联邦学习的深度模型中毒攻击

本文对联邦学习中的此类威胁进行了系统调查，并提出了一种新的基于优化的模型中毒攻击。与现有方法不同，本文主要关注攻击的有效性、持久性和隐蔽性。数值实验表明，所提出的方法不仅可以实现较高的攻击成功率，而且具有足够的隐蔽性，可以绕过现有的两种防御方法。

在现实世界中，联邦学习系统可能包含数百万个客户端，并且每轮只会选择一小部分客户端的更新。因此，模型中毒攻击缺乏持久性可能会导致攻击失败。
为了应对这些挑战，在本文中，我们提出了一种新的基于优化的联邦学习模型中毒攻击。通过利用模型容量并在模型冗余空间中注入中毒神经元以提高攻击的持久性。

主要贡献：

（1）通过分析模型容量，我们提出了一种基于优化的模型中毒攻击，并将对抗性神经元注入神经网络的冗余空间。需要注意的是，这些冗余神经元对于中毒攻击很重要，而它们与联邦学习的主要任务相关性较小。因此，所提出的模型中毒攻击不会降低共享全局模型上主要任务的性能。

（2）我们概括了协作学习系统中用于防御局部模型中毒攻击的两种防御措施。数值实验表明，该方法可以绕过防御方法并获得较高的攻击成功率。

攻击：

冗余空间：在训练阶段只有一小部分神经元发生了变化，而大部分神经元接近于零。那些不变的神经元被称为学习任务的神经网络的冗余空间。

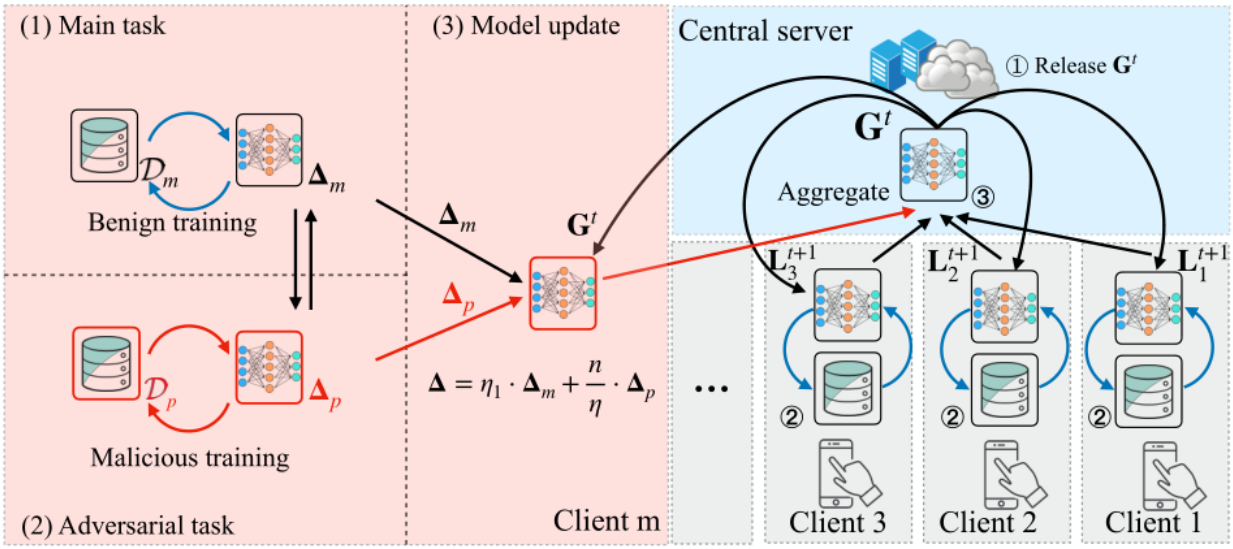
恶意的客户端拥有良性数据集Dm和中毒数据集Dp,攻击者利用 Dm 进行良性训练，在主要任务中保持局部模型的高性能，同时使用 Dp 微调神经网络以进行对抗任务。恶意客户端分别使用 Dm 和 Dp 交替细化主要任务和对抗性任务，这称为交替最小化。

