Title

Subtitle

Author

Date

Abstract

# Heading 1

## Heading 2

### Heading 3

#### Heading 4

##### Heading 5

###### Heading 6

Heading 7

Heading 8

Heading 9

First Paragraph.

Body Text. Body Text Char. Verbatim Char .  [Hyperlink](http://example.com)  . Footnote. [[1]](#footnote-0)

Block Text.

Table caption.

|  |  |
| --- | --- |
| Table | Table |
| 1 | 2 |

Image Caption

DefinitionTerm

Definition

DefinitionTerm

Definition

# 对攻击和错误的鲁棒性

现代机器学习系统比较容易受到各种错误的影响。这些错误包括一些非恶意性错误比如预处理流程中的漏洞、噪音过强的训练标签和不可靠的用户，还包括一些旨在破坏系统训练过程和部署流程的显示攻击。在这一章节，我们将反复看到由联邦学习的分布性本质、体系设计和数据限制所产生的新型错误模式和攻击表面。除此之外，我们还将看到联邦学习中保护隐私的安全机制使得检测和排除这些错误、攻击成为了一项特别具有挑战性的任务。

虽然这些挑战的影响可能会使系统的鲁棒性难以实现，但我们依然能看到许多非常具有潜力的研究方向。我们将会对这些方向和它们如何在联邦学习环境中进行应用和改进做出讨论。我们还将讨论有关不同类型的攻击和失败之间的联系以及这些关联在联邦学习中重要性的一些普遍问题。

本篇章首先在5.1节对对抗攻击进行讨论，之后在5.2节介绍非恶意错误类型，最后5.3节是对隐私保护和系统鲁棒性之间矛盾的探讨

## 对模型性能的对抗攻击

在这一节，我们首先描述敌人的目标和能力，之后概括了联邦学习中一些主要的攻击模式，最后列出了一些该领域尚未解决的问题。我们用术语“对抗攻击”来描述对联邦学习训练和推理过程的恶性修改，该修改被设计来降低系统性能。任何实施对抗攻击的代理人都被称为“敌人”。我们注意到虽然一般来说术语“对抗攻击”只用来描述推理阶段的攻击（有时与所谓的“对抗样本”互换使用），但我们将“对抗攻击”这一概念进行了拓展。我们还注意到一些敌人的目的并不是降低模型性能，而是试图推断出其他用户的隐私数据。这些数据推断攻击已经在第4章进行了深入讨论。因此，在本篇章中，我们将使用“对抗攻击”来指代对模型性能的攻击，而非数据推断方面的攻击。

对抗攻击的一些例子包括数据中毒 [63, 277]、模型更新中毒 [42, 61]和模型回避攻击[377,63,186]。这些攻击可以大致分为训练阶段攻击（中毒攻击）和推断阶段攻击（回避攻击）。与分布式数据中心机器学习和中心化学习体系相比，联邦学习主要的不同点在于其模型的训练是基于一批具有私密、不可检查数据集的不可靠设备进行的。然而，使用已部署模型的推理在很大程度上保持不变(有关这些和其他差异的更多讨论，请参见表1)。因此，联邦学习可能会在训练阶段引入一些新的攻击表面。训练模型的部署通常依赖应用程序，且一般与所使用的学习范式（中心化、分布式、联邦式等）相互独立。尽管如此，我们依然会在接下来的内容中讨论推断阶段的攻击因为（a）一些在训练阶段的攻击会被用来做推断阶段攻击的垫脚石 [277, 61]，（b）许多对推断阶段攻击的防卫措施是在训练阶段实施的。因此，新型的对联邦学习系统的攻击向量可能会结合一些新型推断阶段对抗攻击。我们将在5.1.4节更详细地讨论这一问题。

### 敌人的目标和能力

在这一小节，我们将研究敌人的目标、动机和不同的能力（一些能力针对联邦学习环境）。我们将研究敌人能力的不同维度，并在不同的联邦学习环境（见第一章表格1）下考虑它们。正如我们将要讨论的，不同的攻击方案和防御方法具有不同程度的适用性和倾向兴趣，这取决于联邦学习上下文。特别的，联邦学习设置的不同特征能够影响敌人的能力。比如，在跨设备设置下，仅仅控制了一个用户的敌人可能是微不足道的，但在跨竖井联邦设置下却会产生重大影响。

