doi:10.11887/j.cn.202303014 http://journal.nudt.edu.cn

模型不可知的联合相互学习*

周 伟,李艺颖,陈曙晖,丁 博
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:主流的联邦学习(federated learning, FL)方法需要梯度的交互和数据同分布的理想假定,这就带来了额外的通信开销、隐私泄露和数据低效性的问题。因此,提出了一种新的FL框架,称为模型不可知的联合相互学习(model agnostic federated mutual learning, MAFML)。MAFML仅利用少量低维的信息(例如,图像分类任务中神经网络输出的软标签)共享实现跨机构间的"互学互教",且MAFML不需要共享一个全局模型,机构用户可以自定制私有模型。同时,MAFML使用简洁的梯度冲突避免方法使每个参与者在不降低自身域数据性能的前提下,能够很好地泛化到其他域的数据。在多个跨域数据集上的实验表明,MAFML可以为面临"竞争与合作"困境的联盟企业提供一种有前景的解决方法。

关键词:联邦学习;域偏移;模型不可知
中图分类号:TP181 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2023)03-118-09

Model agnostic federated mutual learning

ZHOU Wei, *LI Yiying*, *CHEN Shuhui*, *DING Bo*
(College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: The mainstream FL (federated learning) methods require gradient interaction and the ideal assumption of the independently identically distribution, which brings additional communication overhead, privacy leakage, and data inefficiency. Therefore, a new FL framework called MAFML (model agnostic federated mutual learning) was proposed. MAFML only used a small amount of low-dimensional information (for example, the soft labels output of the neural network in the image classification task) for sharing to achieve cross-participants "mutual learning and mutual education". Moreover, MAFML did not need a shared global model, users can customize their own private models without restricting the model structure and parameters. At the same time, MAFML used a general approach for avoiding gradient interference so that each participant's model could be well generalized to other domains without reducing the performance of its own domain data. Experiments on multiple cross-domain datasets show that MAFML can provide a promising solution for alliance business facing the "competition and cooperation" dilemma.

Keywords: federated learning; domain shift; model agnostic

传统的机器学习和深度学习任务,大部分是围绕单个任务的学习(或者可以称为单域学习),如图1(a)所示,其处理的数据样本来自单个域(即独立同分布的数据样本),例如人脸识别、目标检测,或者图像生成等任务。当系统切换到一个新的任务或域时,系统需要更换新的网络模型参数或重新初始化参数。深度学习在模式识别等领域超越了人类的性能,但是以数据驱动为基础的系统模型十分脆弱,泛化能力存在弊端。例如,医院通常仅基于自身数据构建深度学习模型,由于隐私、竞争或管理等因素而无法访问其他医院的同类数据。可以想象当患者可以获得来自其他医院数据信息(例如,心电图或者脑电图样本)的协助诊断是多么有吸引力的事情。

协作是如今数据量爆炸、任务复杂度激增后一个多方渴望的解决理念,知识的共享将有助于提高所有机构的绩效。但是,多机构的协助和共享并不是一件容易的事情,共享架构所带来额外的资源开销以及多机构私有数据本质上的偏移性问题是不可忽视的挑战[1]:

1)机器学习应用存在一个普遍性问题:在运行机器学习应用程序之前,系统将所有数据集中到广域网上的一个数据中心中[2-3],但广域网带宽是一种稀缺资源,因此移动所有数据可

* **收稿日期**:2021-06-30
基金项目:科技创新2030-"新一代人工智能"重大资助项目(2020AAA0104803)
作者简介:周伟(1990—),男,山西忻州人,工程师,博士,E-mail:zhouwei14@nudt.edu.cn;
丁博(通信作者),男,湖南攸县人,研究员,博士,硕士生导师,E-mail:dingbo@nudt.edu.cn

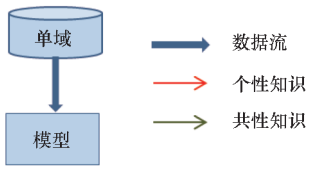
能非常缓慢[4]。此外,图像和视频的快速增长最终会使广域网带宽饱和[5];一些国家的隐私和数据主权法禁止跨越国界或大陆边界传输原始数据[6]。

2)多机构场景中的域偏移或数据集偏差是一个极具挑战性的问题,例如照片艺术卡通素描数据(photo art cartoon sketch database, PACS)。此外,多域表示也可以采取类偏移的形式,简单地以美国国家标准与技术研究所数据库[8](mixed national institute of standards and technology database, MNIST)为例,不同域之间包含不相交的手写数字类别(例如,A 域包含数字[1,2,3],B 域包含数字[5,7,8])。

联邦学习(federated learning, FL)作为一个多机构联合机器学习的应用范式,它充分融合深度学习技术的优势[9-10],在保持数据私有性的同时实现多机构(即非独立同分布的多个域)参与者对全局模型的协同训练,如图 1(b)所示。联邦学习和深度学习技术结合的方式可以充分利用多机构的数据资源和计算力资源,提升深度学习系统的泛化能力,进而提高系统的整体性能,例如图 2中如何判定狗的一种新的形式,如何学习一种新的运动。

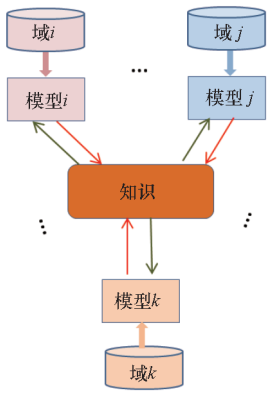
尽管传统 FL 具有良好的监管和经济效益,但仍存在一些缺陷:主流的 FL 利用梯度共享或者参数共享模式(例如,FedAvg [9]),该模式基于一个可信的集中式服务器来聚合梯度或参数,约束所有节点具有相同的网络结构,但会限制参与机构模型个性的诉求。同时,梯度或参数共享仍然面临严重的隐私泄露风险[11-12],尽管有一些技术已弱化这种风险(例如,差分隐私[13]和秘密共享[14]);而且,网络模型梯度的参数空间仍然是巨大的(普通卷积神经网络规模为 10 ~ 100 000 000,甚至更大)。

