Backdoor Federated Learning by Poisoning Backdoor-Critical Layers

Haomin Zhuang[1], Mingxian Yu[1], Hao Wang[2], Yang Hua[3], Jian Li[4], Xu Yuan[5]

[1]South China University of Technology, [2]Louisiana State University,

[3]Queen's University Belfast, UK, [4]Stony Brook University, [5]University of Delaware

IntelliSys Lab

Paper Code

Backdoor-Critical (BC) Layers Observation

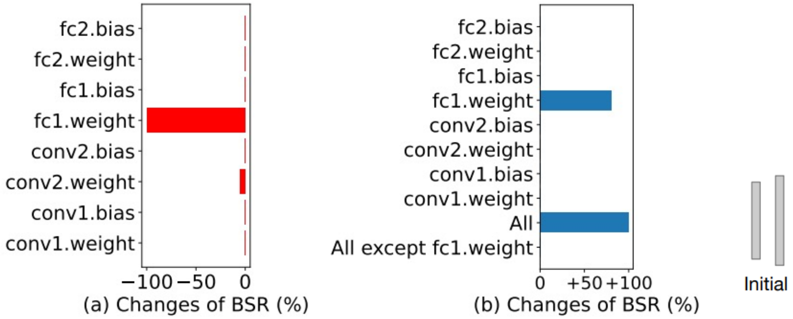


Figure 1. **(a)** The changes in BSR of the malicious model with a layer substituted from the benign model. **(b)** The changes of BSR of the benign model with layer(s) substituted from the malicious

Research question

1) How to identify BC layers?
2) How to utilize BC layers to bypass defenses algorithms?

Contributions

- We propose Layer Substitution Analysis, a novel method that recognizes backdoor-critical layers, which naturally fits into FL attackers' context.
- We design two effective layer-wise backdoor attack methods, that successfully inject backdoor to BC layers and bypass SOTA defense methods without decreasing the main task accuracy.
- Our evaluation on a wide range of models and datasets shows that the proposed layer-wise backdoor attack methods outperform existing backdoor attacks, such as DBA [1], on both main task accuracy and backdoor success rate under SOTA defense methods.

Identifying BC Layers

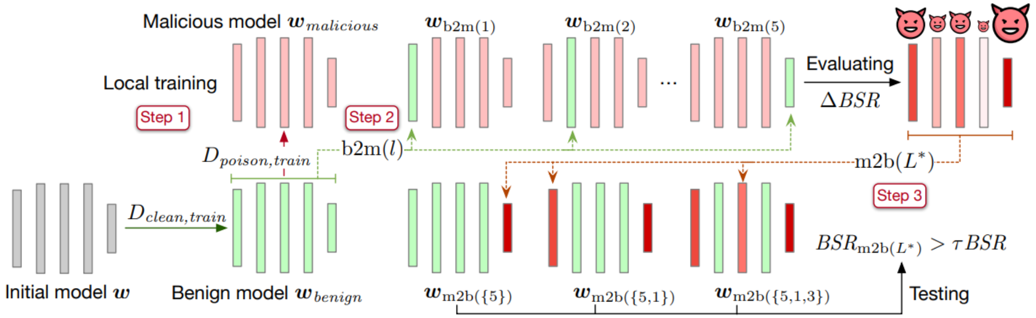


Figure 2. Identifying BC layers with Layer Substitution Analysis.

- Step 1: Train on the clean dataset and retrain on the malicious dataset.
- Step 2: Insert benign layer into the malicious model and evaluate BSR.
- Step 3: Insert malicious layers into the benign model and then evaluate BSR.

Poisoning BC Layers in FL

- For Layer-wise poisoning (LP) attack, we decrease the distance between malicious models and benign models by poisoning BC layers.

$$\widetilde{\boldsymbol{w}}^{(i)} = \lambda \boldsymbol{v} \circ \boldsymbol{u}^{(i)}_{malicious} + ReLU(1-\lambda) \cdot \boldsymbol{v} \circ \boldsymbol{u}_{average} + (1-\boldsymbol{v}) \circ \boldsymbol{u}_{average},$$

where $\boldsymbol{v}$ is the set of BC layers, λ is a hyperparameter for scaling, and $\mathbf{u}_{\text{average}}$ is the mean of simulated benign models.

- For Layer-wise flipping (LF) attack, we flip the signs of parameters in BC layers, where backdoor attack is neutralized by flipping from defenses.