**目标** 在一个较高水平上，对机器学习模型的对抗攻击试图以一种对模型不利的方式修改模型行为。我们发现一个攻击的目标通常是对模型不利的调整范围和调整区域，通常有两个级别的范围：

1. 无针对性攻击，或模型降级攻击，该类型攻击的目标是降低模型的全局精度或全面“摧毁”全局模型[63]。
2. 针对性攻击，或后门攻击，该类型攻击的目标是改变模型对少数例子的行为，同时保持对所有其他例子的良好的整体准确性 [100, 277, 42, 61]。

例如，在图像分类任务中，一个针对性攻击可能会在一组“绿色汽车”的训练图像中添加一个小的视觉工件(后门)，以便让模型将这些图像标记为“鸟类”。然后，训练好的模型将学会将可视化工件与类“bird”相关联。之后可以利用这一点来发动简单的回避攻击，方法是在一辆绿色汽车的任意图像中添加相同的视觉效果，将其归类为“鸟”。模型甚至可以在不需要修改目标推理阶段输入的情况下被植入后门。Bagdasaryan et al [42] 引入了“语义后门”，其中，敌人的模型更新迫使训练过的模型在一小部分数据上学习错误的映射。例如，敌人可以强迫模型将所有绿色的汽车归类为鸟类，从而在推断时产生错误的分类 [42]。

虽然上面的讨论表明了非针对性攻击和有针对性攻击之间的明显区别，但在现实中，这些目标之间存在某种连续性。虽然纯粹的非针对性攻击的目标可能只是降低模型精度，但更细微的非目标攻击可能会将针对除一小部分客户端数据外的所有数据降低模型精度作为目标。这又开始类似于一种有针对性的攻击，其中后门的目的是在少数例子上夸大模型相对于其余评估数据的准确性。类似地，如果一个对手对数据的某个特定特征进行了有针对性的攻击，而该特性恰好出现在所有评估示例中，那么他们(可能是无意中)就设计了一个无针对性的攻击(相对于评估集)。虽然这种连续性对于理解对抗学习的整体面貌非常重要，我们在接下来的部分将只介绍纯粹的针对性攻击和非针对性攻击。

**能力** 与此同时，当对手试图在训练中破坏模型时，他们可能拥有各种不同的能力。我们要注意到联邦学习提出了许多关于敌人会拥有什么能力的问题。清晰地定义这些能力对联邦学习社区权衡各种防守措施的价值十分重要。在表11中，我们提出了一些十分重要的能力轴。

*表11 联邦学习环境下敌人的能力特征*

|  |  |
| --- | --- |
| **特征** | **描述** |
| 攻击向量 | 敌人发起攻击的方式   * *数据中毒*： 敌人修改用来训练的用户数据集 * *模型更新中毒*：敌人更新发送回服务器的模型更新数据 * *回避攻击*：对手改变推断阶段使用的数据 |
| 模型检查 | 敌人是否能够观察到模型参数   * *黑箱*：对手在攻击前和攻击时都没有能力观测到模型参数。在联邦学习环境中一般不是这种情况 * *陈旧白箱*：对手只能观测到一个陈旧的模型。当对手可以接触到参加中间训练回合的客户时，这自然会在联邦环境中出现。 * *白箱*：对手有能力直接观测到模型参数 |
| 参与者串通 | 多个敌人是否可以协同发起攻击   * *无串通*：参与者无法通过串通发起攻击 * *Cross-update串通*：过去的客户端参与者可与未来的参与者协同攻击全局模型在未来的更新 * *Wthin-update串通*：当前客户端参与者可协同发起对模型当前更新的攻击 |
| 参与率 | 在训练期间敌人能多久发动一次攻击   * 在跨设备联邦环境中，一个恶意用户可能只能参与一个模型训练回合 * 在跨竖井联邦环境中，一个敌人可能能持续参与模型的学习过程 |
| 适应性 | 敌人是否能在攻击过程中修改攻击参数   * *静态*：敌人必须在攻击之初确定攻击参数且无法在发起攻击后更改。 * *动态*：敌人能够在模型训练过程中修改攻击 |

在分布式数据中心和集中式设置中，对于各种攻击向量的攻击和防御已经做了大量的工作，即模型更新中毒[69,101,97,294,21]、数据中毒[63,123,368,133]和回避攻击[64,377,187,90,283]。正如我们将看到的，联邦学习增强了许多攻击的效力，并增加了防御这些攻击的挑战。联邦设置与数据中心多机器学习共享一个训练时间中毒攻击向量:从远程工作者发送到共享模型的模型更新。这是一种潜在的强大功能，因为对手可以构建恶意更新来达到精确的预期效果，且在攻击中无需考虑指定的客户端损失函数和训练方案

