引用格式：曹行健, 孙罡, 虞红芳. 基于 Transformer 的多模态个性化联邦学习 [J]. 电子科技大学学报, 2025, 54(2): 242-249.
CAO X J, SUN G, YU H F. Multimodal personalized federated learning based on Transformer[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(2): 242-249.

基于 Transformer 的多模态个性化联邦学习

曹行健，孙 罡*，虞红芳

（电子科技大学 信息与通信工程学院，成都 611731）

摘要：在当前物联网飞速发展的背景下，处理来自各种信息采集设备的多模态数据，尤其是视觉、听觉信号和文本等多元感官信息的数据，对于机器学习落地应用至关重要。Transformer 架构和其衍生的大模型在自然语言处理和计算机视觉中的卓越表现推动了对复杂多模态数据处理能力的追求。然而，这也带来了数据隐私安全和满足个性化需求的挑战。为解决这些挑战，提出一种基于多模态 Transformer 的个性化联邦学习方法，它支持异构数据模态的联邦学习，在保护参与方数据隐私的前提下为其训练更符合其个性化需求的多模态模型。该方法显著提升了多模态个性化模型的性能：相较于对比方法，准确率提高了 15%，这标志着多模态个性化联邦学习在应用场景限制上的突破。

关键词：多模态；Transformer；联邦学习；个性化
中图分类号：TP301 **文献标志码：**A DOI: 10.12178/1001-0548.2024050

Multimodal personalized federated learning based on Transformer

CAO Xingjian, SUN Gang*, and YU Hongfang

(School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: In the context of the current rapid development of the Internet of Things, processing multi-modal data from various information collection devices, especially data from multi-sensory information such as visual, auditory signals and text, is crucial for the applications of machine learning. The outstanding performance of the Transformer architecture and its derived large models in natural language processing and computer vision has promoted the pursuit of complex multi-modal data processing capabilities. However, this also brings the challenges of data privacy security and meeting personalized needs. In order to solve these challenges, this paper proposes a personalized federated learning method based on multi-modal Transformer, which supports federated learning of heterogeneous data modalities, and its training is more consistent with its purpose while protecting the data privacy of the participants. The proposed method significantly improves the performance of the multi-modal personalized model, its accuracy is increased by 15% compared with the comparative method, which marks a breakthrough in the application scenario limitations of multi-modal personalized federated learning.

Key words: multi-modal; Transformer; federated learning; personalized

多模态学习的理念源于大脑能够无缝整合多元感官信息的能力，旨在通过融合不同模态的数据，提供更全面的分析，从而优化任务处理性能[1]。另一方面，物联网（internet of things, IoT）技术的发展极大地提高了数据模态的丰富性。因此，多模态学习方法，作为一种可以利用不同传感器等信号采集设备产生的不同模态数据的机器学习方法，日益受到关注。

近年来，起源于自然语言处理（natural language processing, NLP）领域的 Transformer 模型因其自注意力机制的卓越表现而成为研究焦点[2]。进一步研究发现，这类模型不仅在处理文本数据任务中取得了显著的进展，而且在计算机视觉领域任务如图像和视频识别中也显示出其强大的能力[3]。特别是，视觉变换器（vision transformer, ViT）等架构在计算机视觉任务上取得了突破性的成果[4]。以

收稿日期：2024−03−08
基金项目：国家重点研发计划（2021YFB3101001）
作者简介：曹行健，博士生，主要从事联邦学习算法方面的研究。
*通信作者 E-mail: gangsun@uestc.edu.cn

Transformer 为基础的多模态大模型也在许多现实多模态任务上表现优异。然而，在处理这些多模态信息时，主要面临着两方面挑战：一方面，如何在保护隐私的前提下整合和分析来自不同设备和平台的数据；另一方面，如何为特定用户提供更加符合其个性化需求的任务处理能力。

对于数据隐私挑战，联邦学习（federated learning, FL）是一种可行的方案，其优势在于可以无须分享参与方敏感数据的前提下协同训练机器学习模型。但联邦学习的初衷是利用分布在各参与方的私有数据来训练一个全局模型，因此它在满足个体化需求方面往往显得力不从心。为了克服个性化需求的挑战，FL 社区已提出了一系列解决方案以实现个性化联邦学习（personalized federated learning, PFL），如多任务学习[5]、动态客户参与策略[6]，以及 FedDAT 这类适应性框架[7]。因此，在进行面向多模态数据的机器学习时，要想克服数据隐私泄露和个性化需求的挑战，以 PFL 的形式训练多模态 Transformer 模型显得尤为契合。然而，尽管 Transformer 模型在处理序列任务方面表现出色，但将其应用到多模态 FL 环境中依然存在一定挑战[8]。目前的多模态 FL 方法往往需要数据对齐，这与 FL 保护用户隐私的宗旨不符[9-10]。因此，本文提出了一种基于 Transformer 的全新的多模态个性化联邦学习方法，该方法既能利用 Transformer 对多模态数据的特征提取能力，又结合注重隐私保护的联邦学习，为个性化多模态学习开辟了新的道路。

