兼顾高效与安全的松耦合跨域协作学习平台

苏家兴[1]，赵乙[2*]，李奥[1]，谭崎[1]，刘自轩[1]，松永健宏[1]，徐恪[1*]
（1. 清华大学计算机科学与技术系，北京 100084；
2. 北京理工大学网络空间安全学院，北京 100081）

摘要：协作学习受制于苛刻的通信与部署条件，无法实现高效部署。此外，协作学习面临对抗攻击和隐私泄露等安全威胁。为了促进协作学习在现实世界的部署应用，设计并实现了兼顾高效和安全的松耦合跨域协作学习平台。采用松耦合设计模式，高效并安全地实现协作学习。同时，采用云—边—端协同的三层架构，针对训练任务的合法性与隔离性进行了系统安全设计，从而在系统层面上同时兼顾高效性与安全性。相比于集中式人工智能方案，基于本平台实现的协作学习，在黑产商家识别和恶意流量检测 2 个任务上，预测性能分别提升 35.29% 和 8.30%。在抵御对抗攻击方面，部署对抗训练模块前后，模型鲁棒性在两个任务上分别提升了 570% 和 290%。在抵御成员推理攻击方面，部署差分隐私模块前后，攻击成功率降低了 26.33%。

关键词：协作学习；网络空间安全；高效安全；对抗训练；差分隐私
中图分类号：TP309.2 **文献标志码：**A
DOI: 10.20172/j.issn.2097-3136.240605

Loose coupling cross-domain collaborative learning platform with both efficiency and security

SO Kahing[1], ZHAO Yi[2*], LI Ao[1], TAN Qi[1], LIU Zixuan[1], MATSUNAGA Takehiro[1], XU Ke[1*]
（1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China）

Abstract: Collaborative learning faces challenges in the real-world deployment due to the stringent communication and deployment condi- tions, as well as security threats like adversarial attacks and privacy breaches. To facilitate the practical application of collaborative learning, a loose coupling cross-domain collaborative learning platform that balances efficiency and security was designed and implemented. Specifically, a loosely coupled design pattern was employed to achieve the collaborative learning with efficiency and security. A three-tier architecture encompassing cloud, edge, and endpoint collaboration was adopted, addressing the system's security with regard to the training task legitimacy and isolation, and simultaneously ensuring efficiency and security at the system level. Compared the to centralized artificial intelligence solutions, collaborative learning implemented on this platform exhibited performance improvements of 35.29% and 8.30% in tasks involving the underground business recognition and the malicious traffic detection, respectively. In terms of the defense against adversarial attacks, the model's robustness increased by 570% and 290% in the two tasks after deploying an adversarial training module. Furthermore, the success rate of member inference attacks decreased by 26.33% after deploying a differential privacy module.

Keywords: collaborative learning; cyberspace security; efficient and security; adversarial training; differential privacy

* **通信作者**. E-mail: zhaoyi@bit.edu.cn；xuke@tsinghua.edu.cn

基金项目：国家自然科学基金（62202258，62425201，62132011，U22B2031）

引用格式：苏家兴，赵乙，李奥，等. 兼顾高效与安全的松耦合跨域协作学习平台 [J]. 网络空间安全科学学报，2024，2（6）：74 - 85.

Citation Format: SO K H, ZHAO Y, LI A, et al. Loose coupling cross-domain collaborative learning platform with both efficiency and security[J]. Journal of Cybersecurity，2024，2（6）：74 - 85.

0　引言

随着生成式预训练（Generative Pre-trained Transformer，GPT）[1]等大模型的兴起，数据成为人工智能领域最重要的资产之一。特别地，真实且优质的数据存在巨大缺口。事实上，互联网规模持续增长的同时，互联网所产生的数据量也呈现指数级的增长趋势。但是，这些真实数据分散在互联网的各个设备或终端中，由于涉及用户隐私与终端安全等问题，此类数据资产无法直接运用到人工智能领域，从而形成大量的数据孤岛。

联邦学习（Federated Learning，FL）等协作学习（Collaborative Learning，CL）算法[2-3]作为一种新型的分布式智能模型训练范式，能够在保证数据隐私安全的前提下，实现高效的分布式协同智能，有效解决了数据孤岛问题。具体来讲，协作学习主要涉及聚合节点和多个本地节点①。聚合节点负责训练任务的初始化，并且收集各个本地节点的模型参数。随后，通过聚合算法计算出全局模型，并将其下发到各个本地节点以进行新一轮次训练，循环往复。因此，协作学习允许各个本地节点在不直接共享原始数据的条件下，得到全局最优的模型。

尽管协作学习在学术界与工业界都备受关注，但目前的协作学习实现方案普遍具备节点部署复杂、框架学习成本高、协作算法缺少整合等特点。协作训练框架不统一、通信协议不一致、编码规范不相同等原因，导致相关方案难以复现，阻碍了基于协作学习的技术应用与推广。此外，与其他人工智能技术一样，协作学习容易遭受对抗攻击，具有一定的脆弱性。不仅如此，协作学习涉及多方数据共享，增加了隐私泄露的风险。总体来说，尽管协作学习技术持续发展，但目前已有的框架和平台存在实现难度大、部署流程复杂、功能不完整、使用方式不标准化等缺陷。

为了解决上述问题，本文设计了一个兼顾高效与安全的松耦合跨域协作学习平台（以下简称“平台”）。该平台采用松耦合设计思想，针对协作学习上层算法进行了分解，包括：“数据预处理”“模型定义”“全局聚合器”“本地训练器”和“全局测试器”5 个核心模块。同时，为了提升基于此平台设计的协作学习算法的安全性，本文提出的协作学习平台内置了安全算法集成模块。该模块包含对抗训练及差分隐私算法，以即插即用的方式直接调用。总的来说，本文提出的平台为研究者提供了快速部署基线与对比实验的条件，以及更加安全的协作学习训练范式。

为了验证平台的可用性和相关设计的有效性，本文采用蚂蚁集团组织科技精英赛（ATEC2022）线上赛阶段的真实黑产电商数据集（以下简称“黑产数据集”）和 CICDDoS2019[4]流量数据集。同时，本文分别针对不同数据集设计了基于协作学习的智能模型，并在平台上开展了性能评估。

综上所述，本文主要贡献如下：

（1）针对协作学习上层算法与系统调用耦合性高导致算法设计低效的问题，采用松耦合设计模式对协作学习进行了分解，显著提高了算法设计效率及安全性。

（2）为了解决已有协作学习开源框架使用困难、部署复杂的问题，设计并实现了一个云—边—端三层架构以实现协作学习平台。

（3）为了提升协作学习平台在对抗攻击和隐私泄露方面的安全性，提出的平台内置了即插即用的安全算法集成模块，其中包含 9 种主动对抗训练算法和 2 种差分隐私算子。

（4）基于真实工业场景的黑产商家数据集和开源加密流量数据集的实验评估表明，本文提出的协作学习平台在可用性、高效性与安全性方面均取得良好效果。

本文第 1 节将介绍已有的协作学习系统及相关算法研究；第 2 节针对协作学习流程进行解耦与规范化，使其能够被完全定义；第 3 节主要讨论协作学习平台的系统架构设计；第 4 节描述黑产电商数据集及 CICDDoS2019 数据集在协作学习平台上的实验评估；第 5 节为本文总结与未来工作的讨论。

