中国科学: 信息科学 2023 年 第 53 卷 第 11 期: 2180–2201
SCIENTIA SINICA Informationis
论文

联邦无监督跨模态哈希

朱磊[1,3*], 李京智[1], 王天时[1], 李晶晶[2], 张化祥[1*]

1. 山东师范大学信息科学与工程学院, 济南 250358
2. 电子科技大学计算机科学与工程学院, 成都 611731
3. 鹏城实验室, 深圳 518055
* 通信作者. E-mail: leizhu0608@gmail.com, huaxzhang@sdnu.edu.cn

收稿日期: 2022–09–22; 修回日期: 2022–12–08; 接受日期: 2023–04–06; 网络出版日期: 2023–11–07

国家自然科学基金 (批准号: 62172263)、山东省自然科学基金 (批准号: ZR2020YQ47, ZR2019QF002)、山东省高等学校青年创新团队基金 (批准号: 2019KJN040) 和泰山学者工程 (批准号: ts20190924) 资助项目

摘要 联邦跨模态检索利用分散的客户端学习一个共享跨模态检索模型, 从而降低集中大规模多模态训练数据时高昂的维护成本, 解决分布式数据存储场景下跨模态检索中存在的数据隐私问题. 然而现有的联邦跨模态检索方法大多依赖于大量的语义标注, 这限制了检索模型在大规模应用场景下的扩展性. 与之不同, 本文提出一种无监督的联邦跨模态哈希检索模型, 旨在保护客户端数据隐私的前提下, 学习不依赖语义标注的跨模态检索模型. 由于联邦环境中多模态数据分布不平衡, 局部信息不足以让模型学习到整体数据上的模态间相似性, 从而影响检索性能. 为解决此问题, 本文提出一个全局 – 局部模态间对比正则化方法, 通过使用不同模态的全局哈希模型对单个模态的局部哈希模型进行约束, 使局部哈希模型能够充分感知整体数据的相似性语义, 从而加强对本地跨模态哈希学习过程的引导. 同时, 本文引入一种全局 – 局部模态内知识蒸馏策略来进一步地获取模态内特有的全局知识. 5 个基准跨模态检索数据集上的实验结果验证了本文提出方法的有效性.

关键词 联邦学习, 多模态学习, 无监督学习, 跨模态检索, 无监督跨模态哈希

1 引言

随着网络上文本、图像和视频等数据的爆炸式增长, 人们对在海量多模态数据中进行检索的需求日益增加. 在这其中, 由于人们在进行检索时往往以一种模态的数据检索另一种模态的数据 [1,2], 因此如何能在多模态数据中进行高效的跨模态检索引起了学术界和工业界的广泛关注. 近年来, 深度跨模态哈希 (Hashing) [3] 利用深度神经网络将不同模态数据映射为统一且紧凑的二值哈希码, 显著提高语义相似度计算效率并降低了多模态数据的存储成本, 在大规模跨模态检索上取得了巨大的成功. 然而, 深度跨模态哈希模型的训练通常需要大量标注的多模态数据. 受数据安全与隐私保护相关法律法

引用格式: 朱磊, 李京智, 王天时, 等. 联邦无监督跨模态哈希. 中国科学: 信息科学, 2023, 53: 2180–2201, doi: 10.1360/SSI-2022-0366
Zhu L, Li J Z, Wang T S, et al. Federated unsupervised cross-modal Hashing (in Chinese). Sci Sin Inform, 2023, 53: 2180–2201, doi: 10.1360/SSI-2022-0366

规的限制, 集中大量多模态数据会导致巨大的隐私泄露风险. 此外, 在实际分布式环境下对多模态数据进行统一协调的人工标注是费力耗时的, 这无疑也增加了训练成本. 因此, 如何利用大量的分布式存储的无标注多模态数据进行隐私保护的跨模态哈希学习是非常重要的问题.

