计算机科学
COMPUTER SCIENCE

一种基于CutMix的增强联邦学习框架

王春东, 杜英琦, 莫秀良, 付浩然

引用本文

王春东, 杜英琦, 莫秀良, 付浩然. 一种基于CutMix的增强联邦学习框架[J]. 计算机科学, 2023, 50(11A): 220800021-8.

WANG Chundong, DU Yingqi, MO Xiuliang, FU Haoran. Enhanced Federated Learning Frameworks Based on CutMix [J]. Computer Science, 2023, 50(11A): 220800021-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220800021

一种基于 CutMix 的增强联邦学习框架

王春东　杜英琦　莫秀良　付浩然
"计算机病毒防治技术"国家工程实验室　天津 300384
"学习型智能系统"教育部工程研究中心　天津 300384
天津理工大学计算机科学与工程学院　天津 300384

摘　要　联邦学习(Federated Learning)的出现解决了传统机器学习中存在的"数据孤岛"问题,能够在保护客户端本地数据隐私的前提下进行集体模型的训练。当客户端数据为独立同分布(Independently Identically Distribution,IID)数据时,联邦学习能够达到近似于集中式机器学习的精确度。然而在现实场景下,由于客户端设备、地理位置等差异,往往存在客户端数据含有噪声数据以及非独立同分布(Non-IID)的情况。因此,提出了一种基于 CutMix 的联邦学习框架,即剪切增强联邦学习(CutMix Enhanced Federated Learning,CEFL)。首先通过数据清洗算法过滤掉噪声数据,再通过基于 CutMix 的数据增强方式进行训练,可以有效提高联邦学习模型在真实场景下的学习精度。在 MNIST 和 CIFAR-10 标准数据集上进行了实验,相比传统的联邦学习算法,剪切增强联邦学习在 Non-IID 数据下对模型的准确率分别提升了 23%和 19%。

关键词:联邦学习;非独立同分布数据;数据清洗;数据增强;显著性检测
中图法分类号　TP183

Enhanced Federated Learning Frameworks Based on CutMix

WANG Chundong,DU Yingqi,MO Xiuliang and FU Haoran
National Engineering Laboratory for Computer Virus Prevention and Control Technology,Tianjin 300384,China
Engineering Research Center of Learning-Based Intelligent System,Ministry of Education,Tianjin 300384,China
School of Computer Science and Engineering,Tianjin University of Technology,Tianjin 300384,China

Abstract　The emergence of federated learning solves the problem of "data silos" in traditional machine learning. Federated learning enables the training of collective models while protecting the privacy of the client's local data. When the client's dataset is independently identically distributed(IID) data,federated learning can achieve an accuracy similar to that of centralized machine learning. However,in real scenarios,due to differences in client devices and geographic locations,there are often cases where client's dataset contain noisy data and non-independent identical distribution(Non-IID). Therefore,this paper proposes a CutMix-based federated learning framework,namely CutMix enhanced federated learning(CEFL),which first filters out noisy data through data cleaning algorithms and then trains through CutMix-based data enhancement. Compared with the traditional federated learning algorithm,the accuracy of CutMix enhanced federated learning can be improved by 23% and 19% for the model on Non-IID dataset.

Keywords　Federated learning,Non-independent identically distributed data,Data cleaning,Data augmentation,Saliency detection

1　引言

机器学习(Machine Learning,ML)在人工智能(Artificial Intelligence,AI)领域中迅猛发展,例如计算机视觉、自动语音识别、自然语言处理以及推荐系统等[1]。这些机器学习技术的成功,尤其是深度学习的再次兴起,无一不是建立在大量的数据(亦称大数据)基础之上的[2]。例如 Krizhevsky 等[3]提出了一个涉及手眼协调的机器人抓握学习模型,为了训练他们的网络,共收集了 800000 次抓握数据,使得机器人手臂成功地学习了更多种类的抓握策略。

一般而言,训练人工智能应用模型所需要的数据量都是非常庞大的,但高质量、大数量的训练数据通常很难获得,并且随着各国数据保护条例的颁布,各方数据共享成为了一大难题。为了应对"数据孤岛"的问题,McMahan 等[4]在 2016 年引入了联邦学习的概念。联邦学习是一个基于分布式数据集的机器学习框架,框架中客户端(如移动设备或企业组织)在中央服务器(如服务提供商)的协调下协作训练模型,同时保持私有数据的分散。联邦学习能有效帮助多个机构在满足用户隐私保护、数据安全和政府法规的要求下,进行数据使用和机器学习建模。

联邦学习在拥有巨大商业价值的同时,也面临着很多问题与挑战,其中包括服务器和本地设备之间传输参数所需的通信成本、本地设备所需的计算能力和能耗,以及本文中针对的数据异构性问题[5]。在传统机器学习中,数据分布在同一个机器上,并且假设数据是从同一个分布中独立地采样的,即数据独立同分布(Independently Identically Distribution,IID)。但在现实场景中,由于不同客户端可能属于不同的用户、企业,环境,因此其数据分布往往存在极大差异,即数据

通信作者:王春东(michael3769@163.com)

非独立同分布(Non-IID)。经典联邦学习算法 FedAvg 在 Non-IID 场景下性能会很差,正如 Yang 等[5]所述,由于每个客户都将从不同的数据分布中学习,因此模型更难得到有效的训练。在模型性能上看,Zhu 等[6]证明了当每个设备都缺少特定标签的样本时,MINST 和 CIFAR-10 数据集下的分类准确率与 IID 对应数据集相比,分别降低了 11%和 51%。

因此,为了在保护隐私的同时缓解客户端数据异构性带来的影响,本文提出了一种基于 CutMix 的联邦学习框架,即剪切增强联邦学习 CEFL。框架包括两部分:首先通过数据清洗算法筛选可用数据,之后通过基于 CutMix 的数据增强方式进行模型训练。本文的研究工作主要如下:

1)提出了一种基于 CutMix 的联邦学习框架(CEFL),在此框架中客户端间共享平均数据,并通过优化的 CutMix 技术与私有数据进行混合后计算梯度,通过控制参数值可以在交换信息量与数据隐私之间进行权衡,从而有效提高联邦学习模型在 Non-IID 数据下的性能。

2)提出了一种在训练前数据清洗的算法,通过服务器中的小规模基准数据进行训练,并与客户端共享基准模型,通过这种方式客户端可以筛选出本地的噪声数据,从而保证后续模型训练的准确度。

3)在联邦学习的标准基准数据集上验证了所提出的方法,并将其与标准的联邦学习方法进行了比较。在 Non-IID 环境下,CEFL 的模型精度比联邦平均算法提高了 19%~23%,其能够在保护用户隐私的同时保持模型精度。

2 相关工作

2.1 面向 Non-IID 数据的联邦学习

目前处理联邦学习中 Non-IID 问题的方法可以大致分为 3 种,包括基于算法的方法、基于系统的方法以及基于数据的方法。基于算法的方法中,Fallah 等[7]提出了一种个性化联邦平均算法(Per-FedAvg),作者利用元学习方法构建高质量的初始全局模型,使本地客户更容易在计算成本很低的情况下获得良好的性能;Arivazhagan 等[8]提出了一种用于深度前馈神经网络联邦训练的基本层加个性化层方法(FedPer),个性化层不仅可以提高联邦学习在 Non-IID 数据下的学习性能,还可以降低通信成本。基于系统的方法中,Ghosh 等[9]提出了一种通过比较不同聚类模型的损失值来进行模型相似度评估的方法,通过相似度评估将客户端分为不同的集群进行更新训练,使联邦学习中的全局模型不仅限于一个,而是由用户分簇来决定全局模型个数。本文提出的方法是基于数据的方法,Zhu 等[6]验证了数据共享方法的可用性,通过在服务器上存储全局共享数据集,与客户端共享全局数据,在 CIFAR-10 数据集上,仅使用 5%的全局共享数据,模型测试精度可提高约 30%。但是直接进行数据共享违背了联邦学习的初衷,因此本文提出了一种可以在数据共享的同时保护用户隐私的增强联邦学习框架。

2.2 数据清洗

在联邦学习中,客户端与服务器共享本地训练模型以生成全局模型,相较于本地机器学习,这样做有效保护了用户的数据隐私。但与本地机器学习一样,联邦学习中客户端具有的清晰标注的大规模数据集对于全局模型的精度至关重要,然而目前标注大规模数据集需要消耗大量的人力、物力,并且 Wang 等[10]也论证了即使是高质量的数据集也会包含噪声标签。因此借鉴本地化机器学习对噪声标签的检测技术,提出一种针对联邦学习分布式的数据集清洗算法,非常重要。

目前,集中式机器学习中的噪声数据处理技术已经相对成熟。解决数据噪声的方法包括两种:基于数据层的方法以及基于算法层的方法。在数据层面,现有的方法通常是对噪声数据进行去噪处理。Retu 等[11]使用训练数据的切片生成多个模型,通过这些模型为每个输入生成临时标签,之后通过投票方案来确定是否存在噪声标签。Xie 等[12]设计了拜占庭式鲁棒聚合器,以抵御卷积神经网络上的标签翻转数据中毒攻击。基于算法层面的方法主要是训练抗噪模型。Han 等[13]提出了一种新的深度学习模型 Co-teaching 来减缓噪声标签带来的影响,模型中同时训练两个深度神经网络,并让它们互相通信来交换各自计算得到的损失值较小的数据。

这些方法都有效抑制了噪声数据对训练结果的影响,但是现有的数据清洗算法大多针对的是集中式机器学习,这些方法有一个共同点,即需要在整个训练集上进行计算,但对联邦学习来说,这违背了隐私保护的原则。目前学界较少有针对联邦学习中分布式数据的数据清洗算法。Li 等[14]提出了一种名为"FLDebugger"的自适应框架,其通过量化训练样本对模型预测的影响识别出具有负面影响的错误数据,但这个方法并没有考虑模型在 Non-IID 数据下的鲁棒性。因此本文提出了一种针对联邦学习中分布式数据的数据清洗算法,以在训练开始前清洗数据,提高模型精度并缓解客户端中 Non-IID 数据的影响。

2.3 数据增强

深度学习模型的成功可以归功于数据的数量与多样性。但收集标记数据是一项繁琐而耗时的任务。数据增强旨在通过应用各种变换来增加现有数据的多样性,已被广泛用于训练深度学习模型,显著提高了模型性能和鲁棒性。早期的数据增强大多是手动对数据自身进行变换,例如增加高斯噪声、随机颜色变化、随机剪裁等。除去这些手动设计的数据增强外,Lemley 等[15]提出了一种端到端可学习的增强过程,称为智能增强,他们使用了两个不同的网络,其中一个用于学习合适的增强类型,另一个用于训练实际任务。Cubuk 等[16]提出了一种称为自动增强的有效数据增强方法,该方法定义了各种增强技术的搜索空间,并为每个小批次选择最合适的方法。

Mixup[17]是 ICLR2017 年提出的针对计算机视觉的一项简单的数据增强策略,通过实际数据实例之间的线性插值生成额外数据,可以增加模型的泛化能力,并且能够提高模型对于对抗攻击(Adversarial Attack)的鲁棒性。Mixup 通常被应用于图像分类任务,并被证明可以提高各种数据集(如 CIFAR10 和 ImageNet-2012)的测试精度。但插值增强会导致图像看起来不自然,因此 Yun 等[18]在 2019 年提出了 CutMix,它的初衷是解决 regional dropout 策略中的信息像素丢失的问题。在深度卷积神经网络(CNN)中,为了防止网络过多关注数据的部分小区域的问题,DeVries 等[19]提出了随机特征去除正则化技术,包括随机删除隐藏激活的 dropout 以及用于删除输入上随机区域的 regional dropout,通过随机删除部分输入区域可以提高模型的泛化能力与定位能力。但是这种方法同时也使得模型丢失了关于被删除区域的信息。