1　相关工作

为了解决数据孤岛问题，协作学习在 2016 年被首次提出[2]，协作学习被分成横向协作学习和纵向协作学习[5]。横向协作学习面向的是多方之间的特征空间基本重叠的场景，即多方数据具有相似特征，与分布式机器学习中的数据并行相似。而纵向协作学习主要针对相同空间的不同特征进行聚合与协作。后来逐渐有更多技术与协作学习进行融合，其中包括协作学习分别与迁移学习和区块链所产生的协作

① 本文将交替使用本地节点和协作者，来表达同一含义。

迁移学习[6-7]和群体学习[8]。

文献 [2] 中提出使用 FedAvg 对模型进行聚合，从而实现在不共享数据的前提下进行多方协作学习；FedProx[9] 主要针对协作过程中数据和设备异质性进行优化；Scaffold[10] 的目标在于保证模型能的同时减少通信量；FedPAGE[11] 通过自适应裁剪来应对不同个体在算力等方面的差异，从而提升全局训练效率；FedNova[12] 的核心思想是采用归一化平均方法消除目标不一致性，同时保持快速的误差收敛, 有效地消除由于数据异质性导致的目标不一致性。

为了提高协作学习的效率与安全性，其他领域相关研究被广泛应用到协作学习中。2017 年，Geyer 等[13] 提出了协作学习过程中可能遭受攻击导致隐私数据泄露，通过结合差分隐私的技术可以实现更加安全的协作学习过程。差分隐私逐渐成为协作领域中一个非常重要的组成部分，大量关于差分隐私运用在协作学习中的研究被提出[14-16]。此外，Tan 等[17] 从信息论的角度对协作学习开展了隐私方面的脆弱性分析，在提出量化隐私泄露风险的同时，提出了基于主动增加噪声的隐私控制方法，给参与协作学习的个体提供了灵活的隐私控制能力。2020 年，Zhang 等[18] 和 Bhagojian 等[19] 提出了在协作学习过程中，当协作的其中一方为恶意节点，并对训练数据进行恶意篡改，便能轻易实现投毒攻击，导致全局模型失效，对抗训练可以作为其中一个有效的手段解决这种攻击。2022 年，Jiang 等[20] 为了提高协作过程的通信效率，提出了一种具有自适应和分布式参数修剪的 PruneFL 算法。2023 年，Chen 等[21] 针对网络入侵检测系统场景提出了一种具有理论保证的新型基于优化的输入扰动防御策略 FedDef。Zhang 等[22] 提出了基于区块链的可信公告板，在保证全局模型准确率的前提下，降低了协作过程数据隐私泄露的风险。

近年来，为了简化协作学习的部署和使用，学术界与工业界都实现了大量开源或商业的协作学习框架。FATE[23] 作为目前最主流的开源协作框架之一，是由微众银行开发的一个工业级协作学习框架，旨在提供不同组织之间的协作学习服务。尽管 FATE 在功能上十分完备，但是 FATE 仅提供了算法级接口，研究者必须修改 FATE 的源代码来实现他们自己的联邦算法，这对于非专业人士来说并不容易。TFF（TensorFlow Federated）[24] 为 Google 团队所推出的基于 Tensorflow 的协作学习训练框架，使用方式简单，通过接口方式直接导入库即可快速应用在 Tensorflow 的模型上。但其不支持 PyTorch 或其他框架模型，且只能实现单机模拟多节点的协作环境。PySyft 是由 Ryffel 等[25] 提出的，它相比于 TFF 能更好地兼容 PyTorch 和 Tensorflow 2 种框架模型，既能支持单机模拟环境，也能实现多机多节点的分布式环境。但 PySyft 在提出后缺少维护，缺少详细的使用和接口文档，对研究者的使用并不友好。FedML[26] 更注重提供算法级 api 供用户直接使用，从而简化协作过程的环境部署。它同样能支持 PyTorch 和 Tensorflow 以及单机和多机的情况。但是对于绝大多数使用 FedML 的协作学习实验，尤其是当该实验涉及到不仅是上层协作算法，还有协作学习流程的设计（如网络节点配置、通信交互过程，通信交互内容等）时，仍然需要修改其源代码。近几年还有大量商业框架选择进行开源，包括百度的 PaddleFL[27]、阿里巴巴的 Federatedscope[28]、京东的 Fedlearn[29] 等，但是绝大部分框架在开源后便不再维护，而是在公司内部继续迭代以供商业使用，因此本文将不再进行详细介绍。

2 松耦合协作学习设计

为了简化协作学习流程中繁琐的网络通信配置及复杂的底层调用关系，使得研究者能够将重心放在上层协作学习算法的设计上，必须保证算法与系统调用被解耦，并将上层协作学习算法表示为计算节点可以直接执行的形式，再由平台完成任务分发与管理。这将解决第 1 节中所述难题，即大部分协作学习框架都需要对框架源代码进行修改才可自定义协作学习流程。

为了实现上述目标，平台根据协作学习的特点，将上层协作算法分解成 5 个核心模块：数据预处理模块、模型定义模块、全局聚合器模块、本地训练器模块和全局测试器模块，每个核心模块之间相互呈现松耦合的特性。此外，平台还内置了协作学习算法集成模块，该模块能在全局聚合与本地训练中实现即插即用，以提高协作学习安全性。研究者可以根据自己的实验配置，更改不同模块的算法实现，以保证研究者设计协作算法时能同时具备高效性、灵活性与安全性。而平台则会针对不同模块，相对应地进行任务的打包和下发，以提高协作学习任务在具体实施时的执行效率。各模块及其子模块的协同方式如图 1 所示。