表11中没有讨论的另一个可能的攻击载体是中央聚合器本身。如果一个对手可以入侵聚合器，那么他们就可以很容易地对训练好的模型执行有针对性和无针对性的攻击 [277]。虽然可以通过证明训练过程完整性的方法(如多方计算或零知识证明)检测到恶意聚合器，但是在联邦和分布式数据中心设置中，这一部分工作看起来很相似。因此，我们在续集中省略了对这一攻击载体的讨论。

对手检测模型参数的能力是设计防御方法时的一个重要考虑因素。黑盒模型通常假设对手不能直接访问参数，但可以查看输入-输出对。此假设通常与联邦学习不太相关：因为模型是广播给所有参与者进行本地训练的，所以通常假设对手可以直接访问模型参数（白盒）。此外，开发一种有效的防御白盒、模型更新中毒攻击的方法，也必然能够防御任何黑盒或数据中毒攻击。

在特定的联邦设置(跨设备、跨竖井等)环境中需要评估的一个重要指标是参与者合谋的能力。在训练阶段攻击中，可能会有不同的对手入侵不同数量的客户端。直观来看，如果对手能够协调他们的有毒更新，那么他们可能会比单独行动时更有攻击力。对我们这些可怜的联邦学习防御研究员来说，更糟糕的是，共谋可能不是“实时”发生的（within-update collusion)，它也许是在整个模型更新期间相继发生的（cross-update collusion）

有些联邦设置自然会导致参与率有限:在数亿台设备的情况下，每次更新抽样几千台设备，不太可能在训练过程中对同一个参与者进行多次抽样（除非刻意为之）[74]。因此，仅入侵单个客户端的对手可能只能注入有限次数的中毒更新。一个比较强大的对手可能能够参与每一轮。一个控制多个串通客户的对手也许可以实现持续参与。另外，在表1中的交叉竖井联合设置中，大多数客户机参与每一轮。因此，与在跨设备环境下相比，对手更容易在跨竖井环境下对每一轮发起攻击。

在联邦环境中训练阶段对手的另一能力维度是他们的适应性。在标准的分布式数据中心训练过程中，恶意的数据提供者通常仅能发动静态攻击，他们在训练开始前提供一次中毒数据。相比之下，能够持续参与联邦学习的恶意用户可能会在整个模型训练期间发起中毒攻击，且该恶意用户在训练过程中能够自适应地修改训练数据或模型更新。注意，在联邦学习中，这种适应性通常只有在客户机可以在整个训练过程中多次参与时才有意义。

在下面的章节中，我们将深入了解不同的攻击载体、可能的防御措施以及一些社区可能感兴趣的领域，以推进该领域的发展。

### 模型更新中毒

一个十分自然且强大的攻击类型是模型更新中毒攻击。在这一类攻击中，敌人可以直接操纵对服务提供者的报告。在联邦环境中，这可以通过直接破坏客户机的更新或通过实施某种中间人攻击来执行。在本节我们假设对手可以直接操作更新，因为这可以极大增强对手的能力。因此，我们假设对手（或对手们）直接控制了多个用户，且对手可以直接修改这些用户的输出，以试图使学习中的模型偏向其目标。

**拜占庭式的无目标攻击** 对无目标模型中毒攻击的研究中，有一种特别重要的模型是拜占庭威胁模型，在该模型中，分布式系统中的故障可使系统产生任意输出 [255]。我们对这一概念进行扩展，如果对手可以导致分布式系统中的进程产生任意输出，那么我们就说对该分布式系统中进程的敌对攻击是拜占庭式攻击。因此，拜占庭式攻击可看作是在给定计算节点下无目标攻击的最糟糕形式。因为这种攻击是最糟糕的，我们对无目标攻击的讨论将主要集中在拜占庭式攻击上。然而，我们注意到防守者有更多的手段来应对更初级的无目标威胁模型。

在联邦学习环境中，我们将主要关注对手控制了多个用户的情形。这些拜占庭用户可以给服务器发送任意值，而非发送本地更新后的模型。这会导致全局模型在局部最优处收敛，甚至会导致模型发散 [69]。如果拜占庭式客户端拥有访问模型或非拜占庭式客户端更新的白 盒访问权限，那么他们可能能够根据正常的模型更新调整输出，使之具有相似的方差和量级，从而使其难以被系统检测。拜占庭攻击的灾难性潜力促使人们开始研究用于分布式学习的抗拜占庭聚合机制 [68, 97, 294, 21, 426, 133]。

