Title

Subtitle

Author

Date

Abstract

# 保护用户隐私数据

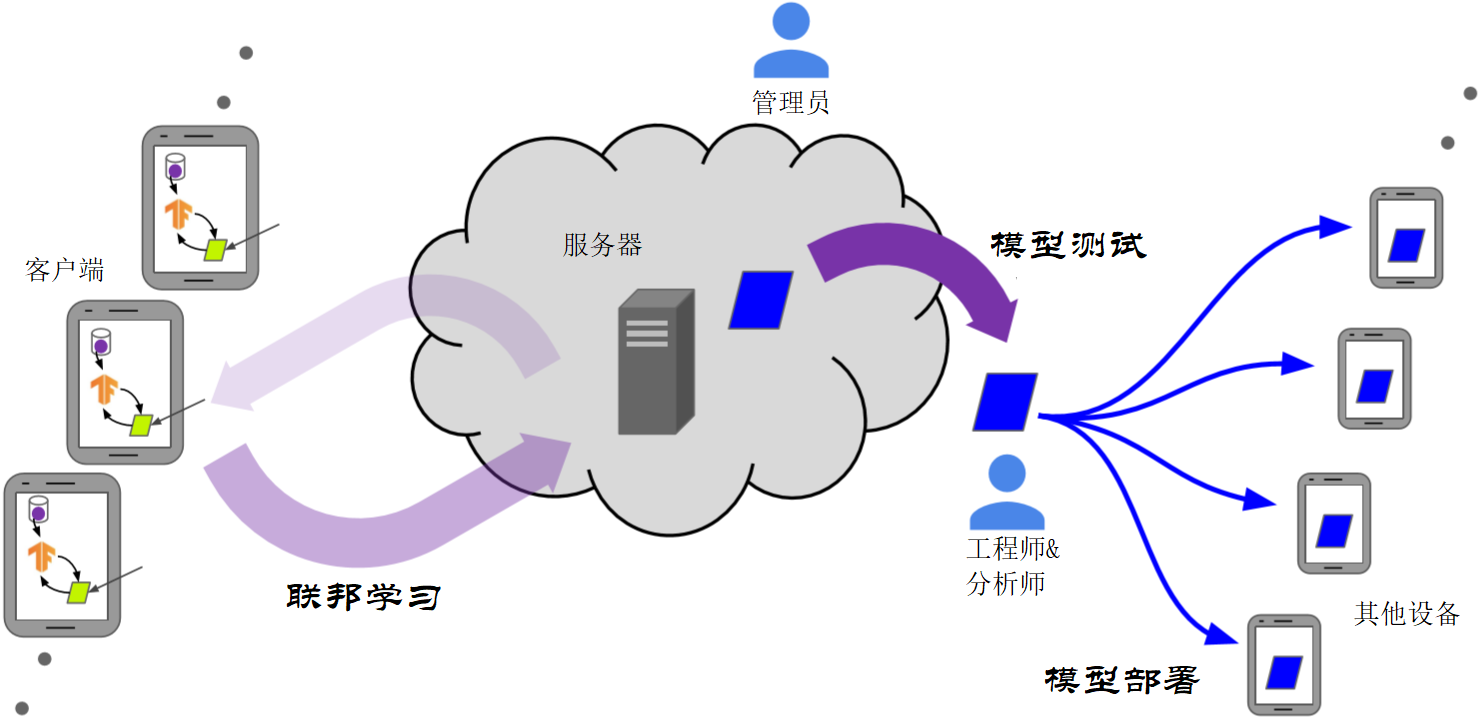
机器学习过程由许多功能不同的角色参与运作。例如，用户可以通过与设备交互来生成训练数据，在机器学习训练过程中，其从这些数据中提取人机交互模式（例如，以训练后模型参数的形式），之后机器学习工程师或分析师可以评估该训练模型的质量，最后可能将该模型部署在最终用户主机上，以支持定制的用户体验（见下图1）。

在一个理想的世界里，系统中的每一个参与者只会学到扮演他们角色所需的信息。例如，如果分析师只需要确定某个特定的质量度量是否超过了所需的阈值，以便授权将模型部署到最终用户，那么在理想化的世界中，该度量值是分析师可以获得的唯一信息；该分析师既不需要访问训练数据，也不需要访问模型参数。类似地，最终用户体验到的可能只需要由经过训练的模型提供的预测，而不需要其他任何内容。

此外，在理想的世界中，系统中的每个参与者都能够轻松、准确地推断出自己和他人的哪些个人信息可能通过参与系统而泄露，参与者将能够利用这一推理结果，就是否参与以及如何参与做出明智的选择。

创建一个具有上述所有理想隐私属性的系统本身将是一项令人望而生畏的壮举，若其还可以实现其他令人满意的属性，则更是难上加难，比如所有参与者的易用性、最终用户体验的质量和公平性（以及影响体验结果的模型），智能地使用通信和计算资源、抵御攻击和失败的能力等。

与其追求无法企及的完美，我们不如另辟蹊径——整个系统由模块化单元组成，这些单元可以相对独立地进行学习和改进，同时我们也要注意，我们最终必须根据上述我们的理想隐私目标，测量整个系统的隐私属性。本节中，我们将提出目前无论是通过单个模块实现方法还是通过整个系统实现方法，还不了解如何同时实现所有目标的领域，作为开放性研究问题。