2.1 数据预处理

协作学习中第一个环节是准备训练数据，相较于一般机器学习领域中的数据预处理，协作学习更

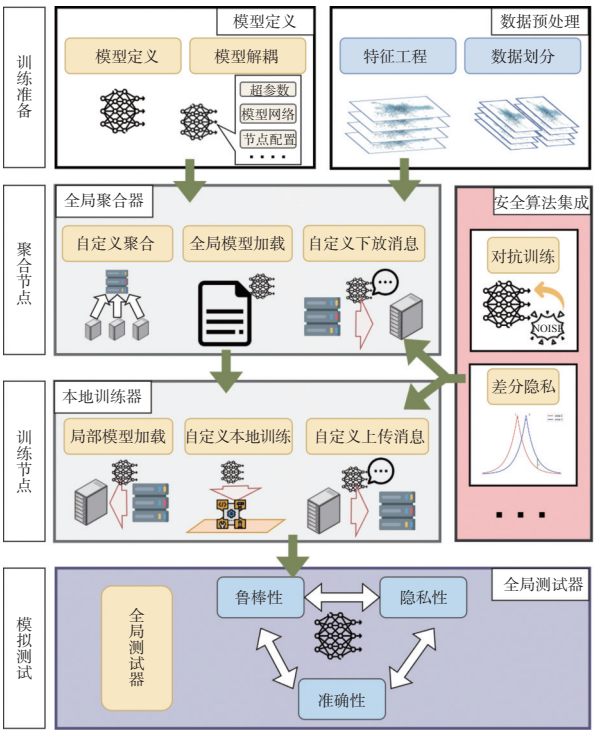


图 1　各模块协同方式

Fig. 1　Interaction of each module

关注每份数据的独立性和隐私性，因此对于该模块必须提供对于每份数据进行独立数据预处理的能力。同时，解决协作各方的数据量差异性与数据特征异构性的问题是现有协作学习领域中的关键问题，因此高效地对数据进行划分对于提高协作算法设计的效率有着显著意义。

平台根据上述协作学习数据的特点，将数据预处理模块抽象成 2 个子模块：

（1）以全量数据为单位的数据划分模块。研究者从宏观上对全量数据进行数据划分。针对实验性质训练任务（该类任务不涉及数据隐私性问题），平台会对每个计算节点分配完整的数据集，数据划分可以在不涉及数据移动的情况下动态调整，极大地减少了因修改数据划分导致数据频繁移动所产生的网络通信开销，提高了协作学习的训练效率。平台也提供了分批上传数据集的能力，即单独对每个节点的数据集进行独立上传，以满足在真实环境下的协作学习需求（即需要保证数据隐私性）。

（2）以局部数据为单位的特征工程模块。研究者从微观上对每个协作节点上的数据进行常见的数据预处理功能，包括数据清洗、挖掘与特征工程。该子模块为研究人员提供能够快速实现差异化数据预处理的方式，以分析数据异构性对协作过程的影响。

2.2　模型定义

当数据完成准备后，需要各方对协作过程的模型进行协商与定义，而平台将模型协商和定义的过程形式化成第二个模块，该模块会在聚合过程、本地训练过程及测试评估过程中被使用。模型定义方式与一般机器学习框架基本一致，为了更加贴合研究者使用习惯，平台针对 PyTorch 与 Tensorflow 2 种常用框架进行了简化，平台仅对模型类名称进行了限制，不对类中实现进行任何限制，以提高模型的灵活度。对于 XGboost、Lightgbm 等其他决策树模型，也均可在平台上实现协作，但这种类型的模型需要使用者在其他部分自行进行模型读取。

2.3　全局聚合器

协作学习的核心在于聚合，大部分的研究都是针对聚合算法而展开，因此第三个模块被设计为全局聚合器。聚合器首先需要保证聚合算法的灵活性，即研究者能够根据自身的模型与数据特征进行聚合算法的调整。其次，由于聚合过程可能涉及除了模型以外的其他额外信息以实现更复杂的聚合策略（如各个计算节点的状态、训练时长、数据分布信息等），因此聚合器还应该能自定义与计算节点的通信内容。为了满足这 2 个特性，平台为聚合器设计了 3 个子模块：

（1）自定义聚合算法模块。每一轮聚合时，会由聚合器先收集各个计算节点完成本地计算后发来的消息，然后调用该模块，研究者通过对该子模块的设计，能够自定义高效安全的聚合算法，该模块完成后会将聚合后的参数发送到下一个子模块。

（2）全局模型加载模块。该子模块接收上一个子模块完成聚合后的全局参数，并对聚合后的模型及其相关信息进行后处理与加载，该模块保证了研究者对聚合后的全局模型参数进行微调和后处理的能力，也为下一个子模块模型下发做准备。

（3）自定义下发消息模块。该子模块用于定义下发给各个节点的通信消息。一方面，聚合器可以携带更丰富的全局参数以协助各个计算节点进行本地训练；另一方面，如何优化协作过程的通信开销是协作学习中十分重要的研究内容，因此通过对该子模块的设计，可以实现模型剪枝等策略，以降低协作学习的通信开销。

2.4　本地训练器

除了对聚合算法的设计以外，各个节点如何进行本地训练并对数据进行隐私保护对于本地训练器的设计是不可或缺的。与 2.3 节全局聚合器相似，在

协作学习中还需要额外考虑本地训练器与全局聚合器之间的通信内容，因此本地训练器与全局聚合器的实现呈现为对偶形式。需要注意训练器与聚合器的子模块顺序不一样，是为了方便读者理解整个过程解耦后的顺序，在实际编程中接口写法顺序不会造成影响。

（1）局部模型加载模块：该子模块接收从聚合器所发送的通信消息，并以此消息的内容对局部模型进行初始化、加载和预处理，该子模块在每一次聚合下发时都会由计算节点所调用。

（2）自定义本地训练模块：该子模块能够实现对局部模型训练算法的设计与训练任务的调试。为了简化平台的使用，平台对该子模块的设计兼容主流的训练框架，因此研究者能够快速把所设计的单机训练代码迁移到平台上并进行协作训练。

（3）自定义上传消息模块：该子模块能够自定义上传到聚合器的通信内容，与全局聚合器中的自定义下发消息模块基本一致。

2.5 全局测试器

测试评估模型性能是协作学习中最后且重要的一环，因此协作学习最后一个模块被设计为全局测试器。全局测试器为研究者提供在训练结束后测试模型的能力。该模块使用训练结束后的全局模型对测试集进行测试，测试集在第 2.1 节数据预处理阶段的数据划分中得到。平台还为研究者提供了测试器日志以查看测试结果，并配备了可视化部件以进行更直观的分析。平台的协作学习安全算法集成模块还提供了攻击模块，以用于评估模型的鲁棒性与隐私性的能力。

2.6 安全算法集成模块

为了兼顾协作学习算法的高效性与安全性，平台还内置了安全算法集成模块，该模块以即插即用的方式直接应用于全局聚合器（见第 2.3 节）与本地训练器（见第 2.4 节），为研究者提供在算法层面上保护数据隐私与模型安全的能力。目前该模块包含对抗训练与差分隐私 2 个子模块。

（1）对抗训练模块。该子模块利用对抗训练技术提升协作学习的鲁棒性和安全性。对抗训练是一种增强模型鲁棒性的技术，其原理是在训练数据中加入经过精心设计的微小扰动（即对抗样本），使得模型在这些扰动数据上也能表现出良好的性能。在协作学习中，由于每个协作节点都有可能遭受外部攻击，因此引入该模块意义重大。基于此模块，研究者可以依据具体的任务需求，选择模块预置的如 PGD、FGSM、JSMA 等 9 种对抗攻击算法，或将它们组合使用来生成对抗样本并进行协作学习模型训练。不仅如此，平台还提供了一套完整的鲁棒性评估工具，专门用于测量协作学习模型的鲁棒性表现。

