Advances and Open Problems in Federated Learning

0 什么是联邦学习

联邦学习是一类分布式机器学习，节点合作生成一个模型却不共享数据

那么，联邦学习与传统分布式机器学习区别是什么?

1. 用户拥有对于设备和数据的控制权
2. 用户节点是不稳定的
3. 通信代价高于计算代价
4. 数据在节点中是non-IID的
5. 数据量是非常不平衡的

由于联邦学习的这些限制，产生了三个主要的研究方向

1. 提高通行效率的算法
2. 抵御隐私泄露
3. 对于恶意节点具有健壮性（恶意节点可能是用户、服务器等等）

1 引言

联邦学习是一类机器学习，在中心服务器（如：服务提供商）的管理下多个客户端（如：移动设备或整个组织）共同训练一个模型，同时保持训练数据去中心化。联邦学习体现了集中数据收集和最小化的原理，并且可以减轻由传统的集中式机器学习和数据科学方法导致的许多系统隐私风险和成本。

联邦学习许多问题的关键特性是它们本质上是跨学科的，要解决这些问题可能不仅需要机器学习，还需要分布式优化，密码学，安全性，差分隐私，公平性，压缩感知，系统，信息论，统计等等。

所谓 跨域的联邦学习就是将联邦学习的某些限制放宽，比如通行效率、用户可靠性、用户节点的数据量等等，下面这张图是更详细的区分了这三者的关系

	存在数据中心的分布式学习	跨域联邦学习	跨设备联邦学习
环境	在大型但“平坦”的数据集上训练模型。 客户群 是单个群集或数据中心中的计算节点。	在孤立的数据上训练模型。 客户是不同的组织（例如医疗或财务）或地理分布的数据中心。	客户是大量的移动或物联网设备。
数据分布	数据被集中存储，可以在客户端之间进行打乱和平衡。 任何客户端都可以读取数据集的任何部分。	数据在本地生成，并保持分散化。 每个客户端都存储自己的数据，而无法读取其他客户端的数据。 数据不是独立或相同分布的。	
管理	集中管理	中心服务器/服务负责组织训练，但从未看到原始数据。	
广域通讯	无(一个数据中心中的完全连接的客户端/集群)。	中心辐射型拓扑，中心代表协调服务提供商（通常不包含数据），分支连接到客户端。	
数据可用性	所有客户几乎总是可用。		在任何时候，只有一小部分客户可用，通常会受昼夜或其他变化影响。
分布规模	通常为1-1000个客户。	通常为2-100位客户。	大规模并行，最多1e10个客户端。
主要瓶颈	计算通常是数据中心的瓶颈，在该中心中，可以假设网络非常快。	可能是计算或通讯。	通讯通常是主要的瓶颈，尽管它取决于任务。 通常，跨设备联邦计算使用wi-fi或更慢的连接。
可溯源性	每个客户端都有一个标识或名称，该标识或名称允许系统专门访问它。		不能直接为客户建立索引（即，不使用客户标识符）。
客户状态	有状态的-每个客户端都可以参与计算的每一轮，并不断携带状态。		无状态的-每个客户可能仅会参与一次任务，因此通常假定在每轮计算中都有一个从未见过的客户的新样本。
客户可靠性	不可靠相对较少。		高度不可靠-预计有5%或更多的客户端参与一轮计算会失败或掉线（例如，由于不符合电量，网络或其他要求而导致设备不合格）。
数据划分	数据可以在客户端之间任意分区/重新分区。	分区是固定的。 可以是样本分区（水平）或特征分区（垂直）。	根据样本固定分区（水平）。