1　问题建模

假设有N个参与方，有M种数据模态$\mathcal{M}_j(j=1,2,\cdots,M)$，对于第$i$个参与方$(i=1,2,\cdots,N)$，其本地数据集为：

$$\mathcal{D}_i=\{(X_k,y_k)\}_{k=1}^{n_i} \tag{1}$$

$$X_k\in C_i \tag{2}$$

$$y\in\mathcal{Y} \tag{3}$$

式中，X_k是输入；y_k是输出；n_i是数据集$\mathcal{D}_i$的大小；$\mathcal{Y}$是输出空间；C_i是第i个参与方的数据模态。

$$C_i\subseteq C \tag{4}$$

$$C=\{\mathcal{M}_1,\mathcal{M}_2,\cdots,\mathcal{M}_M\} \tag{5}$$

由于来自不同模态的数据可能有不同的大小，因此X_k的元素数量可能不同。

不同参与方的C_i可能不同，因为它们可能具有不同的数据模态组成。本文的目标是提出一种多模态联邦学习方法，可以应对这种设置下的联邦学习需求。并在此基础之上，对每个特定参与方进行个性化模型调整，以实现多模态个性化联邦学习模型训练。

现有的多模态联邦学方法之所以不能应用于上述问题，是因为当参与方之间的数据模态组合不同时，即$C_i\neq C$时，现有方法的数据模态融合策略无法应用于现有的模态缺失和输入数据未对齐的情况。在本文提出的框架中，不同模态的数据会首先被分段并进行线性映射和模态信息嵌入，然后对于包含模态信息的线性映射结果使用基于 Transformer 多模态协同训练，来实现面向不同模态组合的多模态联邦学习，最后每个参与方针对其自身的模态组成和任务需求对模型进行调整以得到个性化模型。

1.1　线性嵌入

为了适应多模态数据输入的多样性，本文首先将不同模态的数据分割成多段独立的数据片段；然后对这些片段进行线性映射以标准化它们的表示；随后，为每个转换后的数据片段嵌入一个模态向量，从而将模态特定信息嵌入每个数据片段中；最后，这些附加了模态信息的线性映射结果按原始顺序重新组合形成序列化的线性嵌入结果。

当处理来自任何给定模态的样本X时，本文将其划分为一系列连续的数据片段，遵循该模态数据固有属性：

$$X=[x_1,x_2,\cdots,x_n] \tag{6}$$

每个片段x_i基于该模态的独特特征划分而成。如图像输入可能被划分为图片补丁，而自然语言输入可能被拆分为对应于词嵌入的段落，声波输入可被拆分为多个波形片段。

对于某个给定的模态m，使用一个特定于模态的矩阵$\boldsymbol{M}_m$对该模态的数据片段执行线性映射：

$$\boldsymbol{Y}=[x_1\boldsymbol{M}_m,x_2\boldsymbol{M}_m,\cdots,x_n\boldsymbol{M}_m] \tag{7}$$

式中，$\boldsymbol{M}_m$是针对特定模态m优化的个性化线性映射矩阵，它由两个部分构成：

$$\boldsymbol{M}_m=\lambda\boldsymbol{M}_m^{\mathrm{loc}}+(1-\lambda)\boldsymbol{M}_m^{\mathrm{up}} \tag{8}$$

式中，$\boldsymbol{M}_m^{\mathrm{loc}}$为实现个性化线性嵌入而永久保留在参与方本地的模态$m$的线性映射矩阵，而$\boldsymbol{M}_m^{\mathrm{up}}$则被用作和中心交互聚合。这样设计的目的一方面是希望可以通过$\boldsymbol{M}_m^{\mathrm{up}}$以及中心的聚合和其他各个参与方传递并分享知识，从而使得模型获取全局知识；另一方面又通过$\boldsymbol{M}_m^{\mathrm{loc}}$来保留参与方$i$的个性化知识，避免完全被全局知识同化，这是导致联邦训练模型在个性化任务中表现下降的主要原因。二者之间的比

例由个性化超参数λ控制，其值满足：

$$0<\lambda<1 \tag{9}$$

为了保留模态特定信息并支持任务模型学习，这里将模态向量$\boldsymbol{v}_m$与每个映射后的数据片段$x_i\boldsymbol{M}_m$进行连接：

$$z_i=\text{Concatenate}(\boldsymbol{v}_m,x_i\boldsymbol{M}_m) \tag{10}$$

$$\boldsymbol{Z}=[z_1,z_2,\cdots,z_n] \tag{11}$$

在本文设计的训练框架中，线性映射矩阵$\boldsymbol{M}_m$是可学习参数，可以通过训练进行自动调整，以适应特定模态的线性映射。此外，模态向量$\boldsymbol{v}_m$可以使用固定方式编码，或者同样被设置为可学习参数。