**拜占庭弹式防御** 针对非目标模型更新中毒攻击，特别是拜占庭式攻击，一种流行的防御机制用对平均值的稳健估计取代了服务器上的平均步骤，例如基于中值的聚合器[101，426]、Krum[69]和修剪平均值[426]。过去的研究表明，在适当的假设下，即使在联邦环境中，各种健壮的聚合器对容忍拜占庭攻击的分布式学习[372, 69, 101]也是有效的[326, 415]尽管如此，方等人。[161]在最近的研究中表明，在联邦学习中，多重拜占庭弹性防御对模型中毒攻击的防御作用很小。因此，有必要对拜占庭弹性防御在联邦学习中的有效性进行更多的实证分析，因为这些防御的理论保证可能只在一些假设下成立，而这些假设通常是不符合实际情况的[49, 328]。

另一种模型更新中毒防御机制使用冗余和洗牌数据来减轻拜占庭式攻击[97, 328, 129]。虽然这些机制通常具有严格的理论保证，但它们通常假定服务器可以直接访问数据，或者能够在全局洗牌数据，因此不能直接应用于联邦学习。一个具有挑战性的未解决问题是协调基于冗余的防御机制(可能增加通信成本)和联邦学习系统(旨在降低通信成本)。

**针对性模型更新攻击** 发动针对性模型更新攻击可能比非针对性攻击所需要的敌人数量少，因为前者为敌人所争取的期望结果范围比较小。在这种攻击中，即使是单发攻击也足以将后门引入模型[42]。Bhagojietal[61]表明，如果参与联合学习的设备中有10%受到侵害，则敌人可以通过毒害发送回服务提供商的模型引入后门，即使服务器上存在异常检测器。有趣的是，中毒模型更新的外观和行为（在很大程度上）类似于没有受到目标攻击的模型，这使得单单是检测后门的存在就十分困难。此外，由于对手的目标是只影响少量数据点的分类结果，同时保持全局学习模型的整体准确性，因此针对非目标攻击的防御通常无法解决目标攻击。

现有的针对后门攻击的防御[368、274、387、133、398、355、107]要么需要仔细检查训练数据、访问一组类似分布式数据的保留集，要么需要完全控制服务器上的训练过程，而在联邦学习设置中这些都不可能实现。未来工作可以尝试的一个有趣途径是探索使用零知识证明来确保用户提交的更新属性是预先确定的属性。基于硬件认证的解决方案也可以考虑。例如，用户的手机可能有能力证明共享的模型更新是使用手机摄像头生成的图像正确计算的。

**串通防御** 如果对手能够进行串通，那模型更新中毒攻击的效率将大大提升。这种串通将使得对手能够发动更高效、更难被检测的模型更新攻击[49]。这种模式与黑客攻击密切相关[140]，在这种攻击中，客户可以随意加入或离开系统。由于服务器无法查看客户端数据，因此在联邦学习中检测sybil攻击可能要困难得多。最近的研究表明，联邦学习很容易受到有针对性和无针对性的sybil攻击[168]。联邦学习的潜在挑战包括在不检查节点数据的前提下防御共谋或检测合谋对手。

### 数据中毒攻击

数据中毒是一种比模型更新中毒更具潜在限制性的攻击类型。在这种模式下，对手不能直接损坏到发送到中心节点的信息。相反，对手只能通过替换数据的标签或特定特征来操作客户端数据。与模型更新中毒一样，数据中毒可以分为针对攻击[63, 100, 239]和非针对攻击[277, 42]。

当对手只能影响联邦学习系统边缘的数据收集过程，但不能直接破坏学习系统中的导出量（例如模型更新）时，这种攻击模型对其来说可能更自然。

**数据中毒和拜占庭式健壮聚合** 由于数据中毒攻击会导致模型更新中毒，因此任何针对拜占庭式更新的防御措施都同时可用于防御数据中毒。例如，Xie et al.[416]，Xie[414]和Xie et al[415]提出了拜占庭健壮的聚合器，它成功地防御了卷积神经网络的标签攻击。如第5.1.2节所述，分析和改进联邦学习中的这些方法非常重要。非IID数据和客户的不可靠性都给拜占庭稳健聚合的工作带来了严峻的挑战，并破坏了常见的假设条件。对于数据中毒，拜占庭威胁模型可能过于强大。通过限制数据中毒（而不是一般的模型更新中毒），可以设计一个更具针对性同时也更有效的拜占庭健壮聚合器。我们将在第5.1.3节的末尾对此进行更详细的讨论。