针对上述问题,本文在 FL 框架下提出了一种新的解决方案模型不可知的联合相互学习(model agnostic federated mutual learning, MAFML)。MAFML 借鉴相互学习的思想,仅利用少量低维的“意识”(例如,图像分类任务中神经网络输出的软标签)共享实现跨机构间的“互学互教”。为保留机构模型的个性,每个参与的机构节点都可以拥有自定制的私有模型。在跨域的场景设定下,MAFML 借鉴相互学习的思想[15],使每个机构节点避免或减轻对自身域数据的遗忘,同时允许有益的知识迁移,从而更好地泛化到其他域。



(a) 单域模型

(a) Model of single domain



(b) 多域 - 联邦学习

(b) Multi domains-federated learning

图 1　从单域到多域

Fig. 1　From single domain to multi domains

图 2　域偏移问题

Fig. 2　Issue of domain shift

1　相关工作

1.1　联邦学习

联邦学习的初衷是直接在本地设备上训练统计模型。文献[8]将 FL 的概念扩展到不同数据组织之间的协作。但无论是在垂直 FL[16]还是迁移 FL[17]中,执行特定任务的过程中都要求所有节点一起工作,而且每个节点仅负责任务的部分

要素。受到联邦学习和元学习之间联系的启发，文献[18]将 MAML 结合到联邦框架 Per-FedAvg 来学习初始共享模型，从而实现每个用户快速适应和个性化。文献[19]使用 FedAvg 进行集中模型训练，然后使用 Reptile 继续进行用户的个性化初始模型训练。相关的 FL 研究是文献[20]，它本质上是通过多任务学习方式，且不需要全局模型。但是，每个模型都只关注自己任务的性能，而不考虑将模型泛化到其他任务上。尽管一些 FL 工作的相关研究关注到了异构模型，但文献[21]仍然需要构建类似于 FedAvg 的全局模型结构，之所以称为异构的原因在于它利用了异构的子结构从分布式同构模型上蒸馏知识。FedMD[22]是研究异构 FL 的另外一项工作，但它更加侧重于通过模型蒸馏来获取通信模块。文献[19]研究模型不可知的 FL，但需要优化集中式全局模型并微调局部模型。本文在每个节点上保留了本地定制的同构或异构模型，而无须集中式模型或其他额外模型。

1.2 跨域学习

多域学习[23]旨在最终为多个域创建一个单一模型，尤其是当数据量相对较小且域相似时。如果训练域和测试域已经明确定义好，域自适应(domain adaptation, DA)[24]和域泛化(domain generalization, DG)[25]则是多域学习两个重要的子研究领域。DA 和 DG 的共同点在于它们都关注于到目标域的模型泛化能力，不同则是前者在训练阶段可以访问目标域的数据。不同于 DA 和 DG 的目标，本文希望最终的模型表现形式是：每个节点的模型对于其他域具有良好的泛化能力，同时不会牺牲在节点本地源域的性能。这个思想是同非稳定性分布下的持续学习一致的[26]，持续学习也可以称为终生学习，就是要“最大化迁移能力(泛化)和最小化干扰(忘记)”。

1.3 协同学习

协同学习[27]不同于传统的“知识型老师(一般认为‘老师’比学生知识渊博)”的监督学习，它提供了一种新的学习范式：考虑在“同伴”之间交换信息，即一群“学生”相互学习和相互教授知识。双重学习是其中一个典型的工作[28]，它的两个跨语言翻译模型可以并行训练。文献[29]针对不同域中的同一任务联合培训多个代理，其中语义中层视觉属性用于代理之间的通信，因此代理需要学习特征属性模型以及特征类别模型。共蒸馏[30]用于大规模分布式网络训练，但是所有节点都需要相同的体系结构和数据集。

2 框架的公式化表达

本节将介绍 MAFML 框架的详细信息。假设多机构联盟中包含 N 个节点(每个机构节点可以理解为一个数据域)。

定义 1 节点彼此之间生成和存储具有明显分布差异的数据 $D=\{D_1,D_2,\cdots,D_N\}$，即数据是非独立同分布的，并且每个节点上的数据都包含一组数据标签对($D_i=\{X_i,Y_i\}$)。

基于隐私和通信带宽的考虑，机构间采用软标签的形式进行通信，因此将每个机构节点的数据集拆分为在本地保存的私有数据，可以共享给其他节点的公开数据、验证数据和测试数据，即 $D_i=\{D^i_{\text{pri}},D^i_{\text{pub}},D^i_{\text{val}},D^i_{\text{test}}\}$。不同于 FedAvg 需要共享全局模型的需求，本文框架在充分保留本地私有数据的同时，满足了机构的个性化需求。但是，在相互学习的过程中会出现私有模型性能的冲突干扰，本文框架的最终目标是在不牺牲特定域数据隐私的情况下实现节点之间的相互学习，并且节点间少量公开数据的“意识”交互是一种自然的策略，它可以提高跨节点的数据有效性，同时更好地保护隐私。

在本框架的设计中，假定所有机构间是同构跨域数据，即所有标签 Y_i 都位于 M 类的相同标签空间中。例如，对于相同疾病标签的医学图像，由于不同医疗机构的采集设备、医务人员工作习惯和不同的病症患者的差异性，导致了图像样本的分布差异。在基于深度神经网络的模型基础上，假设节点 i 使用由 $\boldsymbol{\theta}_i$ 参数化的神经网络，模型结构的要求是松散、灵活的，这是本文方法与模型无关的关键所在。此外，节点的大多数数据都保留在本地以更好地保护隐私，节点之间的相互学习是通过少量共享公开数据来实现的。因此，基于上述考虑，MAFML 的工作流程分为两个阶段：本地域局部优化和跨域全局优化。与此同时，为方便实现机构节点间的数据一致性，以及 MAFML 框架少量通信开销的内在优势，在全局优化阶段采取汇聚到中心服务器的数据交互模式。

