针对垂直联邦学习的标签推理攻击

在VFL架构中，只有一个参与者拥有标签，这与HFL不同。

确保私有标签的隐私是VFL提供的基本保证，因为标签可能是参与者的关键资产或高度敏感。因此与HFL相比，VFL的敌对参与者推断私人信息更具挑战性。一般而言，HFL中的敌对参与者控制完整的本地模型，并且可以访问该模型的所有参数的梯度，这可以被滥用来推断私人信息。然而，在VFL中，敌对参与者只控制联合模型的一部分，不能独立运行，只能访问这个不完整模型的梯度。

研究发现由恶意参与者控制的底层模型可能自然地具有推断其他参与者私有标签的能力。
我们证明，在少量辅助标记数据的帮助下，恶意参与者可以以半监督的方式将他/她训练的底层模型微调为完整的标记推理模型。为了增加恶意底层模型的表达能力，从而进一步提高攻击性能，我们进一步引入了主动标签推理攻击来欺骗联邦模型更多地依赖恶意底层模型。实验表明，主动攻击成功地提高了攻击性能。

最后，我们评估了四种可能的防御措施:噪声梯度、梯度压缩、保护隐私的深度学习和离散SGD。我们发现，虽然这些防御措施中的一些可以减轻直接标签推理攻击，但对于我们的被动和主动标签推理攻击却无效。这激发了更好的防御来增强VFL的隐私。

本文贡献：

(1)揭示了VFL的新标签泄漏问题。
(2)提出了三种针对VFL的标签推理攻击，包括直接标签推理攻击、带有模型完成的被动

标签推理攻击和带有恶意局部优化器的主动标签推理攻击。这些攻击涵盖多个实际的VFL设置。
(3)使用真实世界数据集在两个参与者和多个参与者设置下评估了我们对各种任务的攻击，并获得了出色的攻击性能。此外，我们分享了关于主动标签推理攻击的潜在工作机制的见解，并提出了易于理解的证明。我们还评估了针对我们的攻击的四种可能的防御方法，发现它们并不有效，这激励了未来在更好的防御方面的工作。

VFL(垂直联邦学习)：
参与者的数据集共享相同的样本空间，但在特征空间不同。

目前有两种流行的VFL架构:没有模型分裂的VFL和有模型分裂的VFL。
在这两种架构中，都有一个保存标签的可信第三方服务器，以及拥有垂直分区数据的参与者。
VFL的每一次训练迭代都可以分为两步：
第一步是联邦正向传播。所有参与者使用他们的本地数据和底层模型进行本地正向传播，然后向服务器提交本地输出。服务器使用所有参与者的汇总输出来计算最终预测，然后计算相应的损失值。
第二步是联邦向后传播。服务器进行反向传播，并根据每个参与者的输出计算损失的梯度。梯度被发送回每个参与者。然后，每个参与者继续联合反向传播，并更新其底层模型。

对于没有模型分裂的VFL，每个参与者运行一个底层模型，服务器不运行任何模型。每个参与者的底层模型给出一个输出；然后，服务器简单地将所有输出相加得到最终输出。然而，对于无模型分裂的VFL，每个参与者都可以访问输出层，这可能带来重大的标签泄漏风险。

具有模型分裂的VFL是用分裂学习的思想设计的。整个ML模型在特定的中间层被分成顶层模型和一些底层模型，中间层被称为切割层。每个参与者运行一个底层模型，学习本地数据。服务器运行顶层模型来聚集来自每个参与者的隐藏表示，然后计算最终输出。反向传播通过共享切割层的梯度来完成。在带模型分裂的VFL中，参与者无法访问DNN的最后一层。因此，服务器上的标签更安全。

标签推理攻击：

来自训练的底部模型的泄漏和来自梯度的泄漏

1.被动标签推理攻击：

对手可以根据本地拥有的底部模型来推断标签。对手的训练过的底层模型可以将拥有的特征转换成非常有指示性的表示来预测标签。因此，对手可以使用少量的辅助标记数据，通过额外的分类层对底层模型进行微调，以进行标签推断。