（2）差分隐私模块。该子模块利用差分隐私技术来提高系统的数据隐私保护能力，使得协作学习系统进一步满足数据“可用不可见”的基本要求。差分隐私是一种通过扰动模型的方式来限制潜在攻击者的技术，通过裁剪模型并向模型中注入精心设计的噪声来达到这一目的。在协作学习中，包含隐私信息的数据是分散在各边缘计算节点的，而携带数据信息的模型共享会造成数据隐私信息的泄露。通过调用本模块，研究者可以依据具体的任务要求选择高斯噪声或者拉普拉斯噪声对模型进行处理，从而高效、便捷地实现对训练过程的隐私保护。

综上，通过上述协作学习的解耦与模块化，能够将协作学习上层算法设计为计算节点能够直接运行的范式，这为下一节平台的三层架构建立了基础。此时协作学习上层算法能够从整个协作流程中分离出来，研究者可以更加关注协作算法设计本身，而由平台自动化完成复杂的协作环境部署、训练任务分发、训练资源管控等。即插即用的安全算法集成模块也从协作学习算法的层面上保障了平台同时具备高效性与安全性。

3 系统架构

在上一节中，对协作学习算法进行了松耦合分解，以 5 个核心模块实现了协作学习算法层面的高效性与灵活性，并通过安全算法模块保障了其算法安全性。本节将从系统层面分析协作学习平台兼备高效性和安全性所需要面临的挑战和解决方案。第一，需要设计一个高效的平台架构以支持第 2 节所提出的松耦合协作学习算法模块，该架构应以高度自动化的方式执行协作学习任务流，并在维持协作学习灵活性的同时简化研究者部署复杂协作网络的流程。第二，平台需要在系统层面上进行充分的安全性设计，以保障协作学习的数据隐私与模型安全。

3.1 整体架构设计

一般的人工智能系统普遍为二层架构，即控制层和计算层，使用者在控制层进行设计与监控，并由计算层进行实际的训练，一般不涉及大规模的设备异构与数据异构，同时也不涉及隐私安全问题，因此整体的管理相对简单。但对于更强调多方协作的协作学习，仅仅依靠两层架构，无法在维持高灵

活性、高可靠性和高安全性的同时实现高效的协作。因此本文使用三层架构的方式设计并实现了协作学习平台。如图 2 所示，协作学习平台由三层结构实现。

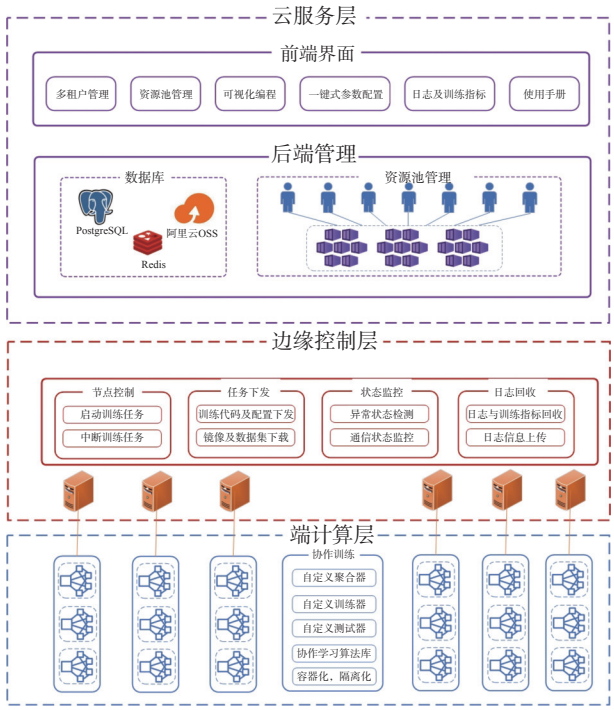


图 2　协作学习平台任务架构图

Fig. 2　Architecture of collaborative learning platform

（1）云服务层。云服务层为三层架构中的最上层，为研究者和管理员提供可视化的交互界面及上层业务接口，后端配备了完备的数据库及资源池管理系统，保证了协作学习平台服务的高可用性。

如第 1 节所述，目前已有的协作学习开源框架缺少易用性的特点，而平台的云服务层就是为了解决这个问题。研究者在该层完成对协作学习任务的所有算法设计及相关网络配置，无须对框架源码进行改动或主动登录到计算节点中进行部署。在训练过程中，研究者也可以在该层实时查看训练状态、训练日志及训练产出。然而，这种简化的设计引发了一个核心问题：如何在简化操作流程的同时，仍然确保协作学习的高度自定义性，或者说，这种简化是否意味着不能处理更为复杂的协作学习任务。对于这个问题，在第 2 节已经提供了详尽的解决方案。

该层涉及资源池的管理，计算资源首先需要在协作学习平台进行注册，完成注册的计算资源会自动在其之上进行计算镜像的下载与部署，并加入资源池中。资源池会定期对计算资源进行状态监测，包括其网络状态（网络连通性）、硬件状态（CPU、GPU、内存、硬盘空间等）和服务状态（镜像是否完备）。对于每个协作学习任务，平台将根据任务参数配置，包括所需计算节点数、计算资源要求进行资源的分配。当计算资源不足时，将进入等待队列，队列将按任务优先级进行分配，若优先级相同，则采用先到先服务（First In First Out，FIFO）策略执行。对于每个提交的协作学习，资源池会自动根据任务情况，为其分配协作过程的聚合节点与本地计算节点，再将每个提交的协作学习任务配置进行打包并向下发送到下层的控制节点。

云服务层还针对业务场景进行了专门的设计和优化，包含科研、竞赛等多种模式，每种模式都具有不同的特点。例如，在竞赛模式下，全局测试器将被锁定为打分规则代码，并在每次训练结束后根据规则为模型自动打分。目前平台已经在实际应用中累计为 400 余名用户提供了 10 000 余次的协作习任务训练。

（2）边缘控制层。边缘控制层为三层架构中的中间层，由多个控制节点组成，控制节点向上负责接收云服务层下发的协作学习任务相关配置与代码，并根据配置与计算资源建立通信链路。建立连接后，控制节点完成以下几项工作。

1）初始化本次协作学习任务。首先需要将本次实验的代码及配置下发到所有涉及的计算节点，并通知计算资源检测本机上是否存在所涉及的数据集，如缺失则向文件型数据库进行请求。随后启动所有计算节点的镜像并开始执行协作学习训练。

2）在协作学习训练过程中，持续监控训练任务状态，作为云服务层及端计算层的通信桥梁。一方面，接收云服务层对于任务的控制信号，如当接收到强制中断信号，则通知下层计算节点立即停止训练，并回收计算资源；另一方面，当计算节点出现异常，立即通知云服务层并发出异常警报。