2.1 本地域局部优化

单个机构域中的数据遵循稳定的独立同分布，本地域优化过程中采取传统的监督学习方式实现模型的收敛过程。节点的局部优化基于梯度的常规网络更新方式。将第 i 个节点的网络定义为 $f_{\boldsymbol{\theta}_i}$，针对框架所应对的分类场景，利用交叉熵

(cross-entropy, CE)损失来表示当前的神经网络对训练数据不拟合的程度:

$$F_{\min(\boldsymbol{\theta}_i)} = \ell^{(\mathrm{CE})}(f_{\boldsymbol{\theta}_i}(x^i_{\mathrm{loc}}), y^i_{\mathrm{loc}}) \tag{1}$$

$$\boldsymbol{g}^i_{\mathrm{loc}} = \nabla_{\boldsymbol{\theta}_i}\ell^{(\mathrm{CE})}(f_{\boldsymbol{\theta}_i}(x^i_{\mathrm{loc}}), y^i_{\mathrm{loc}}) \tag{2}$$

$(x^i_{\mathrm{loc}}, y^i_{\mathrm{loc}}) \in \{D^i_{\mathrm{pri}}, D^i_{\mathrm{pub}}\}$ 是 i 域中用于本地局部优化的批处理数据。上述局部优化过程中,来自其他域包含标签信息的公开数据对于当前训练域更新过程是可见、可用的,利用其他域的公开数据在 FL 研究中是一致采取的[22,31]。式(2)中 $\boldsymbol{g}^i_{\mathrm{loc}}$ 为本地局部优化后的梯度结果。

需要注意的是,$f_{\boldsymbol{\theta}_i}(x^i_{\mathrm{loc}})$ 是样本数据经过深度神经网络最后的 Softmax 层后输出的软标签,即多分类中的类别概率 p^i_{loc}。CE 损失函数是根据预测的软标签和样本的正确标签计算得出的。本地局部优化的梯度更新通常采用常规的监督学习获得,但是为了实现跨机构之间的联邦学习,还需要挑战协作过程中"稳定性 - 可塑性"的困境:稳定性代表了联邦学习过程中,本地节点域在受到其他节点域的影响后,依旧可以保持其性能;可塑性则表示在接受来自其他节点域信息的支持下,本地模型可以很好地泛化到其他节点域。

全局优化的准备工作:节点利用本地批处理数据完成梯度更新计算后,本文利用每个节点共享的公开数据进行下一阶段的准备工作,此时框架随机地从每个域的公开样本中采样 $d^i_{\mathrm{pub}} = (x^i_{\mathrm{pub}}, y^i_{\mathrm{pub}})$,并将其输入对应节点的神经网络模型中计算出软标签 $p^i_{\mathrm{pub}(i)}$(即上文指的低维的"意识"),这里需要注意的是,下标中的 i 表示数据是来自 i 域的,上标中的 i 表示数据的应用网络是 $f_{\boldsymbol{\theta}_i}$。根据计算得出的软标签以及样本对应的正确标签可以计算出批处理公开数据的测试精度 A_i。在本地域局部优化的最后阶段,每个域都会形成对应的 $p^i_{\mathrm{pub}(i)}$ 和 A_i 以用于全局优化。这两种数据相对模型参数而言都是低维的,因此本文方法具有很高的通信效率。

2.2　跨域全局优化

一般意义上的学习过程是受过预训练的老师将其知识迁移到未经专业训练的学生,这可以看作是一个知识蒸馏[32]的过程。蒸馏过程可以是大规模网络转化成一个小规模目标网络,并保留接近于大规模网络的性能;也可以将多个网络的知识转移到目标网络中,使得目标网络的性能接近多个源网络聚合的效果。因此,联合的多个节点可以看作是学生互相教授学习的过程(即每个人都参与到学习和教导的过程中,实现老师和学生角色的交替轮换)。具体而言,通过全局优化的过程实现上述描述的"互教互学"。

作为老师角色的节点:局部优化过程中获得了来自其他域公开数据的 $p^i_{\mathrm{pub}(i)}$ 和 A_i,它们可以看作来自其他域节点的教授经验。$p^i_{\mathrm{pub}(i)}$ 是作为老师的"教学基准",A_i 是老师的"教学信心"。

作为学生角色的节点:学生节点通过域公开数据从所有其他节点学习域经验。为了提高学生节点 $f_{\boldsymbol{\theta}_i}$ 到其他节点域的泛化能力,本文使用其他与之不相交的节点 $f_{\boldsymbol{\theta}_j}$ 计算出的 $p^j_{\mathrm{pub}(j)}$ 和 A_j 作为老师的经验。因此,如果不考虑节点向中心服务器传递的有效时限要求,那么将有 $N-1$ 个老师服务于学生 $f_{\boldsymbol{\theta}_i}$。学习的过程是学生对老师认知的模仿,两者针对同一样本的输出均是多分类问题的类别概率分布。因此,KL(kullback-leibler divergence)的散度损失可用于约束学生和教师预测的匹配。KL 损失函数为:

$$\ell^{(\mathrm{KL})} = \frac{1}{N-1}\sum_{j=1,j\neq i}^{N} A_j \cdot D_{\mathrm{KL}}(P^j_{\mathrm{pub}(j)} \| p^i_{\mathrm{pub}(j)}) \tag{3}$$

