Scope: On Detecting Constrained Backdoor Attacks in Federated Learning

Siquan Huang, Yijiang Li, Xingfu Yan, Ying Gao, *Member, IEEE*, Chong Chen, *Student Member, IEEE*, Leyu Shi, Biao Chen, and Wing W.Y. Ng, *Senior Member, IEEE*

IEEE Transactions on Information Forensics and Security （TIFS 2025）

Motivation

1. Benign更新 和 Backdoor更新 维度划分

Benign更新

Backdoor更新

良性维度 后门维度 冗余维度

冗余维度指的是更新幅度微弱维度

Motivation

1. Backdoor更新 良性维度的绝对变化 掩盖 相对变化大的 后门维度

绝对变化： w_local-w_global

相对变化： (w_local-w_global) / (| w_local | + | w_global |)

Motivation and Method

绝对变化： w_local-w_global

相对变化： (w_local-w_global) / (| w_local | + | w_global |)

W_global	0.03	0.05	0.004	0.1	0.003	0.02	0.005	……
W_backdoor	0.01	0.02	0.001	0.05	0.005	0.06	0.002	……
绝对变化 (update)	-0.02	-0.03	-0.003	-0.05	0.002	0.04	-0.003	……
相对变化	-0.5	-0.43	-0.6	-0.33	0.25	0.5	-0.43	……

Backdoor 客户端的主任务是良性任务， 良性维度数量 > 后门维度数量，-0.02 -0.03 维度数量占据主导掩盖后门维度0.002，-0.003维度的异常, 但0.002，-0.003维度相对变化较大；

update转换为相对变化进行后门检测，后门维度相对变化 0.002，-0.003 维度转换为 0.25, -0.43

Motivation and Method

2. 攻击者 操纵 冗余维度，增加微弱变化 掩盖 后门维度异常

Backdoor更新

相对变化： -1/2 -3/7 -3/5 -1/3 1/4 1/2 3/7 1/10 1/11 1/12 1/15 ……

冗余维度只进行了微弱更新，在高维参数空间更加明显。维度诅咒给基于指标计算的各类防御机制带来了比较大的挑战。攻击者对部分冗余参数进行微弱修改，通过微弱变化累计，掩盖后门维度的异常变化。

$$(g_i^j)^* = (g_i^j)^\varphi$$ 各个维度的相对变化进行自乘，放大不同维度之间的相对差异，排除冗余维度干扰

$$(g_i^j)^* = (g_i^j)^\varphi \cdot \mathrm{sgn}(g_i^j)$$ 保留原始相对变化方向，$\varphi = 2$

Motivation and Method

3. 服务器对各客户端上传的更新经过 1.归一化，2. 放缩 处理之后计算客户端之间的余弦相似性得分

$$\delta_i = \sum_{p=1}^{K}(1 - \frac{\langle \boldsymbol{g_i}^*, \boldsymbol{g_p}^* \rangle}{|\boldsymbol{g_i}^*| \cdot |\boldsymbol{g_p}^*|})$$

具体为每个客户端 i 分别与剩余的客户端 p 计算余弦距离并进行求和获得最终得分 δ_i

服务器对 δ_i 进行聚类，将良性id 进行聚合更新本轮全局模型。

Detecting Backdoor Attacks in Federated Learning via Direction Alignment Inspection

Jiahao Xu Zikai Zhang Rui Hu
University of Nevada, Reno
{jiahaox,zikaiz,ruihu}@unr.edu

CVPR 2025

Method

1. Temporal Direction Alignment (TDA)

$$\omega_i := \langle \Delta_i^t, \theta^t \rangle / (\|\Delta_i^t\| \|\theta^t\|).$$

服务器为各个客户端上传的的 update 与当前的全局模型参数 θ^t 计算余弦相似性,获得TDA指标

2. Masked principal sign alignment checking(MPSA)

服务器提取各个客户端 update 在每个维度的方向，然后对每个维度进行方向聚合Fedsign

$$p \leftarrow \operatorname{sgn}\left(\sum_{i=1}^{n} \operatorname{sgn}(\Delta_i^t)\right)$$

$$\rho_i := 1 - \left\| \left(\operatorname{sgn}(\Delta_i^t) - p\right) \odot \operatorname{Top}_k(\Delta_i^t) \right\|_0 / k$$

服务器将客户端的TDA指标用中位数和标准差计算异常(MZ-Score)；将MPSA指标同样使用中位数标准差计算异常；过滤异常id，剩余良性id聚合更新全局模型。

Method

Algorithm 1: AlignIns

Input: Set of n local model updates $\{\Delta_i^t\}_{i=1}^n$ where m of them are malicious, global model θ^t, TDA radius λ_c, MPSA radius λ_s, extraction parameter k