3）训练过程中持续回收日志，并在训练结束时将训练产出（模型文件、训练指标、竞赛得分等）上传到云服务层。为了保证研究者在云服务层能持续看到训练任务的实时输出，控制节点会回收本次任务所有计算节点的输出日志文件，并将其存储到文件型数据库中，用户每一次查看日志都会向文件型数据库进行请求获取。而对于训练产出，将部分向上直接回传到云服务层，部分存储到文件型数据库中。

（3）端计算层。端计算层为三层架构中的底层，其以所有计算节点作为主体，在该层中进行协作学习的实际训练，其训练状态、日志及模型产出都会

上传到控制层，因此该层为无状态，即仅负责进行计算相关操作，不保留任务信息，以最大化计算节点的计算性能，降低对存储性能的要求。

端计算层所有计算节点使用 Docker 容器进行训练，容器仅在训练过程中被创建，训练结束后即销毁，以确保每个任务的独立性、安全性和计算资源的可重用性。为了提高节点使用率，在平台任务负载量较高时，一般会在同一个节点上同时安排多个协作任务（具体根据资源池情况分配），这得益于协作学习本地训练过程开销一般不会过大，且中间存在大量网络传输及等待聚合时间，同时利用容器隔离性能保证任务之间相互不受影响，在保证训练任务独立性的同时提高了平台整体的效率。

训练以改良后的 FedML[26] 协作框架进行，主要在 FedML 的框架基础上增加了协作流程解耦（见第 2 节）以及协作学习平台的安全算法集成模块（见第 2.6 节）。FedML 提供了多种通信协议接口，包括 PyTorch RPC、Google RPC、MQTT 等，对协作学习过程中聚合节点与本地节点之间的通信消息进行了封装，以同时支持 PyTorch 和 Tensorflow 2 种训练框架。协作过程也支持使用 GPU 进行训练，并会采用 cuda RPC 协议提高 GPU 环境下的通信效率，对于 GPU 的使用配置也可以在 3.1 节中云服务层进行配置，支持设置异构节点协作训练。

完整的协作训练过程如图 3 所示。研究者在云服务层上通过网页服务进行可视化的代码编写及训练配置，随后由平台对任务进行调度、打包及下发。再由任务管理器，即边缘控制节点对计算节点进行控制、监视、下发与回收，其间由端计算节点以 FedML 为框架执行协作学习的训练。训练完成后自底向上回收所有的日志及模型产出，并在云服务端向研究者提供可视化的分析工具。这个架构在保留协作学习过程高度自定义性的同时，为研究者提供了一个更加高效灵活的协作学习平台。

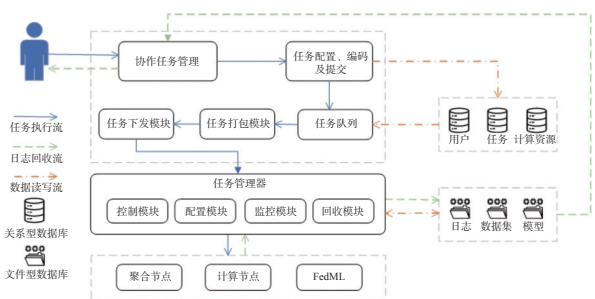


图 3 协作学习平台任务流程图

Fig. 3 Workflow of collaborative learning task

3.2 平台系统安全性

协作学习是为了保护用户数据隐私而被提出，因此协作学习平台对于安全性的设计也需要十分谨慎。算法层面我们提供了安全算法集成模块（见第 2.6 节），而针对系统级的安全保障，主要从以下方面进行详细设计。

（1）权限。在云服务层，所有用户的所有请求都首先需要通过身份认证，且所有的训练任务、数据集、提交代码都会被默认设置为私有，用户若希望公开其训练数据或代码，也需要先通过申请，得到管理员审核才可以公开，而管理员也无权直接获取用户上传的数据集，以保护用户的数据隐私。管理员具备运维控制节点与计算节点的权限，但无权对服务器进行直接访问，仅可对服务器进行任务中断、容器重启、镜像更新等基础操作，以避免管理员直接获取用户的提交记录，最大化保证了用户在云端的数据隐私性。

在边缘控制层，每个控制节点仅能对当前训练任务中所涉及的节点进行管理。在每次训练任务结束后，都会删除该控制节点管理本次计算节点的权限，且清空本次与计算节点的所有通信配置，以保证即使控制节点遭受攻击也不会对其他控制节点或所涉及的计算节点造成影响。

在端计算层，运行代码不具备对操作系统中其他文件目录的访问权限，以确保使用者无法在自定义代码块中嵌入恶意代码对计算节点进行攻击。而且，所有训练任务均使用容器进行隔离，使用者同样无法通过代码块的方式访问或控制其他同时运行的训练任务，以保证每个用户的训练任务隐私性。

（2）网络连接。云服务层与一般网络服务一样，均配置了请求限制及异常警报，当检测到大量异常提交，如分布式拒绝服务（Distributed Denial of Service，DDoS）攻击短时间内提交了大量训练任务，平台首先会对任务提交进行限制，并且将任务安排在训练队列中，以确保正在运行的任务不会被干扰，同时通知网络管理员检查训练任务队列中的任务是否正常。

为了保证控制节点与计算节点不会被外界发起网络攻击，除了云服务层节点以外，平台所有节点均处于内网状态，仅能通过云服务层进行任务下发。所有节点在注册进入资源池时，必须先完成网络配置，以确保整个内网处于安全状态，攻击者无法通过直接攻击边缘控制层或端计算层获取数据集或训练模型。

（3）代码合法性。作为一个实验性质的平台，利用用户的实验代码是最容易对平台进行攻击的方式，因此如何防范恶意代码的植入对于平台和用户数据的安全性至关重要。由于存在上述（1）和（2）两种保护措施，平台对于实验代码采用软防御的策略，以尽可能维持平台的灵活性及可用性，即不直接禁止或中断风险代码，而是在出现异常时进行安全警告与人为处理。

在每次训练任务提交时，平台会先对 5 个算法模块进行静态分析，如是否使用了管理操作系统的第三方库、是否存在获取服务器路径等操作，如果存在则在任务执行下发前立刻向管理员进行提示，并警告任务提交人，再由管理员检查代码内容是否可以继续执行。

在执行过程及结束时会对输出日志进行二次检查，如日志中存在异常输出、打印了服务器信息或尝试访问其他非法路径，均会被记录并警告。当警告累积到一定次数后，将提示管理员需要对风险用户进行提交限制或封禁。

4　实验评估

本文针对协作学习平台及其算法集成库中的对抗攻击与差分隐私库进行了对比实验。实验分别使用了一个百万级黑产电商数据集与 CICD-DoS2019 数据集对平台进行了性能评估，以证明研究者能够简易、高效且安全地使用平台实现协作学习训练任务。