如图 1 所示，线性嵌入的过程首先将不同模态数据（图像、文本和音频）划分为不同的片段。然后通过它们各自的面向模态的线性映射矩阵$\boldsymbol{M}_m$来处理这些片段。之后，每个片段的线性映射结果都与一个模态向量连接，该模态向量用以表征对应数据片段的模态信息。经过上述处理，原始不同模态数据被转化为统一的包含其数据模态信息的片段序列。

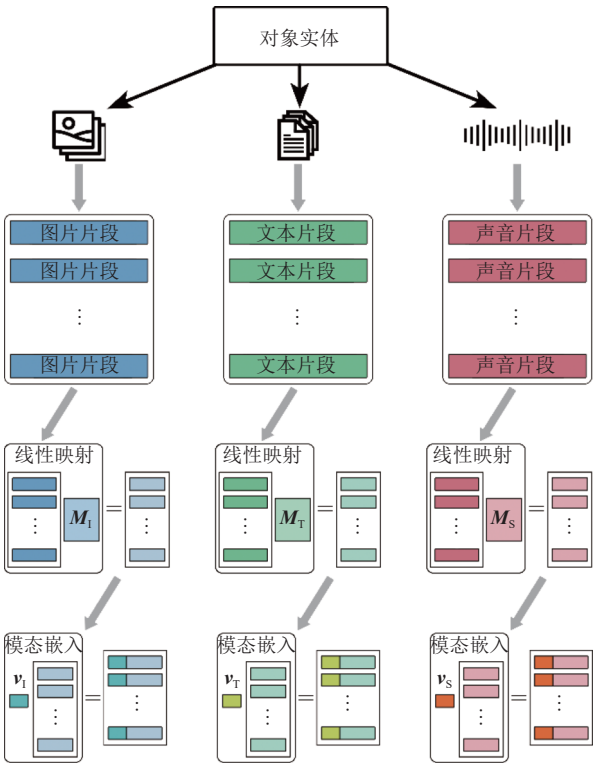

图 1 线性映射与模态嵌入

经过线性映射和模态嵌入的处理之后，后续的模型学习训练过程可以将上述过程的结果当作相同模态数据对待，且无须担心模型无法感知不同模态数据的差异。因为模态差异通过针对每种模态训练的线性映射矩阵和嵌入的模态信息向量所感知，此时，不同模态的数据对于后续模型训练的过程就类似于两种差异较大的单一模态数据。因此，后续可以使用单一模态的联邦学习进行处理。

1.2 单模态联邦平均

总体全局模型的训练可以通过联邦平均算法实现。具体来说，参与方使用各自线性映射和模态嵌入的结果单独训练他们的本地模型。经过一定次数的训练迭代后，本地训练暂停，并将训练得到的模型参数上传至中心同步。随后，中央服务器对从各个参与方收到的模型进行加权平均聚合，权重与每个参与方的贡献成比例。之后，将聚合得到的全局模型重新分发给各个参与方。这个迭代过程一直持续到训练收敛，最终得到多模态联邦学习全局模型。

在联邦平均中，参与方i在第t轮的局部模型训练涉及使用当前全局模型参数θ_t更新局部模型参数$\theta_{i,t}$，其更新方式如下：

$$\theta_{i,t}=\theta_{i,t-1}-\eta\nabla L_i(\theta_{i,t-1}) \tag{12}$$

式中，η表示学习率；$L_i(\theta_{i,t-1})$表示第$t-1$轮中参与方i的损失函数；$\nabla L_i(\theta_{i,t-1})$是损失函数相对于迭代（$t-1$）时模型参数的梯度。

全局模型参数θ_t的聚合在每轮训练结束时执行，更新的全局模型参数θ_{t+1}是所有本地更新的模型参数的加权平均值，即：

$$\theta_{t+1}=\sum_{i=1}^{N}w_i\theta_{i,t}/\sum_{i=1}^{N}w_i \tag{13}$$

式中，N表示参与同步的参与方的总数；w_i表示参与方i的模型在聚合过程中的对应权重。

训练过程的终止取决于每轮训练结束时是否满足预先设定的收敛条件。这些条件可能包括损失函数的收敛或全局模型参数的稳定。如果满足这些条件，则停止训练；否则，进入下一轮训练。

在提出的架构中，第t轮的联邦平均需要聚合的全局参数包括的每种模态的线性映射矩阵$\boldsymbol{M}_{j,g}^t$和任务模型参数W_g^t，其更新方式如下：

$$\boldsymbol{M}_{j,g}^t\leftarrow\sum_{i\in C_j}w_i\boldsymbol{M}_{j,i}^t \tag{14}$$

$$W_g^t\leftarrow\sum_{i=1}^{N}w_iW_i^t \tag{15}$$

$$\sum_{i=1}^{N}w_i=1 \tag{16}$$

式中，C_j表示拥有第j个数据模态的参与方集合；$\boldsymbol{M}_{j,i}^t$是由参与方i上传的第j个模态的本地映射矩阵；W_i^t是参与方i在第t轮次上传的本地任务模型参

