计算机前沿介绍结课报告

人工智能及应用-基于联邦学习的数据隐私保护

院（系）：计算机学院

专　　业：计算机科学与技术

姓　　名：常文瀚

班级学号：191181

指导教师：唐厂

2021 年 4 月 20 日

目录

第一章 联邦学习的起源与出处

1.1 联邦学习的背景介绍

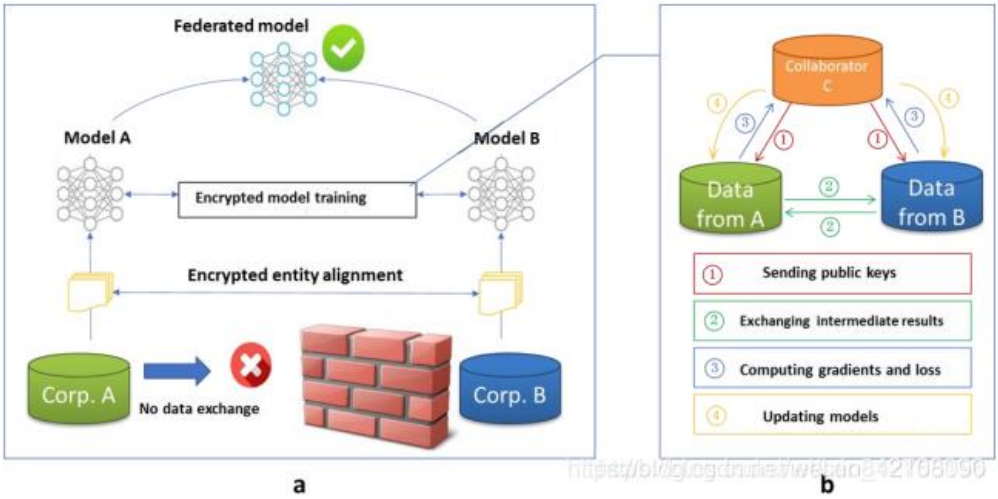
现实生活中，除了少数巨头公司能够满足，绝大多数企业都存在数据量少，数据质量差的问题，不足以支撑人工智能技术的实现；同时国内外监管环境也在逐步加强数据保护，陆续出台相关政策，如欧盟最近引入的新法案《通用数据保护条例》(GDPR)，我国国家互联网信息办公室起草的《数据安全管理办法(征求意见稿)》，因此数据在安全合规的前提下自由流动，成了大势所趋；在用户和企业角度下，商业公司所拥有的数据往往都有巨大的潜在价值。两个公司甚至公司间的部门都要考虑利益的交换，往往这些机构不会提供各自数据与其他公司做与单的聚合，导致即使在同一个公司内，数据也往往以孤岛形式出现。

基于以上不足以支撑实现、不允许粗暴交换、不愿意贡献价值三点，导致了现在大量存在的数据孤岛，以及隐私保护问题，联邦学习应运而生。

1.2 联邦学习的出处

联邦学习的出处是金融机构的痛点，尤其是像“微众银行”这样的互联网银行。一个实用的例子是检测多方借贷。这在银行业，尤其是互联网金融一直是很头疼的一个问题。多方借贷是指某不良用户在一个金融机构借贷后还钱给另一个借贷机构，这种非法行为会让整个金融系统崩溃。要发现这样的用户，传统的做法是金融机构去某中心数据库查询用户信息，而各个机构必须上传他们所有用户，但这样做等于暴露金融机构的所有重要用户隐私和数据安全，这在 GDPR 下就不被允许。