4.1　数据集介绍

黑产电商数据集：该数据集为蚂蚁集团所提供的 ATEC2022 科技精英赛线上赛阶段的真实黑产电商数据集（已进行脱敏操作）。该数据集目标为对测试集中的商家进行二分类，判断其是否为黑产商家。数据集主要包含以下 3 个部分。

（1）交易信息。共计 9 889 830 条记录，包含交易时间、交易金额、交易双方编号等 7 个特征。

（2）买家信息。共计 6 687 853 条记录，包含交易中买家编号、性别、年龄等 4 个特征。

（3）商家标签。共计 30 503 条记录，包含交易中的商家编号及其黑产商家标志位，若标志位为 1，则确定为黑产商家，若标志位为 0，则代表其风险未知，不能确定其是否为黑产商家。

CICDDoS2019 数据集[4]：该数据集由加拿大网络安全研究所公开②。数据集共 50 063 112 个记录，其中 DDoS 攻击共有 50 006 249 个，正常样本共有 56 863 个，使用 CICFlowMeter 从数据集中提取出 80 多种特征，13 种攻击标签。

4.2　实验设计

实验均在协作学习平台上实现，以上 2 个数据集均被划分成了 11 份（10 份用于训练，1 份用于测试），以用于进行分布式的协作训练，因此共使用 11 台计算节点进行协作训练。实验均使用 CPU 服务器进行测试，且每个计算节点均为 8 核 CPU，32 G 内存，通信带宽为 3 Mbps，在 2 个数据集上均设计了对应的全连接模型进行实验。实验分为以下 2 部分。

（1）对比传统单机训练任务和平台上实现的协作学习训练任务，以证明平台具有高可用性与灵活性，能在其之上实现高效的协作学习任务，模型在测试集上表现的推理能力更优。其中因为协作训练对比单机训练在理论上具有更强的算力以及更大的数据集，因此为了对比实验的公平性，采用数据独立同分布的划分方式，以便单机训练也能获得较好的性能。

在 CICDDoS2019 数据集的多分类任务上，以分类精度作为模型准确性的评价指标。在黑产电商数据集的二分类任务上，由于仅靠精度难以区分模型的准确性，因此使用归一化折损累计增益值（Normalized Discounted Cumulative Gain，NDCG）作为评价指标。NDCG 的计算公式如式（1）和式（2）所示，rel（i）表示排名第 i 个项目的相关性，$\log_2(i)$ 表示排名 i 的对数，IDCG 是理想情况下的 DCG 值，通常是对结果进行按相关性降序排列时的 DCG 值，k 设置为 300。

$$\mathrm{DCG}=\mathrm{rel}(1)+\frac{\mathrm{rel}(2)}{\log_2(2)}+\frac{\mathrm{rel}(3)}{\log_2(3)}+\cdots+\frac{\mathrm{rel}(k)}{\log_2(k)} \tag{1}$$

$$\mathrm{NDCG}=\frac{\mathrm{DCG}}{\mathrm{IDCG}} \tag{2}$$

（2）对比一般协作学习训练与使用平台安全算法集成模块的协作学习任务，以证明平台安全算法集成模块能高效地帮助使用者设计并优化训练过程，提升训练产出模型的鲁棒性与隐私性。

分别在黑产二分类任务和流量多分类任务上使用投影梯度下降算法（Projected Dradient Descent，PGD）[30] 和快速梯度下降法（Fast Gradient Sign Method，FGSM）[31] 2 种对抗攻击，并对比在使用平

② 来源：http://www.unb.ca/cic/datasets /CICDDoS2019

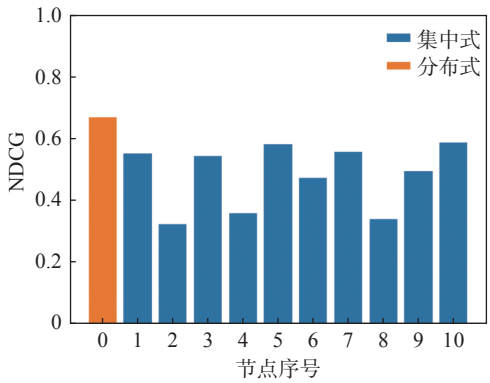


图 4 黑产任务的 NDCG 值对比

Fig. 4 Comparison of NDCG values of underground business recognition tasks

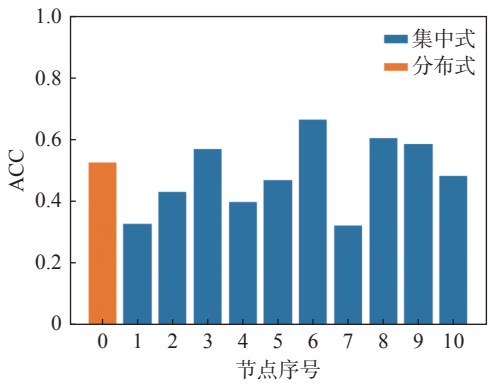


图 5 流量任务的精度对比

Fig. 5 Comparison of accuracy of malicious traffic detection tasks

台主动对抗训练算法前后的模型鲁棒性，即抵御对抗攻击的能力。

还通过 Ml-Docter[32] 实现了对模型的黑盒成员推理攻击，在黑产二分类任务中对平台的差分隐私模块进行了评估，对比添加高斯噪声前后模型的被攻击成功率以评估模型的隐私安全性。

4.3 实验结果

（1）准确性

2 个分类任务的单机训练与协作训练的对比实验结果分别如图 4 和图 5 所示。图中横坐标的 0 号节点对应协作训练，1~10 号节点分别对应 10 个计算节点各自的单机训练，纵坐标分别为 NDCG 值与精度值。

如图 4 所示的黑产任务，每一个训练节点单机训练的 NDCG 值都低于使用协作学习平台进行训练的 NDCG 值，10 个节点单机训练的平均 NDCG 值为 0.476，而协作学习的 NDCG 值为 0.644，提高了 35%。如图 5 所示的流量任务，尽管存在部分计算节点的单机性能优于协作学习，其主要原因是采用均匀分布的数据划分，但是在平均精度上单机节点的精度