对应地,

$$D_{\mathrm{KL}}(P^j_{\mathrm{pub}(j)} \| p^{\mathrm{pub}(j)}_i) = E_{p_j}[\log p^j_{\mathrm{pub}(j)} - \log p^i_{\mathrm{pub}(j)}] \tag{4}$$

在全局优化阶段,节点利用公开数据进行了学习模仿的 KL 损失。同时,可以充分利用这部分数据计算传统的监督损失(例如,CE 损失函数)

$$\ell^{(\mathrm{CE})} = \frac{1}{N-1}\sum_{j=1,j\neq i}^{N} \ell^{(\mathrm{CE})}(f_{\boldsymbol{\theta}_i}(x^j_{\mathrm{pub}}), y^j_{\mathrm{pub}}) \tag{5}$$

利用两个损失函数,本文采取反向传播可以计算对应的梯度为:

$$\boldsymbol{g}^i_{\mathrm{pub}} = \nabla_{\boldsymbol{\theta}_i}(\ell^i_{(\mathrm{KL})} + \ell^i_{(\mathrm{CE})}) \tag{6}$$

因此,全局优化过程中参与的节点轮流充当学生和老师的角色。每个节点将其他 $N-1$ 个节点作为教师,通过其公开数据提供学习经验和基准。那么,综合上述两个阶段获得的两个梯度:局部优化梯度 $\boldsymbol{g}^i_{\mathrm{loc}}$ 和全局优化梯度 $\boldsymbol{g}^i_{\mathrm{pub}}$。需要考虑的问题是:是否直接将全局梯度更新到模型参数 $\boldsymbol{\theta}_i$ 上?回顾本文跨节点的学习目标,即不仅关注"可塑性"(提高对其他域的泛化),而且关注"稳定性"(减少其自身域的遗忘)。跨机构联邦学习带来了诸多优化挑战,尤其以梯度更新作为典型的模型优化方式,冲突的梯度方向以及过度估计的全局梯度更新可能会严重影响本地域模型,导

致更差的系统性能和数据使用效率。因此,需要构建一种简洁高效的多机构梯度冲突消解方法,以缓解全局梯度对于局部梯度 $\boldsymbol{g}^i_{\text{loc}}$ 的干扰。

假设联邦学习中一个关键的优化问题是由相互矛盾的梯度引起的,梯度的更新方向不一致,这由负内积来衡量。

定义 2 将两个阶段梯度更新 $\boldsymbol{g}^i_{\text{pub}}$ 和 $\boldsymbol{g}^i_{\text{loc}}$ 的夹角定义为 σ。当夹角的余弦值 $\cos\sigma < 0$ 时,称梯度发生了冲突,反之没有。

本文的目标是通过直接改变梯度本身来防止冲突。这里框架采用了一个简洁有效的方式:如果梯度发生冲突,即它们的余弦相似度为负,则将全局梯度投影到局部梯度的法线平面上,这等于消除了全局梯度的冲突分量,从而减少了机构联邦学习中的破坏性梯度干扰。当余弦相似度为正时,直接采用全局梯度结果,利用一般性的优化器 Adam[33] 来更新模型参数。

当冲突条件满足时,采取映射替换的方式来消解,即 $\tilde{\boldsymbol{g}}_i = f_{\text{projection}}(\boldsymbol{g}^i_{\text{loc}}, \boldsymbol{g}^i_{\text{pub}})$,$f_{\text{projection}}$ 为映射替换函数。具体过程是:通过从 $\boldsymbol{g}^i_{\text{pub}}$ 中减去 $\boldsymbol{g}^i_{\text{loc}}$ 与平面正交的分量来计算其投影。如果平面是水平的,则意味着计算减去 $\boldsymbol{g}^i_{\text{pub}}$ 的垂直分量,剩下水平分量。该垂直分量计算为 $\boldsymbol{g}^i_{\text{pub}}$ 在平面法线向量 $\boldsymbol{g}^i_{\text{loc}}$ 上的投影,具体如下:

$$\tilde{\boldsymbol{g}}_i = \boldsymbol{g}^i_{\text{pub}} - \frac{\boldsymbol{g}^i_{\text{loc}} \cdot \boldsymbol{g}^i_{\text{pub}}}{\|\boldsymbol{g}^i_{\text{loc}}\|^2} \boldsymbol{g}^i_{\text{loc}} \tag{7}$$

算法 1 中细节描述了整个算法流程。总而言之,在每个节点域中,首先对本地保留数据和公开数据进行采样,然后在本地计算:使用 $\ell^{(\text{CE})}$ 在本地保留数据批量上执行 Adam 更新梯度 $\boldsymbol{g}^i_{\text{loc}}$,然后在跨域的公开数据上形成“意识”信息 $p^i_{\text{pub}(i)}$ 和 A_i 作为相互学习的准备。在全局优化阶段,在多域的共同学习中引入了 $\ell^{(\text{KL})}$ 和常规的 $\ell^{(\text{CE})}$ 损失函数,以计算跨域公共梯度 $\boldsymbol{g}^i_{\text{pub}}$。然后根据梯度冲突条件判断,计算适当的梯度 $\tilde{\boldsymbol{g}}_i$ 作为最终的全局梯度来更新节点模型的网络。这不仅是向其他域积极的前向迁移,而且是积极的后向迁移以保持模型在其原始域上的性能。