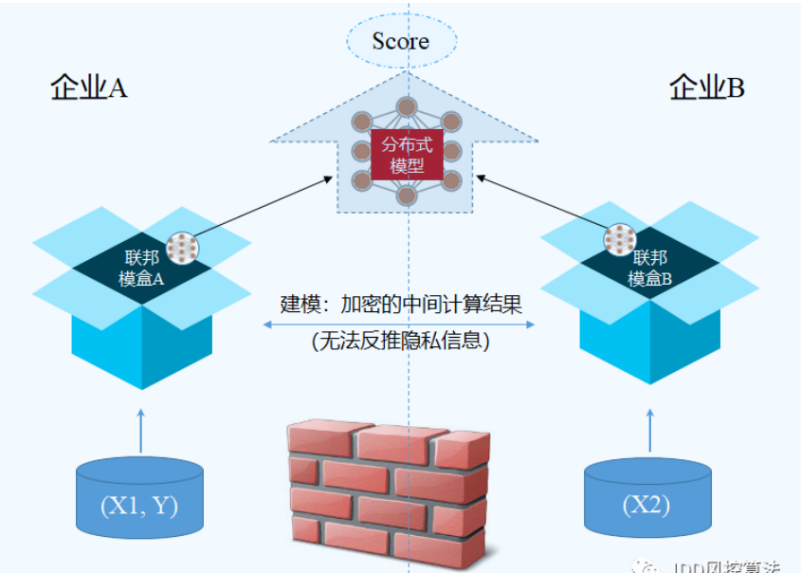
在联邦学习的条件下，没有必要建立一个中心数据库，而 任何参与联邦学习的金融机构可以利用联邦机制向联邦内的其他机构发出新用户的查询，其他机构在不知道这个用户具体信息的前提下，回答在本地借贷的提问。这样做既能保护已有 用户在各个金融机构的隐私和数据完整性，同时也能完成查询多头借贷的这个重要问题。



图（1） 联邦学习图解

1.3 联邦学习的概念

联邦学习本质上是一种分布式机器学习技术，或机器学习框架。它的目标是在保证数据隐私安全及合法合规的基础上，实现共同建模，提升 AI 模型的效果。联邦学习最早在 2016 年由谷歌提出，原本用于解决安卓手机终端用户在本地更新模型的问题。



图（2） 典型的纵向联邦学习案例

第二章 联邦学习与隐私保护

2.1 联邦学习与隐私保护

关于隐私问题，目前在联邦学习中经常使用的是安全多方计算（SMC）、同态加密、差分隐私（DP），本文中主要讨论差分隐私的研究进展。

在差分隐私中，数据通常是由可信第三方进行加噪声来实现隐私。在联邦学习中，服务器作为 DP 机制的可信任实现者，确保隐私输出。然而，我们通常需要减少可信任参与者的参与，最近几年提出了一些方法。

1、 本地差分隐私（Local differentital privacy）

本地差分隐私通过让每个客户机在与服务器共享数据之前对其数据应用差分隐私，可以在不需要信任集中服务器的情况下实现差异隐私。LDP 相对于 DP 而言，不需要可信第三方的参与，而且数据在发送之前就已经实现了隐私保护。不幸的是，LDP 实现隐私保护的同时数据的可用性会降低，而 DP 对小数据集来说可提高可用性。

2、 分布式差分隐私（Distributed differential privacy）

在分布式差分隐私模型中，客户机首先计算并编码一个最小的（特定于应用程序的）报

告，然后将编码的报告发送到一个安全的计算函数，其输出对于中心服务器来说是可用的且满足隐私要求。编码是为了在客户端维持隐私，安全计算函数有多种形式，可能是多方计算（MPC）协议，也可能是 TEE 上的一个标准计算或者两者的结合。分布式差分隐私有两种实现方式：基于安全聚合和安全洗牌。

3、 混合差分隐私（Hybrid differential privacy）

混合模型根据用户的信任模型偏好划分用户的多个信任模型。有两种方法：一种使用最少信任，提供了最低实用性但可以在所有用户上应用。第二种是使用最信任的模型，提供了很高的实用性，但只能应用在值得信任的用户上。通过允许多个模型共存，混合模型机制可以从给定用户群中获得更高的效用。（例如，在一个系统中，大多数用户在本地隐私模型中贡献他们的数据，一小部分用户贡献他们的数据在可信第三方。）