核心思想：

在优化器的指导下将中毒神经元嵌入到神经网络的冗余空间中，以保持攻击的隐蔽性和持久性。

恶意的客户端拥有良性数据集Dm和中毒数据集Dp,攻击者利用 Dm 进行良性训练，在主要任务中保持局部模型的高性能，同时使用 Dp 微调神经网络以进行对抗任务。恶意客户端分别使用 Dm 和 Dp 交替细化主要任务和对抗性任务，这称为交替最小化。

对手交替训练 Dm 和 Dp 的小批量，分别用于主要任务和对抗任务。最后，攻击者将中毒模型更新 Δp 提升到中央服务器。



模型中毒攻击的目标是在共享全局模型上的任何指定点上造成有针对性的错误分类，同时保持测试数据集 Dtest 上的预测准确性。

深度模型中毒攻击的流程，它由主任务、对抗任务和模型更新模块组成。
在主要任务中，我们利用良性训练来保持局部模型 的高性能，并约束神经网络权重以使中毒模型类似于良性客户端。

在对抗任务中，我们打算将对抗特征嵌入到神经网络的冗余空间中，以提高攻击的持久性。遵循交替最小化策略，我们在模型收敛之前迭代优化主要任务和对抗性任务。

（1）主要任务：

安全的中央服务器可以利用其辅助知识来检查提交的模型更新的两个关键属性。

首先，安全中央服务器可以评估在验证集上提交的模型更新的准确性。

其次，安全中心服务器可以通过简单的统计方法验证模型更新，以拒绝异常模型。

我们为主要任务提出了约束损失函数，以绕过中央服务器中的那些防御方法。
对于第一个挑战，恶意客户端交替训练一小批清洁样本 Dm 和中毒样本 Dp，以保持提交的模型更新的高性能。
对于第二个挑战，在损失函数中添加一个正则化项，以使恶意客户端的模型更新与统计数据中的良性相似。
总体而言，主要任务的对抗目标变为：

$$\arg\min_{\Theta^\star} = \mathcal{L}_M(\mathcal{D}_m;\Theta^\star) + \rho_1 \left\| \Delta_m^{t+1} - \overline{\Delta}_{\text{ben}}^{t} \right\|_2$$

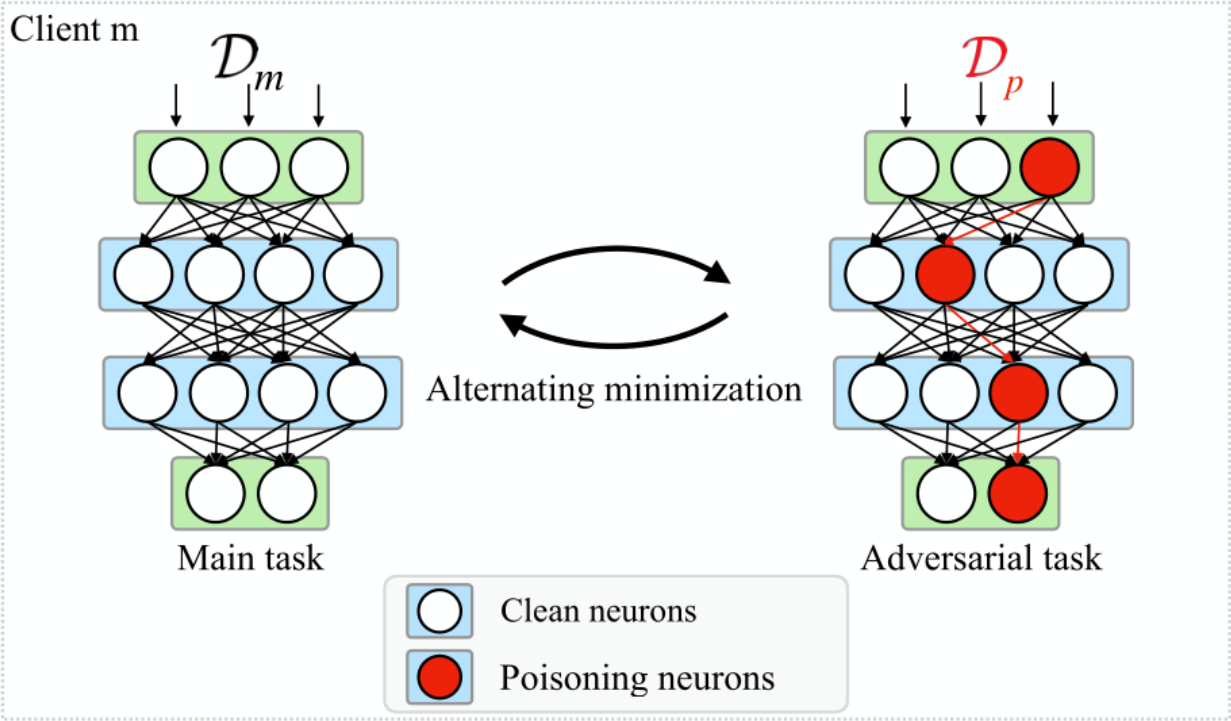
其中 ρ1 是超参数，LM 表示主要任务的交叉熵损失，Δt ben 表示良性客户端的平均模型更新。Δt ben 是使用共享全局模型估计的。
假设聚合模型类似于良性客户端的本地模型，因为共享全局模型收敛到一个测试精度高的点。