3 模型的收敛和性能评估

3.1 模型的收敛性

假设式(1)对应的 $\ell^{(\text{CE})}$,和式(6)对应的 $\ell^{(\text{CE})} + \ell^{(\text{KL})}$ 是凸函数且可微分的。$\boldsymbol{g}^{\text{loc}}$ 为前者产生的梯度,$\boldsymbol{g}^{\text{pub}}$ 为后者产生的梯度。那么,根据全局优化描述,在两个梯度冲突判断的过程中,存在两种情况:$\cos\sigma < 0$ 或 $\cos\sigma \geqslant 0$。前者代表了全局梯度对于局部梯度的正向作用,此时对于模型参数的更新,本文采用标准的 Adam 优化器;后者则表示两个梯度向量存在影响性能的冲突挑战。因此,根据梯度映射对于损失函数所带来的影响,本文扩展了其损失函数的构成,即:

$$\ell(\boldsymbol{\theta}^+) \leqslant \ell(\boldsymbol{\theta}) + \nabla\ell(\boldsymbol{\theta})^{\text{T}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}) \tag{8}$$

定理 1 框架构建的关于全局优化和局部优化结合的映射方式会收敛到:当两个梯度更新的夹角 $\cos\sigma = -1$,或者最优的损失函数值 $\ell(\boldsymbol{\theta}^*)$。

算法 1 模型不可知的联合相互学习

Alg. 1 Model agnostic federated mutual learning

已知:N 个节点域 $D = \{D_1, D_2, \cdots, D_N\}$,$D_i = \{D^i_{\text{pri}}, D^i_{\text{pub}}, D^i_{\text{val}}, D^i_{\text{test}}\}$

输入:初始化神经网络 $\{f_{\theta_1}, f_{\theta_2}, \cdots, f_{\theta_N}\}$,学习率 β, η

输出:优化后的神经网络 $\{f_{\theta_1}, f_{\theta_2}, \cdots, f_{\theta_N}\}$

while not 模型收敛或者达到最大迭代次数 **do**
  for $i \in [1,2,\cdots N]$ **do**
    随机采样 d^i_{loc} 和 d^i_{pub}
    根据式(2)计算 $\boldsymbol{g}^i_{\text{loc}}$
    更新参数 $\theta_i \leftarrow \theta_i - \beta \cdot \boldsymbol{g}^i_{\text{loc}}$
    在 d^i_{pub} 上计算 $p^i_{\text{pub}(i)}$、A_i,并传递到中心服务器
  end for
  for $i \in [1,2,\cdots N]$ **do**
    for $j \in [1,2,\cdots N]$, $j \neq i$ **do**
      利用 f_{θ_i} 计算 $p^i_{\text{pub}(j)}$
      根据式(6)计算 $\boldsymbol{g}^i_{\text{pub}}$
      if 冲突满足 **then**
        $\tilde{\boldsymbol{g}}_i = \boldsymbol{g}^i_{\text{pub}} - \frac{\boldsymbol{g}^i_{\text{loc}} \cdot \boldsymbol{g}^i_{\text{pub}}}{\|\boldsymbol{g}^i_{\text{loc}}\|^2} \boldsymbol{g}^i_{\text{loc}}$
      else
        $\tilde{\boldsymbol{g}}_i = \boldsymbol{g}^i_{\text{pub}}$
      end if
      更新参数 $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \eta \cdot \tilde{\boldsymbol{g}}_i$
    end for
  end for
end

证明:根据式(8)中对于损失函数的扩展,梯度运算结果 $\nabla\ell(\boldsymbol{\theta})^{\text{T}} = (\boldsymbol{g}^{\text{pub}})^{\text{T}}$,其中 $\boldsymbol{\theta}^+$ 代表了采用梯度映射运算后的模型参数结果:

$$\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta\left(\boldsymbol{g}^{\text{pub}} - \frac{\boldsymbol{g}^{\text{loc}} \cdot \boldsymbol{g}^{\text{pub}}}{\|\boldsymbol{g}^{\text{loc}}\|^2} \boldsymbol{g}^{\text{loc}}\right) \tag{9}$$

将式(9)代入式(8),并将其继续展开,其运算结果如下:

$$\ell(\boldsymbol{\theta}^+) \leqslant \ell(\boldsymbol{\theta}) + \nabla\ell(\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta})$$
$$= \ell(\boldsymbol{\theta}) - \eta(\boldsymbol{g}^{\mathrm{pub}})^{\mathrm{T}}\left(\boldsymbol{g}^{\mathrm{pub}} - \frac{\boldsymbol{g}^{\mathrm{loc}} \cdot \boldsymbol{g}^{\mathrm{pub}}}{\|\boldsymbol{g}^{\mathrm{loc}}\|}\boldsymbol{g}^{\mathrm{loc}}\right)$$
$$= \ell(\boldsymbol{\theta}) - \eta\|\boldsymbol{g}^{\mathrm{pub}}\|^2\left(1 - \frac{(\boldsymbol{g}^{\mathrm{loc}} \cdot \boldsymbol{g}^{\mathrm{pub}})^2}{\|\boldsymbol{g}^{\mathrm{loc}}\|^2\|\boldsymbol{g}^{\mathrm{pub}}\|^2}\right) \quad (10)$$

将两个梯度冲突判断中的余弦函数展开:

$$\cos\sigma = \frac{\boldsymbol{g}^{\mathrm{loc}} \cdot \boldsymbol{g}^{\mathrm{pub}}}{\|\boldsymbol{g}^{\mathrm{loc}}\|\|\boldsymbol{g}^{\mathrm{pub}}\|} \quad (11)$$

那么,把式(11)代入式(10),可以获得该不等式的最终形式:

$$\ell(\boldsymbol{\theta}^+) \leqslant \ell(\boldsymbol{\theta}) + \nabla\ell(\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta})$$
$$\leqslant \ell(\boldsymbol{\theta}) - \eta\|\boldsymbol{g}^{\mathrm{pub}}\|^2(1 - \cos\sigma^2) \quad (12)$$

因此,重复上述优化过程,模型将最终收敛在:当损失函数 $\ell(\boldsymbol{\theta})$ 到达最优值 $\ell(\boldsymbol{\theta}^*)$;或当式(12)中后一项为 0 的情况下达到收敛。 □