专门为数据中毒攻击设计的数据净化和网络修剪防御通常依赖于“数据净化”方法[123]，该方法旨在移除中毒或其他异常数据。最近的研究开发了使用稳健统计的数据净化方法[368, 355, 387, 133]，该方法被证明能在少量异常值下保持鲁棒性[133]。这种方法既可以应用于有针对性的攻击，也可以应用于无针对性的攻击，并都取得了一定程度的经验成功[355]。

用于防御后门攻击的一类相关防御是“修剪”防御。修剪防御不是移除异常数据，而是试图移除干净数据上不活动的激活单元 [274, 398]。该方法是由先前的研究所推动的，这些研究表明，在经验上，被设计来引入后门的中毒数据经常会触发“后门神经元”[189]。虽然此类方法不需要直接访问所有客户端数据，但它们需要代表全局数据集的“干净”测试数据。

数据清理和网络修剪都不能直接在联邦学习环境中工作，因为它们通常都需要访问客户端数据，或者类似于客户端数据的其他数据。因此，是否可以在联邦环境中使用数据清理方法和网络修剪方法而不丢失隐私，或者针对数据中毒的防御是否需要新的方法，这还是一个悬而未决的问题。此外，Koh et al. [240]最近的研究表明，许多基于数据清理的启发式防御仍然容易受到自适应中毒攻击，这表明即使是在联邦学习环境中实现了数据清理或许也不足以抵御数据中毒。

在联邦学习中，即便只是检测有毒数据的存在（不要求对其纠正或用有毒数据标识被入侵的客户端）也是一项挑战。当该数据中毒攻击企图安装后门时，该困难还将进一步增大，因为就算是全局训练精度或单用户训练精度这些性能指标也不足以探测出后门的存在。

**模型更新中毒与数据中毒的关系** 由于数据中毒攻击最终会导致客户端输出到服务器信息的更改，所以数据中毒攻击是模型更新中毒攻击的特殊情况。另一方面，还不清楚哪些类型的模型更新中毒攻击可以由数据中毒攻击实现或近似。 Bhagoji et al. [61] 最近的工作表明，数据中毒攻击的强度可以会弱一些，特别是在参与率较低的情况下（见表11）。一项有趣的研究是量化这两种攻击之间的差距，并将这一差距与对手在这些攻击模式下的相对实力联系起来。虽然这个问题可以独立于联邦学习之外提出，但是由于对手能力的差异（见表11），它在联邦学习中特别重要。例如，可以执行数据中毒攻击的客户端的最大数量可能比能执行模型更新中毒攻击的数量高得多，特别是在跨设备设置中。因此，理解这两种攻击类型之间的关系，尤其是它们与客户端数量之间的关系，将极大地帮助我们理解联邦学习中的威胁。

这个问题可以通过多种方式来解决。 根据经验，人们可以研究各种攻击行为的差异。或研究各种模型更新中毒攻击是否可以近似为数据中毒攻击，并为此开发方法。理论上，尽管我们推测模型更新中毒比数据中毒更强，但我们还没有就此下最终的定论。一种可能的方法是使用机器教学工作中的见解和技术(参考文献[438])来理解“最佳”数据中毒攻击，如[292]。任何正式的声明都可能依赖于数量，如损坏的客户端数量和相关的函数类。直观地说，模型更新中毒和数据中毒之间的关系应该依赖于模型相对于数据的超参数化。

### 推断阶段攻击

在回避攻击中，对手可能试图通过小心操作输入模型的样本来绕过已部署的模型。一种经过充分研究的回避攻击形式是所谓的“对抗性攻击”。这些是测试输入的扰动版本，对于人类来说，它们与原始的测试输入几乎无法区分，但是却愚弄了经过训练的模型[64,377]。在图像和音频领域，对抗性的例子通常是通过在测试例子中加入范数有界的扰动来构建的，尽管最近的一些研究探讨了其他的失真方法[155, 408, 223]。在白盒环境中，上述扰动可以通过约束优化方法(如投影梯度上升法[247,283])来产生，该方法试图最大化范数约束下的损失函数。这种攻击经常会导致自然训练的模型在cifar10或ImageNet等图像分类基准上的精度为0[90]。在黑盒设置中，模型也被证明容易受到基于对模型查询访问[99,82]或基于替代模型 [377, 315, 386]. （由相似数据训练得到）的攻击。虽然在数据中心环境中考虑黑盒攻击可能更为自然，但联合学习中的模型广播步骤意味着任何恶意客户端都可以访问该模型。因此，联合学习增加了防御白盒回避攻击的需要。