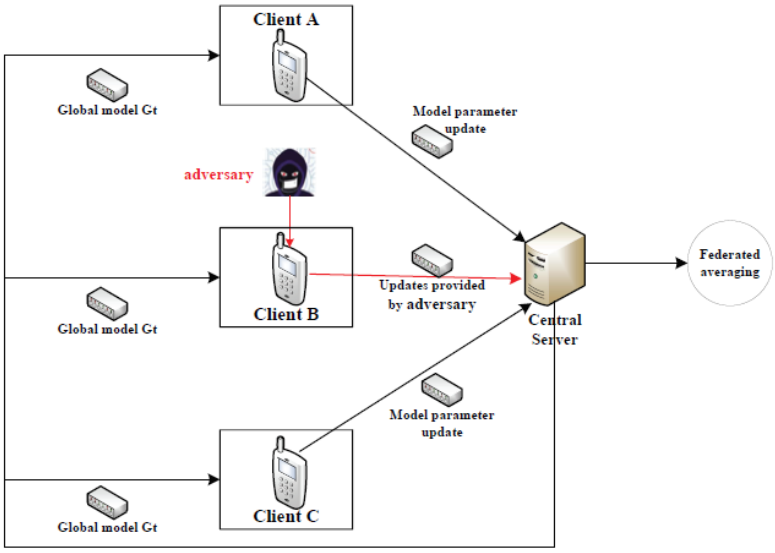
2.2 基于联邦学习的隐私保护实验复现

本文中的实验复现了文章《Model Poisoning Defense on Federated Learning:A Validation Based Approach》中的模型攻击实验

2.2.1 联邦学习中的模型中毒攻击

模型中毒攻击的目标是使全局模型对攻击者期望的事物类别进行错误分类。与针对训练数据集的完整性的数据中毒攻击不同，联邦学习中的对手执行模型中毒攻击将目标数据类别的标签操纵为错误的标签。错误的标签会导致目标模型从本地模型更新中获得错误的预测结果，该预测结果将被发送到中央服务器以进行汇总，然后对此类事物的全局模型性能产生负面影响。

攻击者可能具有强大的功能，例如直接控制被攻击客户端的本地训练数据并修改值。这种攻击可以达到模型的主要任务和子任务上的高性能，这正是攻击者的期望。如上所述，集中式服务器无法观察到客户端的训练数据更新，也无法注意到攻击者的恶意更新，这给防御此类攻击带来了严峻的挑战。当前，没有太多有效的工作来解决联邦学习针对模型中毒攻击的脆弱性。



图（3） 图解模型攻击

2.2.2 数据中毒攻击实验复现

为了在 FL 系统中模拟 N 个参与者（其中 m%为恶意）的标签翻转攻击，在实验开始时，我们从所有参与者中随机指定 N * m%个参与者为恶意，其余的都是诚实的。为了解决随机选择恶意参与者的影响，默认情况下，我们将每个实验重复 10 次并报告平均结果。 除非另有说明，否则我们使用 m = 10%。