3.2　性能评估函数

模型的性能评估函数:后向迁移(backward transfer, BWT)、前向迁移(forward transfer, FWT)和平均准确率(average accuracy, ACC)。

后向迁移:$P^i_{\mathrm{BWT}} = F_i(D^i_{\mathrm{test}})$。$P^i_{\mathrm{BWT}}$ 是 i 节点的模型 $f_{\boldsymbol{\theta}_i}$ 在该节点测试数据上的域内性能。联合学习过程中存在积极的后向迁移,即来自其他节点的学习经验可以使得本域的性能提升。同时,还存在消极后向迁移。

前向迁移:$P^i_{\mathrm{FWT}} = F_i(\sum_{n=1,n\neq i}^{N} D^n_{\mathrm{test}})$。$P^i_{\mathrm{FWT}}$ 是 i 节点模型 $f_{\boldsymbol{\theta}_i}$ 在所有其他节点测试数据上的跨域性能。FWT 显示模型的泛化性能。

平均准确率:$P^i_{\mathrm{ACC}} = F_i(\sum_{n=1}^{N} D^n_{\mathrm{test}})$。$P^i_{\mathrm{ACC}}$ 是 i 节点模型 $f_{\boldsymbol{\theta}_i}$ 在所有节点测试数据上的全域性能。

这里 F 是用于计算测试数据准确性的常规函数,F 的下标 i 表示使用第 i 个节点的网络模型,括号中的数据为测试数据。这些指标越大,则节点的模型性能越好。

4　实验分析

在跨域任务环境中评估 MAFML 方法的效果:数字分类(旋转 MNIST)和图像识别(PACS)。数据集具有域偏移特性。本文使用 Ray[34] 框架对 MAFML 方法进行分布式的实现,MAFML 的程序代码采用 PyTorch 框架构建深度神经网络模型,并在具有 4 个 GPU(NVIDIA GeForce GTX 1080 Ti)的服务器上训练运行。MAFML 主要与以下基准方案进行比较:

独立(independent, IND):每个节点在基于 CE 损失函数的 Adam 优化过程中,仅使用自己域的私有和公开数据,训练过程中完全避免了其他节点可能会带来的干扰因素。

汇聚(aggregation, AGG):每个节点在基于 CE 损失函数的 Adam 优化过程中,使用的数据不仅仅是自己域的私有和公开数据,同时还汇聚了来自其他域的公开数据。AGG 方法通常是跨域情况下一个很强烈的基准对比。

FedMD[22]:适用于通过模型蒸馏实现的异构 FL。

FedAvg[8]:仅适用于具有同构网络的 FL 常规方法,它采用集中式服务器将节点的梯度聚合,然后将相同的参数分配给节点。

4.1　旋转 MNIST 实验

4.1.1　数据集和实验设定

旋转 MNIST(RMNIST、Rotated MNIST)包含不同的域,每个域对应经典 MNIST 数据集中不同旋转程度的样本。因此,不同的节点具有不同的数据统计特性。基本视图(M0)是通过从原始 MNIST 数据集中随机选择十个类别,每个类别 100 张图像形成的,然后根据 M0 创建 3 个旋转域,每个旋转域均沿顺时针方向旋转 20°,分别命名为 M20、M40、M60。根据私有数据、公开数据、验证数据和测试数据,将每个节点域的数据默认拆分为 65%、10%、10%、15% 的比例。

节点上简单地部署同构网络 LeNet[35] 进行实验。训练过程采用 Adam 优化器(学习率为 0.001,权重衰减为 0.000 1),整个训练的迭代次数为10 000,批处理大小为 32。一些超参因素将会影响模型的性能如 α、域共享数据分片 D^i_{pub} 所占的比例。本文固定设置 $D^i_{\mathrm{pub}} + D^i_{\mathrm{pri}}$ 的比例为 75%,D^i_{val} 和 D^i_{test} 的比例则稳定地设置为 10% 和 15%。需要注意的是,IND 和 FedAvg 的性能与 α 的值无关。

4.1.2　结果分析

表 1 显示了上述所描述方法之间的比较,其中部分方法是 α 的函数。每 50 轮使用验证数据进行评估,并根据训练过程中最大 ACC 保存的模型对三个指标进行最终测试。表 1 中每个指标的最大值标注为粗体。对于大多数 α 的设置情况,MAFML 总是优于其他方法。通常,设置中 α 值的增加会使 MAFML 性能有所改善。IND 通常在 BWT 上胜过 AGG 和其他方法,但是在 FWT 的性能要比 AGG 低得多,原因在于 IND 仅在每个节点域内部数据中训练,而没有其他域的信息。AGG