表1：联邦学习环境与数据中心中分布式学习的典型特征

1.1 跨设备联邦学习环境

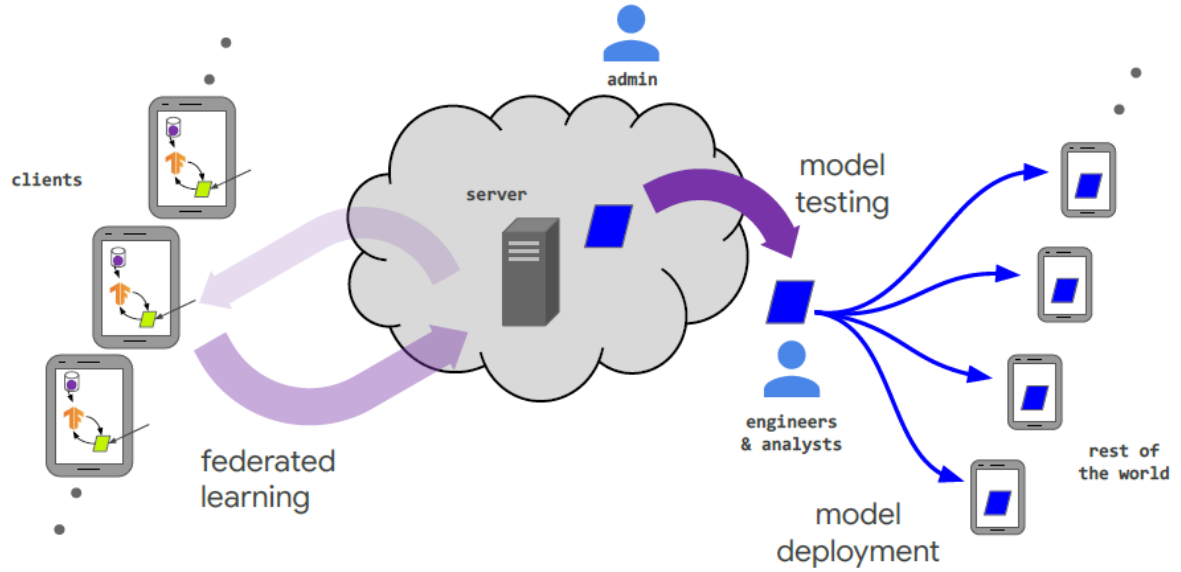


图1：联邦学习训练的模型以及系统中各个参与者的生命周期

1.1.1 联邦学习中模型的生命周期

1. **问题辨别**：模型工程师判断问题是否要用联邦学习解决。
2. **客户端检测**：如果需要，可以对客户端（例如，在手机上运行的应用）进行检测，在本地存储必要的训练数据。
3. **仿真原型（可选）**：模型工程师可以使用代理数据集在联邦学习仿真中对模型体系结构进行原型设计并测试学习超参数。

4. **联邦学习模型训练**：开始执行多个联邦学习训练任务以训练模型的不同变体，或使用不同的优化超参数。
5. **（联邦）模型评估**：在对任务进行了充分的训练之后，将对模型进行分析并选择好的候选。
6. **部署**：最后，一旦选择好模型，它将经历一个标准模型启动过程，以及分阶段推出（以便在发现较差性能和影响太多用户之前回滚）。

总数	10^6-10^{10} 个设备
一轮训练使用的设备	50 – 5000
参与训练一个模型的设备总数	10^5-10^7
模型收敛的轮数	500 – 10000
实际训练时间	1 – 10 天