在实验的复现中，我们选择将 Fashion-MNIST 训练集中的“trouser”标签替换为“dress”标签。

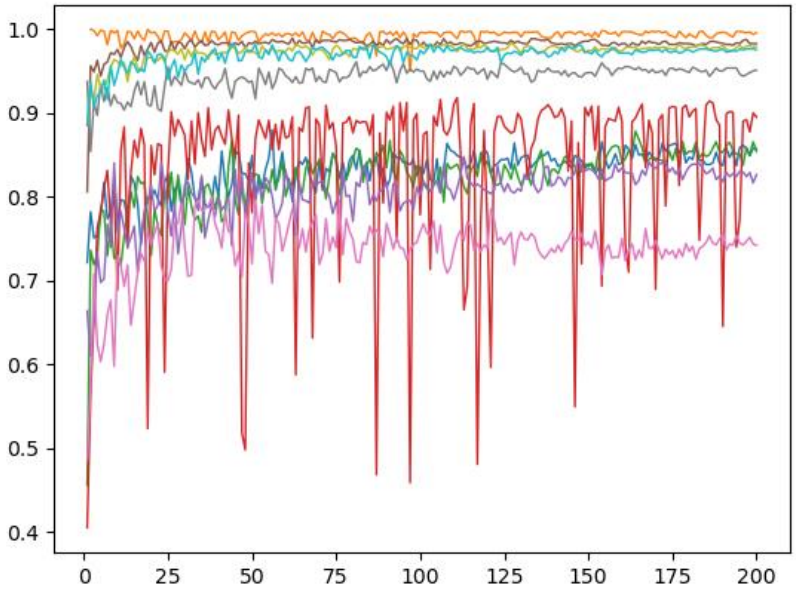
1、标签反转部分代码如下：

```
def poison_data(logger, distributed_dataset, num_workers, poisoned_worker_ids, replacement_method):
    # TODO: Add support for multiple replacement methods?
    poisoned_dataset = []
    class_labels = list(set(distributed_dataset[0][1]))
    logger.info("Poisoning data for workers: {}".format(str(poisoned_worker_ids)))
    for worker_idx in range(num_workers):
        if worker_idx in poisoned_worker_ids:    #如果是选中的被毒害客户端，那么进行标签反转

poisoned_dataset.append(apply_class_label_replacement(distributed_dataset[worker_idx][0], distributed_dataset[worker_idx][1], replacement_method))
        else:                                    #正常的直接分配数据集
            poisoned_dataset.append(distributed_dataset[worker_idx])

    log_client_data_statistics(logger, class_labels, poisoned_dataset)

    return poisoned_dataset

def replace_1_with_3(targets, target_set):
    for idx in range(len(targets)):
        if targets[idx] == 1:
            targets[idx] = 3

    return targets
```

2、联邦学习聚合部分代码如下：

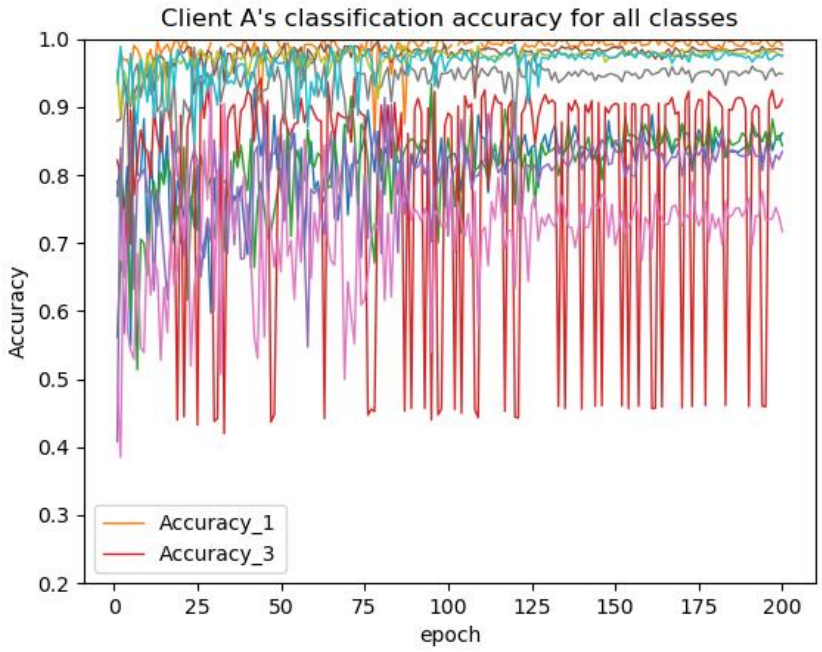
```
#把训练的客户端参数用平均的策略生成新的模型参数
args.get_logger().info("Averaging client parameters")
parameters = [clients[client_idx].get_nn_parameters() for client_idx in random_workers]
new_nn_params = average_nn_parameters(parameters)
#对每一个客户端更新参数
for client in clients:
    args.get_logger().info("Updating parameters on client #{}", str(client.get_client_index()))
    client.update_nn_parameters(new_nn_params)
```