表1 旋转 MNIST 上三个度量的测试结果

Tab. 1 Test result on three metrics on Rotated MNIST

%

方法	M0-LeNet			M20-LeNet			M40-LeNet			M60-LeNet			平均结果		
	ACC	BWT	FWT	ACC	BWT	FWT	ACC	BWT	FWT	ACC	BWT	FWT	ACC	BWT	FWT
MAFML($\alpha=5\%$)	86.33	**93.33**	84.22	89.00	94.00	87.33	**89.00**	94.67	87.11	86.83	**94.00**	85.33	87.79	**94.00**	86.00
AGG($\alpha=5\%$)	85.50	92.67	83.11	87.50	93.33	85.56	83.67	90.00	81.56	83.83	93.33	80.67	85.13	92.33	82.73
FedMD($\alpha=5\%$)	84.17	87.33	83.11	85.33	91.33	83.33	86.67	**96.00**	83.56	84.17	91.33	81.78	85.09	91.50	82.95
MAFML($\alpha=10\%$)	**90.67**	**93.33**	**89.78**	**90.00**	**95.33**	**88.67**	85.50	90.67	86.89	86.67	91.33	85.33	88.21	92.67	87.67
IND	66.39	91.33	58.08	78.11	94.00	72.82	72.39	93.11	65.48	56.89	91.78	45.48	68.45	92.56	60.47
AGG($\alpha=10\%$)	86.50	90.00	85.33	87.17	92.67	85.33	86.67	94.00	84.22	80.67	91.33	77.11	85.25	92.00	83.00
FedMD($\alpha=10\%$)	85.00	88.67	83.78	87.67	**95.33**	85.11	82.00	90.67	79.11	85.67	90.00	84.22	85.09	91.17	83.06
FedAvg	86.50	77.33	89.56	86.50	86.67	86.44	86.50	92.67	84.44	86.50	89.33	85.56	86.50	86.50	86.50
MAFML($\alpha=15\%$)	89.33	91.33	88.67	89.67	92.67	**88.67**	88.83	93.33	**87.33**	**89.00**	92.00	**88.00**	**89.21**	92.33	**88.17**
AGG($\alpha=15\%$)	87.83	92.00	86.44	89.67	92.10	88.44	87.83	94.00	85.78	86.00	91.33	84.22	87.83	92.36	86.22
FedMD($\alpha=15\%$)	88.67	89.33	88.44	89.00	93.33	87.56	85.00	90.00	83.33	84.33	92.67	81.56	86.75	91.33	85.22

表现出比 FedMD 更好的性能。对于 FedAvg，其基于梯度的通信带宽成本是基于“意识”方法的1 000倍以上。本文保持其通信强度与 MAFML 基本相同的情况下记录结果。在同构网络设定下，MAFML 在所有指标上的表现明显更好。

4.2 PACS 实验

4.2.1 数据集和实验设定

PACS 是用于域泛化和域适应研究的对象识别数据集。它包含来自 4 个不同域（照片、艺术绘画、卡通和素描）的 9 991 张图像（剪切大小为 224 像素 ×224 像素），具有 7 个类别：狗、大象、长颈鹿、吉他、房子、马和人。原始的 PACS 数据集已被固定拆分为训练、验证和测试三部分，因此为了满足 MAFML 的数据结构，将 PACS 测试部分的 10% 作为节点的公开数据，测试部分剩余的 90% 作为节点实际的测试数据，PACS 验证部分作为节点的验证数据，并直接使用其训练部分作为节点的私有数据。PAC 上三个度量的测试结果如表 2 所示。

首先，将 ResNet18 作为 MAFML 的同构网络部署到所有节点。更重要的是，当模型异构时，MAFML 具有天然优势，分别将网络 ResNet18、ResNet34、AlexNet 和 VGG11 随机部署为多域节点的网络结构。训练阶段，模型使用 Adam 优化器（学习率为 0.000 1，权重衰减为 0.000 01）训练 10 000 次迭代，批处理大小设置为 32。

4.2.2 结果分析

从表 2 的结果可以看出：①MAFML 通常会体现优于其他方法的效果，尽管对于同构模型 IND 在 BWT 上的表现要好一些。当同构节点部署的网络使用 ResNet18 架构时，FedAvg 的原始通信带宽约是本文的 10^6 倍，考虑到实际训练过程的控制，将 FedAvg 的通信带宽控制为本文的 100 倍，但它仍然无法赶上 MAFML 的性能。②特别是，当不同节点模型为异构类型时，FedAvg 本质上不适用于这种情况。从平均结果看，MAFML 仍胜过其他方法，AGG 依旧是一个强基准方法。此外，还有一个有趣的发现，虽然采用 VGG11 网络的节点在 Sketch 域中的性能不佳（请参阅其对应的 IND、AGG、BWT 结果），但是通过相互学习的过程它也不会拖累其他节点反而会使其受益（MAFML 在这些域的表现优于 IND 和 FedMD）。因此，还可以通过对模型能力的自我评估来反映式（3）中 A_j 作为“教学信心”的合理性。

表 2　PACS 上三个度量的测试结果

Tab. 2　Test result on three metrics on PACS

%

方法	Photo-ResNet18			Art_painting-ResNet18			Cartoon-ResNet18			Sketch-ResNet18			平均结果		
	ACC	BWT	FWT	ACC	BWT	FWT	ACC	BWT	FWT	ACC	BWT	FWT	ACC	BWT	FWT
MAFML	**85.25**	**100**	82.30	87.82	99.79	84.74	**86.90**	**100**	82.91	**91.10**	**99.97**	**85.66**	**87.77**	99.94	83.90
IND	51.45	**100**	41.70	70.53	99.95	62.94	73.48	99.95	65.36	62.95	99.89	39.05	64.60	**99.95**	52.26
AGG	84.52	99.93	81.42	86.30	99.89	82.79	85.46	**100**	81.00	89.35	99.89	82.53	86.41	99.93	81.94
FedMD	82.39	99.87	78.88	85.75	99.62	82.17	83.93	99.91	79.03	88.52	98.56	82.02	85.15	99.49	80.53
FedAvg	84.93	95.62	**82.78**	84.93	72.06	**88.25**	84.93	72.23	**88.82**	84.93	94.68	78.62	84.93	83.65	**84.62**
MAFML	84.34	**100**	81.21	**89.59**	**100**	87.25	82.27	**100**	76.82	52.74	78.83	37.08	77.24	94.71	70.59
IND	51.08	99.57	41.29	77.72	99.30	72.15	68.52	99.39	59.05	44.79	78.75	22.83	60.53	94.25	48.83
AGG	84.90	**100**	81.90	89.50	**100**	86.85	80.80	98.77	75.28	52.81	78.01	36.51	77.00	94.20	70.14
FedMD	80.05	**100**	76.05	86.90	99.08	83.75	78.07	95.65	72.67	51.40	75.47	35.83	74.11	92.55	67.08

5　结论

本文提出了模型不可知的联合相互学习:一种可以在节点之间协作且高效通信的“互教互学”的 FL 框架。MAFML 不限制机构的深度神经网络模型的结构和参数,并且在跨域情况下具有更好的可塑性和稳定性。实验表明,MAFML 为行业联盟提供了一种有前途的方式来解决“竞争与协作”的困境。

参考文献(References)

[1] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[2] CANO I, WEIMER M, MAHAJAN D, et al. Towards geo-distributed machine learning[EB/OL]. (2016 - 03 - 30)[2021 - 05 - 10]. https://arxiv.org/abs/1603.09035.