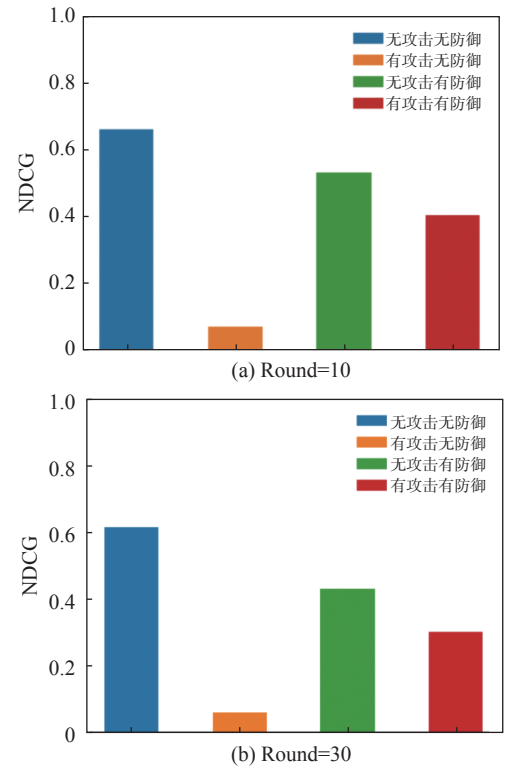


图 6 黑产任务的模型鲁棒性对比

Fig. 6 Comparison of model robustness of underground business recognition tasks

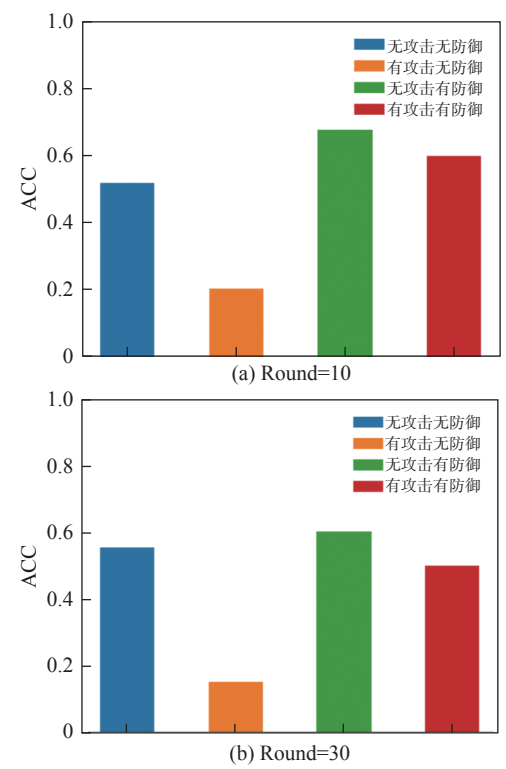


图 7 流量任务的模型鲁棒性对比

Fig. 7 Comparison of model robustness of malicious traffic detection tasks

为 0.482，而协作学习的精度为 0.522，精度提高了 8%，准确性仍然有一定的提升。

实验证明，在实验设计的过程中，即使采用均匀分布这一类对单机训练更加友好的数据划分方式，协作学习在不同的任务上仍然能高效地提升预测准确性。而面对数据分布更加复杂且异构的情况（更符合现实），协作学习能更有效地提升模型的预测性能。

（2）模型鲁棒性

2 个分类任务的模型鲁棒性对比实验结果如图 6 和图 7 所示，其中 round 为协作聚合轮次，图例中“攻击”指测试过程中使用了对抗样本进行攻击，“防御”指训练阶段使用了对抗训练算法。

实验结果显示，在一般的协作学习训练中，对抗攻击在 2 个数据集上都具有十分明显的效果，导致模型在一般样本上有较好的表现，但是在对抗样本上预测准确性极差。在黑产任务上，不同聚合轮次的 NDCG 值分别为 0.071 和 0.057，而流量模型则为 0.205 和 0.153，此时模型基本无法使用。而由于协作学习的数据分散且独立的特点，攻击者能十分容易地实现对抗攻击。

在使用了对抗训练算法后，模型在对抗攻击样本上也能维持较好的准确性，在黑产数据集上，NDCG 值从 0.071 增加到 0.406，提升了 570%，而在流量数据集上则从 0.205 增加到了 0.605，相对于无防御情况提升了 290%，在 2 个训练任务上都十分有效地提高了模型的鲁棒性。

在协作学习中，对抗训练的研究十分重要，平台目前提供了 9 种对抗训练算法，极大地提高了研究者设计算法的效率及安全性，也能有效推进对抗攻击相关的研究。

（3）数据隐私性

在黑产二分类任务中使用了高斯差分隐私进行了训练，并使用了 Ml-Docter[32] 实现了黑盒成员推理攻击，攻击的实验结果如表 1 所示。

实验结果显示，在没有添加差分隐私的模型上，模型的隐私性十分有限，即使使用的是黑盒的成员推理攻击，仍然能达到 70% 以上的成功率。而当模型添加了 ε=10，δ=1e-6 的高斯差分隐私后，成员推理攻击上的成功推测精度减少了 26.33%，F1-Score 减少了 60.16%，模型抵御攻击的能力明显有所提升，尽管模型因添加噪声导致准确性有所下降，但是其下降的幅度在可承受的范围。

平台所提供的差分隐私模块提供了高斯和拉普拉斯 2 种噪声，其中的差分隐私超参数如 ε 和 δ 都可通过对安全算法集成模块的参数调整快速部署并进行对比实验。如表 1 中，通过调整超参数能够快速对比不同模型下的准确性与隐私性，为研究者实现协作学习提供了极大的便利。

表 1　数据隐私性对比实验结果

Table 1　Comparison of data privacy in underground business recognition tasks

	无差分隐私	ε=10, δ=1e-5	ε=10, δ=1e-6
模型NDCG	0.618	0.553	0.547
攻击精度	0.736	0.553	0.542
攻击AUC	0.802	0.589	0.57
攻击F1-Score	0.773	0.623	0.308

5　结束语

为了解决目前协作学习中分布式环境部署困难、算法耦合性高、协作学习算法缺少整合等问题，我们提出并实现了兼顾高效与安全的松耦合跨域协作学习平台。为了保证协作算法的实现具备高效性与灵活性，将协作学习的算法进行解耦，分解成了“数据预处理”“模型定义”“全局聚合器”“本地训练器”与“全局测试器”5 个松耦合的核心模块，以此提高协作学习算法设计和实现的效率，解决因异构框架导致的实验难以复现的问题。同时内置了即插即用的安全算法集成模块，该模块包含对抗训练与差分隐私算法，可以保证协作学习算法层面的安全性。

在系统层面上，平台采用云—边—端分层架构以保障其具备高效性、灵活性与可扩展性，在云服务层进行协作算法设计、训练相关配置及节点资源管理，在边缘控制层进行训练过程中节点的管理、监控和回收，在端计算层进行实际的协作训练。为了保证协作学习实验中数据集、代码和设备的安全，我们还针对平台安全性进行了详细设计。与一般的软件系统不同，平台除了要防御云服务层的攻击，更要防范恶意用户的训练代码植入以从控制层与计算层发起攻击，平台通过对设备权限、网络连接及代码合法性上的严格管理，提高了训练系统的安全性。目前平台已经为 400 余名研究者提供实验环境，并运行了 10 000 余次的协作学习训练任务。

在电商黑产数据集和 CICDDoS2019 数据集上的对比实验中，本平台展现出其在真实分布式环境进

行协作学习的能力，产出的模型准确度超过单机训练。此外，通过整合对抗训练模块和差分隐私模块，平台提高了协作学习训练的鲁棒性和安全性。实验结果表明，在抵御对抗攻击方面，平台部署对抗训练模块前后，模型鲁棒性在 2 个任务上分别提升了 570% 和 290%；在抵御成员推理攻击方面，平台部署差分隐私模块前后，攻击成功率降低了 26.33%。

在未来的工作中，我们将从 3 个方面继续对平台进行改进：第一，丰富平台的算法集成模块，将更多前沿研究纳入平台中，如模型裁剪模块等；第二，针对不同的硬件设备，如可编程交换机和 FPGA 等，增加端设备的接口规范；第三，在云服务层对更多业务场景进行优化，在教育、医疗等多领域提供协作学习服务，进一步降低协作学习的实施难度。

致谢：此工作的实验部分基于 ATEC 前沿科技探索社区 ATEC 科技精英赛项目展开。清华大学作为社区的发起单位之一，与蚂蚁集团合作完成了多项实践型技术人才培养工作。谨就该项工作中蚂蚁科技集团股份有限公司的支持表示感谢！

参考文献：

[1] BROWN T，MANN B，RYDER N，et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems，2020，33：1877-1901.