在实验代码中，本人添加了记录实验精确度的函数，并在实验结束的时候绘制图像，以下为本地客户端与服务端全局模型的子任务精确度随训练轮数变化的图像



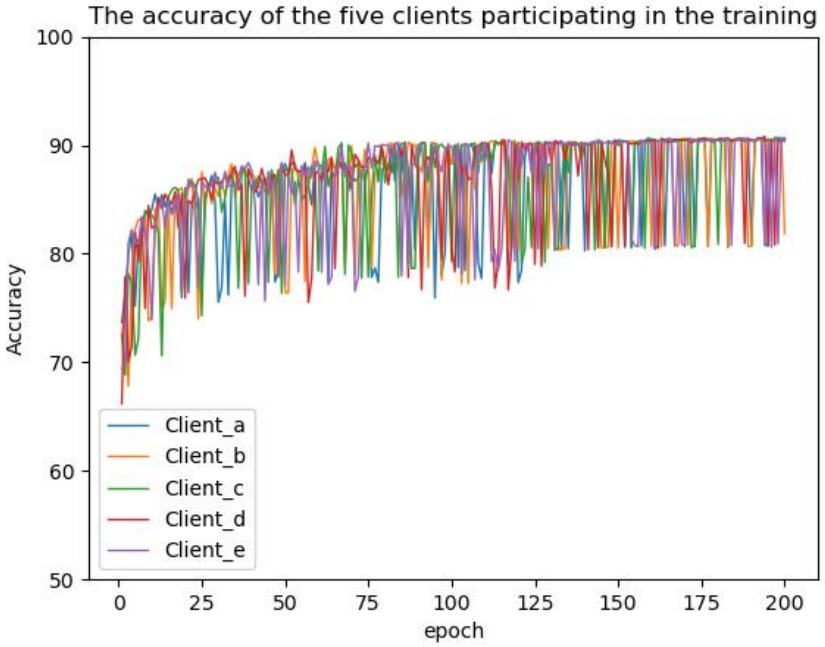
图（4） 全局模型的子任务精确度随训练轮数变化图

上图即为全局模型子任务随轮数变化的精确度变化，可以看到红色线段代表的精确度会不定期降低很多，这是因为在某一轮随机选中了攻击者客户端，在被攻击后，使得全局模型中标签反转攻击的目标子任务精确度降到了很低。下图是某一客户端端随轮数变化子任务精确度的变化。