CutMix 不从数值角度对两个样本插值，而是从图像的空间角度考虑，把一张图片上的某个随机矩形区域剪裁到另一张图片上生成新图片。标签的处理和 Mixup 保持一致，都是按照新样本中两个原样本的比例确定新的混合标签的比例。这种新的处理可以解决 dropout 带来的问题，也更适合图像中信息连续这个特点。

3 CEFL：剪切增强联邦学习

3.1 系统定义

本文考虑一个具有多个客户端以及一个服务器的横向联邦学习系统。可以应用于联邦学习中各个参与方的数据集有相同的特征空间和不同的样本空间的情况，主要研究带标记训练数据的监督学习，每个客户端拥有各自的本地隐私数据集，其中包含各类数据。联邦学习任务由服务器下发，并且服务器拥有一个小型的基准数据集。基准数据集的特点是准确度高，可以近似认为不含有错误标签，但是数据量较小以至于无法单独训练整个任务模型，因此需要与客户端合作进行模型训练。本文不关注客户端的不可信的问题，假设客户端是诚实并自愿参模型优化任务。其中客户端中的噪声可能由以下原因造成：(1)数据并非针对本次训练任务；(2)数据的标签标记错误；(3)数据的标签表示错误。客户端数据的 Non-IID 问题可能由以下原因造成：(1)客户端所在地区不同；(2)采样设备差异；(3)数据量不均衡。

假设有 K 名参与方(也可称为数据拥有者或客户端)在一个横向联邦学习系统中。设 D_k 表示第 k 名参与方所拥有的数据集，$D_k=(X_k,Y_k)$，$k=1,\cdots,K$；D 表示所有客户端的总体数据，即 $D=\bigcup D_k$。每个客户端通过最小化本地数据集上的损失函数来训练其本地模型 M^k，本文使用卷积神经网络(CNN)结构作为基础来训练联邦学习分类模型。由于在所有数据中含有部分噪声数据，本文用 T_k 来表示第 k 名参与方数据集中的可用数据(即不含噪声标签的数据)，$T_k\subseteq D_k$；T 表示所有客户端中的可用数据，即 $T=\bigcup T_k$。这部分数据由 3.2 节所提出的数据清洗算法选出。

训练阶段使用 3.3 节提出的剪切增强算法进行，聚合阶段使用联邦平均算法，不同的是聚合时客户端模型的权重由本地可用数据量决定。在通信轮次 $t=0,\cdots,T-1$ 中，客户端 $k\in\{1,\cdots,N\}(N<K)$ 被选中参与本地训练并在完成后将模型 ω_t^k 发送至服务器。每轮本地更新结束后，服务器接收到本地模型后进行加权平均，在 t 轮通讯后，全局模型更新为 $\omega_t=\sum_{k=1}^{N}p_k\omega_t^k$，其中 $p_k=|T_K|/|T|$。更新的全局模型被发送回下一轮的客户端，通过 3.3 节提出的数据增强算法进行更新，其中本地训练次数 $i=0,1,\cdots,E-1$，batch 大小为 B，本地学习率为 η。l 是模型的损失函数，$f(x,\omega_t)$ 给定模型权重 ω_t 的输入 x 的模型输出。系统的总体框架流程如图 1 所示。

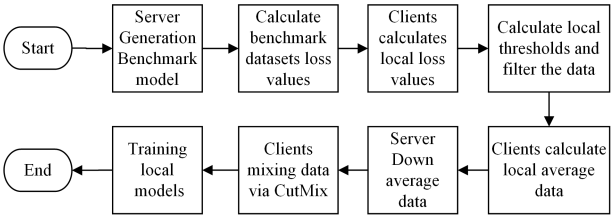

图 1 整体框架流程

Fig. 1 Overall framework process

3.2 数据清洗算法

数据清洗流程是在联邦学习任务发布后，本地模型训练前进行，目的是通过去除具有错误标签的数据，提高本地模型的准确度。在整个联邦学习过程中，数据清洗算法只需要进行一次，并且过程中只交换少量模型数据以及损失值，不会在客户端间交换原始数据。在保证联邦学习安全和通信效率的前提下，数据清洗提高了模型准确性，并且能在一定程度上减轻客户端数据异质性的影响。与文献[10]类似，本文将客户端数据噪声大致分为两类：一类是在数据集中混合了其他类型的数据，例如在物体识别任务中混合了多张手写数字的图片；另一类则是数据集中某些数据被错误标记，例如在手写数字识别数据中“1”的图片被标记为 2。算法流程如图 2 所示。

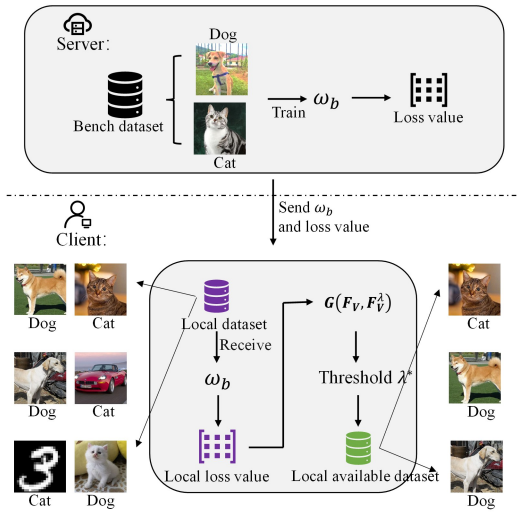

图 2 数据清洗流程图

Fig. 2 Data cleaning flow chart

算法的中心思想是：服务器通过使用本地的基准数据集训练得到一个基准模型，并计算测试集数据的损失值；随后将基准模型以及损失值发送给客户端。基准模型由于训练数据量少，因此精度会比较差，但是它对与本次任务相关的正确数据仍然具有更高的准确度。因此客户端在收到基准模型后，通过基准模型来计算每个数据样本的损失，并与基准测试数据集的损失值的分布进行比较。为了避免出现个别客户端由于数据筛选出现样本数量偏差的情况，算法在不同客户端间确定一组动态阈值，损失高于阈值的数据样本被视为噪声并被排除在训练之外，从而过滤掉噪声数据。算法细节如算法 1 所示。