在训练过程之后，对手得到训练过的底部模型。然后，对手在经过训练的底层模型之上添加额外的随机初始化层，以形成用于标签推断的“完整模型”。这些附加层称为推理头。

一旦半监督训练完成，对手就得到了一个完全训练的完整模型。该模型可以为对手的每个数据预测一个标签。这样，对手成功地用可用特征推断出任何感兴趣的样本的标签，而不管该样本是否在训练数据集中。在整个攻击过程中，对手保持“诚实但好奇”。

2.恶意局部优化器的主动标签推理攻击

对手可以主动欺骗联邦模型，使其更多地依赖其底层模型，从而增加其表达能力，即对手在训练阶段积极地做一些恶意的动作。利用底层模型更好的表达能力，对手可以训练出更准确的完整模型进行标签推理。

对手使用专门设计的恶意本地优化器，对手可以恶意地提高学习率，使其底层模型加速梯度下降，从而在每次迭代中向服务器提交更好的特征。最后，它使得顶层模型更多地依赖于对手的底层模型，而不是其他参与者。

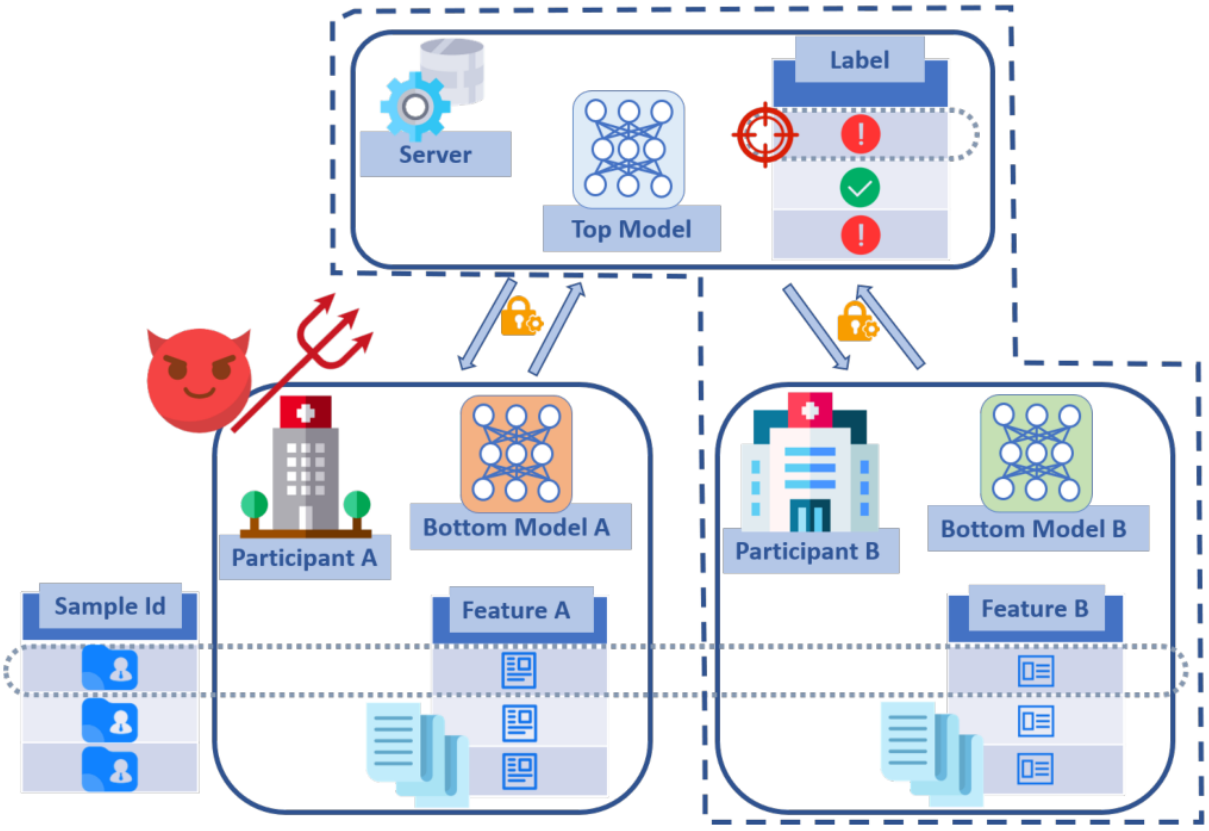
通过在训练阶段使用恶意的局部优化器，对手可以获得具有关于标签的更多隐藏信息的训练的底层模型。