表2：典型的跨设备联合学习应用程序的数量级大小

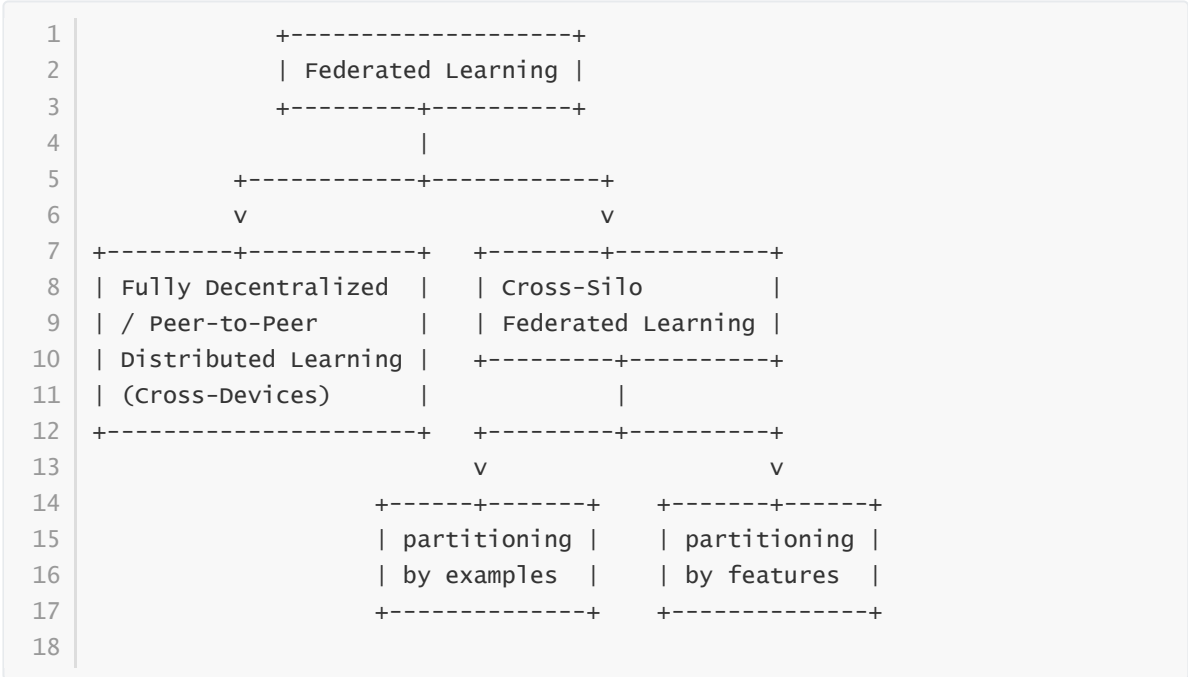
联邦学习系统面临的主要实际挑战之一是使上述工作流程尽可能简单，理想地接近机器学习系统实现集中训练的易用性。

1.1.2 典型的联邦学习训练过程

1. **客户端选择**：服务器从满足资格要求的一组客户端中采样。 例如，移动电话可能仅在连接到未计费的wi-fi且处于空闲状态时才连入服务器，以避免影响设备用户。
2. **广播**：选定的客户端从服务器下载当前模型权重和训练程序。
3. **客户端计算**：每个选定的设备都通过执行训练程序在本地计算对模型的更新。
4. **聚合**：服务器收集设备更新的聚合。该阶段也是许多其他技术的集成点，稍后将讨论这些技术，其中可能包括：用于增加隐私的安全聚合，为了通信效率而对聚合进行有损压缩，以及针对差分隐私的噪声添加和更新限幅。
5. **模型更新**：服务器根据从参与当前轮次的客户端计算出的聚合更新，更新在服务器本地的共享模型。

1.2 跨域联邦学习

现实情境：许多公司或组织希望基于其他们的数据来训练模型，但不能直接共享其数据。



横向联邦学习

在两个数据集的用户特征重叠较多而用户重叠较少的情况下，我们把数据集按照横向 (即用户维度)切分，并取出双方用户特征相同而用户不完全相同的那部分数据进行训练。这 种方法叫做横向联邦学习。比如有两家不同地区银行，它们的用户群体分别来自各自所在的地区，相互的交集很小。

纵向联邦学习

在两个数据集的用户重叠较多而用户特征重叠较少的情况下，我们把数据集按照纵向 （即特征维度）切分，并取出双方用户相同而用户特征不完全相同的那部分数据进行训练。 这种方法叫做纵向联邦学习。比如有两个不同机构，一家是某地的银行，另一家是同一个地方的电商。它们的用户群体很有可能包含该地的大部分居民，因此用户的交集较大。

联邦迁移学习

在两个数据集的用户与用户特征重叠都较少的情况下，我们不对数据进行切分，而可以 利用迁移学习来克服数据或标签不足的情况。这种方法叫做联邦迁移学习。 比如有两个不同机构，一家是位于中国的银行，另一家是位于美国的电商。由于受到地域限制，这两家机构的用户群体交集很小。同时，由于机构类型的不同，二者的数据特征也只有小部分重合。在这种情况下，要想进行有效的联邦学习，就必须引入迁移学习， 来解决单边数据规模小和标签样本少的问题，从而提升模型的效果。