（2）对抗性任务：

灾难性遗忘是神经网络模型的必然特征，特别是当网络在多个任务上按顺序训练时。在联邦学习系统中，由于对抗任务学习到的知识会随着中央服务器聚合模型更新而突然微调和丢失，很难同时保持模型中毒攻击的持久性和有效性。

然而，当模型收敛时，大多数神经元在微调过程中几乎没有变化，这些神经元是神经网络中主要任务的冗余空间，可以作为注入对抗任务的最佳位置。它可以用于我们的对抗训练。

通过在冗余空间中嵌入中毒神经元来设计我们的攻击策略。攻击者将这些神经元连接起来并将它们加入到对抗路径中，该路径在模型聚合下是持久且稳健的。



中毒神经元选择和注射示意图。

攻击者精心选择在主要任务期间几乎没有变化的神经元序列（红色圆圈）来执行对抗任务。

如何找到冗余空间：

在一个变量函数中，由于二阶导数衡量的是损失函数的曲率，二阶导数可以找到随机梯度下降（SGD）的优化方向。由于这一观察，通过计算二阶导数，可以找到对主要损失函数有相当大影响的神经元任务。

通过捕获多变量函数的所有二阶导数信息，Hessian 矩阵通常起到类似于单变量微积分中普通二阶导数的作用。Hessian 矩阵可以测量主任务更新的距离和方向。

因此攻击者使用其本地干净数据集 Dm 简单地计算 Hessian 矩阵（即等式（3）中的 hi）以获得损失函数的二阶导数。

Hessian 矩阵中的值越高，表示主任务中的神经元越“重要”。

通过避免在对主要任务特别有影响的位置注入中毒神经元，以减轻对抗性任务的灾难性遗忘。

在本文中，使用结构正则化项来惩罚优化器，以避免在训练对抗任务时更新这些神经元：

$$\arg\min_{\Theta^\star} = \mathcal{L}_A\left(\mathcal{D}_p; \Theta^\star\right) + \rho_2 \sum_{\theta_i \in \Theta^\star} h_i(\Delta\theta_i)^2$$

