《Poisoning Attack in Federated Learning using Generative Adversarial Nets》

文章的主要思想是：

设计并构建了一种基于 GAN 的新型中毒攻击。 攻击者首先充当良性参与者，然后秘密训练 GAN 来模仿不属于攻击者的其他参与者的训练集的原型样本。 然后，这些生成的样本将由攻击者完全控制以生成中毒更新，并且攻击者会将已缩放的中毒更新上载到服务器时破坏全局模型。 本文实验表明，攻击者可以使用 GAN 成功生成其他良性参与者的样本，并且全局模型在中毒任务和主要任务上的准确率均超过 80%。

攻击目标：

1. 样本生成：攻击者无需直接访问其本地数据集即可成功模仿来自良性参与者训练数据的原型样本。
2. 中毒任务的精度。
3. 模型主要任务的精度。

攻击流程：

1. 首先基于生成对抗网络（GAN）模型生成中毒样本，并将这些样本添加到本地训练数据集中。
2. 将生成的样本注入错误的标签，进行训练以计算中毒的局部参数。
3. 将放缩后的参数上传到中央服务器。

Algorithm 1: Poison Attack in Federated Learning.

Input: Global model M_t; Participants' updates ΔL_t^i;
Loss function ℓ; Learning rate η.
Output: Poisoned updates $\Delta \hat{L}_t^i$.
Initialize generator G and discriminator D
for $t \in (1, 2, \cdots, T)$ **do**
  // Server execution
  Send M_t to the participants
  Recive updates from participants: ΔL_{t+1}^i
  Update the globa model: M_{t+1}
  // Participants execution
  Replace the local model: $L_t^i \leftarrow M_t$
  if *the user type is* $\mathcal{A}$ **then**
    Initialize D by the new local model L_t^i
    for *each epoch* $e \in (1, \cdots, E)$ **do**
      Run G on D for targeted class
      Using D to update G
      Generate samples of targeted class by G
      Assign wrong label to generated samples
      Insert poison data to the local dataset $\mathcal{D}$
      for *each batch* $b_p \in \mathcal{D}_{poison}$ **do**
        $L_{t+1}^p = L_{t+1}^p - \eta_{adv} \nabla \ell(L_t^p, b_p)$
      end
    end
    Calculate poisoned update: $\Delta L_{t+1}^p = L_{t+1}^p - L_t^p$
    Scale up the update: $\Delta \hat{L}_{t+1}^p = \lambda \Delta L_{t+1}^p$;
  end
  else
    Calculate benign update: $\Delta L_{t+1}^i = L_{t+1}^i - L_t^i$;
  end
  Upload the local update ΔL_{t+1}^i (including $\Delta \hat{L}_{t+1}^p$) to the central server
end

GAN：鉴别器 D 和生成器 G。

1. 在攻击端，采用 GAN 架构创建一个全局模型的副本作为鉴别器。随着联合学习训练过程的进行，鉴别器同步更新。
2. 生成器将生成的图像发送给鉴别器，以判断生成图像的真伪。
3. 攻击者加入联邦学习并获取最新的全局模型以更新鉴别器。
4. 重复上述步骤，直到生成器能够生成高质量的模拟样本。

Fig. 3. Results generated when attacker runs a GAN trained on MNIST and AT&T datasets.

（MNIST 数据集、AT&T 灰度人脸数据集）

文章特点：

在攻击端部署 GAN，从而生成良性客户端所具有样本的模拟样本，进而进行投毒攻击。特点是，随着联邦学习的进行，全局模型精度增加，GAN 的鉴别器也随之变得更强大，这使得生成器得到了有效的学习，进而生成精度较高的模拟样本。