近年来, 随着社会各界对数据隐私保护关注度的不断上升, 联邦学习 [4] 应运而生. 作为一种新兴的分布式隐私保护机器学习方法, 联邦学习可以在不暴露每个参与方数据的前提下学习共享模型, 从而有效避免数据隐私泄漏. 据笔者所知, 当前针对联邦跨模态检索的研究工作较少. 文献 [5] 融合了有监督跨模态检索方法和联邦学习框架. 但是联邦有监督跨模态检索方法严重依赖客户端中多模态数据中的监督信息, 难以适用于实际应用中客户端标注信息不足的情况. 同时, 基于实值特征的跨模态检索模型降低了语义相似度计算的效率. 为了解决上述问题, 本文重点探究无监督环境下的联邦跨模态哈希检索, 实现高效的跨模态检索.

本文提出一种新的联邦无监督跨模态哈希框架 (federated unsupervised cross-modal Hashing, FedUCH), 用于在分布式数据存储下学习无监督跨模态哈希检索模型, 同时保护数据隐私. 无监督跨模态哈希学习主要从不同模态数据出发, 学习不同模态数据间的相似性. 但在联邦环境下, 客户端仅依靠挖掘本地数据的模态间相似性作为模型训练引导. 客户端的本地训练容易受到客户端之间多模态数据分布不平衡的影响, 这会使得局部跨模态哈希模型无法感知不同模态在整体数据上的相似性, 对检索性能产生不利影响. 受到现有工作 [4,6~8] 的启发, 本文提出全局 – 局部模态间对比正则化, 通过使用不同模态的全局哈希模型对单个模态的局部哈希模型进行约束, 让局部哈希模型能够充分学习到整体多模态数据知识, 从而加强对本地跨模态哈希学习过程的引导. 此外, 不同模态的局部哈希模型仅学习到了局部数据和全局模型中的跨模态知识, 并没有学习到全局的模态内知识. 这样训练出的模型可能会过度拟合局部数据, 从而使得模型聚合的过程变得困难 [9]. 为解决此问题, 本文还提出一个全局 – 局部模态内知识蒸馏策略, 进一步地获取全局模态内知识. 本文的主要贡献如下:

(1) 本文提出一种新的联邦无监督跨模态哈希学习模型, 这是在联邦环境下进行无监督跨模态哈希学习的第 1 篇工作. 该模型可以在保护数据隐私的条件下有效地进行无监督跨模态哈希检索模型的学习.

(2) 本文提出利用全局模型产生的表示在不同模态间对本地训练进行对比正则化, 并在同一模态内进行知识蒸馏, 从而在模态内与模态间规范本地模型的训练, 缓解了多模态数据不平衡性带来的局部模型偏差. 对比实验和消融实验验证了本文提出方法的有效性.

本文的其余部分组织如下. 在第 2 节中, 回顾了跨模态哈希检索和联邦学习的相关工作. 在第 3 节中, 介绍了提出的联邦无监督跨模态哈希检索模型 FedUCH. 在第 4 节中, 进行相关实验并展示实验结果. 在第 5 节中, 总结全文.

2 相关工作

2.1 跨模态哈希

跨模态哈希旨在将不同模态数据映射为统一且紧凑的二值哈希码 [10], 并根据哈希码之间的汉明距离 (Hamming distance) 进行高效检索. 现有的跨模态哈希方法大致可分为有监督和无监督两类, 同时这两类方法又可进一步分为基于浅层机器学习和深度学习的方法.

早期的有监督跨模态哈希方法基于标签信息探索不同模态在汉明空间中的语义相关性, 例如: 语义保持哈希 (semantics-preserving Hashing, SePH) [11]、监督矩阵分解哈希 (supervised matrix factorization

Hashing, SMFH) [12] 等. 这些浅层方法无法捕获不同模态之间的深层语义关联, 因此在检索性能上难以优于基于深度学习的检索方法. 借助于深度神经网络强大的表征能力, 深度跨模态哈希 (deep cross-modal Hashing, DCMH) [3] 和深度视觉语义哈希 (deep visual-semantic Hashing, DVSH) [13] 等方法能够实现准确高效的跨模态检索.