Output: Aggregated model update $\widetilde{\Delta}$

Initialize benign set $\mathcal{S} \leftarrow \emptyset$
$\omega \leftarrow \{\mathrm{TDA}(\Delta_i^t, \theta)\}_{i=1}^n$ ◁ by Equation (1).
$p \leftarrow \mathrm{sgn}(\sum_{i=1}^n \mathrm{sgn}(\Delta_i^t))$
$\rho \leftarrow \{\mathrm{MPSA}(\Delta_i^t, p, k)\}_{i=1}^n$ ◁ by Equation (2).
for $i \in [n]$ **do**
  $\lambda_{i,c} \leftarrow \mathrm{MZ\_score}(\omega_i, \omega)$ ◁ by Equation (3).
  $\lambda_{i,s} \leftarrow \mathrm{MZ\_score}(\rho_i, \rho)$ ◁ by Equation (3).
  if $|\lambda_{i,c}| \leq \lambda_c$ *and* $|\lambda_{i,s}| \leq \lambda_s$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{i\}$
  end
end
$c \leftarrow \mathrm{med}(\{\|\Delta_i^t\|\}_{i \in \mathcal{S}})$
$\widetilde{\Delta} \leftarrow (1/|\mathcal{S}|) \sum_{i \in \mathcal{S}} (\Delta_i^t \cdot \min\{1, c/\|\Delta_i^t\|\})$
return $\widetilde{\Delta}$;

Experiments

Dataset (Model)	Methods	Clean MA↑	Badnet				DBA				Neurotoxin				Avg. BA↓	Avg. RA↑
			BA↓		RA↑		BA↓		RA↑		BA↓		RA↑			
			r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5		
CIFAR-10 (ResNet9 [16])	FedAvg	89.47	51.56	67.61	45.79	31.24	56.21	70.42	40.62	27.92	44.89	79.40	50.41	19.60	61.68	35.93
	FedAvg*	89.47	2.06	2.06	85.60	85.60	2.06	2.06	85.60	85.60	2.06	2.06	85.60	85.60	2.06	85.60
	RLR	79.16	2.32	**2.00**	76.72	73.33	3.01	3.04	77.09	77.13	3.12	3.87	73.98	73.29	2.89	35.93
	RFA	87.73	70.67	90.24	27.74	9.26	47.67	66.97	47.29	30.14	81.27	96.13	17.11	3.69	75.49	22.54
	MKrum	87.02	81.10	97.47	18.11	2.51	2.17	4.33	83.89	79.10	65.28	89.18	31.81	10.01	56.59	37.57
	Foolsgold	89.49	69.14	68.84	29.64	30.10	51.18	60.73	44.83	36.08	2.91	2.82	85.27	84.76	42.60	51.78
	MM	89.15	41.19	93.88	53.88	6.01	52.24	51.30	43.54	45.08	43.92	83.92	51.12	15.11	61.08	35.79
	Lockdown	88.56	6.31	10.82	81.88	79.50	11.63	6.03	78.82	75.77	3.40	3.27	82.73	83.14	6.91	80.31
	AlignIns	88.64	**1.91**	2.21	**86.03**	**85.57**	**2.13**	**2.14**	**85.77**	**85.88**	**2.66**	**2.20**	**85.46**	**85.31**	**2.21**	**85.67**
CIFAR-100 (VGG9 [44])	FedAvg	64.29	99.20	99.54	0.68	0.35	99.25	99.36	0.64	0.54	94.41	93.36	4.36	5.28	97.52	1.98
	FedAvg*	64.29	0.62	0.62	53.03	53.03	0.62	0.62	53.03	53.03	0.62	0.62	53.03	53.03	0.62	53.03
	RLR	44.34	96.57	99.85	1.81	0.12	24.41	94.08	24.97	3.22	**0.04**	**0.00**	29.07	29.73	52.49	14.82
	RFA	53.92	4.32	1.45	37.60	39.88	2.15	0.78	39.73	41.51	99.74	89.59	0.21	6.59	33.01	27.59
	MKrum	51.28	1.33	1.54	38.13	38.49	1.36	1.54	37.85	37.91	99.82	99.87	0.12	0.10	36.21	25.49
	Foolsgold	64.13	99.02	99.30	0.83	0.57	99.15	99.39	0.74	0.51	21.79	6.21	42.06	46.40	70.81	15.19
	MM	63.26	99.51	99.87	0.37	0.11	99.53	99.70	0.35	0.19	98.48	98.97	1.32	0.83	99.34	0.53
	Lockdown	62.88	55.21	24.14	28.45	43.06	34.37	49.02	34.06	27.93	0.85	0.67	42.66	47.04	27.38	37.20
	AlignIns	63.45	**0.79**	**0.71**	**50.45**	**51.53**	**0.45**	**0.57**	**50.81**	**52.08**	0.49	0.53	**51.11**	**50.66**	**0.59**	**51.11**

THANKS