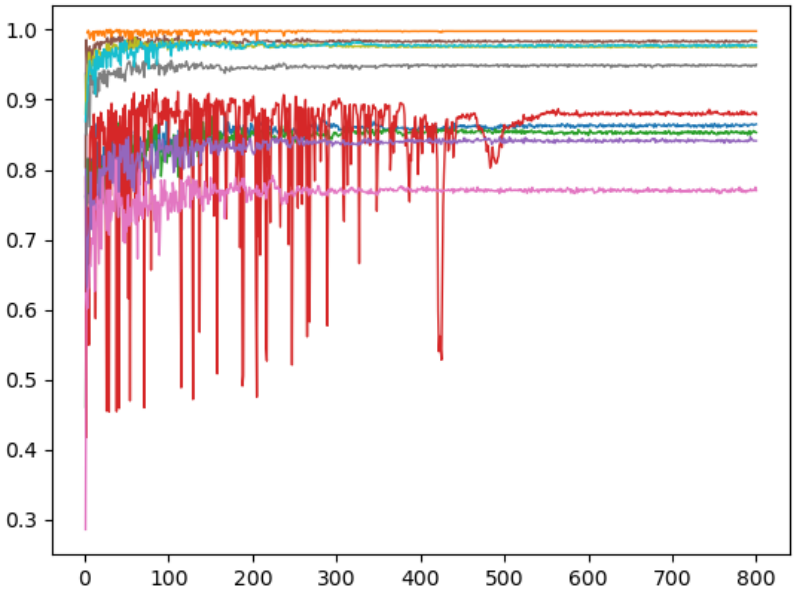
图（5） 随机选中的客户端 A 模型子任务精确度随轮数变化图

下图是随机选择的五个客户端的全局任务精确度随时间变化的图像，可以看到，当有攻击者进行攻击的时候，子任务的精确度会下降，客户端全局任务的精确度也会随之下降，为聚合带来不稳定因素。

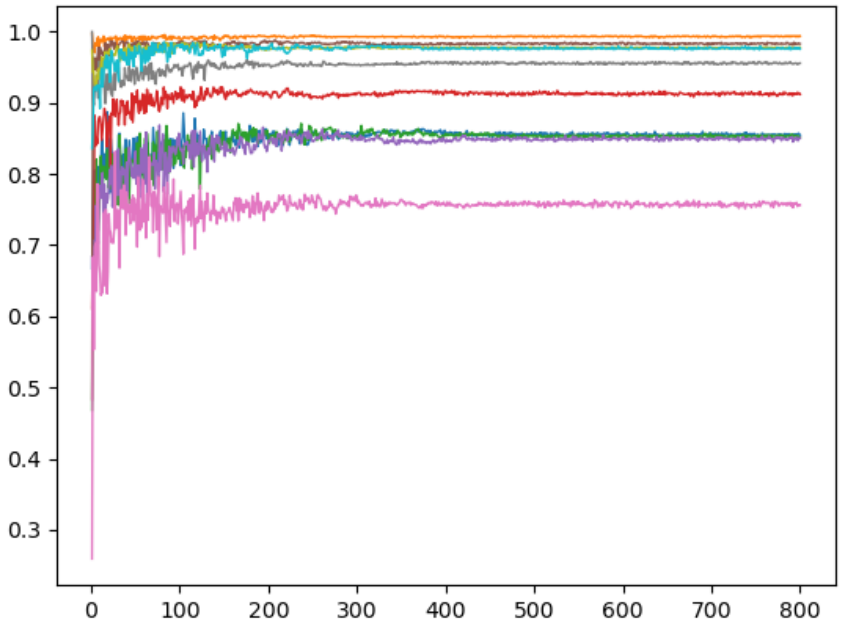


图（6） 随机选中的客户端 A 模型全局任务精确度随轮数变化图

上述实验只训练了 200 轮，这是我们可以看出联邦学习还没有收敛稳定，我们将训练轮数延长到了 800 轮，并且做一组没有攻击者的正常联邦学习训练的实验，再做一组有攻击者的实验进行对比，结果如下。

