MPAF:基于虚假客户端的联合学习模型中毒攻击

本文提出了第一个基于假客户的模型中毒攻击，称为MPAF。

具体来说，我们假设攻击者将假客户端注入到联合学习系统中，并在训练期间将精心制作的假本地模型更新发送到云服务器，使得所学习的全局模型对于许多不加选择的测试输入具有较低的准确性。
为了实现这个目标，我们的攻击将全局模型拖向攻击者选择的低精度基础模型。具体来说，在每一轮联合学习中，假客户端都会制作指向基本模型的假本地模型更新，并在将它们发送到云服务器之前放大它们的影响。我们的实验表明，即使采用经典防御和范数剪裁，MPAF也会显著降低全局模型的测试精度，这凸显了对更高级防御的需求。

我们提出了MPAF，它只根据全球模型来制作虚假的本地模型更新。

具体来说，在MPAF，攻击者选择一个任意的模型(称为基础模型),该模型与全局模型共享相同的体系结构，并且具有较低的测试准确性。例如，攻击者可以随机初始化一个模型作为基础模型。
因此，在每一轮FL中，伪客户端通过从基础模型中减去当前全局模型来生成伪局部模型更新的方向。然后，假冒客户端扩大假冒本地模型更新的幅度，以扩大它们在全局模型更新中的影响

贡献：

(1):首次研究了基于虚假客户端的模型中毒攻击。
(2):提出了MPAF，这是一种新颖的无目标模型中毒攻击，它基于虚假客户端，除了在训练期间接收到的全局模型之外，不需要关于FL系统的额外知识。
(3):在多个数据集和多种FL方法上评估MPAF。结果表明，MPAF是有效的，表明我们的攻击即使在应用经典防御和规范裁剪时也是有效的。

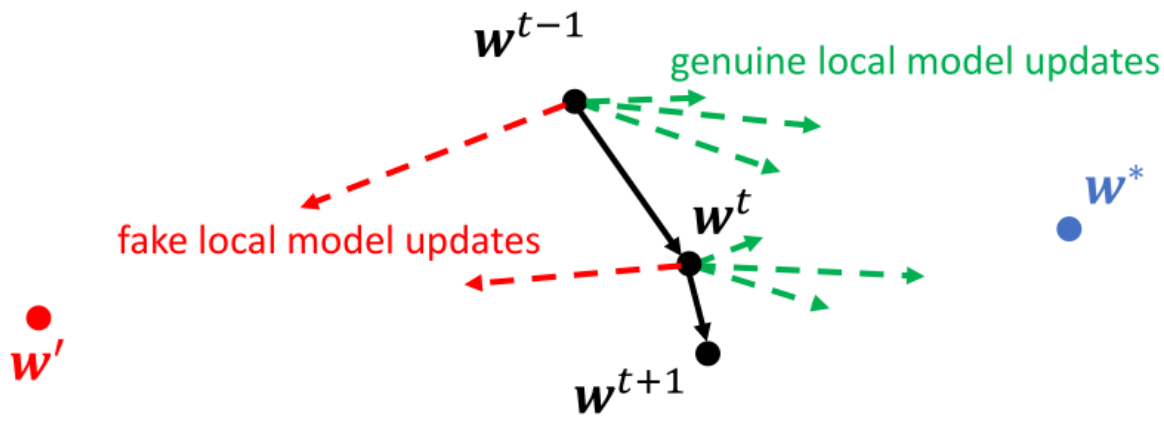
MPAF:

攻击者可以选择一个随机初始化的模型作为基础模型，其测试精度接近随机猜测。

在MPAF，假客户精心制作他们的本地模型更新，将全球模型拖向基础模型。具体而言，在第t轮FL中，假冒客户端生成假冒的局部模型更新，其方向通过从基本模型参数中减去当前全局模型参数来确定。然后，假客户将它们的假本地模型更新放大一个因子λ，以放大它们的影响。