为了使模型对规避攻击具有更强的鲁棒性，人们提出了各种方法。在这里，健壮性通常是由白盒对抗示例上的模型性能来度量的。不幸的是，被学者提出的许多防御被证明只能提供一种表面上的安全感[31]。另一方面，对抗性训练(用对抗性样本训练的一个健壮模型)通常对白盒回避攻击具有一定的健壮性[283, 413, 352]。对抗性训练通常表示为一个极小极大优化问题，其中对抗性样本和模型权值交替更新。我们注意到，对抗性训练没有一个规范化的公式，诸如极小极大优化问题和学习率等这些超参数的选择会显著影响模型的稳健性，特别是对于像ImageNet这样的大规模数据集。此外，对抗性训练通常只提高对训练中包含的对抗性示例的特定类型的稳健性，这可能使训练后的模型容易受到其他形式的对抗性噪声的影响[155,383,354]。

将对抗性学习方法引入到联邦学习环境中的需求引发了许多新的问题。例如，对抗训练在获得显著的稳健性之前可能需要许多时间。然而，在联邦式学习，尤其是跨设备的联邦式学习中，每一种训练样本的学习次数是有限的。一般来说，对抗性训练主要是针对IID数据开发的，尚不清楚它在非IID环境下的表现。例如，在不能在训练之前检查训练数据的联邦学习环境中，在扰动范数上设置适当的界限以执行对抗性训练（即使在IID设置中也是一个具有挑战性的问题[212]）变得更加困难。另一个问题是生成对抗性样本相对昂贵。虽然一些对抗性训练框架试图通过重用对抗性样本来最小化这一成本[352]，但这些方法仍然需要大量客户端计算资源。这在跨设备设置中可能存在问题，在这种情况下，对抗性样本生成可能会加剧内存或电量的使用。因此，在联邦学习环境中中可能需要新的设备鲁棒优化技术。

**训练阶段攻击和推断阶段攻击的关系** 上述对规避攻击的讨论通常假设对手在推理阶段具有白盒访问权限（可能由于联邦学习的系统级特征）。这忽略了这样一个事实:对手可能会破坏训练过程，从而制造或增强模型的推理时间漏洞，如[100]。对手可以通过无目标和有目标的攻击来实现这一点；对手可以使用有目标的攻击来创建特定类型的对抗性的漏洞[100，189]或使用无目标的攻击来降低对手训练的有效性。

一种可能的对联合训练和推理阶段对手的防御是检测后门攻击 [387, 96, 398, 107]。我们在第5.1.3节中详细地讨论了将先前的防御措施（如上文所述）应用于联邦设置的困难。然而，在许多联邦学习环境中，只检测后门可能不够好，因为我们还希望保证输出模型在推断阶段的稳健性。更复杂的解决方案可能是将训练时间防御(如健壮聚合或差异隐私)与对抗训练结合起来。该领域其他尚未解决的工作可能涉及量化各种类型训练时间攻击是如何影响模型推断时间脆弱性的。考虑到在防御纯训练时间或纯推理时间攻击方面存在的挑战，这一部分的工作必然更具探索性。

### 隐私保障方面的防御能力

联邦学习系统中的许多挑战可以被视为确保一定程度的健壮性：不管是否恶意，干净的数据被破坏或以其他方式被篡改。最近关于数据隐私的工作，特别是差分隐私（DP）[147]，从稳健性的角度定义了隐私。简而言之，在训练或测试时加入随机噪声，以减少特定数据点的影响。有关差异隐私的详细说明，请参阅第4.2.2节。作为一种防御技术，差异隐私有几个令人信服的优点。首先，它提供了强大的、在最坏情况下抵御各种攻击的保护措施。其次，有许多已知的差分私有算法，并且防御可以应用于许多机器学习任务。最后，已知差分隐私在组合下是封闭的，在这种情况下，后期算法的输入是在观察早期算法的结果后确定的。

