基于生成式对抗网络的联邦学习后门攻击方案

作者：陈大卫、付安民、周纯毅、陈珍珠 时间：2021

期刊：计算机研究与发展 第11卷58期

前言：

- 联邦学习使用户在数据不出本地的情形下参与协作式的模型训练，降低了用户数据隐私泄露风险，广泛地应用于智慧金融、智慧医疗等领域．但联邦学习对后门攻击表现出固有的脆弱性，攻击者通过上传模型参数植入后门，一旦全局模型识别带有触发器的输入时，会按照攻击者指定的标签进行误分类。因此针对联邦学习提出了一种新型后门攻击方案Bac_GAN，通过结合生成式对抗网络技术将触发器以水印的形式植入干净样本，降低了触发器特征与干净样本特征之间的差异，提升了触发器的隐蔽性，并通过缩放后门模型，避免了参数聚合过程中后门贡献被抵消的问题，使得后门模型在短时间内达到收敛，从而显著提升了后门攻击成功率。

主要内容：

- 联邦学习将深度学习模型与分布式训练相结合，使得多方用户在不共享数据的情况下，协同参与训练全局模型，降低了传统集中式学习中的用户隐私泄露风险和通信开箱，从技术层面可以打破数据孤岛，明显提高深度学习的性能，能够实现多个领域的落地应用，比如智慧医疗、智慧金融、智慧零售和智慧交通等。但联邦学习数以万计的用户中可能存在恶意用户，并且用户的本地训练过程对于服务器不可见，服务器无法验证用户更新的正确性，特别是服务器采用加权平均算法对参数进行更新，限制了异常检测的使用，这些缺陷的存在使得联邦学习框架极易遭受投毒攻击、对抗样本攻击和后门攻击。
- 针对后门攻击中收敛速率较慢以及触发器与干净样本差异较大而易被检测问题，本文提出了一种新型联邦学习后门攻击方案Bac_GAN，通过采用生成式对抗网络技术设计了１个触发器生成算法Trig_GAN，能够降低触发器样本与训练样本间的差异，从而提升触发器的隐蔽性．并且通过良性特征混合训练以及缩放后门模型大幅缩短模型的收敛速率，从而显著提升后门攻击成功率。
- 该文的主要贡献为：
    - 设计了１个新的基于生成式对抗网络的触发器生成算法Trig_GAN，从联邦学习样本中直接生成触发器，将触发器以水印的形式植入干净样本，明显降低了触发器样本与训练样本的差异，从而提升了触发器的隐蔽性。
    - 基于设计的触发器算法，提出了一种新型联邦学习后门攻击方案Bac_GAN，该方案通过良性特征混合训练保证了后门任务与正常分类任务的精度．特别是，通过缩放后门模型，避免了参数聚合过程中后门贡献被抵消的问题，使得后门模型能够短时间内达到收敛，进而提升后门攻击成功率。
    - 通过从触发器生成、水印系数、缩放系数等后门攻击核心要素进行了实验测试，给出了影响后门攻击性能的最佳参数，并与在数据集上对比现有典型后门攻击方案，实验证明Bac_GAN 能够显著提升后门攻击的收敛速率与攻击成功率。

- 联邦学习后门攻击模型：

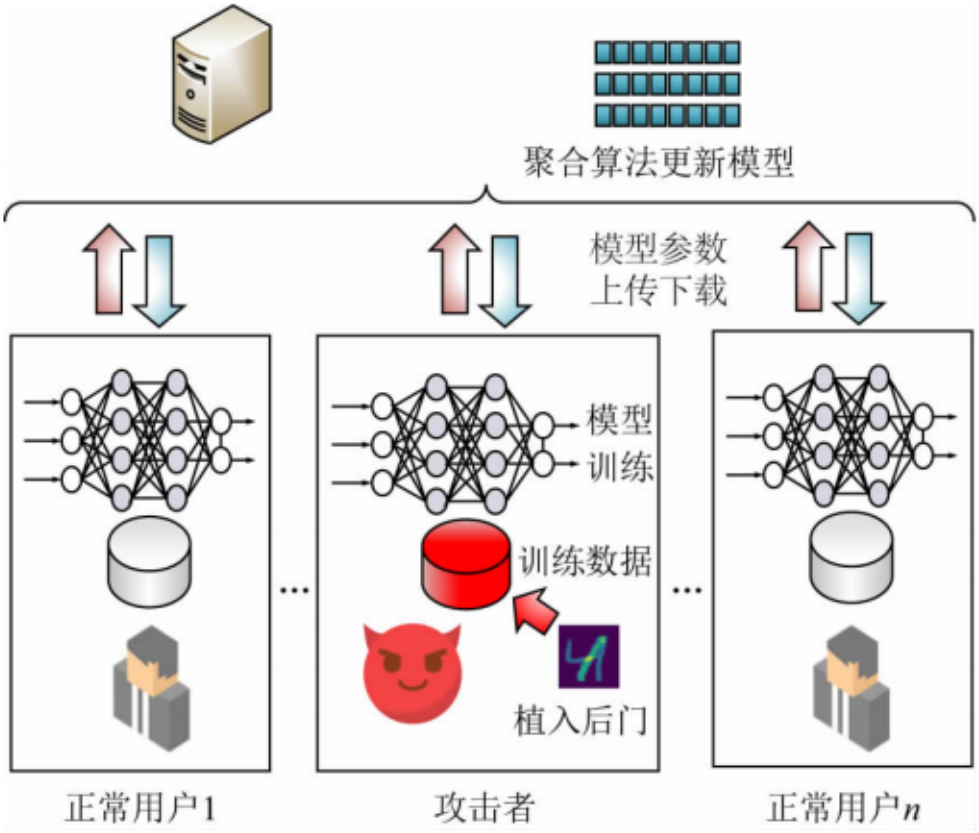

Fig. 2　The Backdoor attack model of federated learning

图 2　联邦学习后门攻击模型

- 联邦学习模式下的GAN模型：

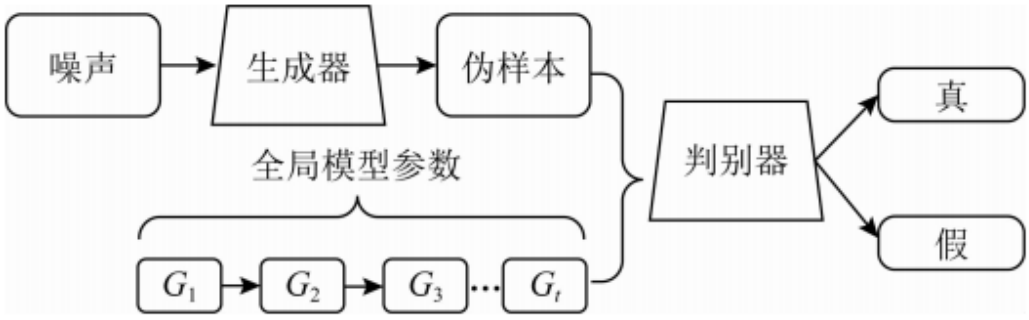

Fig. 3　GAN model of federated learning mode

图 3　联邦学习模式下 GAN 模型

结论：

- 该文提出的攻击方案，降低了触发器特征与干净样本特征之间的差异，并通过以水印的方式添加至干净样本，提升了触发器的隐蔽性。同时，通过缩放模型技术使得后门模型在短时间内达到收敛，从而提升了攻击成功率。