数。然后，参与方使用中心返回的聚合参数 $\boldsymbol{M}_{j,g}^{t}$ 和 W_g^t 分别更新其本地线性映射矩阵和任务模型参数。

2　框架设计

在设计与个性化联邦学习协同工作的多模态 Transformer 架构时，需要考虑应对各种数据模态的复杂性，并满足各参与方的特定需求。本节详细描述本文提出的基于 Transformer 的多模态个性化联邦学习框架的设计，并强调通过整合个性化迁移方法来增强模型在全局训练之后的个性化任务性能。

根据上一节的分析，本文提出的框架需要首先对各种模态的数据进行分割，并对其进行线性映射，将分割后的数据转换为统一的向量表示形式。其次，利用包含模态信息的模态向量的线性映射的结果，并按原始序列排列形成处理后的数据片段序列，以便 Transformer 模型的后续处理。

需要注意的是，为应对不同模态间数据片段尺寸的不一致，本架构为不同模态设计了特定的映射矩阵大小来确保线性映射后结果向量的长度一致性。由式 (7) 可知，不同模态数据的线性映射矩阵的行数不同，而是与其模态数据片段的长度相同，同时这些线性矩阵的列数相同，这样可以保证不同数据模态的数据片段经过线性映射后的结果长度相同，进而便于在后续 Transformer 模型中统一进行处理。

在框架内的参数优化主要分为两大类。一是针对线性映射和模态嵌入参数的优化，至少涵盖了为各种数据模态定制的线性变换矩阵，同时也可能包括模态信息向量。这些参数不再依赖预设编码，而是可以通过自动学习算法进一步训练。二是基于 Transformer 的任务模型参数优化。

全局训练一个轮次结束后，参与方将其本地各模态线性映射矩阵、模态嵌入向量和 Transformer 任务模型参数上传至中央服务器进行汇总聚合。服务器运用联邦平均方法来计算全局的模态线性映射矩阵、模态嵌入向量和 Transformer 模型的全局参数。

各参与方（无论是物联网设备还是大型网络中的节点）都在本地数据上进行初始模型参数的训练，从而将特征与偏好融入模型中。全局模型训练完成后，每个参与方可以在全局模型的基础上对其特定的任务需求微调以得到局部模型。经过个性化调整的局部模型对比全局模型更精准地适应本地任务。

这里以 3 个参与方为例介绍多模态个性化联邦学习框架。如图 2 所示，3 个不同的参与方 A、B 和 C 各自拥有不同模态的数据，用不同的图形（三角形、正方形、圆形）来表示。每个参与方使用自己的数据在本地训练模型，生成局部参数，包含自身具有的数据模态的线性映射矩阵以及本地任务模型参数。这些参数代表了 3 个参与方本地模型从本地数据中独立学习到的知识，因为它们基于各自的数据训练得来。随后，各参与方将局部训练好的模型参数发送到中心服务器。中心服务器使用联邦平均算法聚合这些分布式生成的模型参数。聚合过程完成后，中心服务器得到了一个全局模型参数 $\boldsymbol{W}$ 和对应每种模态的全局线性映射矩阵，它们综合了所有参与方的局部知识。全局模型参数 $\boldsymbol{W}$ 和各个模态的全局线性映射矩阵 $\boldsymbol{M}_{\triangle/\square/\circ}$之后被发回给各个参与方，用于更新各参与方的本地模型和不同模态的线性映射矩阵，使得各个模型都能从其他参与方那里获得全局知识，增强各自模型的性能。这种迭代训练及微调过程一直进行到模型性能达到预设的终止条件，这通常需要在全局一致性与局部性能之间找到一个平衡点。在得到全局模型之后，每个参与方可以根据自身模态数据和特定任务对全局模型进行微调以获得更适合其任务的个性化多模态模型。