无监督跨模态哈希旨在无需语义标注的前提下保持多模态数据模态内与模态间的语义相关性. 在基于浅层机器学习的无监督方法中, Ding 等 [14] 提出协同矩阵分解哈希 (collective matrix factorization Hashing, CMFH), 通过潜在因子的协同矩阵分解学习公共的汉明空间. Liu 等 [15] 提出融合相似性哈希 (fusion similarity Hashing, FSH), 将基于图的模态间融合相似性嵌入到汉明空间中学习哈希码. 最近的无监督方法借助深度表征, 可以更精确地描述多模态数据的模态间相关性, 并在语义对齐方面表现良好. 无监督深度跨模态哈希 (unsupervised deep cross-modal Hashing, UDCMH) [16] 结合深度学习与矩阵分解, 将特征映射与哈希码学习进行联合优化. 深度联合语义重构哈希 (deep joint-semantics reconstructing Hashing, DJSRH) [17] 通过哈希码学习来最大程度地重构多模态联合语义结构, 从而挖掘实例之间潜在的语义关联. 基于联合模态分布的相似性哈希 (joint-modal distribution-based similarity Hashing, JDSH) [18] 通过构建保持跨模态语义相关性的多模态联合相似性矩阵, 更好地缓解了多模态数据的语义鸿沟. Yu 等 [19] 提出深度图近邻一致性保持网络 (deep graph-neighbor coherence preserving network, DGCPN), 通过图学习探索数据及其近邻之间的语义关系, 解决了多模态数据模态间的相似性不准确问题. 然而, 这些方法需要聚集大量数据来训练深度神经网络以捕获模态间的相似性, 这会带来潜在的数据泄露风险并且不适用于实际应用中数据量较少的情况. 在本文中, 我们基于现有的无监督深度跨模态哈希方法, 设计了一种新的联邦无监督跨模态哈希框架, 其能够在数据无需聚集的分布式存储场景下, 对每个客户端进行隐私保护的联邦无监督跨模态哈希学习.

2.2 联邦学习

联邦学习是谷歌 (Google) 于 2016 年提出的一种保护数据隐私的分布式机器学习技术 [20], 它使多方客户端能够在服务器协调下共同训练机器学习模型, 并保持数据始终存储在客户端, 从而保护数据隐私. 近年来, 客户端之间数据分布的不平衡性是联邦学习的关键挑战之一 [4, 7, 8, 21, 22], 目前大多数工作都是基于联邦平均算法 (federated averaging algorithm, FedAvg) [4] 并对其进行改进来缓解客户端之间数据分布不平衡的问题. Li 等 [7] 提出 FedProx 框架, 在客户端更新时引入针对局部模型优化的修正项, 从而提高模型收敛的稳定性; Wang 等 [8] 提出一种标准化的联邦平均方法 FedNova, 其充分考虑每个客户端局部模型对全局模型的贡献, 从而优化服务器对客户端局部模型的聚合过程. 此外, 一些学者认为应该为客户端建立个性化模型来解决数据的不平衡问题, 例如基于元学习 [23] 和多任务学习 [24] 的模型等. 但是, 这些方法都假设客户端的数据拥有标注完善的类别标签, 难以应用于大规模无标注数据的实际场景. 最近, 一些无监督联邦学习框架 [25, 26] 也被提出, 但这些方法主要关注于数据的通用表征学习, 未能充分考虑到多模态数据间的语义相关性, 因此联邦学习后的客户端模型难以实现高效准确的相似性近邻搜索.

近期, 有关多模态联邦学习的工作也取得了一定的进展. Liu 等 [27] 通过将本地多模态数据的表示上传到服务器进行联邦学习. Xiong 等 [28] 提出基于个性化的联邦学习并通过共同注意力 (co-attention) 机制融合本地不同模态的互补信息. Zhao 等 [29] 通过训练自动编码器从客户端的不同模态数据学习模态匹配信息, 能够支持在单模态和多模态数据上训练全局模型. Chen 等 [30] 针对客户端中模态不一致问题, 使用动态和多视图的图结构自适应地捕获多模态客户端模型之间的相关性. Chen 等 [31] 提出层次梯度混合 (hierarchical gradient blending, HGB), 自适应度量局部模型的过拟合和泛化行为, 同时计

算模态的最佳融合和局部模型的最佳聚合权重. 这些多模态联邦学习的方法中存在共享数据表示导致隐私泄露, 中央服务器计算过程复杂等问题. 此外, 这些方法均针对于有监督或半监督的分类任务, 需要客户端对数据进行耗时费力的统一标注, 不适用于无监督场景下的多模态联邦学习.