[2] MCMAHAN B，MOORE E，RAMAGE D，et al. Communication- efficient learning of deep networks from decentralized data[C]//Proceedings of International Conference on Artificial Intelligence and Statistics （ AISTATS ）. PMLR， 2017：1273-1282.

[3] ZHAO Y，XU K，CHEN J，et al. Collaboration-enabled intelligent internet architecture：opportunities and challenges[J]. IEEE Network，2022，36（5）：98-105.

[4] SHARAFALDIN I，LASHKARI AH，HAKAK S，et al. Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy[C]//2019 international carnahan conference on security technology (ICCST). IEEE，2019: 1-8.

[5] WEI K， LI J， MA C， et al. Vertical federated learning：challenges，methodologies and experiments[J]. ArXiv preprint ArXiv：2202. 04309，2022.

[6] LIU Y，KANG Y，XING C，et al. A secure federated transfer learning framework[J]. IEEE Intelligent Systems， 2020，35（4）：70-82.

[7] LIU Y，KANG Y，XING C，et al. A secure federated transfer learning framework[J]. IEEE Intelligent Systems，2020，35（4）：70-82.

[8] WARNAT-HERRESTHAL S，SCHULTZE H，SHASTRY K L， et al. Swarm learning for decentralized and confidential clinical machine learning[J]. Nature， 2021， 594（7862）：265-270.

[9] LI T, SAHU A K, ZAHEER M, et al. Federated optimization in heterogeneous networks[J]. Proceedings of Machine Learning and Systems，2020， 2：429-450.

[10] KARIMIREDDY S P，KALE S，MOHRI M，et al. Scaffold：stochastic controlled averaging for federated learning[C]//International Conference on Machine Learning. PMLR， 2020：5132-5143.

[11] ZHOU G，LI Q，LIU Y，et al. edPAGE：pruning adaptively toward global efficiency of heterogeneous federated learning[J]. IEEE/ACM Transactions on Networking, 2024， 32（3）：1873-1887.

[12] WANG J，LIU Q，LIANG H，et al. Tackling the objective inconsistency problem in heterogeneous federated optimization[J]. Advances in Neural Information Processing Systems，2020，33：7611-7623.

[13] GEYER R C，KLEIN T，NABI M. Differentially private federated learning：a client level perspective[J]. ArXiv preprint ArXiv：1712. 07557，2017.

[14] WEI K，LI J，DING M，et al. Federated learning with differential privacy：algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security，2020，15：3454-3469.

[15] TRUEX S，LIU L，CHOW K H，et al. LDP-Fed：federated learning with local differential privacy[C]//Proceedings of the Third ACM Inter- national Workshop on Edge Systems，Analytics and Networking. 2020：61-66.

[16] CHUANXIN Z，YI S，DEGANG W. Federated learning with Gaussian differential privacy[C]//Proceedings of the 2020 2nd International Conference on Robotics，intelligent Control and Artificial Intelligence. 2020：296-301.

[17] TAN Q, LI Q, ZHAO Y, LIU Z, et al. Defending against data reconstruction attacks in federated learning: an information theory approach[C]//33rd USENIX Security Symposium （USENIX Security 24）. 2024：325-342

[18] ZHANG J, CHEN J, WU D, et al. Poisoning attack in federated learning using generative adversarial nets[C]//2019 18th IEEE international conference on trust，security and privacy in computing and communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE). IEEE，2019：374-380.

[19] BHAGOJI A N，CHAKRABORTY S，MITTAL P，et al. Analyzing federated learning through an adversarial lens[C]//International Conference on Machine Learning. PMLR，2019：634-643.

[20] JIANG Y，WANG S，VALLS V，et al. Model pruning enables efficient federated learning on edge devices[J]. IEEE Transactions on Neural Networks and Learning Systems，2023（34）：10374-10386.

[21] CHEN J，ZHAO Y，LI Q，et al. FedDef：defense against gradient leakage in federated learning-based network intrusion detection systems [J]. IEEE Transactions on Information Forensics and Security，2023（18）：4561-4576.

[22] ZHANG Z，XU K，LI Q，et al. Seccl：securing collaborative learning systems via trusted bulletin boards[J]. IEEE Communications Magazine，2020，58（1）：47-53.

[23] LIU Y，FAN T，CHEN T，et al. Fate：an industrial grade platform for collaborative learning with data protection[J]. The Journal of Machine Learning Research，2021，22（1）：10320-10325.

[24] HARD A，RAO K，MATHEWS R，et al. Federated learning for mobile keyboard prediction[J]. ArXiv preprint ArXiv：1811. 03604，2018.

[25] RYFFEL T，TRASK A，DAHL M，et al. A generic framework for privacy preserving deep learning[J]. ArXiv preprint ArXiv：1811 . 04017，2018.

[26] HE C，LI S，SO J，et al. Fedml：a research library and benchmark for federated machine learning[J]. ArXiv preprint ArXiv：2007. 13518，2020.

[27] MOHASSEL P, RINDAL P. ABY3：a mixed protocol framework for machine learning[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer And Communications Security. 2018: 35-52.

[28] XIE Y，WANG Z，GAO D，et al. Federatedscope：a flexible federated learning platform for heterogeneity[J]. ArXiv preprint ArXiv：2204. 05011，2022.

[29] LIU B，TAN C，WANG J，et al. Fedlearn-algo：a flexible open- source privacy-preserving machine learning platform[J]. ArXiv preprint ArXiv：2107. 04129，2021.

[30] MADRY A，MAKELOV A，SCHMIDT L，et al. Towards deep learning models resistant to adversarial attacks[J]. ArXiv preprint ArXiv：1706. 06083，2017.

[31] GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and harnessing adversarial examples[J]. ArXiv preprint ArXiv：1412. 6572，2014.

[32] LIU Y，WEN R，HE X，et al. ML-Doctor：holistic risk assessment of inference attacks against machine learning models[C]//31st USENIX Security Symposium（USENIX Security 22）. 2022：4525- 4542.