使用生成式对抗网络的联邦学习投毒方法

2024 年 5 月 3 日

目录

报告要求: 从以下课题中任选 1 个课题，调研、学习、并撰写一个课程技术报告。
可选课题:

- AI 模型攻击方法及实验实践
- AI 模型防御方法及实验实践
- 去隐私技术及实验实践

1 引言

1.1 联邦学习

传统的中心化的机器学习模型的训练需要用户上传自己所有需要训练的数据, 用户的隐私存在泄露的风险. 相反, **联邦学习 (Federated Learning)** 是一种新型的分布式学习框架, 通过多个协作者共同参与完成模型的学习. 联邦学习的本质是分布式的机器学习, 服务器和参与者 (模型学习者) 共享模型的参数, 服务端无法获得参与者的训练数据, 不同参与者的训练数据也是不同的, 参与者的训练数据和训练过程是本地化的, 对服务端保密. 服务端的任务是维护全局模型, 接收来自内部参与者对模型参数的更新, 并随机选择一定数量的参与者运用均值化算法更新全局模型, 直至收敛.

设有 n 个用户参与联邦学习, 且学习目标都为这 n 个参与者所知, 这 n 个用户本地的数据集都是互不相同的. 对模型的的每次迭代, 用户从服务端下载全局模型的参数, 并根据本地的数据集训练模型. 每个用户在训练阶段会将梯度上传至服务端, 服务端将来自多个用户的梯度进行平均并累积到当前的全局模型上. 下面是对这一过程的形式化描述:

$$m_{i+1} = m_i + \frac{1}{n}\sum_{k=1}^{n} g_i^k \tag{1}$$

其中, m_{i+1} 表示在第 i 轮迭代时共享的模型 (参数), g_i^k 表示第 k 个用户在第 i 轮迭代上传的梯度.

联邦学习的主要目的是在本地化数据集的基础上, 构建一个联合的机器学习模型, 同时保护数据集的隐私. 联邦学习在边缘计算和众包系统中有很重要的应用.

1.2 投毒攻击

在联邦学习的架构中, 由于参与者的训练数据和训练过程对服务端是不可见的, 全局模型的直接性能受内部的参与者的影响. 内部参与者可以发起主动的**投毒攻击 (Poisoning Attack)**, 即对模型的参数进行不正确的更新并上传至服务器, 影响全局模型的性能.

在这里, 我们研究和评估联邦学习系统中基于**生成式对抗网络 (Generative Adversarial Nets, GAN)** 的投毒攻击.

在训练阶段, 攻击者首先作为内部参与者, 在本地训练 GAN 模拟其它参与者的训练样本 (不属于此攻击者), 攻击者使用这些生成的模拟样本数据对模型进行投毒更新, 通过增大投毒训练的规模并将训练后的模型上传至服务器可影响全局模型的性能, 使全局模型拥有某些攻击者所期望的性质. 在推理阶段, 受到投毒攻击的模型则可能对某些给定的输入获得攻击者想要的输出而非实际上应该正确的输出.

以下几点原因使得投毒攻击是有可能的:

- 联邦学习系统中经常有大量的参与者, 而有些参与者可能成为攻击者;
- 参与者的本地化的训练数据和训练过程对服务端是不可见的, 服务端无法验证来自参与者对模型参数更新的正确性;
- 来自多个不同的参与者的对模型的本地更新可能是完全不同的, 并且**安全聚合协议 (Secure Aggregation Protocol)** 会使得这些来自参与者的本地更新对服务端来说不可审计.

1.3 生成对抗网络

生成对抗网络 (GAN) 在计算机视觉研究领域取得了极大的成功, 能够基于原始图像集生成高质量的假图像 (`fake image`).

GAN 结构中有两个神经网络:

- **生成器 (Generator)**: 生成器 (G) 生成图像.
- **鉴别器 (Discriminator)**: 鉴别器 (D) 判别图像是来自生成器还是原始图像集, 它们可以表示为 0/1(假/真).

为了训练一个 GAN, G 从遵循高斯或均匀分布的先验分布的随机向量 z 中生成图像。生成的图像是 D 的输入。同时, D 通过原始图像集进行预训练, 可用于区分输入是真实的还是虚假的。通过进行对抗训练的博弈, G 和 D 的性能都可以得到提高。下面是 GAN 的训练目标:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2}$$