拜占庭鲁棒联邦学习的局部模型中毒攻击

本文中，我们首次对联邦学习的局部模型中毒攻击进行了系统的研究。我们假设攻击者已经损害了一些客户端设备，攻击者在学习过程中操纵被损害的客户端设备上的局部模型参数，使得全局模型具有较大的测试错误率。

我们将我们的攻击公式化为优化问题，并将我们的攻击应用于最近的四种拜占庭健壮的联邦学习方法。

通过在四个真实世界数据集上的实验结果表明，我们的攻击可以大大增加由联合学习方法学习的模型的错误率。

我们概括了两种针对数据中毒攻击的防御方法，以防御我们的本地模型中毒攻击。我们的评估结果表明，在某些情况下，一种防御方法可以有效地防御我们的攻击，但在其他情况下，这些防御方法不够有效，这突出了对联合学习的本地模型中毒攻击的新防御方法的需求。

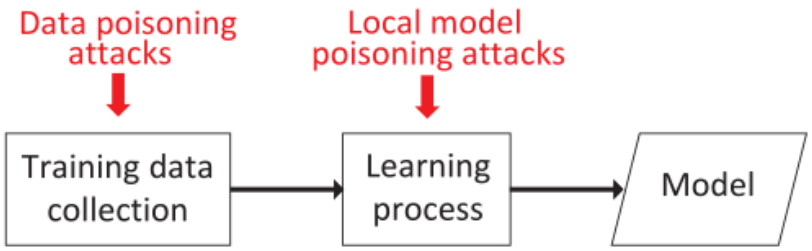


Figure 1: *Data* vs. *local model* poisoning attacks.

本文对拜占庭健壮联邦学习的局部模型中毒攻击进行了首次研究。

攻击：

本地模型中毒攻击的一个关键挑战是如何将本地模型从受损的工作设备发送到主设备。

因此我们将制作局部模型公式化为在联邦学习的每次迭代中解决一个优化问题。具体地，如果没有攻击，主设备可以在迭代中计算全局模型，我们称之为攻击前全局模型。我们的目标是在受损的工作设备上制作本地模型，使全局模型最大程度地偏离攻击前全局模型变化的相反方向。多次迭代累积的偏差会使学习到的全局模型与攻击前的模型显著不同。

防御：

为了应对这一挑战，我们推广了RONI和TRIM来抵御本地模型中毒攻击。
在每次迭代中使用拜占庭健壮聚合规则计算全局模型之前，两种防御都移除潜在恶意的局部模型。
一种防御是移除对全局模型的错误率具有较大负面影响的局部模型(受RONI的启发，移除对模型的错误率具有较大负面影响的训练样本)，而另一种防御是移除导致较大损失的局部模型(受TRIM的启发，移除对损失具有较大负面影响的训练样本)，其中在验证数据集上评估错误率和损失。
我们分别称这两种防御为基于错误率的拒绝(ERR)和基于损失函数的拒绝(LFR)。此外，我们结合了错误和LFR，即，我们删除被错误或LFR删除的局部模型。
我们的实证评估结果显示，LFR优于ERR而且大部分情况下联合防守堪比LFR。此外，LFR在某些情况下可以抵御我们的攻击，但LFR在其他情况下不够有效。

主要贡献:

(1)对攻击拜占庭鲁棒联邦学习进行了第一次系统的研究。
(2)提出对拜占庭鲁棒性联邦学习的局部模型中毒攻击。我们的攻击会在学习过程中操纵受损工作设备上的本地模型参数。
(3)我们概括了两种针对数据中毒攻击的防御方法，以防御本地模型中毒攻击。我们的结果表明，虽然其中一种在某些情况下是有效的，但它们在其他情况下的成功是有限的。

拜占庭健壮聚合规则

Krum：Krum选择与其他m个模型相似的一个作为全局模型。直觉是，即使选定的局部模型来自被控制的工作设备，其影响也可能受到限制，因为它类似于来自良性工作设备的其他局部模型。模型的相似是采用的欧氏距离判定。

Bulyan:Bulyan采用修剪平均值的方式去剪枝掉这些异常参数，因此Bulyan是对krum的一次优化。

Trimmed mean:将模型的每个参数独立出来进行选择，排序后去掉最大最小值，在计算均值来作为该参数的聚合值。

Median:同样将参数独立出来进行选择，只不过选择的是中位数。

聚合规则：

根据攻击者是否知道聚合规则来考虑两种情况。特别是，攻击者可能知道各种场景中的聚合规则。例如，服务提供者可以公开聚集规则，以便增加联合学习系统的透明度和信任度。

(1) 当攻击者知道聚合规则时，将利用已知的聚合规则为受威胁的工作设备创建本地模型参数。
(2)当攻击者不知道聚合规则时，将根据特定的聚合规则为受威胁的工作设备创建本地模型参数。实证结果表明，这种精心制作的局部模型也可能攻击其他聚集规则。同时观察到本地模型中毒攻击在不同聚集规则之间的不同水平的可转移性。

模型攻击的优化问题：

定义一个方向量，1表示当前梯度增加，-1表示当前梯度减小，其次定义攻击前的梯度与攻击后的梯度，那么优化问题的实质就是，使得攻击后的梯度与攻击前的梯度差别尽量大。
其中s是所有全局模型参数变化方向的列向量，w是攻击前的全局模型，w'是袭击后的全球模型。

$$\max_{\mathbf{w}'_1,\cdots,\mathbf{w}'_c} \mathbf{s}^T(\mathbf{w}-\mathbf{w}'),$$
$$\text{subject to } \mathbf{w} = \mathcal{A}(\mathbf{w}_1,\cdots,\mathbf{w}_c,\mathbf{w}_{c+1},\cdots,\mathbf{w}_m),$$
$$\mathbf{w}' = \mathcal{A}(\mathbf{w}'_1,\cdots,\mathbf{w}'_c,\mathbf{w}_{c+1},\cdots,\mathbf{w}_m),$$