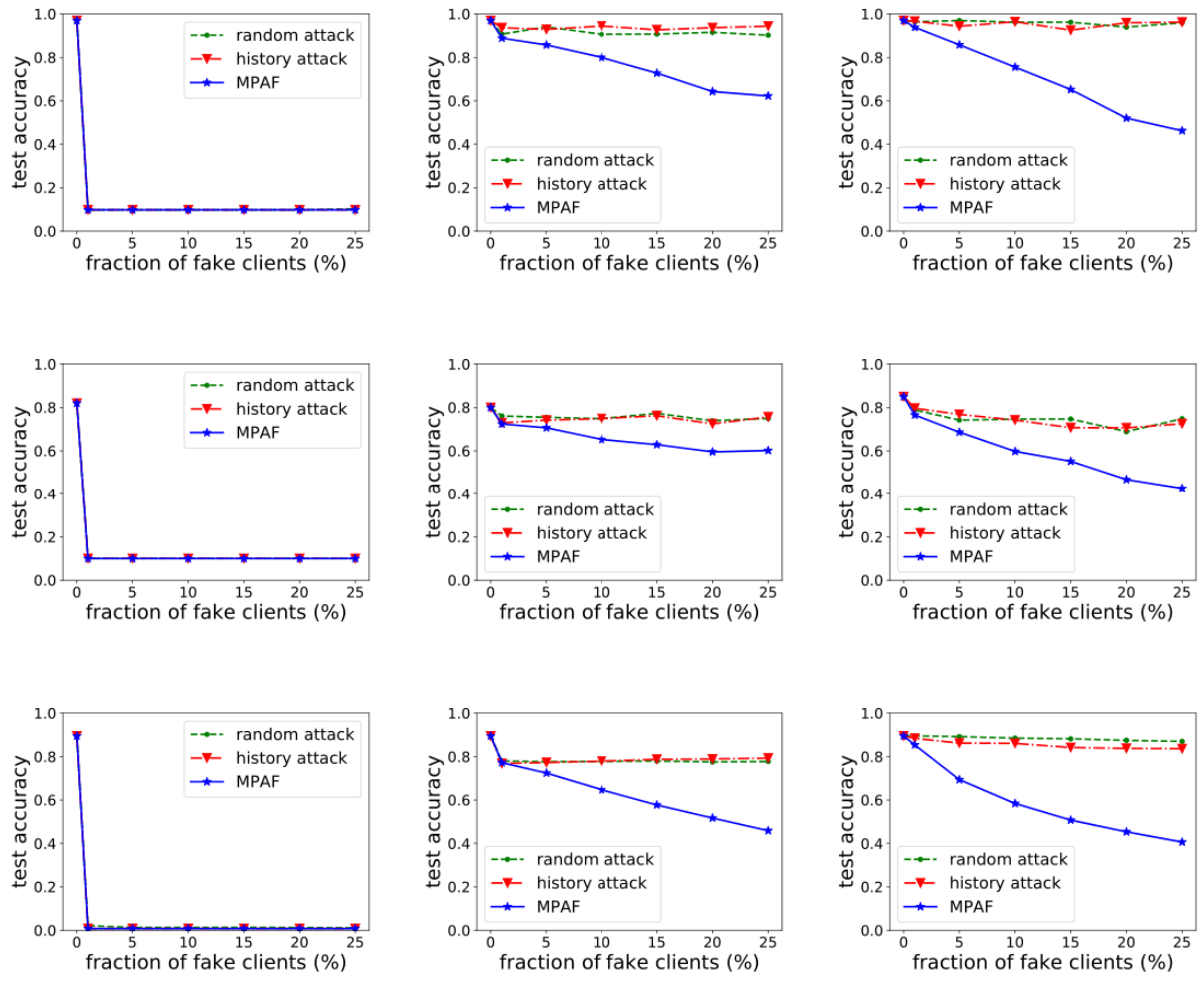
在FL的第t轮中，伪客户端通过从基础模型中减去当前全局模型来计算伪局部模型更新的方向，如果偏离这个方向，全局模型更接近基础模型。然后，假客户将放大一个因子λ，以放大幅度。攻击者可以选择较大的λ，以保证在云服务器聚合了来自假冒客户端的假冒本地模型更新和来自正版客户端的正版本地模型更新之后，攻击仍然有效。