图 1 联邦学习的分类

2 提高效率和有效性

开发更好的优化算法； 为不同的客户提供不同的模型； 在联邦学习上下文中使诸如超参数搜索，架构搜索和调试之类的； 提高沟通效率；等等。

2.1 Non-IID Data in Federated Learning

1.不相同的客户分布

- 特征分布不平衡
- 标签分布不平衡
- 标签不同，特征相同
- 特征不同，标签相同
- 数量偏分布不平衡

non-IID数据集的大多数工作都集中在标签分布不平衡上。

2.数据不独立

3.数据集移位

2.1.1 Strategies for Dealing with Non-IID Data

一种自然的方法是修改现有算法（例如，通过选择不同的超参数）

对于某些应用程序，可能有可能扩充数据，以使跨客户端的数据更加相似。

不再将所有的用户数据给与相同的权重，替代方法包括限制来自任何一个用户的数据贡献。

除了解决不相同的客户分布之外，使用多种模型还可以解决由于客户可用性变化而引起的对独立性的损害。

2.2 联邦学习算法优化

在典型的联合学习任务中，目标是要学习一个单一的全局模型，该模型可以使整个训练数据集上的经验风险函数最小化，以实现优化，non-IID和不平衡数据，有限的通信带宽以及不可靠和有限的设备可用性尤其重要。
设备总数巨大的联邦学习环境（例如，跨移动设备）需要每轮只需要几个客户端参与的算法（客户端采样）。 此外，每个设备可能最多只参与一次给定模型的训练，因此无状态算法是必需的。 这排除了直接在数据中心环境中非常有效的各种方法的直接应用。
与其他技术的可组合性。例如密码安全聚合协议，差分隐私以及模型和更新压缩。

N	Total number of clients
M	Clients per round
T	Total communication rounds
K	Local steps per round.

Table 4: Notation for the discussion of FL algorithms including Federated Averaging.

Server executes:
initialize x_0
for each round $t = 1, 2, \ldots,$ T **do**
  $S_t \leftarrow$ (random set of M clients)
  for each client $i \in S_t$ **in parallel do**
    $x_{t+1}^i \leftarrow \text{ClientUpdate}(i, x_t)$
  $x_{t+1} \leftarrow \sum_{k=1}^{M} \frac{1}{M} x_{t+1}^i$

ClientUpdate(i, x)**:**
for local step $j = 1, \ldots, K$ **do**
  $x \leftarrow x - \eta \nabla f(x; z)$ for $z \sim \mathcal{P}_i$
return x to server

Algorithm 1: Federated Averaging (local SGD), when all clients have the same amount of data.

联邦平均算法

2.2.1 超参数调整

在机器学习中超参数调整主要涉及如何提高模型的准确性，而不是针对移动设备的通信和计算效率。

2.2.2 神经网络设计

神经网络结构搜索（NAS）NAS有三种主要方法，它们利用进化算法，强化学习或梯度下降来搜索特定数据集上特定任务的最佳体系结构。

2.2.3 联邦学习的查错和解释

对于一个联邦学习模型，训练是在本地进行的，如果模型出现问题，查找问题的根源将极为困难

2.2.4 压缩

- 梯度压缩 – 减小了从客户端到服务器通信的对象的大小，该对象用于更新全局模型
- 模型广播压缩– 减小从服务器向客户端广播的模型的大小，客户端从该模型开始本地训练。
- 缩减本地计算 – 对整体训练算法的任何修改，以使本地训练过程在计算上更加有效。