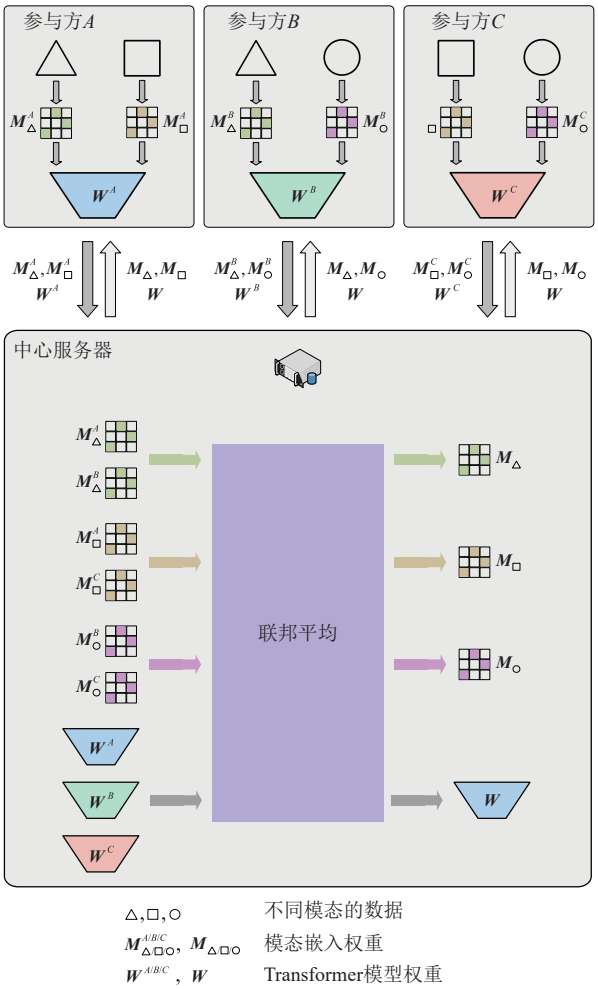

图 2　具有 3 个参与方的多模态个性化联邦学习框架

从根本上讲，个性化联邦学习策略融合了全局学习与本地学习的精华，既赋予了模型泛化能力，又保留了个性化的应对能力，提供了一个既强大又精确贴合个性化任务需求的模型。

3 实验与分析

本文对模型基于丰富的多模态数据进行了训练验证，这包括了文本、图像标题对以及交错的图像文本数据。使用了 Pile[11]（多个来源的庞大英语文本数据集，常用于训练大型语言模型）和 Common Crawl 快照[12]，后者包括了 CC-Stories 和 RealNews 数据库。对数据集进行了去重和过滤处理，移除了可能对下游任务造成干扰的近似重复和无关数据。

图像与标题对的数据集主要包括英文的 LAION-2B[13]、LAION-400M[14]、COYO-700M[15] 等数据集，这些数据集构成了图像–字幕对数据集的基础。这些数据是通过对 Common Crawl 的挖掘得到的，方式是挑选出其中带有相应替代文本的图像源。字幕生成是在 MS COCO Caption[16] 和 Flickr30k[17] 数据集上进行测试的，利用 COCO Karpathy 分割出测试集并且分析得到 Flickr30k。其中，所有图像都被标准化至 224×224 的像素分辨率，并且在进行字幕生成时采用大小为 5 的波束搜索技术，并从训练集中随机挑选样本进行少镜头（few-shot）场景的训练。

本文的模型架构由一个 12 层的 Transformer 网络构成，每层包含 2 048 个隐藏单元，并由大小为 4 096 的前馈网络（feedforward neural network, FNN）和 16 个多头注意力头部组成。模型参数采用 Magneto 方法进行初始化，以提高优化的稳定性。预训练的 CLIP ViT-L/14 模型，带有 1 024 维的图像特征，为图像的表示提供了强有力的支持，有助于模型快速收敛。训练过程中，图像大小被调整至 224×224 像素，并且除了最后一层之外，所有的 CLIP 模型参数都处于冻结状态。

训练过程经历了 30 000 次迭代，大约处理了 9 600 万个令牌。每个批次包含 32 000 个令牌，其中 14 000 个来自文本数据集，14 000 个源自图像标题对，而另外 4 000 个来自图像与文本结合的数据集。本文采用了 AdamW 优化器进行训练，其 β 参数设置为 0.90 和 0.98，并应用了 0.01 的权重衰减率和 0.10 的丢包率。学习率则是从 2×10^{-4} 的峰值在预热期间开始，随后线性递减至零。文本的标记化是通过 SentencePiece 在数据准备阶段进行的，以保障从一个或多个源中输入的序列能够维持完整的句子结构。

3.1 验证实验

按照设计的框架和上述参数进行训练，训练的损失和精度曲线如图 3 和图 4 所示，在最初的 10 个全局训练轮次内，观察到训练和验证损失指标明显且快速减少。这种明显的下降证明了模型加速适应训练数据集的复杂性。当经过第 20 个全局训练通信轮次之后，发现描述损失的曲线趋于平缓，这表明模型在学习方面已经成熟，并且接近最佳性能状态，多模态全局模型的性能开始收敛。