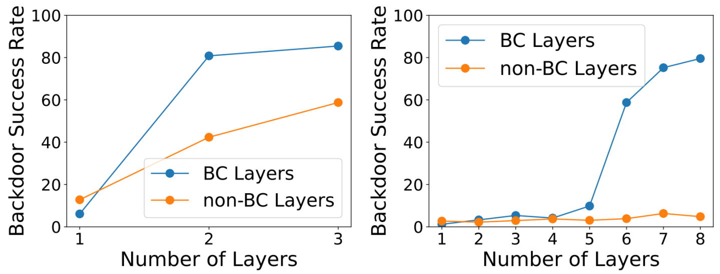
$$\boldsymbol{w}^{(i)}_{LFA} := -(\boldsymbol{w}^{(i)}_{m2b(L^*)} - \boldsymbol{w}) + \boldsymbol{w}.$$

[1] Chulin Xie et al. "DBA: Distributed Backdoor Attacks Against Federated Learning." In Proc. of ICLR, 2019.

Experiment Result Highlights

Model (Dataset)		VGG19 (CIFAR-10)			ResNet18 (CIFAR-10)			CNN (Fashion-MNIST)		
Attack		Baseline	LP Attack (LF Attack)	DBA	Baseline	LP Attack (LF Attack)	DBA	Baseline	LP Attack (LF Attack)	DBA
FedAvg (non-IID)	Best BSR	84.88	**92.8**±0.99	41.15	85.19	**94.19**±0.99	21.19	**99.97**	87.69±4.3	**99.97**
	Avg BSR	74.69	**83.55**±0.43	25.88	70.53	**89.12**±1.4	10.94	**99.9**	78.84±9.16	**99.9**
	Acc	78.89	**79.95**±0.46	78.97	77.58	77.89±0.43	**77.99**	88.28	**88.42**±0.23	87.95
FLTrust (non-IID)	Best BSR	**92.91**	76.56±34.38	42.14	**92.43**	82.05±25.34	37.16	74.17	89.44±3.44	**100.0**
	Avg BSR	**67.3**	65.44±31.56	15.88	**75.84**	71.52±29.17	15.11	68.97	77.05±4.67	**100.0**
	Acc	75.1	74.03±4.06	**75.11**	75.72	69.9±5.74	**77.51**	**89.51**	89.48±0.1	89.31
FLAME (non-IID)	Best BSR	47.03	**88.68**±4.98	38.25	23.04	**95.41**±0.93	9.77	0.18	**84.33**±3.12	0.58
	Avg BSR	7.78	**60.72**±2.44	7.33	7.22	**90.15**±3.51	3.88	0.1	**74.91**±2.66	0.4
	Acc	62.91	56.92±1.12	**63.3**	**76.04**	71.48±0.36	75.27	87.78	87.05±0.21	**87.89**
RLR (non-IID)	Best BSR	79.37	**92.17**±1.81 (2.79±0.81)	43.79	81.61	**93.16**±0.85 (1.37±0.02)	13.85	20.27	0.0 ± 0.0 (**70.52**±3.13)	38.25
	Avg BSR	74.01	**89.24**±2.09 (0.6±0.09)	33.69	60.83	**82.14**±7.46 (0.7±0.1)	7.8	15.09	0.0 ± 0.0 (**66.12**±2.94)	7.33
	Acc	67.33	**72.1**±0.58 (63.2±3.94)	64.3	75.07	73.44±0.95 (**76.48**±0.32)	75.04	85.56	86.09 ± 0.13 (**86.45**±0.41)	63.3
MultiKrum (non-IID)	Best BSR	22.93	**95.87**±0.51	29.44	12.72	**95.94**±0.97	10.63	1.09	**89.95**±2.74	0.28
	Avg BSR	7.84	**75.93**±2.49	8.44	3.95	**90.12**±1.38	5.61	0.39	**74.94**±6.97	0.1
	Acc	58.93	**69.28**±3.29	64.81	**74.49**	72.26±1.34	73.02	87.31	**87.58**±0.21	**87.58**
FLDetector (non-IID)	Best BSR	**95.49**	87.28±0.69	16.28	5.23	**90.31**±2.04	5.89	74.64	99.45±0.13	**99.93**
	Avg BSR	**95.42**	86.71±0.54	16.14	5.21	**86.56**±1.32	5.87	66.11	96.32±0.41	**99.9**
	Acc	55.25	**57.95**±1.37	56.67	**64.39**	63.89±0.91	65.25	79.16	75.96±0.81	**79.78**
FLARE (non-IID)	Best BSR	**96.67**	93.47±4.32	25.48	17.16	**79.94**±4.06	26.96	2.02	82.64±4.16	**100**
	Avg BSR	**94.45**	70.23±5.83	8.18	6.24	**53.72**±7.73	6.62	1.54	78.18±2.41	**100**
	Acc	70.25	**77.28**±1.46	69.95	**71.39**	70.84±1.63	64.22	**88.29**	88.07±0.46	88.01

Table 2. Main task accuracy and BSR on Non-IID datasets.



(a) ResNet18 (b) VGG19

Figure 7. : Attacking a fixed number of BC layers or non-BC layers under FLAME training ResNet18 on IID CIFAR-10 dataset.

Contact: z365460860@gmail.com, haowang@lsu.edu