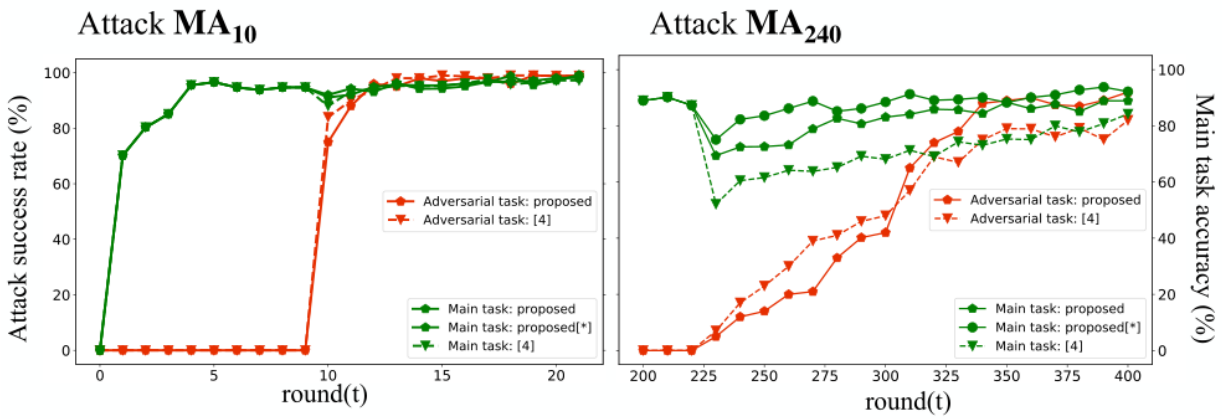
其中 ρ2 是超参数，hi 是与主要任务（即 LM）相关的目标的二阶导数，Δθi 表示对抗性任务的参数更新。通过利用正则化项，攻击者最终可以构建一条对抗路径，在交替最小化下保持对抗和主要任务的准确性。

模型更新：

交替使用优化方程（2）和（3）。对于每一步 i，对抗任务的目标函数首先使用方程（2），然后用方程（3）最小化。最后，攻击者将具有因子 n/η 的恶意模型更新 Δp 提升到下一个 epoch：

其中 Δm 是从方程 (2) 获得的模型更新，Δp 是从方程 (3) 获得的模型更新

$$\boldsymbol{\Delta} = \eta_1 \boldsymbol{\Delta}_m + \frac{n}{\eta} \boldsymbol{\Delta}_p$$



(a) MNIST (b) CIFAR-10

实验结果：

在实验中，我们展示了所提出的模型中毒攻击的性能评估。所提出的攻击方法在多次攻击（MAt）和单次攻击（SAt）场景下进行了评估。

在 MAt 场景中，恶意客户端在前 t 轮表现得像良性客户端，然后每轮发起模型中毒攻击。相比之下，在 SAt 场景中，恶意软件仅在 t 轮发起单次攻击。在本实验中，假设每轮都选择恶意客户端。
在 MAt 场景中评估攻击的有效性，同时在 SAt 场景中验证攻击的持久性。

在多发攻击场景中，所提出的攻击可以有效地执行对抗任务，而不会降低主要任务的准确性。即使对于像 CIFAR-10 这样的大型数据集，所提出的模型中毒攻击对于对抗性任务的攻击成功率也可以达到 90% 以上，提出的深度模型中毒攻击可以对 Dp 实现高攻击成功率，同时保持其对测试图像的准确性。

在单次攻击场景中，攻击成功率下降缓慢，这意味着所提出的攻击在减轻对抗性任务的灾难性遗忘方面是有效的，说明所提出的深度模型中毒攻击能够在长时间的聚合中保持较高的攻击成功率。所提出攻击的高性能背后的主要原因是中毒神经元嵌入在神经网络的冗余空间中。这降低了在主要任务训练期间微调中毒神经元的概率。

结论：

由于每轮只会选择一小部分客户端更新，因此联邦学习中的模型中毒攻击的有效性、持久性和隐身性变得越来越具有挑战性。

为了克服这些困难，我们在本文中提出了一种新的联邦学习深度模型中毒攻击。通过利用目标函数中的正则化项，我们将恶意神经元注入神经网络的冗余空间。

广泛的评估结果经验证明，我们提出的攻击策略通过显着提高有效性、持久性和鲁棒性方面的性能而优于后门攻击。在我们未来的工作中，我们打算进一步研究深度神经网络的范式并探索一种有效的机制，以找到注入对抗任务的神经元。