我们简要描述使用差异隐私来防御之前提到的三种攻击。

**防御模型更新中毒攻击** 服务提供者可以通过(1)对客户端模型更新实施一个规范约束(例如，通过剪切客户端更新)，(2)聚合剪切的更新，(3)向聚合添加高斯噪声来约束任何单个客户端对整个模型的贡献。这种方法可防止对任何个人更新(或一小群恶意的个人)进行过度拟合，与差异隐私的训练相同(在第4.3.2节中讨论)。Sunetal[374]最近对这种方法进行了研究，且该研究显示了将差异隐私应用于防御目标攻击的初步成功。然而，Sun等人分析的实验和目标攻击的范围可以扩展到包括更一般的对抗攻击。因此，DP是否确实是一种有效的防御手段还有待进一步验证。更重要的是，目前还不清楚DP的超参数（`2范数界和噪声方差）是如何作为模型大小/架构和恶意设备分数的函数来选择的。

**防御数据中毒攻击** 数据中毒可以被认为是学习算法鲁棒性的失败:一些被攻击的训练实例可能会严重影响学习模型。因此，抵御这些攻击的一种自然方法是使学习算法具有差异私有性，从而提高鲁棒性。最近的研究探索了用来防止数据中毒的差异隐私方法[281]，特别是在联邦学习环境中[177]。直观地说，一个只能够修改几个训练样本的对手不可能在学习模型的分布上造成很大的变化。

虽然差分隐私是一种灵活的防御数据中毒的方法，但它也有一些缺点。其主要缺点是必须在学习过程中加入噪声。虽然这不一定是一个困难，像随机梯度下降这样的常见学习算法已经注入了噪声，但添加的噪声会损害学习模型的性能。除此之外，对手只能控制少数设备。因此，差异隐私可以被视为对数据中毒的一种强有力的防御，同时也可以被视为一种较弱的防御。它的强大之处在于，无论对手的目标是什么，都提供考虑最坏情况的保护方案；它的弱处在于，必须限制对手，并且必须在联邦学习过程中添加噪声。

**防御推断阶段的回避攻击** 差异隐私也被研究用来作为对推理阶段攻击的防御，在推理阶段攻击中，对手可能通过修改测试样本来操纵模型。一个直接的方法是使预测器本身具有差异隐私性；然而，这有一个缺点，即预测变得随机，这通常是一个不受欢迎的特性，也会损害可解释性。更复杂的方法[258]添加噪音，然后以最高的概率发布预测。我们认为，在这方面还有进一步探索的机会。

## 非恶意故障模式

与数据中心训练相比，联邦学习特别容易受到来自服务提供商控制之外的不可靠客户端的非恶意故障的影响。与对抗性攻击一样，系统因素和数据约束也会加剧数据中心设置中存在的非恶意故障。我们还注意到，设计用于解决最坏情况下对抗性健壮性的技术（将在以下各节中加以描述）也能够有效地解决非恶意故障。虽然非恶意故障通常比恶意攻击的破坏性小，但它们可能更常见，并且与恶意攻击有共同的根源和复杂性。因此，我们期望在理解和防范非恶意故障方面取得进展，同时也为防范恶意攻击提供信息。

虽然为分布式计算开发的通用技术可以有效地提高联邦学习的系统级健壮性，但是由于跨设备和跨数据孤岛联邦学习的独特特性，我们对专门用于联邦学习的技术更感兴趣。下面我们讨论联邦学习环境中三种可能的非恶意故障模式：客户端报告故障、数据管道故障和带噪模型更新。我们还讨论了使联邦学习对此类失败更加健壮的潜在方法。

**客户机汇报失败**：还记得，在联邦学习中，每轮训练都涉及向客户机广播模型、本地客户机计算和向中央聚合器进行客户机汇报。对于任何参与的客户机，系统因素可能会导致这些步骤中的任何一个出现故障。这种故障在跨设备联合学习中尤其可能发生，在这种情况下，网络带宽变得更为受限，客户端设备更可能是计算能力有限的边缘设备。即使没有显式失败，也可能存在分散的客户机，它们比同一轮中的其他节点需要更长的时间来汇报其输出。如果掉队者需要足够长的时间来汇报，为了提高效率，可能会从通信回合中忽略它们，从而有效地减少参与客户的数量。在“普通”联邦学习中，这不需要真正的算法更改，因为联邦平均可以应用于任何客户端汇报模型更新。

