联邦学习

联邦学习简介

```
Federated learning is a machine learning setting where multiple entities
(clients) collaborate in solving a machine learning problem, under the
coordination of a central server or service provider. Each client’s raw data
is stored locally and not exchanged or transferred; instead, Focused updates
are updates narrowly scoped to contain the minimum information necessary for
the specific learning task at hand; aggregation is performed as earlier as
possible in the service of data minimization.
```

以上定义来自参考[1]. 可以总结为联邦学习本质上是一种分布式机器学习，目标是在保证各方数据隐私的情况下，完成联合建模的一种分布式机器学习方案。

参考文献[1],[6]，形式化定义如下：

定义N个数据提供方$F_1, \ldots, F_n$以及他们对应的数据$D_1, \ldots, D_n, D = \bigcup_{i=1}^{n} D_i$，假设在数据$D$下真实训练的模型结果为$M_{sum}$, 准确值为$V_{sum}$, 基于联邦学习模型计算的模型结果为$M_{fed}$,准确值为$V_{fed}$,要求对于足够小的非负数$\delta$, 满足：

$$|V_{sum} - V_{fed}| <$$

将其称之为$\delta - accuracy$损失。

分类

针对参与方的**数据分布**的不同，文献[2]将联邦学习学习常分为3类， 假设2方计算， X表示特征，Y为标签，I为样例的ID。

- 横向联邦学习：两个数据集的用户特征重叠较多，而用户(也可以称作样本或者ID空间等)重叠较少的情况下， 对数据按照用户纬度切分；例如两方均有用户的所有属性数据，但是所在的范围不一样。

  形式定义如下：

$$X_i = X_j, Y_i = Y_j, I_i \neq I_j, \forall D_i, D_j, i \neq j$$

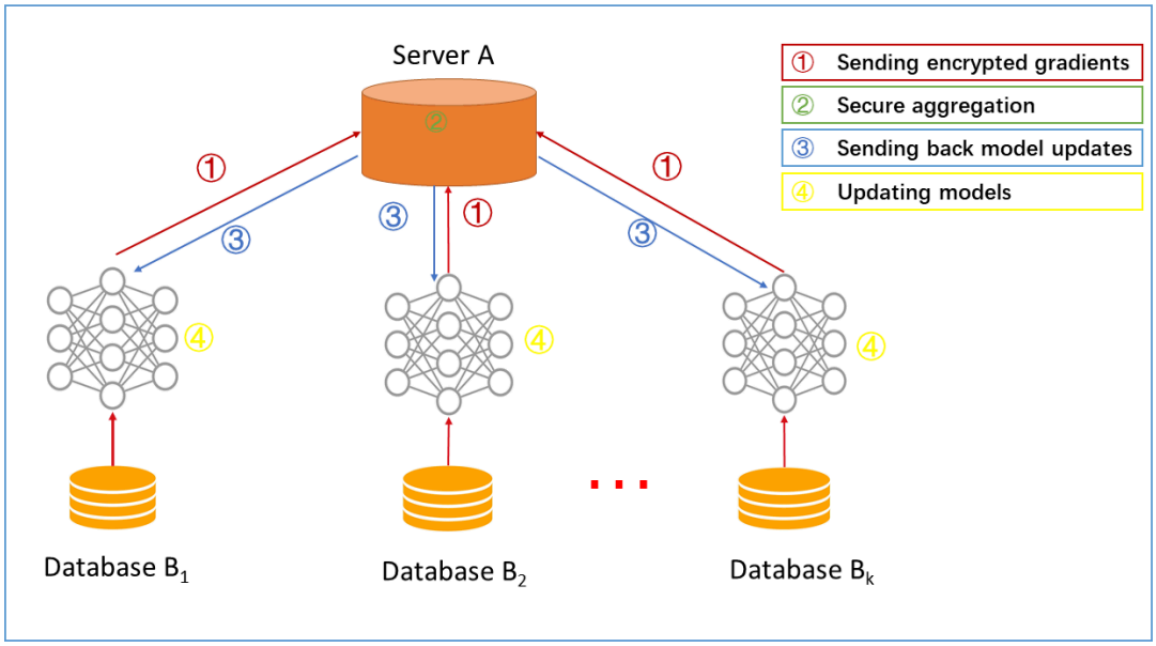

图1： Architecture for a horizontal federated learning system 图来自[6]