算法 1 数据清洗算法

输入：$D_k=\{X_k,Y_k\}$ for $k=1,\cdots,N$

输出：$T_k=\{X_k,Y_k\}$ for $k=1,\cdots,N$

1. Server：
2. Initialize ω_B for benchmark model
3. $\mathbb{S}_t\leftarrow$K clients selected at random
4. Split the bench dataset B into B_{test}, B_{train}
5. for local epoch from 1 to E do
6. for batch b$\in$(1,B) do
7. $\omega_B\leftarrow\omega_B-\eta\nabla l(f(x_i,\omega_B),y_i)$
8. end for
9. end for
10. $V=\{l(f(x_i,\omega_B),y_i):\forall(x_i,y_i)\in B_{test}\}$/ * 计算基准模型损失值 * /

11. Send V, ω_B to clients $k\in\mathbb{S}_t$

12. Client k:

13. Receive V, ω_B from Server

14. $C_k=\{l(f(x_i,\omega_B),y_i):\forall(x_i,y_i)\in D_k\}$/ * 计算本地数据损失 * /

15. $F_v(x)\leftarrow Pr\{X\leqslant x:X\in V\}$

16. $F_C(x)\leftarrow Pr\{X\leqslant x:X\in C_k\}$

17. $F_C^{\lambda}(x)\leftarrow Pr\{X\leqslant x|X\leqslant\lambda:X\in C_k\}$

18. $G(F_V,F_V^{\lambda})=\sup_x|F_V-F_c^{\lambda}|$

19. $\lambda^*=\arg\min_{\lambda}[G(F_V,F_c^{\lambda})-F_C(\lambda)]$/ * 计算阈值 * /

20. $T_k=(x_i,y_i)\in D_n:l(f(x_i,\omega_B),y_I)\leqslant\lambda^*$

步骤 1—步骤 11 为服务器端进行的操作，首先初始化基准模型，然后划分基准数据集以及选择参与训练的客户端，接下来通过划分的基准训练集训练基准模型，并计算测试集的损失值，最后将基准模型与损失值发送给客户端。步骤 12—步骤 20 为客户端进行的操作，在收到服务器发送的基准模型后，计算本地数据在基准模型上的损失值，得到本地损失值以及基准数据损失值的累积分布函数$F_c(x)$，$F_v(x)$。当算法中将阈值设为变量λ后，就可以得到在λ限制下的累积分布函数F_c^{λ}。

现在回顾数据清洗算法的目标：使客户端数据与基准数据的分布相似性最大化。当使用在基准模型上的损失值来近似数据分布后，任务就简化为了：使客户端数据的模型损失值与基准数据的模型损失值分布相似性最大化。本文使用 KS 距离来计算两个分布的相似性：

$$G(F_V,F_V^{\lambda})=\sup_x|F_V-F_c^{\lambda}| \tag{1}$$

这时只需要确定当两个分布距离最小时阈值λ的值即可。但在实际情况中，若存在某些客户端的地理位置、设备等外界因素条件与服务器较为不同的情况，并且数据量较少时，为所有客户端设置统一阈值可能会导致客户端数据出现数据量倾斜的情况，这会导致模型泛化性能下降。因此，在最小化分布距离时加入$F_C(\lambda)$项，用来平衡数据量的流失。

$$\lambda_i^*=\arg\min_{\lambda}[G(F_V,F_c^{\lambda})-F_C(\lambda)] \tag{2}$$

最后，每个客户端通过阈值λ_i^*过滤数据就可以去除噪声数据，从而提高系统模型的准确度。

3.3 剪切增强算法

直观地说，在 Non-IID 数据存在的情况下，联邦学习模型的性能下降是由异构数据分布引起的，基于数据的方法旨在通过修改分布来解决这个问题。数据共享和数据扩充是两种较为主流的方法。数据共享最为简单有效的方法就是从其他客户端或者服务器的基准数据集接收原始个人数据，但是这违背了联邦学习的隐私设置。本文所提出的剪切增强算法，为了在保护客户隐私的前提下进行数据共享，为用户提供了一种可调节的细粒度隐私，只需要在客户端间共享混合数据，就可以达到共享原始数据的效果。

剪切增强算法主要分为两步，首先是客户端之间进行数据共享，接着将共享数据与本地数据通过 CutMix 方法进行混合增强。在数据共享阶段，拥有数据的客户端将自身的平均数据发往服务器，其中平均值的计算由参数N_k决定，它同时也控制了剪切增强算法的隐私与通信成本。当$N_k=1$时，客户端间将直接共享原始数据，这违反了联邦学习对隐私保护的要求，也会提高通信成本。服务器在接收到各数据客户端所发送的平均数据后，将其聚合为一组数据发送给参与模型更新的客户端。参与模型更新的客户端在接收到平均值数据后，就开始进行本地模型的训练。在训练中使用其他客户端的平均数据的最简单直接的方法就是像本地数据一样直接使用，但通过后续实验可以看出，混合过程会大大降低平均数据的可用性，从而使模型的泛化性无法得到提高。因此考虑使用 CutMix 方法进行数据的混合增强，CutMix 通过使用一张图像的部分连续信息去代替另一张图像的连续信息，可以在图像数据上获得比 Mixup 更好的混合增强效果。