3.直接标签推理攻击

对手也可以直接利用接收到的梯度来推断训练样本的标签，而不是依赖于训练的底部模型。

对手可以通过分析从服务器接收的梯度的符号来直接推断标签。直接标签推理攻击适用于用于分类任务的主流损失函数，包括交叉熵损失、加权交叉熵损失和负对数似然损失。

对于没有模型分裂的VFL，对手可以直接从服务器发回的梯度符号中推断出标签。缺点是它只能推断训练样本的标签，因为在推断时没有可用的梯度。然而，为了推断任意样本的标签，对手可以使用获得的训练样本的标签作为辅助标记数据，进一步进行设计的被动标签推断攻击。



攻击效果：

Dataset	Train Set Size	Test Set Size	Number of Classes	Known Label Quantity Per Class	Metric	Attack Performance			
						Train Set		Test Set	
						Passive	Active	Passive	Active
CIFAR-10	50,000	10,000	10	4	Top-1 Acc	0.8024	**0.8484**	0.6299	**0.6342**
CIFAR-100	50,000	10,000	100	4	Top-5 Acc	0.6267	**0.6732**	0.4319	**0.4700**
CINIC-10	180,000	90,000	10	4	Top-1 Acc	0.7206	**0.7818**	0.5440	**0.5995**
Yahoo Answers	50,000	20,000	10	10	Top-1 Acc	0.6335	**0.6424**	0.6370	**0.6419**
Criteo	80,000	20,000	2	50	Top-1 Acc	0.6828	**0.6879**	0.6785	**0.6830**
BHI	69,181	17,296	2	35	F1 Score	0.7614	**0.7824**	0.7519	**0.7673**

防御：

针对标签推理攻击的四种可能的防御方法:噪声梯度、梯度压缩、保护隐私的深度学习和离散SGD

噪声梯度：在梯度中添加噪声是FL中一种常见的防御策略。在VFL，服务器可以在将梯度发送给参与者之前添加拉普拉斯噪声。

梯度压缩：梯度压缩是一种为通信效率和隐私保护而设计的策略。核心思想是只共享一部分绝对值最大的梯度。即使只有10%的梯度值被共享，HFL仍然可以产生高精度的全球模

型。

保护隐私的深度学习：隐私保护深度学习是中介绍的一种全面的隐私增强方法。它包括三种防御策略:差分隐私、梯度压缩和随机选择。在每次迭代中，服务器执行以下步骤来保护梯度:
(1)随机选择一个梯度值，生成噪声，并将噪声添加到梯度值中；
(2)如果加入噪声后的梯度值大于阈值τ，则保持该值，否则将其设置为零；
(3)循环前两步，直到收集到梯度值的θu分数。θu和τ都是超参数，用于平衡模型性能和防御性能之间的权衡。

离散SGD：由于它只保留梯度的迹象，它进一步提高通信效率和加强隐私保护。具体地，服务器首先观察第一时段中共享梯度的分布。分布的均值和标准差分别表示为μ和σ。服务器将间隔设置为[2σ，+ 2σ]。区间之外的梯度被视为异常值并被丢弃。然后，服务器将该时间间隔分成N个子时间间隔。在下面的训练过程中，在向参与者发送梯度之前，服务器首先将每个梯度值舍入到子区间的最近端点。超参数N控制保留多少共享梯度的幅度信息。

结论：

实验表明，在真实世界的大规模数据集上，所提出的标签推理攻击对VFL是有效的。我们还评估了可能的防御措施，包括梯度压缩、噪声梯度、隐私保护深度学习和离散SGD。虽然这些防御措施中的一些可以有效地减轻直接标签推理攻击的威胁，但是它们对于我们的被动和主动标签推理攻击是无效的。这将激励未来更好的防御工作。