[3] THUSOO A, SHAO Z, ANTHONY S, et al. Data warehousing and analytics infrastructure at facebook[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data, 2010: 1013 - 1020.

[4] PU Q, ANANTHANARAYANAN G, BODIK P, et al. Low latency geo-distributed data analytics[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 421 - 434.

[5] LANDA R, CLEGG R G, ARAUJO J T, et al. Measuring the relationships between internet geography and RTT[C]//Proceedings of 22nd International Conference on Computer Communication and Networks (ICCCN), 2013.

[6] VULIMIRI A, CURINO C, GODFREY P B, et al. WANalytics: geo-distributed analytics for a data intensive world[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data, 2015: 1087 - 1092.

[7] LI D, YANG Y X, SONG Y Z, et al. Deeper, broader and artier domain generalization[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV), 2017.

[8] GHIFARY M, KLEIJN W B, ZHANG M J, et al. Domain generalization for object recognition with multi-task autoencoders[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015.

[9] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency[EB/OL]. (2016 - 10 - 18)[2021 - 05 - 10]. https://arxiv.org/abs/1610.05492.

[10] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1 - 19.

[11] MELIS L, SONG C, DE CRISTOFARO E, et al. Exploiting unintended feature leakage in collaborative learning. [C]//Proceedings of 40th IEEE Symposium on Security & Privacy, 2019.

[12] ZHU L, LIU Z, HAN S. Deep leakage from gradients[EB/OL]. (2019 - 12 - 19)[2021 - 05 - 10]. https://arxiv.org/abs/1906.08935.

[13] SHOKRI R, SHMATIKOV V. Privacy-preserving deep learning[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015: 1310 - 1321.

[14] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2017: 1175 - 1191.

[15] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[16] HARDY S, HENECKA W, IVEY-LAW H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption[EB/OL]. (2017 - 11 - 29)[2021 - 05 - 10]. https://arxiv.org/abs/1711.10677.

[17] LIU Y, KANG Y, XING C, et al. Secure federated transfer learning[EB/OL]. (2020 - 06 - 24)[2021 - 05 - 10]. https://arxiv.org/abs/1812.03337.

[18] FALLAH A, MOKHTARI A, OZDAGLAR A E. Personalized

federated learning with theoretical guarantees: a model-agnostic meta-learning approach[C]// Proceedings of 34th Conference on Neural Information Processing Systems, 2020.

[19] JIANG Y, KONENČNÝ J, RUSH K, et al. Improving federated learning personalization via model agnostic meta learning[EB/OL]. (2019 - 09 - 27)[2021 - 05 - 10]. https://arxiv.org/abs/1909.12488v1.

[20] SMITH V, CHIANG, SANJABI M, et al. Federated multi-task learning[EB/OL]. (2018 - 02 - 27)[2021 - 05 - 10]. https://arxiv.org/abs/1705.10467.

[21] SHEN T, ZHANG J, JIA X, et al. Federated mutual learning [EB/OL]. (2020 - 09 - 17)[2021 - 05 - 10]. https://arxiv.org/abs/2006.16765.

[22] LI D, WANG J. FedMD: heterogenous federated learning via model distillation[C]//Proceedings of NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality, 2019.

[23] YANG Y X, HOSPEDALES T M. A unified perspective on multi-domain and multi-task learning[C]//Proceedings of 3rd International Conference on Learning Representations, 2015.

[24] LONG M S, ZHU H, WANG J M, et al. Unsupervised domain adaptation with residual transfer networks [C]// Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016.

[25] LI Y Y, YANG Y X, ZHOU W, et al. Feature-critic networks for heterogeneous domain generalization [C]// Proceedings of the 36th International Conference on Machine Learning, 2019.

[26] MCCLOSKEY M, COHEN N J. Catastrophic interference in connectionist networks: the sequential learning problem[M]// Psychology of Learning and Motivation. Amsterdam: Elsevier, 1989: 109 - 165.

[27] LEE W S. Collaborative learning for recommender systems[C]// Proceedings of International Conference on Machine Learning (ICML), 2001: 314 - 321.

[28] XIA Y C, HE D, QIN T, et al. Dual learning for machine translation [C]//Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016.

[29] BATRA T, PARIKH D. Cooperative learning with visual attributes [EB/OL]. (2017 - 05 - 16)[2021 - 05 - 10]. https://arxiv.org/abs/1705.05512.

[30] ANIL R, PEREYRA G, PASSOS A, et al. Large scale distributed neural network training through online distillation[EB/OL]. (2020 - 08 - 20)[2021 - 05 - 10]. https://arxiv.org/abs/1804.03235.

[31] ZHAO Y, LI M, LAI L, et al. Federated learning with non-IID data [EB/OL]. (2018 - 06 - 02)[2021 - 05 - 10]. https://arxiv.org/abs/1806.00582.

[32] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[C]// Proceedings of NISP Deep Learning Workshop, 2014.

[33] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]// Proceedings of ICLR, 2015.

[34] MORITZ P, NISHIHARA R, WANG S, et al. Ray: a distributed framework for emerging AI applications [C]// Proceedings of 13th USENIX Symposium on Operating Systems Design and Implementation, 2018.

[35] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278 - 2324.