在经典 CutMix 算法中，假设$x\in R^{W\times H\times C}$与$y$为训练样本及其标签，算法的目标是通过组合两个训练样本(x_A,y_A)和(x_B,y_B)来生成一个新的训练样本$(\tilde{x},\tilde{y})$，然后使用生成后的训练样本训练具有原始损失函数的模型。

$$\tilde{x}=M\odot x_A+(1-M)\odot x_B \tag{3}$$

$$\tilde{y}=\lambda y_A+(1-\lambda)y_B \tag{4}$$

其中，$M\in\{0,1\}^{W\times H}$为一个 0-1 矩阵，通过矩阵乘法可以提取和保留原始样本的部分区域。与 Mixup 类似，CutMix 算法中两个样本的组合比λ在(0,1)均匀分布中采样，主要区别在于 CutMix 使用另一个训练图像的面片替换图像区域，并生成比 Mixup 更局部自然的图像。

通过 CutMix 来混合平均数据以及本地数据，可以有效提高模型的泛化性能。但通过观察实验时的样本数据发现 CutMix 存在一定的缺陷，当剪裁部分为两张图的背景部分时，图片本身的真实标签并没有发生改变，但算法仍然会使用 Mixup 的方法计算标签，这就将提供错误的训练样本(见图 2)。因此在本框架中加入了显著性检测算法，首先检测出图像中信息密度较高的区域，再进行 CutMix，从而改善实验结果。本文中使用 Qin 等[20]提出的显著性检测算法，具体流程如图 3 所示。

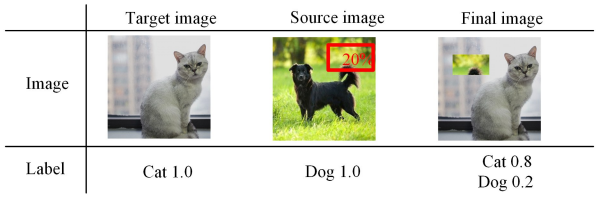

图 3 CutMix 中的错误混合

Fig. 3 Wrong mix in CutMix

客户端每一轮训练时都随机抽取一个本地样本和平均样本进行 CutMix 混合，之后执行随机梯度下降算法。算法 2 为剪切增强算法的具体步骤，算法 3 为剪切增强算法中本地客户端的更新算法。算法 2 将客户端的本地平均数据收集并发送给参与训练的客户端，客户端使用算法 3 中的混合增强算法进行模型训练。

算法 2 剪切增强算法

输入：$\mathbb{D}_k=\{X_k,Y_k\}$ for k=1,…,N

N_k: number of data instances used for computing average $\bar{x},\bar{y}$

1. Initialize ω_0 for global server

2. for t=0,…,T−1 do

3. for client k with updated local data do

4. Split local data into N_k sized batches
5. Compute $\bar{x}, \bar{y}$ for each batch
6. Send all $\bar{x}, \bar{y}$ to server
7. end for
8. $\mathbb{S}_t \leftarrow$K clients selected at random
9. Send ω_t to client $k \in \mathbb{S}_t$
10. if updated then
11. Aggregate all $\bar{x}, \bar{y}$ to X_m, Y_m
12. Send X_m, Y_m to clients $k \in \mathbb{S}_t$
13. end if
14. for $k \in \mathbb{S}_t$ do
15. $\omega_{t+1}^k \leftarrow \text{LocalUpdate}(k, \omega_t; X_m, Y_m)$
16. end for
17. $\omega_{t+1}^k \leftarrow \frac{1}{K} \sum_{k \in \mathbb{S}_t} p_k \omega_{t+1}^k$

end for

算法 3 本地更新算法

LocalUpdate($k, \omega_t; X_m, Y_m$) under Algorithm 2

$\omega \leftarrow \omega_t$

1. for e=0,…,E−1 do
2. Split $\mathbb{D}_k$ into batches of size B
3. for batch b(X,Y) do
4. Select an entry x_m, y_m from X_m, Y_m
5. $x_{vm} = \text{SaliencyDetect}(x_m)$
6. $r_x, r_y = \text{argmax}(x_{vm})$
7. $\lambda = \text{Unif}(0,1)$
8. $r_w = \text{sqrt}(1-\lambda)$
9. $r_h = \text{sqrt}(1-\lambda)$
10. $x_1 = \text{Round}(\text{Clip}(r_x - r_w/2, \text{min}=0))$
11. $x_2 = \text{Round}(\text{Clip}(r_x + r_w/2, \text{min}=W))$
12. $y_1 = \text{Round}(\text{Clip}(r_y - r_w/2, \text{min}=0))$
13. $y_2 = \text{Round}(\text{Clip}(r_y + r_w/2, \text{min}=H))$
14. for each x_i, y_i in X,Y do
15. $x_i[:,:,x_1:x_2, y_1:y_2] = x_m[:,:,x_1:x_2, y_1:y_2]$
16. $\lambda = 1-(x_2-x_1)\times(y_2-y_1)/(W\times H)$
17. $y_i = \lambda \times y_i + (1-\lambda) \times y_m$
18. end for
19. $\omega \leftarrow \omega - \eta \nabla l(\omega, b)$
20. end for
21. end for
22. return ω

4 实验验证

4.1 实验设置

本文中实验模拟了经典的联邦学习设置，使用大量模拟客户端在联邦学习系统中进行实验（节点数 $N=100$），其中每个客户端均拥有自己的本地数据。系统拥有一个中心服务器，中心服务器上含有少量基准数据集。客户端在本地训练模型并上传到服务器，服务器将模型参数聚合并发送给下一轮参与训练的客户端。

