《网络空间安全系统设计》
课程报告

基于参数稀疏化的 FedAvg 算法的横向联邦学习图像分类

小组成员： 张立飞 余天航 李冲 李润锦 刘皓予

日期： 2024 年 5 月 22 日

西北工业大学网络空间安全学院

目录

一、研究背景

1.1 项目简介

本次实验，我们学习了联邦学习的基础理论，了解相关领域研究者的相关工作与前沿的一些发展方向。此外，在第二部分我们介绍了联邦学习的相关算法与参数稀疏化的实现。最后，基于 cifar-10 数据集，我们在 1 个中央服务器，N 个用户的场景下复现了基于参数稀疏化的 FedAvg 算法的横向联邦学习过程。

1.2 相关工作

1.2.1 联邦学习

联邦学习（Federated Leanring，FL）是一种前沿的分布式学习方法，使多个用户能够共享训练结果，同时维护其个人数据的隐私。随着数据安全被日益重视，从不同数据所有者收集数据以进行机器学习预测变得越来越具有挑战性。联邦学习除了增加训练数据之外，还可以保护用户的隐私，同时克服机器学习和深度学习模型面临的挑战。由于数据隐私和安全受到全世界的关注，联邦学习的概念从理论层面日益上升到实践层面。

谷歌引入联邦学习的概念，其明确目标是打破数据孤岛现象[1]。其设计以信息安全、终端数据隐私和个人数据隐私为基础，优先考虑海量用户之间的大数据有效交换和机器学习的执行[2]。除了神经网络之外，联邦学习还可以利用其他流行的机器学习技术，例如随机森林模型，该模型除了加强企业间数据交换的安全性以及因交换而训练模型的精度外，还成功解决了跨组织共享数据时的隐私保护问题[3]随着联邦学习的发展，目前，该模型的实用性已扩展到各种新环境。