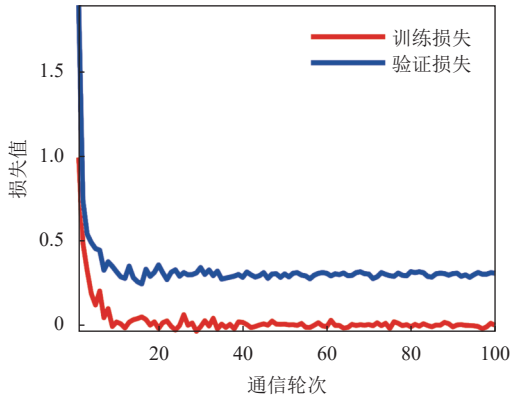

图 3 训练过程损失值

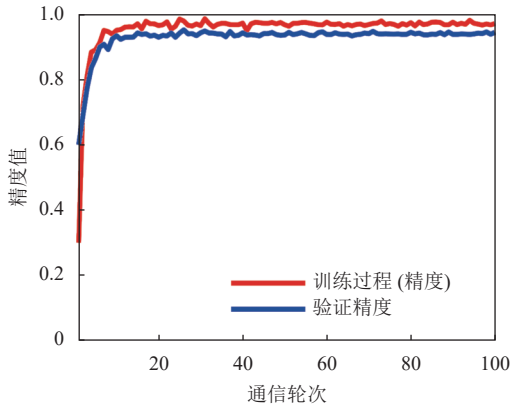

图 4 训练过程精度值

在训练损失减少的同时，模型的准确性也在显著地提高，这在训练和验证场景的图形曲线上都得到了明确的量化。在前 10 个通信轮次内，准确性出现了突出的飞跃，这表明模型对数据特征和标签的快速同化。经过最初的显著增强后，精度轨迹进入稳定阶段，在第 20 个通信轮次之后小幅波动变得明显。这表明该模型随着时间的推移实现了稳定可靠的预测准确性。

模型性能的表现如图 5 所示。该图提供了相较

于本地多模态学习模型的比较，说明本文提出的模型在联邦学习框架下相比于本地多模态基线模型更加有效。该图的横轴列举了参与 FL 过程的参与方，而纵轴则代表本文模型的测试性能相对于基线模型的提升幅度。不同参与方的性能改进存在明显差异，本文提出的模型的性能比基线模型的性能至少高出 15%，在一些情况下高出 25% 以上。本文提出的模型提供的平均性能提升为 20% 以上。

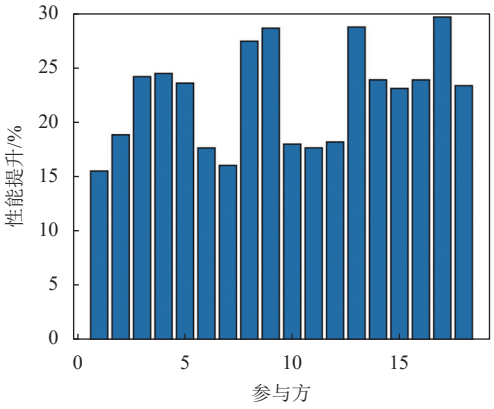

图 5　多模态个性化联邦模型性能

这组实证研究验证了基于 Transformer 的多模态个性化联邦学习可以显著提高多模态学习系统的性能。这种性能提升归因于大量用户的参与而带来的知识的增加。因此，相比于单个参与方内部孤立的、可能过度拟合的训练数据，本文提出的模型显现出明显的优势。

此外，联邦学习策略整合了一个不仅更大而且更多样化的数据集，有效地整合了广泛的数据变化，为模型提供了更全面、更广泛适用的训练机制。因此，与传统的方法相比，后者可能会受到数据范围狭窄的影响，可能导致模型过度拟合且无法泛化。而前者通过利用更多样化的数据，联邦学习模型更有望提供更加细致和适应性更强的模型性能，更好地反映现实场景的复杂性。

总之，从本节实验中发现多模态个性化联邦学习方法显著提高了相较于本地训练模型的性能，它能够利用多模态数据集的复杂性和多样性，并针对参与方个性化提供更优的模型性能。

3.2　超参数实验

本节验证用于线性映射矩阵个性化超参数λ对所提出算法的影响。实验的数据配置分为独立同分布（independent and identically distributed, IID）设置和非独立同分布（non-independent and identically distributed, Non-IID）设置。实验中令λ取值 0.1~0.9，间隔为 0.2，观察所有参与方平均模型性能的变化比值，结果如图 6 所示。

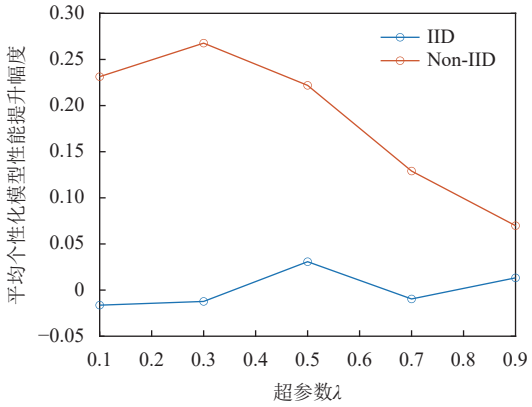

图 6　超参数λ对模型性能的影响

超参数λ用于调控个性化程度，其取值范围为 0~1，其中$\lambda=1$对应于完全的本地训练，而 $\lambda=0$ 对应于纯粹的全局训练。图中折线分别表示 IID 数据和 Non-IID 数据下模型相对于本地训练模型的性能提升幅度。

