.5293.8510.9399.5177.9251.1215.0687.8414.8896.4688.5163.818.7290.2393.6311.1399.4877.9249.9516.5787.5013.6380.9788.7686.595.8590.0191.4310.2791.0577.8742.9714.3585.9523.1950.6386.9185.228.3689.3493.5711.3895.9677.7052.0514.2086.1853.1222.3987.7585.46 <td>No DefenseSancdifiBDMAEZIPSampDetox (CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)93.84</td>	No DefenseSancdifiBDMAEZIPSampDetox (CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%) CA(%) PA(%) ASR(%)93.84

Our SampDetox achieves the best CA, PA and ASR compared to other SOTA defenses against various attacks

Experiments: Ablation Study and Discussion

• Impacts of Different Stages, Denoising, and Hyperparameters $\overline{t}_1, \overline{t}_2$

Attack BadNets SIG LC TrojanNN Dynamic	Visibility	Noise*				Stage 1		SampDetox (Stage 1 + Stage 2)				
	v	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)		
BadNets	0.052	56.25	49.71	3.93	90.12	88.15	6.61	89.57	90.15	2.11		
SIG	0.185	51.81	17.16	15.89	83.10	9.48	92.33	83.71	65.06	11.03		
LC	0.121	47.75	28.04	1.07	81.50	61.45	34.75	80.72	74.36	1.55		
TrojanNN	0.137	58.03	45.89	5.68	92.54	35.51	46.86	92.78	89.95	1.86		
Dynamic	0.098	56.26	41.18	1.09	87.58	85.62	3.82	87.52	88.62	1.45		
Blended	0.067	59.42	45.22	1.53	88.65	86.65	1.86	90.23	86.65	1.96		
LF	0.005	56.12	40.95	2.55	89.43	87.50	3.02	90.01	87.40	3.02		
WaNet	0.005	57.71	43.73	5.38	89.43	88.91	5.60	89.34	88.92	5.59		
ISSBA	0.006	57.10	37.35	1.14	90.92	86.50	1.61	90.74	86.51	1.60		
BPP	0.009	58.40	42.03	5.75	89.09	84.84	6.15	90.59	84.83	6.15		

	Fix	$\operatorname{ed} \bar{t}_2 = 0$)	Fixed $\bar{t}_1 = 20$							
\overline{t}_1	CA(%)	PA(%)	ASR(%)	\overline{t}_2	CA(%)	PA(%)	ASR(%)				
5	92.07	62.48	27.58	40	92.19	60.90	30.07				
10	91.22	78.69	12.96	60	92.02	73.32	19.32				
15	90.92	86.35	5.39	80	91.86	79.38	14.11				
20	90.65	86.43	1.73	100	91 72	83.68	613				
25	88.26	85.13	1.80	120	92.02	85.22	2.34				
30	86.77	84.56	1.71	150	91.85	84.92	2.28				
35	84.91	83.78	1.75	200	92.26	81.87	2.30				
40	82.30	83.01	1.72	250	92.39	77.03	2.29				

Extra time overhead and effectiveness using DDIM

14	13.91	7			Attack	Samp	Detox+	DDPM	Samp	Detox+I	DDIM
12.0				SampDetox's inference time is	Attack	CA(%)	PA(%)	ASR(%)	CA(%)	PA(%)	ASR(%)
13.8				comparable to that of no defense	BadNets	89.57	90.15	2.11	89.49	90.13	2.12
pu of					SIG	83.71	65.06	11.03	83.82	65.13	10.98
0.0	0.4516 0.4735	10	0.4953		LC	80.72	74.36	1.55	80.62	74.22	1.53
(Se	0.4516 0.4755	0.2526			TrojanNN	92.78	89.95	1.86	92.83	89.87	1.69
e 0.4		0.3526			Dynamic	88.52	88.62	1.45	88.52	88.72	1.42
1. I					Blended	90.23	86.65	1.96	90.15	86.54	2.02
0.2				Using DDIM does not reduce the	LF	90.01	87.40	3.02	90.09	87.61	3.10
	0.0435		0.0651		WaNet	89.34	88.92	5.59	89.48	88.82	5.54
0				effectiveness of SampDetox	ISSBA	90.74	86 51	1.60	90.76	86.65	1.55
v	No Defense Sancdifi BDMAE ZIP	ZIP+DDIM	Ours Ours+DDIM		BPP	90.59	84.83	6.15	90.42	84.91	6.17
	Defense M	lethods			3 						

Conclusion

Problems of Existing Black-box Backdoor Defenses

- Detection-based defenses greatly reduce the usability of tasks
- Purification-based methods are based on the impractical assumption

Contributions of our work

- Reveal the correlation between the visibility of triggers and the robustness of poisoned samples
- Present a novel perturbation-based sample detoxification method together with its theoretical foundations

Experimental results

 Extensive experimental results show the applicability and superiority of our approach over state-of-the-art (SOTA) backdoor defense methods

Thank You !

8