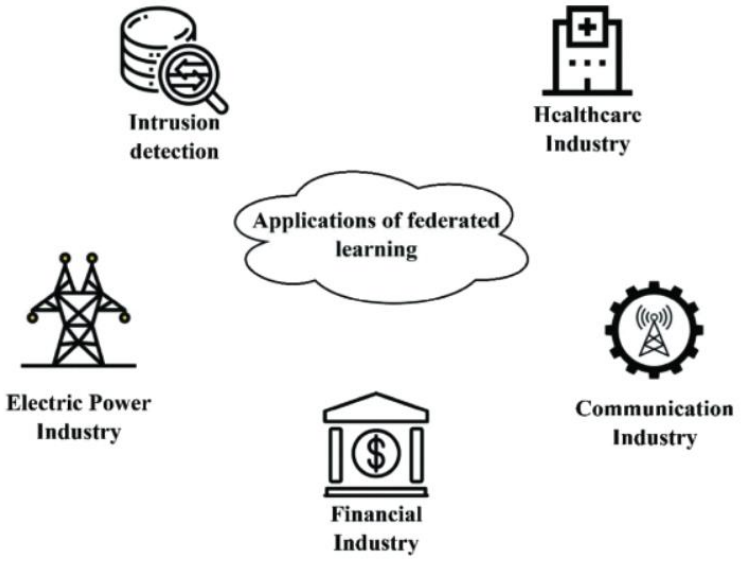

图 1 联邦学习的应用

1.2.2 参数稀疏化

稀疏化策略本质上模型压缩的一种，同样是通过传输少量的参数，一方面减少服务端与客户端之间的网络带宽；另一方面也能够防止全局模型参数泄露。

二、研究内容

2.1 研究方案

2.1.1 系统建模

我们采取实验的系统中，共有 1 个中央服务器，N 个用户，他们之间的交互过程如图 2 所示：

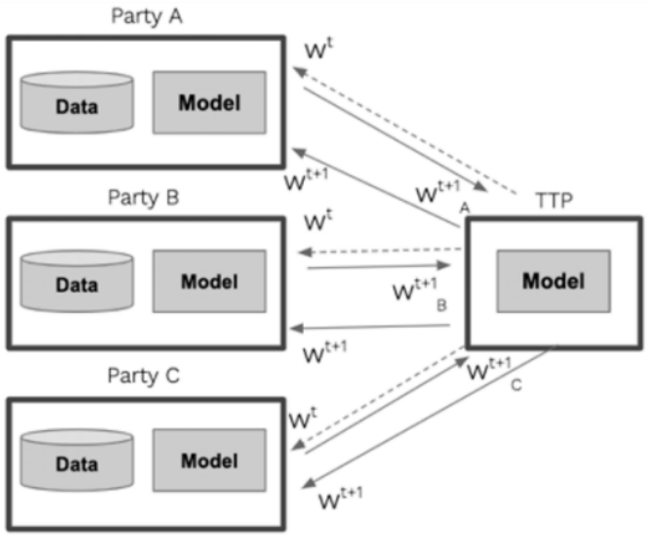

图 2 系统模型

我们采取了以下参数，利用 cifar-10 数据集来进行基于参数稀疏化的 FedAvg 算法的横向联邦学习过程。

训练轮数：20（模型上传和下载次数）

本地更新：3 次

用户数量：10（N）

随机选择用户数量：2（每次更新随机选 2 个用户）

批次大小：32

掩码矩阵：bernoulli 分布随机生成 0-1 矩阵

此外，我们做出了当 bernoulli 分布概率不同时，各自的收敛曲线图。

2.1.2 相关算法

本次实验中，主要使用了 Fedavg 算法，FedAvg 是一种联邦学习算法，允许多个用户同时训练一个机器学习模型。

算法步骤如下：（1）在每一轮迭代的步骤 t，服务端发送当前全局模型参数给客户端 ；（2）非抽样子集中的客户端根据 t，通过 SGD 更新本地模型；（3）抽样子集中每个客户端上传更新后的本地参数 t+1 ；（4）在迭代步骤 t+1,服务端根据全局模型参数计算加权平均值。

通过这种加权平均方法，服务端可以融合来自不同客户端的知识，更新全局模型。在这个过程中，每个客户端的权重通常由其数据量决定，即数据多的客户端对全局模型的影响更大。随后，更新后的全局模型参数再次被分发到各个客户端，用于下一轮的本地训练。这个循环持续进行，直到模型达到预定的性能标准或者经过了设定的迭代轮数。

FedAvg 提供了一种有效的方法来联合多个客户端共同训练机器学习模型而不需要中心化的数据存储，有助于缓解数据孤岛问题，并能在遵守数据隐私法规的前提下，推动机器学习技术的应用和发展。在处理大规模分布式数据时，FedAvg 在提高模型性能和数据利用率方面具有很大潜力。

算法伪代码如下：

Algorithm 1 Federated Averaging (FedAvg)

Server executes:
procedure SERVEREXECUTION
  for each round $t = 1, 2, \dots, T$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k \leftarrow ClientUpdate(k, w_t)$
    end for
    $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{n} w_{t+1}^k$
  end for
end procedure

ClientUpdate(k, w): ▷ Run on client k
procedure CLIENTUPDATE(k, w)
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
    end for
  end for
  return w
end procedure