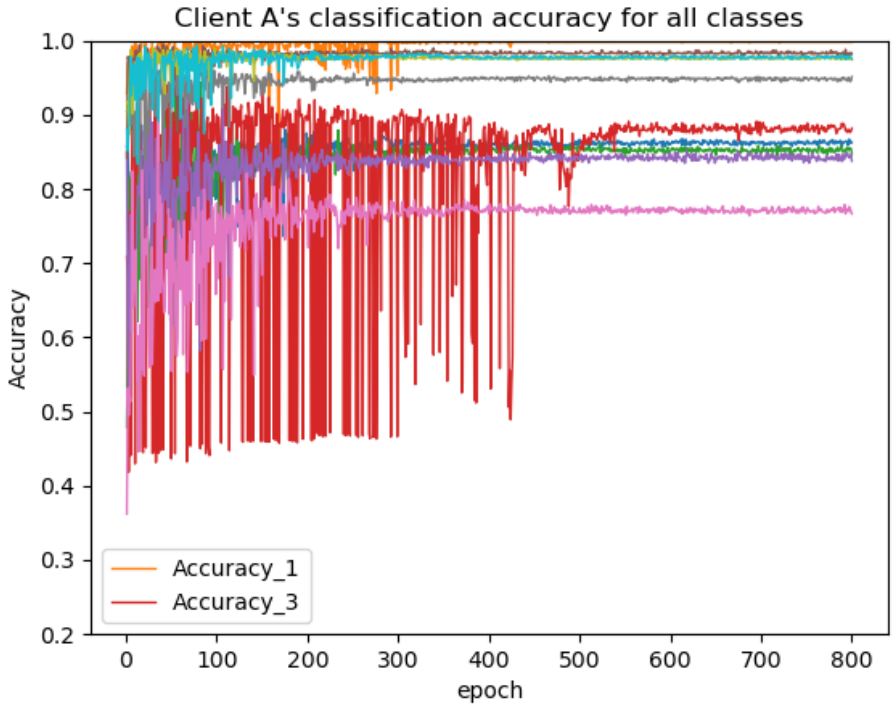


图（7） 全局模型的子任务精确度随训练轮数变化图

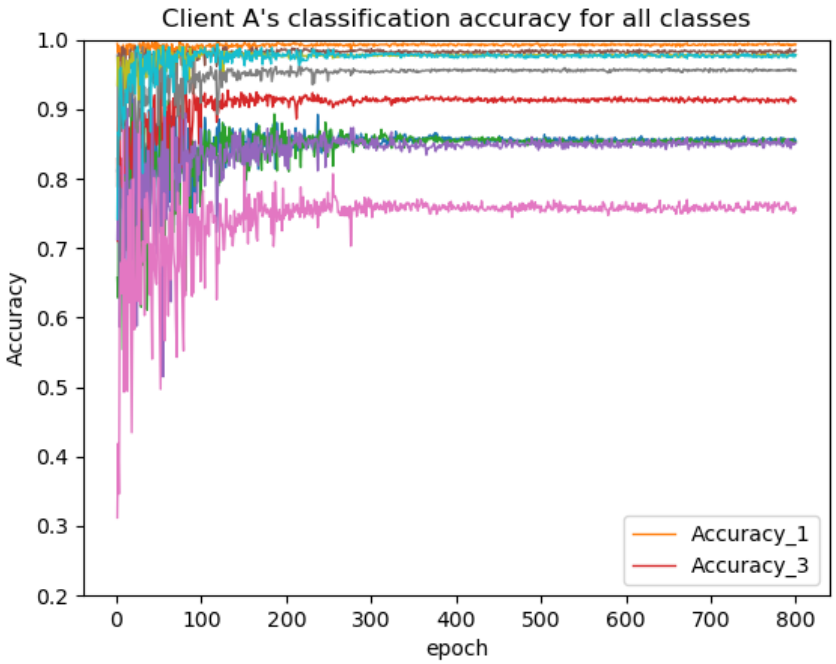


图（7） 没有攻击者时全局模型的子任务精确度随训练轮数变化图

通过对两次试验的对比，我们可以看出在没有攻击者的时候，联邦学习收敛的速度更快，而且原本被攻击的子任务的分类精确度更高，以下本地客户端的子任务分类精确度对比图更加验证了这一现象。



图（8）客户端 A 的子任务精确度随训练轮数变化图



图（9） 没有攻击者时客户端 A 的子任务精确度随训练轮数变化图

第三章 联邦学习的研究意义与现状

3.1 联邦学习的实际应用价值与现实意义

联邦学习具有保护隐私和多方本地数据安全的极大优势。避免集中式存储数据，安全合规地从多源不互通的数据中创造新的价值，充分利用各方数据资源，优化机器学习训练结果，学习参与方可以在联合形成协同合作的联邦大数据环境，形成联邦学习生态。对金融、互联网、通信、零售、交通运输、工业生产等行业提供计算服务支持。我们可以从以下四方面窥探联邦学习的价值和意义。