从图中可以观察到，随着个性化程度的增加（λ 取值由 0.1 增加到 0.9），在 IID 数据情形下，模型性能的变化量相对平稳，围绕着零线波动，这表明在 IID 数据分布下，个性化程度对模型性能的影响较小。

相反，在 Non-IID 数据情形下，随着λ取值的增加，性能改变量呈明显的下降趋势。当λ取值为 0.3 时，Non-IID 数据下的性能改善量达到峰值，随后随着 λ 增加而快速下降。这表示在个性化程度较低（即更接近全局训练）时，Non-IID 数据下模型性能提升逐渐增加。但是，随着个性化程度的继续增强，模型性能开始下降，表明在 Non-IID 数据分布的情况下，适度的个性化有助于提升性能，而过度的个性化则可能导致性能下降。

3.3　对比实验

本节进行对比试验，验证相较于其他联邦多模态方法本方案的性能表现。图 7 展现了以本文的架构使用 PFL 训练的多模态 Transformer 模型相较于 4 种现有方法的全面性能对比。X 轴代表不同的参与方，而 Y 轴则定量展示了本文方法在每一对比实验中的性能提升百分比。

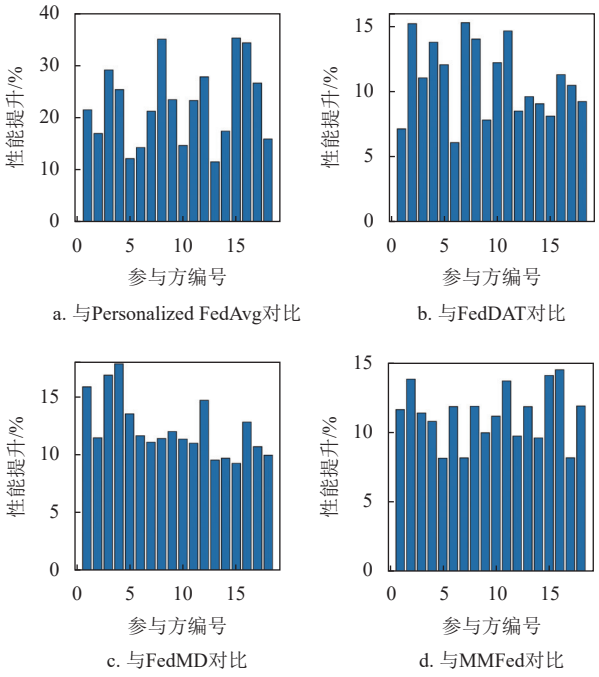

图 7 与其他方案对比性能提升

Personalized FedAvg[18] 通过为每一模态独立训练联邦平均模型，再针对不同模态的数据在参与方上进行精细微调，以此运作。观察图 7a，可以看到本文方法实现了大约 10% 至 35% 的显著性能提升。这一显著的提升，得益于本文的框架可以同时利用多个模态对模型进行训练，而 Personalized FedAvg 只能从单一模态训练再对单模态模型的结果进行再次整合，这一过程中如果出现某一对象的模态缺失则无法进行。因此本文方法具有更高的数据知识利用率。

FedDAT[7] 致力于实现高效的参数优化和参与方间的知识转移，以促进不同视觉语言任务的联合微调。如图 7b 所示，本文方法始终保持了较好的性能，增益率介于 5% 至 15%。本文模型的优异性能，可归因于对数据进行分段以及建立线性映射，这些策略产生了更为精细的特定模态嵌入，并通过 Transformer 集成进行了高效处理，实现了比 FedDAT 更为高效的知识整合与传递。

FedMD[19] 通过调整模型输出的 Logits 来实现个性化学习。如图 7c 所示，本文方法相对于 FedMD 的性能提升 5% 至 15%，展示了其性能优势。这些性能优势源自本文方法中精细的参数优化路径，它涵盖了线性映射、模态嵌入和任务相关细节的自适应调整。这些调整超越了 Logits 对齐带来的效果，它们在表征多模态数据的同时，更有效地捕捉了多模态数据的多样性和复杂性，且不需要像 FedMD 那样利用外部数据集——这对于多模态场景通常是难以满足的。

MMFed[20] 整合了多模态特征融合的注意力机制，并允许参与方间的知识共享。从图 7d 可以看到，本文的架构相较于 MMFed 为各参与方提升了约 5% 至 15% 的性能。本文方法的成功关键在于集成了定制化的线性映射矩阵以及模态信息向量，相较于 MMFed 的方法，提供了更为丰富和更具上下文感知的特征融合及知识交换效率。

总之，实验验证了本文提出的基于 Transformer 的多模态个性化联邦学习框架的稳定性和健壮性，该框架量身定制以高效地服务于不同参与方，无论是拥有丰富多模态数据的用户，还是那些数据资源受限于单一模态的用户。框架的设计理念在于构筑一个协作共赢的生态系统，使得所有参与方能互相分享各自不同模态数据的知识，同时不牺牲个体数据的隐私性和个性化需求。