本文所使用的数据集包括：MNIST 数据集以及 CIFAR-10 数据集。实验中以 3 种方式划分数据并分配给客户端。(1)IID 数据划分：将每个标签的数据平均划分到各个客户端，这样每个客户端的数据的分布相同。(2)Non-IID 数据划分：为了模拟真实场景，实验中将数据以 Non-IID 的方式划分到各客户端。在 MNIST 数据集中，本文使用 Hsieh 等[21]提出的划分方式，将数据按标签大小排列后，每 300 个样本划分为一组数据，每个客户端各拥有两组数据，这样大多数客户端只能获得两类数字。CIFAR-10 数据集与 MINIST 数据集的划分方式大致相同，每个客户端也只拥有两类图片。(3)含有噪声的数据划分：本文在 ID 数据的划分基础上，为每个客户端增加了两类噪声，一类是其他不相关数据集的数据（例如在 MNIST 数据集中加入部分 CIFAR-10 数据集的数据），另一类是改变原始本地数据集的部分标签值，通过含有噪声的数据集可以验证数据过滤算法的性能。对于服务器端的小型基准数据集，直接从原始数据集中随机采样获得，包含训练任务的所有类图像。

本文使用卷积神经网络结构作为基础来训练联邦学习分类模型。CNN 结构为 9 层：5×5×32 卷积层→2×2 池化层→LRN→5×5×32 卷积层→LRN→2×2 池化层→z×256 全连接层→256×10 全连接层→Softmax 层。其中 z 取决于输入图像大小，对于 MNIST，$z=1568$，对于 CIFAR-10，$z=2048$。

4.2 数据清洗算法实验

在数据清洗实验中，为了评估算法性能，分别在 MNIST 数据集和 CIFAR-10 基准数据集上与 4 种算法进行比较：(1)系统的基准模型（Benchmark Model）；(2)联邦平均算法（FedAvg）；(3)FLDebugger 算法；(4)结合了数据清洗的联邦平均算法（FedAvg+Data cleaning）。

4.2.1 不同数据划分下的对比实验

针对 MNIST 数据集，使用 3 种数据划分进行对比实验：(1)IID 数据划分；(2)Non-IID 数据划分；(3)含有噪声的数据划分。其中前两类直接按前文所述进行划分即可；对于含有噪声的数据划分，首先将从 SVHN 数据集中采样的数据混合到目标数据集，混合的噪声数量为原始数据集的 20%，同时将 20%的客户端数据中的真实数据标签＋2 来错误标记标签。

对于 MNIST 数据集，图 4 给出了在使用上述 3 种数据划分方式时，4 个模型在测试集上的精确度。其中基准模型的精确度为常数，因为它是在联邦学习任务开始前就已经完成训练。在 IID 数据上，由于基准模型只拥有少量数据，导致精确度较低，其他 3 种模型的精确度相差不大，这是由于在理想的数据划分情况下，各个客户端拥有几乎与基准模型相同分布的数据，从而使得模型准确度很高（联邦平均算法达到了 96.7%，FLDebugger 算法达到了 98.1%，FedAvg＋Data cleaning 算法达到了 94.8%）。在 Non-IID 数据上，可以看出联邦平均算法与 FLDebugger 算法的模型准度率都降低了 10%左右，但加入了数据清洗算法的联邦平均算法一定程度上减小了 Non-IID 数据对模型准确度的影响，这是由于在数据筛选后客户端的数据分布更趋于相似。在添加噪声的数据集中，可以看出联邦平均算法的模型准确度发生了骤降，同时很难收敛，而数据清洗算法则只有少量的降低，并且明显优于基准模型和联邦平均算法，对比 FLDebugger 算法也有少量的提升。这表明了本文方法对含噪数据集有较好的鲁棒性。

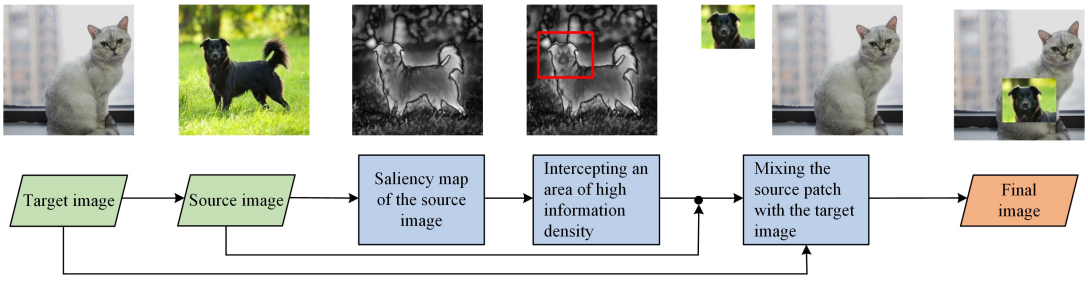

图 4 具有显著性检测的 CutMix

Fig. 4 CutMix with significance detection

对于 CIFAR10 数据集，采用了相同的方法进行了实验（在噪声混合时，使用了 FEMNIST 数据集的部分数据作为噪声）。可以看出，在 IID 环境下图像与 MNIST 数据集类似，但是在 Non-IID 环境中，由于 CIFAR-10 数据集类别较手写数字多，且图像噪声也大，因此比 MNIST 数据集上的性能下降更多（联邦平均算法的模型精确度下降了几乎 40%），CEFL 也有了较大的性能下降，但是较之其他两种方法还是有近 23% 的提升。在噪声数据下与 MNIST 数据集类似。

4.2.2 不同比例基准数据集影响实验

在上一节实验中，将基准数据集的数据量定为整体数据的 2%。为了得出不同比例基准数据集对联邦模型的影响，对比了在不同数量的基准数据集下的模型准确度。使用了 CIFAR-10 上的含噪声划分数据集，迭代了 100 轮，结果如表 1 所列。结果表明，随着数据量的增加，基准模型的准确度逐渐提高（数据量由 1%增加到 4%，准确度增加了大约 20%），但几乎不影响本文提出的数据清洗算法的准确度（数据量由 1%增加到 4%，准确度只增加了 2%），这意味着数据清晰算法只需要少量的基准数据集就可以对噪声数据拥有良好的鲁棒性。