攻击：

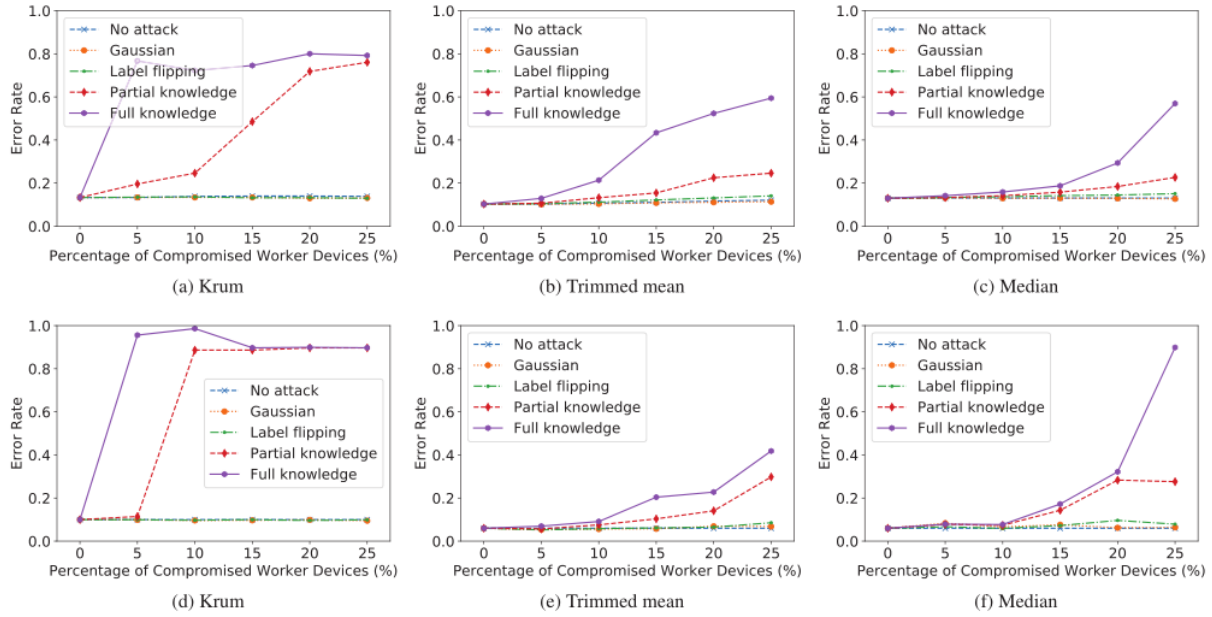
攻击 Krum

由于Krum是选择一个最接近的局部模型作为返回结果，那么就尽量在攻击时让FL选择被控制的设备的局部模型。这种情况下一共分为两步。首先，要让控制的设备的局部模型偏差尽量大，其次，要让其他被控制的设备的设置的局部模型都接近这个偏差值，这样方便FL依照Krum原则对选择该模型作为全局模型，从而达到攻击目的。

攻击Trimmed Mean

由于Trimmed Mean将模型的每个参数独立出来进行选择，排序后去掉最大最小值，在计算均值来作为该参数的聚合值。因此可以设置控制的设备的模型的梯度尽量取中间值，然后逐步偏移，达到攻击的目的。

攻击效果：



(a) Krum (b) Trimmed mean (c) Median

(d) Krum (e) Trimmed mean (f) Median

防御：

ERR

基于错误率的拒绝(ERR):在这种防御中，我们计算每个局部模型对验证数据集的错误率的影响，并移除对错误率具有较大负面影响的局部模型。具体来说，假设我们有一个聚合规则。

LFR

基于损失函数的拒绝(LFR):在这种防御中，是基于局部模型对损失函数的影响而不是验证数据集的错误率来移除局部模型。通过计算验证数据集上模型A和B的交叉熵损失函数值，移除具有最大损失影响的c个局部模型，聚集剩余的局部模型以更新全局模型。

	No attack	Krum	Trimmed mean
Krum	0.14	0.72	0.13
Krum + ERR	0.14	0.62	0.13
Krum + LFR	0.14	0.58	0.14
Krum + Union	0.14	0.48	0.14
Trimmed mean	0.12	0.15	0.23
Trimmed mean + ERR	0.12	0.17	0.21
Trimmed mean + LFR	0.12	0.18	0.12
Trimmed mean + Union	0.12	0.18	0.12
Median	0.13	0.17	0.19
Median + ERR	0.13	0.21	0.25
Median + LFR	0.13	0.20	0.13
Median + Union	0.13	0.19	0.14

结论：

实验证明机器学习团体声称对一些工作设备的拜占庭故障具有鲁棒性的联合学习方法，容易受到我们的本地模型中毒攻击，这些攻击在学习过程中操纵从受损的工作设备发送到主设备的本地模型。

特别地，为了增加所学习的全局模型的错误率，攻击者可以在受损的工作者设备上制作本地模型，使得聚集的全局模型最大程度地偏离当没有攻击时全局模型将改变的相反方向。我们可以概括现有的对数据中毒攻击的防御，以防御我们的本地模型中毒攻击。这种广义的防御在某些情况下有效，但在其他情况下不够有效。我们的结果强调，我们需要新的防御措施来抵御本地模型中毒攻击。