此外，即使在参与方贡献的数据分布不均匀时，本文方法也能够跨不同模态保持稳定的测试准确度，这种性能凸显了该框架利用多模态数据的潜力。对比实验结果进一步印证了本文框架的有效性，证明了它能够让拥有多模态数据的参与方成为学习改善的核心动力，从而提升所有参与方的学习效果，包括那些仅具备单模态数据的参与方。这种知识的传递是联邦学习体系的核心，尤其是在处理分散和异质数据特征的场景中尤为关键。

4 结束语

本研究致力于解决多模态个性化联邦学习中固有的复杂性问题，特别关注不同参与方间数据模态的差异性。在这些情景下，传统的联邦学习范式可能会遇到困难。为了克服这一问题，本文设计了一个创新的框架，它利用 Transformer 模型对多模态数据的处理优势，以促进高效的多模态个性化联邦学习。

利用 Transformer 在多模态数据的特征提取和表示方面的强大能力，本文框架能够高效地处理异构模态数据，增强联邦训练的效果。本文方法使用线性映射和模态嵌入的方法整合多源异构模态数据，从而实现了模型性能的显著提升。本文合并不同的数据模态不仅提高了准确度，而且还为构建更健壮的机器学习模型进行了有效探索，当存在模态无法对齐或模态缺失时，本文方法优于标准的单模态个性化联邦学习和孤立的多模态学习模型，扩展了多模态个性化联邦学习的适用场景。

参 考 文 献

[1] XU P, ZHU X T, CLIFTON D A. Multimodal learning with transformers: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113-12132.

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010.

[3] HAN K, XIAO A, WU E, et al. Transformer in transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.

[4] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. [2024-03-01]. https://doi.org/10.48550/arXiv.2010.11929.

[5] CAO X J, LI Z H, SUN G, et al. Cross-silo heterogeneous model federated multitask learning[J]. Knowledge-Based Systems, 2023, 265: 110347.

[6] LI R, MA F L, JIANG W J, et al. Online federated multitask learning[C]//Proceedings of the IEEE International Conference on Big Data. New York: IEEE, 2019: 215-220.

[7] CHEN H K, ZHANG Y, KROMPASS D, et al. FedDAT: An approach for foundation model finetuning in multi-modal heterogeneous federated learning[EB/OL]. (2023-08-21)[2024-02-12]. https://arxiv.org/abs/2308.12305.

[8] CHE L W, WANG J Q, ZHOU Y, et al. Multimodal federated learning: A survey[J]. Sensors, 2023, 23(15): 6986.

[9] CAO X J, SUN G, YU H F, et al. PerFED-GAN: Personalized federated learning via generative adversarial networks[J]. IEEE Internet of Things Journal, 2023, 10(5): 3749-3762.

[10] LI Q B, WEN Z Y, WU Z M, et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(4): 3347-3366.

[11] GAO L, BIDERMAN S, BLACK S, et al. The pile: An 800GB dataset of diverse text for language modeling [EB/OL]. [2024-02-12]. https://ui.adsabs.harvard.edu/abs/2021arXiv210100027G/abstract.

[12] BERGSMA S, PITLER E, LIN D K. Creating robust supervised classifiers via web-scale n-gram data[C]// Proceedings of the Annual Meeting of the Association for Computational Linguistics. Uppsala: Association for Computational Linguistics, 2010: 865-874.

[13] SCHUHMANN C, BEAUMONT R, VENCU R, et al. LAION-5B: An open large-scale dataset for training next generation image-text models[J]. Advances in Neural Information Processing Systems, 2022, 35: 25278-25294.

[14] SCHUHMANN C, VENCU R, BEAUMONT R, et al. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs[EB/OL]. (2021-11-03)[2024-02-12]. https://arxiv.org/abs/2111.02114.

[15] BYEON M, PARK B, KIM H, et al. Coyo-700m: Image-text pair dataset[EB/OL]. (2015-04-01)[2024-02-12]. https://arxiv.org/abs/1504.00325.

[16] CHEN X L, FANG H, LIN T Y, et al. Microsoft COCO captions: Data collection and evaluation server[EB/OL]. (2015-04-01)[2024-02-12]. https://arxiv.org/abs/1504.00325.

[17] PLUMMER B A, WANG L W, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models [C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2015: 2641-2649.

[18] FALLAH A, MOKHTARI A, OZDAGLAR A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach[EB/OL]. [2024-02-12]. https://arxiv.org/abs/2002.07948.

[19] LI D, WANG J. FedMd: Heterogenous federated learning via model distillation[EB/OL]. [2024-03-01]. https://doi.org/10.48550/arXiv.1910.03581.

[20] XIONG B C, YANG X S, QI F, et al. A unified framework for multi-modal federated learning[J]. Neurocomputing, 2022, 480: 110-118.

编　辑　张　莉