3 保护用户数据隐私

我们提倡一种策略，其中整个系统由可以相对独立地研究和改进的模块化单元组成 。

3.1 相关技术

差分隐私

Secure Multi-Party Computation (MPC) protocol 安全多方计算

Trusted Execution Environment (TEE) 可信执行平台

Private Disclosure techniques 隐私披露技术

remote attestation and zero-knowledge proofs 远程证明和零知识证明

以上的多种技术可用于提供可验证性：零知识证明（ZKP），可信执行环境（TEE）或远程证明。

（零知识证明 指的是证明者能够在不向验证者提供任何有用的信息的情况下，使验证者相信某个论断是正确的。）

Technology	Characteristics
Differential Privacy (local, central, shuffled, aggregated, and hybrid models)	A quantification of how much information could be learned about an individual from the output of an analysis on a dataset that includes the user. Algorithms with differential privacy necessarily incorporate some amount of randomness or noise, which can be tuned to mask the influence of the user on the output.
Secure Multi-Party Computation	Two or more participants collaborate to simulate, though cryptography, a fully trusted third party who can: • Compute a function of inputs provided by all the participants; • Reveal the computed value to a chosen subset of the participants, with no party learning anything further.
Homomorphic Encryption	Enables a party to compute functions of data to which they do not have plain-text access, by allowing mathematical operations to be performed on ciphertexts without decrypting them. Arbitrarily complicated functions of the data can be computed this way ("Fully Homomorphic Encryption") though at greater computational cost.
Trusted Execution Environments (secure enclaves)	TEEs provide the ability to trustably run code on a remote machine, even if you do not trust the machine's owner/administrator. This is achieved by limiting the capabilities of any party, including the administrator. In particular, TEEs may provide the following properties [373]: • Confidentiality: The state of the code's execution remains secret, unless the code explicitly publishes a message; • Integrity: The code's execution cannot be affected, except by the code explicitly receiving an input; • Measurement/Attestation: The TEE can prove to a remote party what code (binary) is executing and what its starting state was, defining the initial conditions for confidentiality and integrity.

Table 8: Various technologies along with their characteristics.

3.2 潜在威胁

Data/Access Point	Actor	Threat Model
Clients	Someone who has root access to the client device, either by design or by compromising the device	Malicious clients can inspect all messages received from the server (including the model iterates) in the rounds they participate in and can tamper with the training process. An honest-but-curious client can inspect all messages received from the server but cannot tamper with the training process. In some cases, technologies such as secure enclaves/TEEs may be able to limit the influence and visibility of such an attacker, representing a meaningfully weaker threat model.
Server	Someone who has root access to the server, either by design or by compromising the device	A malicious server can inspect all messages sent to the server (including the gradient updates) in all rounds and can tamper with the training process. An honest-but-curious server can inspect all messages sent to the server but cannot tamper with the training process. In some cases, technologies such as secure enclaves/TEEs may be able to limit the influence and visibility of such an attacker, representing a meaningfully weaker threat model.
Output Models	Engineers & analysts	A malicious analyst or model engineer may have access to multiple outputs from the system, e.g. sequences of model iterates from multiple training runs with different hyperparameters. Exactly what information is released to this actor is an important system design question.
Deployed Models	The rest of the world	In cross-device FL, the final model may be deployed to hundreds of millions of devices. A partially compromised device can have black-box access to the learned model, and a fully compromised device can have a white-box access to the learned model.

Table 7: Various threat models for different adversarial actors.

3.2.1 应对外部恶意攻击者的威胁

审核迭代过程和最终模型

集中差异隐私训练

隐藏迭代过程

对不断发展的数据进行重复分析

防止模型盗用和滥用

3.2.2 内部威胁

以下只是内部威胁的一部分。

在跨设备联邦学习环境中，我们拥有一台具有大量计算资源的服务器和大量客户端，这些客户端（i）仅能与该服务器通信（如星型网络拓扑），并且（ii）连通性可能受到限制和带宽。
控制服务器的主动恶意攻击者可能会模拟大量伪造的客户端设备（“ Sybil攻击”），或者可以从可用设备池中优先选择以前受到破坏的设备。

3.2.3 现有解决方案的局限性

Local differential privacy

LDP假定用户的隐私完全来自该用户自己的随机性； 因此，用户的隐私保证独立于所有其他用户添加的其他随机性。
造成这种困难的部分原因是，引入的随机噪声的幅度必须与数据中信号的幅度相当，这可能需要合并客户端之间的报告。
因此，要获得与中心差分隐私相同的效果，就需要相对较大的用户群或较大的参数选择。

Hybrid differential privacy

它不提供用户本地噪声的隐私放大功能。
目前尚不清楚哪个应用程序领域和算法可以最有效地利用混合信任模型数据，目前在混合模型上的工作通常假设无论用户信任偏好如何，其数据都来自相同的分布

The shuffle model

首先是可信中间层的要求

第二个缺点是，The shuffle model的差分隐私保证与参与计算的对手用户数量成比例地降低

Secure aggregation

这种方法有几个局限性：（a）假设一个半诚实的服务器（仅在私钥基础结构阶段），（b）允许服务器查看所有的聚合数据（可能仍会泄漏信息），（c ）对于稀疏向量聚合而言效率不高，并且（d）缺乏强制客户输入格式正确的能力。