表 1 不同比例基准数据集的模型准确度

Table 1 Model accuracy on benchmark datasets with different scales

	1%	2%	3%	4%
Benchmark Model	51.2	57.1	65.4	71.2
FedAvg+Data cleaning	80.5	81.1	81.4	82.3

4.3 剪切增强算法实验

4.3.1 在 Non-IID 数据集下的对比实验

为了验证剪切增强算法在 Non-IID 环境下的性能，本节实验都采用 Non-IID 的数据划分。将 FedAvg 和 FedProx 作为对比模型，并且为了证明通过使用 CutMix 混合两种数据能够有效提高模型精准度，在对比实验中加入了直接使用平均数据的算法（下文称之为 OriginalMix）。分别在 MNIST 和 CIFAR-10 中进行实验，结果如图 5 所示。

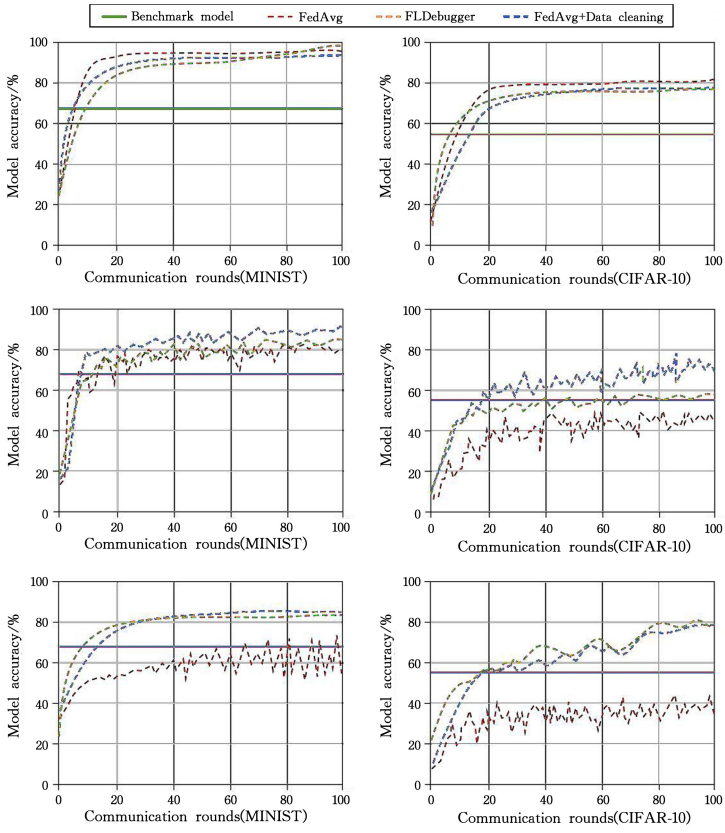

图 5 不同数据划分下的对比实验

Fig. 5 Comparison experiments with different data divisions

在相同的联邦学习设置下，比较了这 4 种方法随着通信轮次增加的模型准确度曲线，如图 6 所示。可以看出，直接使用平均数据的 OriginalMix 性能略高于 FedAvg 算法以及 FedProx 算法，CEFL 的性能又优于这 3 种方法，并且在两个测试数据集上达到了更快的收敛速度。这说明通过客户端之间交换平均数据可以有效缓解 Non-IID 数据带来的影响，同时使用优化后的 CutMix 算法将平均数据与本地数据混合，能进一步提高模型的鲁棒性。

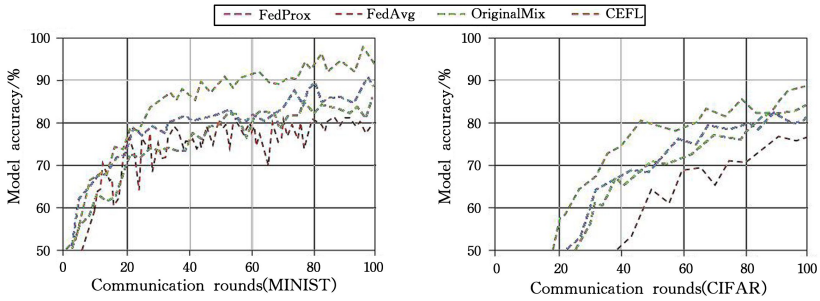

图 6　在 Non-IID 数据集下的对比实验

Fig. 6　Comparison experiments on Non-IID dataset

4.3.2　CutMix 中不同混合比对模型精度的影响

在经典 CutMix 算法中，混合比 λ 在(0，1)区间内均匀取样，但是可以通过固定混合比的方式优化模型性能。表 2 列出了在不同混合比下，CEFL 在 MNIST 与 CIFAR-10 数据集上分别迭代 100 轮和 500 轮的模型性能比较(数据划分为 Non-IID 划分)。在混合比增加时，本地数据获得了更多外部数据信息，模型的准确度开始上升，当继续增加混合比时，数据可用性大大降低，从而导致了模型精度的下降，因此应通过超参数 λ 的选择来平衡这两者的关系。

表 2　不同混合比对模型精度的影响

Table 2　Effect of different mixing ratios on model accuracy

(单位：%)

λ	0.05%	0.1%	0.2%	0.5%
MNIST	90.1	93.5	95.6	71.2
CIFAR-10	78.2	80.8	81.5	69.3

4.3.3　均值计算时 N_k 对模型精度的影响

在混合增强算法中，选择使用每个客户端的所有本地数据计算 X_m 和 Y_m。通过改变每个客户端在计算均值时所使用的 N_k 值，来观察 N_k 的潜在影响，结果如表 3 所列。直观来看，N_k 值的减小将导致需要计算更多条平均数据，从而增加额外的计算负担并降低隐私保护。而从实验结果看来，对于 MNIST 和 CIFAR-10 数据集，随着 N_k 值增加，性能只有小幅下降，如表 3 所列。这说明本文算法可以在保证隐私的同时降低通信量，并且不会损失过多的模型精度。