其步骤如下：

1. 参与方在本地进行梯度计算，然后将梯度进行加密、添加差分噪音或者基于秘钥共享机制加密本地梯度，然后将加密梯度传递给中心化服务器A；
2. 服务端进行多方安全计算，计算梯度聚合；
3. 将聚合计算的梯度结果返回给不同的参与方；
4. 参与方解密梯度结果，并且更新本地梯度；

例如对于LR，文献[5]、[6]采用模型平均聚合算法，在协调服务器上进行将参数结果进行平均或者不采用协调服务器，直接利用半同态[7]进行双方参数交换。

横向联邦建立在半诚实模型的基础上，对于权重信息可能导致信息泄露，可以引入差分隐私[8]等技术进一步将结果进行模糊处理。

- 纵向联邦学习： 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，对数据按照特征纬度切分。纵向联邦核心解决的是**“模型并行”场景下的本地数据隐私保护**。

  形式定义如下：

$$X_i \neq X_j, Y_i \neq Y_j, I_i = I_j, \forall D_i, D_j, i \neq j$$

典型的架构图下。

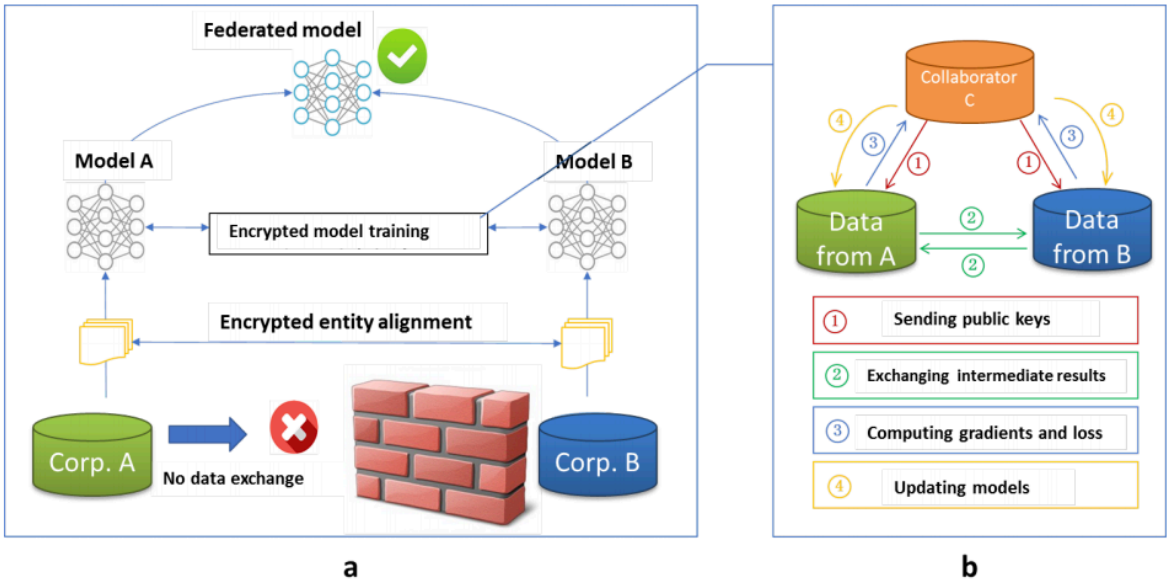

图1： Architecture for a vertical federated learning system 图来自[6]

训练的基本步骤如下：

1. 样本对齐。这一部分借助于PSI[9,10,11]（隐私保护求交）计算各方的用户ID交集；
2. 加密模型训练， 使用交集用户进行训练：

    2.1. 由第三方C常见一堆公私钥，并向A和B发送公钥，用来加密需要传输的数据；

    2.2. A和B分别计算和自己相关的特征中间结果，并加密交互，用来求得各自梯度和损失；

    2.3. A和B分别计算各自加密后的梯度并添加掩码(additional mask)发送给C，同时B计算加密后的损失发送给C；

    2.4. C解密梯度和损失后回传给A和B，A、B去除掩码并更新模型。

例如只有一方有Y，另外一方有X， 要在不暴露的情况下计算权重矩阵W。常借助于同态、多方安全计算等，实现联合梯度运算, 文献[5] P26-37以及文献[6]给出了LR算法的具体实现, 包括依赖第三方和不依赖第三方的方案。

- 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，而利用迁移学习[3],[4]来克服数据或标签不足的情况。