1、丰富的数据资源是联邦学习最大的金矿。

原本分散在各规模企业的数据，通过联邦学习生态达成，可以发挥其自身作用，有了更好的用武之地。利用“联邦学习+人工智能”真正的赋能大数据并反哺个人和企业业务，用数据和科学提升业务效益。

2、打破传统企业机构的数据边界，利用联邦学习提升智能化效果。

改变过去商务智能和政府仅仅依靠机构内部数据的局面。

3、达成各行业联手，共建全行业的联邦学习生态。

联邦学习的出现已经开始改变大数据在各行各业的应用方式，联邦大数据生态的构建也离不开学界和工业界的共同探索和推动，使用联邦学习技术的各方应当携手，联合制定数据联邦行业规范，促成多方联邦数据协议，达成标准化、协同化、规范化的联邦学习环境。

在信息流通日益渗透到企业和个人的今天，联邦学习将逐渐成为金融、保险、投资、医

疗等众多行业领域实现商业价值和隐私安全保护的最佳途径。

3.2 联邦学习国内外研究现状

联邦学习的技术框架建设方面。谷歌首先提出开源的离散数据联邦学习应用框架 TensorFlow Federated (TFF)。TensorFlow Federated 主要支持利用如今数量众多的移动智能终端设备和边缘端计算设备的计算能力，保证数据不离开本地的同时训练本地机器学习模型，通过 Google 开发的 Federated Averaging 算法，即使在较差的通信环境下，也能实现保密、高效、高质量的模型汇总和迭代流程，且移动端和边缘端用户体验上不做任何牺牲和妥协。目前 Google 已经将联邦学习应用在移动设备键盘输入预测上。

在学术研究与行业应用上，腾讯发起的中国首家互联网银行——微众银行正在积极探索。在国际人工智能专家、微众银行首席人工智能官杨强教授带领下的 AI 团队开源了首个联邦学习“FATE（FederatedAI Technology Enabler)” 工业框架，作为安全计算框架支持联合 AI 生态系统，该框架可以实现基于同态加密和多方计算的安全计算协议，在信贷风控、客户权益定价、监管科技等领域推出了相应的商用方案。

目前，联邦学习的国际标准化工作已经完成，随着 IEEE 联邦学习基础架构与应用标准工作组的第二次会议在美国洛杉矶的召开，海内外 13 家来自科技、金融、教育、医疗等不同行业的知名研究机构及企业从多角度探讨联邦学习技术的应用案例，对联邦学习标准草案的制定提出建设性意见，该标准草案于 2020 年出台，意味着将为立法和监管提供更多技术依据。

第四章 总结

4.1 学习总结

通过对计算机前沿介绍课程中人工智能专题的学习，使我学习了人工智能的各种基本算法和思想，了解了各种方法的应用领域和适用范围。未来的编程开发中，也需要对采集的数据进行处理和做出判断，因此必然涉及人工智能的相关知识。课程包含内容很多，涵盖的领域非常广泛，虽然学习深度有限，但是正是对人工智能知识的广泛了解，才能扩展我的研究思路，选定方向和研究算法，进行更深层次的研究。

4.2 引用与参考

[1] Wang Y , Zhu T , Chang W , et al. Model Poisoning Defense on Federated Learning: A Validation Based Approach[M]. 2020.

[2] Tolpegin V , Truex S , Gursoy M E , et al. Data Poisoning Attacks Against Federated Learning Systems[J]. 2020.

[3] https://zhuanlan.zhihu.com/p/79284686

[4] https://zhuanlan.zhihu.com/p/100688371