*图4-1 联邦学习模型的生命周期及联邦学习系统中的不同参与者*

*（与第7页图相同）*

联邦学习提供了一个有吸引力的结构，可以将整个机器学习工作流程分解成我们想要的可实现的模块单元。联邦学习模型的一个主要优点是它可以通过数据最小化为参与的用户提供一定程度的隐私：设备从不发送原始用户数据，只将对模型的更新（例如梯度更新）发送到中央服务器。这些模型更新更侧重于要完成的学习任务而非关注原始数据（即，与原始数据相比，它们严格不包含关于用户的附加信息，而且通常不包含其他意义），并且单个更新只需要由服务器暂时保存。

虽然这些特性可以在集中所有训练数据的基础上提供显著的实用性隐私改进，但是在这个基线联邦学习模型中仍然没有隐私的正式保证。例如，可以构造这样的场景，在该场景中，原始数据的信息从客户端泄漏到服务器。比如，知道以前的模型和用户的梯度更新将允许某一方推断该用户持有的训练示例。因此，本节调查现有的结果，并概述了设计可以提供严格的隐私保障的联邦学习系统的开放性的挑战。我们更专注于联邦学习和分析设置中特定的问题，而不考虑了在一般的机器学习设置中也会出现的问题。

除了针对用户隐私的攻击之外，还有其他种类的对联邦学习的攻击；例如，对方可能会试图阻止模型被学习，或者他们可能会试图使模型产生偏向对方的训练结果。

我们稍后在第5章讨论这些类型的攻击的讨论。本章的其余小节梗概如下。第4.1节讨论了我们希望提供抵御的各种威胁模型。第4.2节列出了一套核心工具和技术，可用于针对第4.1节中讨论的威胁模型提供严格的保护。第4.3节假定可信服务器的存在，并讨论在对抗对手和/或分析师提供保护方面的公开问题和挑战。第4.4节讨论了在没有完全可信服务器的情况下的开放性问题和挑战。最后，第4.5节讨论了关于用户感知的开放性问题。

## 参与者，威胁模型与深层隐私

在联邦学习中，对隐私风险的规范处理需要一种整体的、跨学科的方法。对于一些风险类型，可以通过将现有技术扩展到指定场景中从而保护隐私和减轻风险，而其他更复杂的风险类型则需要跨学科的协同努力。

隐私不是二进制量，甚至不是标量。这种规范处理的第一步是准确描述不同的参与者（见第1节图1，为方便起见，在第35页重复）及其在模型中扮演的角色，最终确定相关的威胁模型（见表7）。例如，我们希望将服务器管理员的视图与使用所学模型的分析师的视图区分开来，因为可以想象，设计用于针对恶意分析师提供强大隐私保证的系统可能不会提供任何恶意攻击的机会。这些参与者行为模式很好映射到其他文献中讨论的威胁模型上；例如，在Bittau等人中。[67，第3.1节]，“编码器”对应于客户机，“洗牌者”通常对应于服务器，“分析器”可能对应于服务器或分析师完成的后处理。

例如，一个特定的系统可能提供差异性的隐私保证，比如向服务器管理员提供的特定参数为ε，而分析师观察到的参数结果可能具有更高的保护性ε’<ε。

此外，这一保证可能仅适用于能力受到特定限制的对手，例如可以观察服务器上发生的所有事情（但不能影响服务器的行为）的对手，这类对手同时完全控制占客户端总数比例为γ的客户端（“完全控制”即可以观察他们可获知的所有数据并以任意方式影响其行为）；对手也可能被认为无法破解在特定安全级别σ下，实例化的加密机制。为了对抗实力突破这些限制的对手，在服务器管理员的看来可能仍然需要差异性隐私保证，但认为分析师观察的结果在较弱的隐私保护级别ε’>ε。

正如我们在本例中看到的，精确地指定系统的假设和隐私目标，以及诸如差异性隐私保证、诚信但好奇行为等安全性概念，可以很容易地通过到几个参数（ε、ε’、ε0、γ、σ等）进行具体实例化。

实现联邦学习所需的所有隐私属性通常需要将下述的许多工具和技术组合到端到端系统中，可能两种分层策略都是为了保护系统的同一部分（例如，在可信执行环境（TEE）中运行安全多方计算（MPC）协议的一部分，使对手更难对该组件产生足够大的损害）以及使用不同的策略来保护系统中不同的部分（例如，使用MPC保护模型更新的聚合，然后在服务器之外共享聚合更新之前使用隐私披露技术）。

因此，我们提倡构建联邦系统，在一种或另一种技术无法提供其预期隐私贡献的情况下，尽可能优雅地降低隐私属性。例如，在TEE中运行MPC协议的服务器组件可能允许维护隐私，即使TEE安全性或MPC安全性假设中的任何一个（但不是两个）在实践中都不成立。另一个例子是，要求客户端将原始训练示例发送到服务器端TEE将被强烈地称为让客户端向服务器端TEE发送渐变更新，因为如果TEE的安全性将失败，后者的隐私期望将更优雅地降低。我们将这种优雅降级的原则称为“深度隐私”，类似于成熟的深度防御网络安全原则[311]。

As such, we advocate for building federated systems wherein the privacy properties degrade as gracefully as possible in cases where one technique or another fails to provide its intended privacy contribution. For example, running the server component of an MPC protocol inside a TEE might allow privacy to be maintained even in the case where either (but not both) of the TEE security or MPC security assumptions fails to hold in practice. As another example, requiring clients to send raw training examples to a server-side TEE would be strongly dispreferred to having clients send gradient updates to a server-side TEE, as the latter’s privacy expectations degrade much more gracefully if the TEE’s security were to fail. We refer to this principle of graceful degradation as “Privacy in Depth,” in analogy to the well-established network security principle of defense in depth [311].

### 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-1)

Block Text.

Table caption.

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

Image Caption

cross-population pattern

人机交互模式

DefinitionTerm

Definition

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