不幸的是，当使用安全聚合（SecAgg）[73]时，无响应的客户机变得更具挑战性，特别是当客户机在SecAgg协议期间退出时。虽然SecAgg被设计为对大量掉队者具有鲁棒性[74]，但仍然存在失败的可能性。故障的可能性可以通过各种补充的方式降低。一个简单的方法是在每一轮中选择更多的设备。这有助于确保掉队者和失败的设备对整体收敛的影响最小[74]。但是，在不可靠的网络设置中，这可能不够。降低失效概率的一个更复杂的方法是提高SecAgg的效率。这缩短了客户退出对SecAgg产生不利影响的时间窗口。另一种可能是开发一个异步版本的SecAgg，它不需要客户端在固定的时间段内参与，可能是通过采用通用异步安全多方分布式计算协议[366]中的技术。更大胆的推测是，可能会采用在多轮计算中聚合的SecAgg版本。这将允许掉队节点被包含在后续回合中，而不是完全退出当前回合。

**数据管道故障**：而联邦学习中的数据管道只存在于每个客户端中，管道仍然会面对许多潜在的问题。特别是，任何联邦学习系统仍然必须定义如何访问原始用户数据并将其预处理为训练数据。此管道中的错误或意外操作可能会极大地改变联邦学习过程。虽然数据管道错误通常可以通过数据中心设置中的标准数据分析工具发现，但联邦学习中的数据限制使检测变得非常困难。例如，服务器无法直接检测到特征级预处理问题（例如反转像素、拼接单词等）[32]。一种可能的解决方案是使用具有差异隐私的联邦方法训练生成模型，然后使用这些方法合成新的数据样本，这些样本可用于调试底层数据管道[32]。为机器学习开发不直接检查原始数据的通用调试方法仍然是一个挑战。

**带噪模型更新**：在上面的5.1节中，我们讨论了攻击者从一些客户端向服务器发送恶意模型更新的可能性。即使不存在攻击者，发送到服务器的模型更新也可能由于网络和体系结构因素而失真。这在跨客户机设置中尤其可能，在这些设置中，单独的实体控制服务器、客户机和网络。由于客户端数据，可能会发生类似的失真。即使客户端上的数据不是故意恶意的，它也可能具有噪声特征[301]（例如，在视觉应用中，客户端可能具有输出缩放到更高分辨率的低分辨率相机）或噪声标签[307]（例如，如果用户表示应用程序的推荐并非偶然相关）。虽然跨数据孤岛联邦学习系统（见表1）中的客户端可以执行数据清理以消除此类污染，但由于数据隐私限制，在跨设备设置中不太可能发生此类处理。最后，无论是由于网络因素还是噪声数据，上述的污染都可能损害联邦学习过程的收敛性。

由于这些污染可被视为模型更新和数据中毒攻击的温和形式，因此一种缓解策略将是使用防御措施来对抗模型更新和数据中毒攻击。鉴于目前在联邦环境下缺乏明显的健壮训练方法，这可能不是一个实际的选择。此外，即使存在这样的技术，它们对于许多联邦学习应用来说可能过于计算密集。因此，这里的开放性工作涉及开发对小到中等水平的噪声具有鲁棒性的训练方法。另一种可能性是，标准联邦训练方法（如联邦平均法[289]）对少量噪声具有内在的鲁棒性。研究各种联邦训练方法对不同噪声水平的鲁棒性，将有助于了解如何确保联邦学习系统对非恶意故障模式的鲁棒性。

## 探索隐私和鲁棒性之间的矛盾关系

确保隐私的一个主要技术是安全聚合（SecAgg）（见4.2.1）。简而言之，SecAgg是一个工具，用于确保服务器只看到客户机更新的聚合，而不是任何单独的客户机更新。虽然SecAgg在确保隐私方面很有用，但它通常会使对抗攻击的防御更难实现，因为中央服务器只能看到客户机更新的聚合。因此，研究如何在使用安全聚合时防御对抗攻击是非常重要的。现有的基于距离证明的方法（例如BuffDebug[84]）可以保证上述基于DP的剪辑防御与SECAGG兼容，但是开发节省计算和通信的距离证明仍然是一个积极的研究方向。

SecAgg还为其他防御方法引入了挑战。例如，许多现有的拜占庭鲁棒聚合方法利用服务器XeETEL[415]上的非线性操作，尚不知道这些方法是否与最初设计用于线性聚合的安全聚合兼容。最近的工作已经找到了在SecAgg[326]下，在一次一般的聚合循环中使用少量SecAgg调用近似几何中值的方法。但是，在使用SecAgg的情况下，一般不清楚可以计算哪些聚合器。

1. Footnote Text. [↑](#footnote-ref-0)