  形式定义如下：

$$X_i \neq X_j, Y_i \neq Y_j, I_i \neq I_j, \forall D_i, D_j, i \neq j$$

迁移学习的核心是，找到源领域和目标领域之间的相似性。联邦迁移学习的步骤与纵向联邦学习相似，只是中间传递结果不同（实际上每个模型的中间传递结果都不同）。文献[12]给出了一种设计思路。

由上面的介绍可以看到，传统的分布式机器学习跟水平联邦学习比较类似。

威胁模型

联邦学习基本上建立在半诚实模型的基础上。

隐私保护技术

技术j链接

TEE：可信执行环境

基于芯片的扩展指令集，提供硬件**可信基(TCB)**，以及**内存安全访问**机制，提供安全API跟操作系统交互，通过远程认证完成跨安全容器访问；

典型实现有Intel SGX/ARM TrustZone等。很容易实现通用机密计算，工程化程度非常高。

同态

$$Dec(En(a) \odot En(b)) = a \oplus b <=> f(En(x), En(y)) == En(f(x, y))$$

主要实现有Pallier/RSA/Lattice-based等实现。

一种特殊形式：双线性对(Bilinear map)映射e存在多项式时间算法进行计算。双线性对在BLS、ZKP、ABE等较多应用。

同态广泛应用在信息隐藏、外包运算、文件存储、密文检索等。

MPC

MPC是一系列多方安全计算协议的统称。在保护多方输入数据隐私的情况下，完成联合计算。

对于(p, d)， p是参与者，d是该参与者的输入数据，如下计算：

$$Given\ (p_1, d_i), \ldots, (p_n, d_n)$$
$$compute\ f(d_1, \ldots, d_n)$$

比较著名的协议有OT/GC（2方），SPDZ（多方）等。

广泛应用在融合计算、联邦学习、匿名投票等。

差分隐私

$$D : database,\ ||D| - |D^{'}|| = 1$$
$$\forall S \in imA$$
$$Pr\{q(D) \in S\} \leq e^{\epsilon} \times Pr\{q(D^{'}) \in S\}$$

应用广泛，主要用在统计查询、数据脱敏隐私保护等。

总结联邦学习过程中主要有以下流派：

1. 半同态流派： 例如微众FATE，字节fedlearner；
2. MPC流派： 百度BFC, 阿里摩斯等
3. TEE流派： 百度mesatee等

涉及到的隐私计算技术包括：

1. TEE 可信计算环境： 基于硬件可信基(TCB) ，以及内存安全访问机制，提供安全系统API，通过远

程认证完成跨安全容器访问；常见有intel sgx等。
2. MPC 多方安全计算： 针对无可信第三方的且保护输入数据隐私的情况下完成联合计算，包含加密电路、不经意传输以及秘钥共享等多种协议以及相互之间组合实现；特别是最近借助batched OT在解决psi问题，效率极大的提升。
3. 同态加密： 密文计算的输出解密等于其对应明文计算; 例如基于rsa盲签名实现psi等。
4. DP 差分隐私： 保留统计学特征的前提下去除个体特征以保护用户隐私

实例

LR

SecureBoost

参考

1. Peter Kairouz. et.al. Advances and Open Problems in Federated Learning, 2019
2. 杨强, et.al《GDPR对AI的挑战和基于联邦迁移学习的对策》， CCAI 2018
3. Sinno Jialin Pan and Qiang Yang Fellow, IEEE, A survey on transfer learning, 2009
4. 机器之心， 《迁移学习全面概述：从基本概念到相关研究》， 链接
5. 刘洋 范涛 微众银行高级研究员《联邦学习的研究与应用》CCF-TF 14，链接
6. Q. Yang, Y. Liu, T. Chen & Y. Tong, Federated machine learning: Concepts and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10(2), 12:1-12:19, 2019
7. Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai. 2018. Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. IEEE Trans. Information Forensics and Security，13, 5 (2018),1333–1345
8. Reza Shokri and Vitaly Shmatikov. 2015. Privacy-Preserving Deep Learning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15). ACM, New York, NY, USA, 1310–1321.
9. https://anquan.baidu.com/article/860
10. Chen, H., Laine, K., and Rindal, P. Fast private set intersection from homomorphic encryption. Cryptology ePrint Archive, Report 2017/299, 2017.https://eprint.iacr.org/2017/299.
11. https://zhuanlan.zhihu.com/p/85422763
12. Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, Qiang Yang, Fellow, IEEE, A Secure Federated Transfer Learning Framework, 2018