表 3　N_k 对模型精度的影响

Table 3　Effect of N_k on model accuracy

(单位：%)

N_k	5	10	20	50
MNIST	92.4	94.6	94.2	93.5
CIFAR-10	78.2	80.8	81.5	69.3

4.3.4　显著性检测对模型精度的影响

为了验证通过加入显著性检测，能够有效提高 CutMix 的性能，设置了一组消融实验。在 MNIST 与 CIFAR-10 数据集上，对比了加入显著性检查与没有显著性检测时 CEFL 的模型精度，分别迭代了 100 轮与 500 轮(数据划分为 Non-IID 划分)，结果如表 4 所列。通过实验可以看出，在 MNIST 数据集以及 CIFAR-10 数据集上，显著性检测的加入，一定程度上解决了 CutMix 中错误混合的情况，提高了大约 3%～4% 的性能。

表 4　显著性检测对模型精度的影响

Table 4　Effect of Saliency detection on model accuracy

(单位：%)

	MNIST	CIFAR-10
－Saliency Detection	91.5	80.3
＋Saliency Detection	94.3	84.5

结束语　本文针对联邦学习中客户端数据非独立同分布且存在噪声的问题，提出了一种基于 CutMix 的增强联邦学习框架。首先通过数据清洗降低客户端数据噪声，之后在客户端间共享平均数据，并通过优化的 CutMix 技术与私有数据进行混合后计算梯度，从而有效提高联邦学习模型在 Non-IID 数据下的性能。最后通过实验证明了 CEFL 框架在噪声及 Non-IID 环境下具有良好的鲁棒性。

但是提出的方法仍有不足，因为尽管客户端间通过平均值算法可以在一定程度上保证客户的隐私，但它可能会导致新类型的隐私问题。除了可以通过调整 N_k 来增加隐私外，后续的工作将会考虑加入差分隐私的方式来保证数据的安全，通过将噪声级别与 N_k 值结合，在 N_k 值较小时提供较高的噪声级别，在 N_k 值较大时提供较小的噪声级别，从而平衡数据可用性和安全性。

参 考 文 献

[1]　RAGHU M，SCHMIDT E. A Survey of Deep Learning for Scientific Discovery[J]. arXiv：2003.11755，2020.

[2]　POUYANFAR S，SADI S，YAN Q Y，et al. A Survey on Deep Learning：Algorithms，Techniques，and Applications[J]. ACM Computing Surveys，2019，51(5)：1-36.

[3]　LEVINE S，PASTOR P，KRIZHEVSKY A，et al. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection[J]. arXiv：1603.02199，2016.

[4]　MCMAHAN B，MOORE E，RAMAGE D，et al. Communication-Efficient Learning of Deep Networks from Decentralized Data[C]// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017：1273-128.

[5]　YANG Q，LIU Y，CHEN T. Federated Machine Learning：Concept and Applications[J]. ACM Transactions on Intelligent Sys-

tems and Technology, 2019, 10(2): 1-19.
[6] ZHU H, XU J, LIU S, et al. Federated learning on non-IID data: A survey[J]. Neurocomputing, 2021, 465: 371-390.
[7] FALLAH A, MOKHTARI A A, OZDAGLAR A A, et al. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada, 2020: 3557-3568.
[8] ARIVAZHAGAN M G, AGGARWAL V, SINGH A K, et al. Federated Learning with Personalization Layers[J]. arXiv: 1912.00818, 2019.
[9] GHOSH A, CHUNG A A, YIN J A, etal. An Efficient Framework for Clustered Federated Learning[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada, 2020: 19586-19597.
[10] WANG Y S, LIU Y A, MA W A, et al. Iterative Learning with Open-set Noisy Labels[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8688-8696.
[11] RETU, STAVROU G F A, LOCASTO A A, et al. Casting out Demons: Sanitizing Training Data for Anomaly Sensors[C]// 2008 IEEE Symposium on Security and Privacy (sp 2008). 2008: 81-95.
[12] XIE C, KOYEJO O, GUPTA I. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance[C]// 36th International Conference on Machine Learning (ICML 2019). 2019: 11928-11944.
[13] HAN B, YAO B A, YU Q A, et al. Co-Teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). 2018.
[14] LI A R, ZHANG A A, WANG L A, et al. Privacy-Preserving Efficient Federated-Learning Model Debugging[J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(10): 2291-2303.
[15] LEMLEY, BAZRAFKAN J A, SHABAB. Smart Augmentation Learning an Optimal Data Augmentation Strategy[J]. IEEE Access, 2017, 5: 5858-586.
[16] CUBUK E D, ZOPH B, MANÉ D, et al. AutoAugment: Learning Augmentation Strategies From Data[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 113-123.
[17] ZHANG H, CISSE M, DAUPHIN Y N, et al. mixup: Beyond Empirical Risk Minimization[J]. arXiv: 1710.09412, 2018.
[18] YUN S, HAN D, OH S J, et al. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 6022-6031.
[19] DEVRIES T, TAYLOR G W. Improved Regularization of Convolutional Neural Networks with Cutout[J]. arXiv: 1708.04552, 2017.
[20] QIN X, ZHANG Z, HUANG C, et al. BASNet: Boundary-Aware Salient Object Detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 7471-7481.
[21] HSIEH K, PHANISHAYEE A, MUTLU O. The Non-IID Data Quagmire of Decentralized Machine Learning[C]// Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020: 4387-4398.

WANG Chundong, born in 1969, Ph.D, professor, is a senior member of China Computer Federation. His main research interests include cyberspace security, blockchain technology, etc.