来自伪客户端的伪本地模型更新将全局模型拖向基础模型。



$$\min_{\boldsymbol{g}_i^t, i\in[n+1,n+m], t\in[0,T-1]} \|\boldsymbol{w}^T - \boldsymbol{w}'\|,$$

攻击效果：



(a) FedAvg (b) Median (c) Trimmed-mean

规范裁剪：

规范裁剪作为联邦学习中对抗后门攻击的对策。

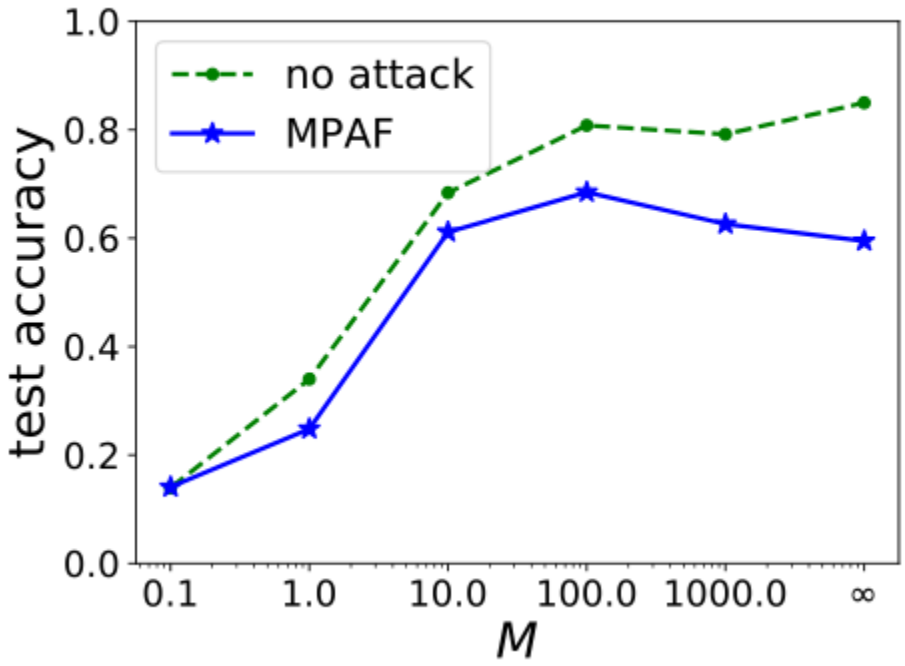
服务器选择范数阈值M，并且裁剪所有范数大于M的局部模型更新，使得它们的范数变为M。数不大于M的局部模型更新保持不变。

剪切的局部模型更新的最大范数是m。因此，恶意局部模型更新的影响将是有限的。所

以，当采用规范裁剪作为对策时，依赖于缩放的本地模型更新的后门攻击将具有较低的攻击成功率。

规范裁剪作为对策的想法并不局限于后门攻击。还可以作为一种对策来应对涉及扩展的无目标攻击。在MPAF，我们使用比例因子λ来增加聚集过程中假本地模型更新的影响。因此，应用规范裁剪作为针对MPAF的对策是直观的。

用M → ∞来表示没有范数剪裁的情况。，当使用范数剪裁时，MPAF仍然可以有效地降低全局模型的测试精度。
具体来说，在没有攻击的情况下，全局模型在M → ∞时达到最大的测试精度0.85。然而，在MPAF下，当M在100左右时，全局模型实现了0.68的最大测试精度，这表示0.17的精度损失。
全球MPAF模型和无攻击模型随着M的减小而变小。这是因为随着M减小，更多的假局部模型更新被剪切。然而，随着M的降低，无攻击下的测试精度也降低，M < 100导致测试精度比M → ∞时低得多。这是因为当M减小时，更良性的局部模型更新也被剪切，这导致更不精确的全局模型。



结果表明，MPAF仍然有效地降低了全球模型的测试精度，即使经典的防御(例如，修剪平均)和规范剪辑都被采用。

结论和讨论：

在这项工作中，我们提出了MPAF，第一个基于虚假客户端的模型中毒攻击。我们考虑了攻击者的最低知识设置，并表明我们的攻击即使在应用经典防御和规范裁剪时也是有效的，这突出了对基于假冒客户端的模型中毒攻击的更高级防御的需要。