算法流程如 3 所示：

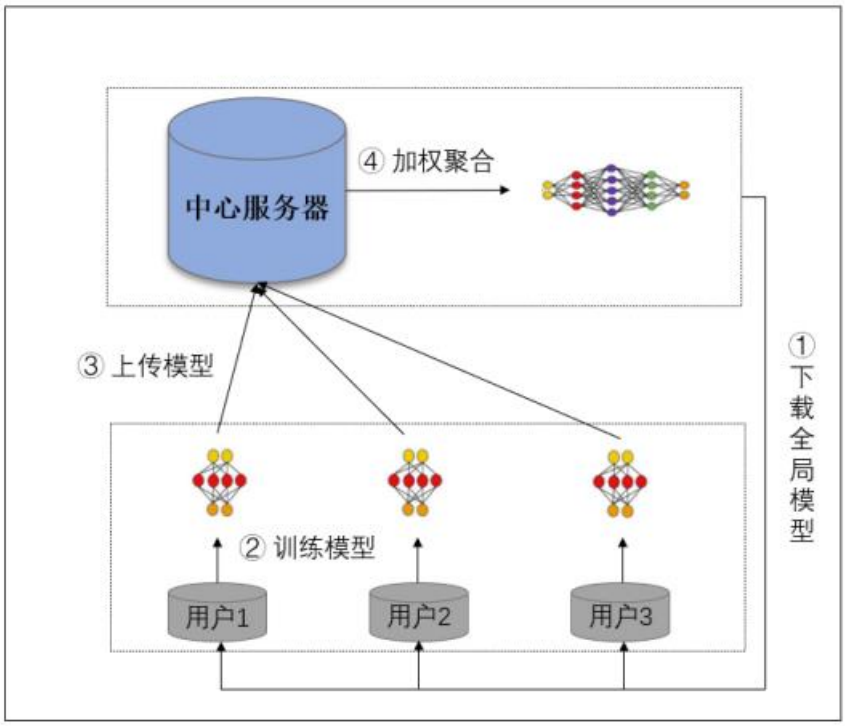

图 3 联邦学习流程图

2.2 实验结论

2.2.1 数据与性能分析

在 bernoulli 分布概率 p=0.6，0.8，0.9 时，模型收敛曲线如下图 4,5,6 所示。

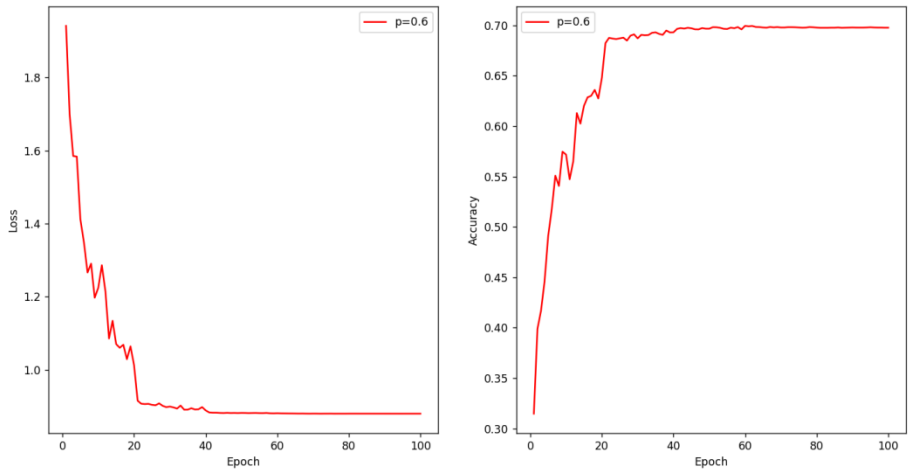

图 4 p=0.6 时收敛曲线

p=0.6 时，损失在前期迅速下降，并趋于稳定。准确率在前期迅速上升，最终达到约 0.7。

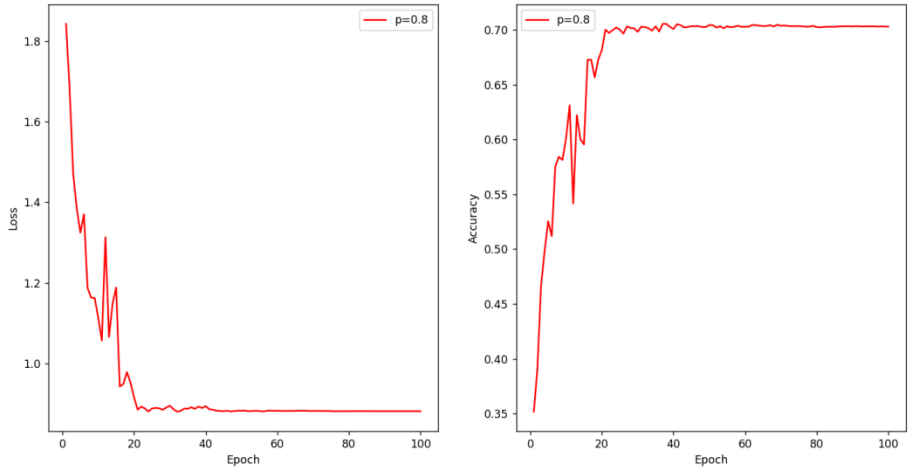

图 5 p=0.8 时收敛曲线

p=0.8 时，损失下降速度较快，并且波动较小。准确率在前期上升迅速，达到稳定状态后的准确率约为 0.7。

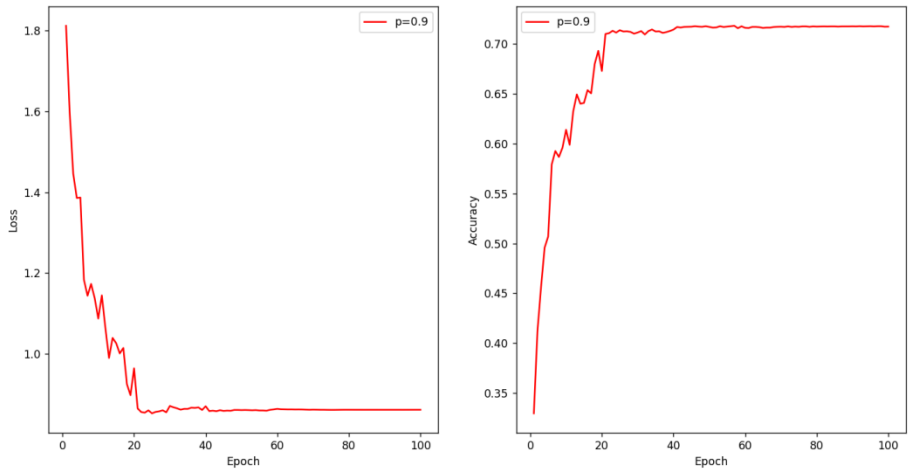

图 6 p=0.9 时收敛曲线

p=0.9 时，损失下降较平稳，且最终达到较低的损失值。准确率上升较快，并且波动较小，达到约 0.7 的准确率。

2.2.2 结论

从以上结果可以看出：随着 Bernoulli 分布概率的增加，模型的损失值趋于稳定更快，同时准确率也能达到较高水平。三种不同的 Bernoulli 分布概率的情况下，模型最终的准确率都趋近于 0.7，说明 Bernoulli 分布概率对最终准确率的影响相对较小，但对损失的收敛速度有一定影响。

在实际应用中，可以选择较高的 Bernoulli 分布概率以加快模型的收敛速度，并在训练初期取得较低的损失值和较高的准确率。在某些对模型稳定性要求较高的应用场景中，选择较高的 Bernoulli 分布概率有助于提高模型训练的稳定性和可靠性。

三、成员分工及个人课程目标达成情况

成员分工情况如下：

负责部分
项目整体协调，联邦学习框架搭建, 文档撰写
代码 debug，参数调整
收集资料，代码修改
代码修改，数据分析，文档撰写
初期规划，提出建议

参考文献：

[1] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li and H. Vincent Poor, "Federated Learning for Internet of Things: A Comprehensive Survey", IEEE Commun. Surv. Tutorials, vol. 23, no. 3, pp. 1622-1658, Jul. 2021.

[2] M. Abdel-Basset, H. Hawash, N. Moustafa, I. Razzak and M. Abd Elfattah, "Privacy-preserved learning from non-iid data in fog-assisted IoT: A federated learning approach", Digit. Commun. Networks, Dec. 2022.

[3] X. Gu, Z. Tianqing, J. Li, T. Zhang, W. Ren and K. K. R. Choo, "Privacy accuracy and model fairness trade-offs in federated learning", Comput. Secur., vol. 122, pp. 102907, Nov. 2022.