在信息检索与联邦学习的结合上, 目前已有一些工作取得相当可观的成果. Xu 等 [32] 提出联邦患者哈希 (federated patient Hashing, FPH) 实现跨机构的医疗合作, 其分别在客户 – 服务器架构和对等网络架构两种联邦设置中, 通过构建相似性保留损失和异构性挖掘损失保留模态内和模态间的语义关联, 同时证明了 FPH 在跨机构患者匹配中的有效性. 联邦跨模态检索 (federated cross-modal retrieval, FedCMR) [5] 首次将深度跨模态检索与联邦学习结合, 从而避免数据分布式存储场景下跨模态检索模型训练可能遇到的数据隐私问题, 降低聚集大规模训练数据时高额的维护成本. 最近, Yang 等 [33] 引入全局原型哈希码来保持与客户端局部哈希码的语义一致性, 从而促进服务器端的全局模型聚合. 然而, 上述联邦检索方法均依赖于客户端中多模态数据的监督信息, 难以适用于实际应用中数据标注信息不足的情况.

因此, 针对目前多源无监督数据下跨模态检索的研究缺陷, 本文提出一种联邦无监督跨模态哈希模型进行隐私保护下的分布式跨模态哈希学习. 同时, 利用全局 – 局部模态间对比正则化和模态内知识蒸馏, 在模态内与模态间对本地跨模态哈希模型进行约束, 加强对本地跨模态哈希学习过程的引导, 缓解不同客户端之间多模态数据不平衡的问题.

3 方法

3.1 系统概述

本小节以图像和文本两种模态为例介绍本文提出的联邦无监督跨模态哈希模型 (FedUCH), 应当注意的是 FedUCH 也能够被用于任意多模态数据的联邦无监督跨模态哈希检索. 本文提出的 FedUCH 遵循了标准的横向联邦学习架构, 包括一个云服务器和多个客户端, 具体系统架构如图 1 所示. FedUCH 中每个客户端分别拥有部分多模态数据, 它们均将多模态数据存储在本地且不与其他客户端进行数据传输, 从而保护自身的多模态数据隐私.

在联邦无监督跨模态哈希学习开始时, 云服务器首先初始化跨模态哈希模型并将其下发给每个客户端. 之后对于每一轮次的联邦学习, 服务器与客户端均执行如下联邦优化过程:

• 客户端跨模态哈希学习. 首先, 每个客户端 k ($k = 1, 2, \ldots, K$) 分别利用特征提取器获取本地多模态数据的特征表示, 这里假设每个客户端的特征提取器均相同. 然后, 通过无监督跨模态哈希方法挖掘本地多模态数据间的语义相似性并构建关联矩阵作为引导信息. 接着, 图像哈希模型 ImgNet_k^L 和文本哈希模型 TxtNet_k^L 将多模态数据特征映射为哈希码. 最后, 通过对两种模态产生的哈希码进行相似性度量来计算损失函数, 从而达到对哈希模型的优化.

• 服务器模型聚合. 每个客户端首先将跨模态检索模型中的图像哈希模型 ImgNet_k^L 的参数 Θ_k^V 和文本哈希模型 TxtNet_k^L 的参数 Θ_k^T 进行加密, 然后将加密后的哈希模型参数上传到服务器. 云服务器首先解密跨模态哈希模型的参数, 然后根据 FedAvg [4] 算法对来自各个客户端的图像和文本哈希模型分别进行聚合, 从而获得全局哈希模型 ImgNet^G, TxtNet^G.

• 客户端模型更新. 云服务器加密全局跨模态哈希模型后将其分发给各个客户端模型, 各个客户端根据解密后的全局跨模态哈希模型和本地已有的局部跨模态哈希模型生成适用于本地多模态数据的哈希模型.

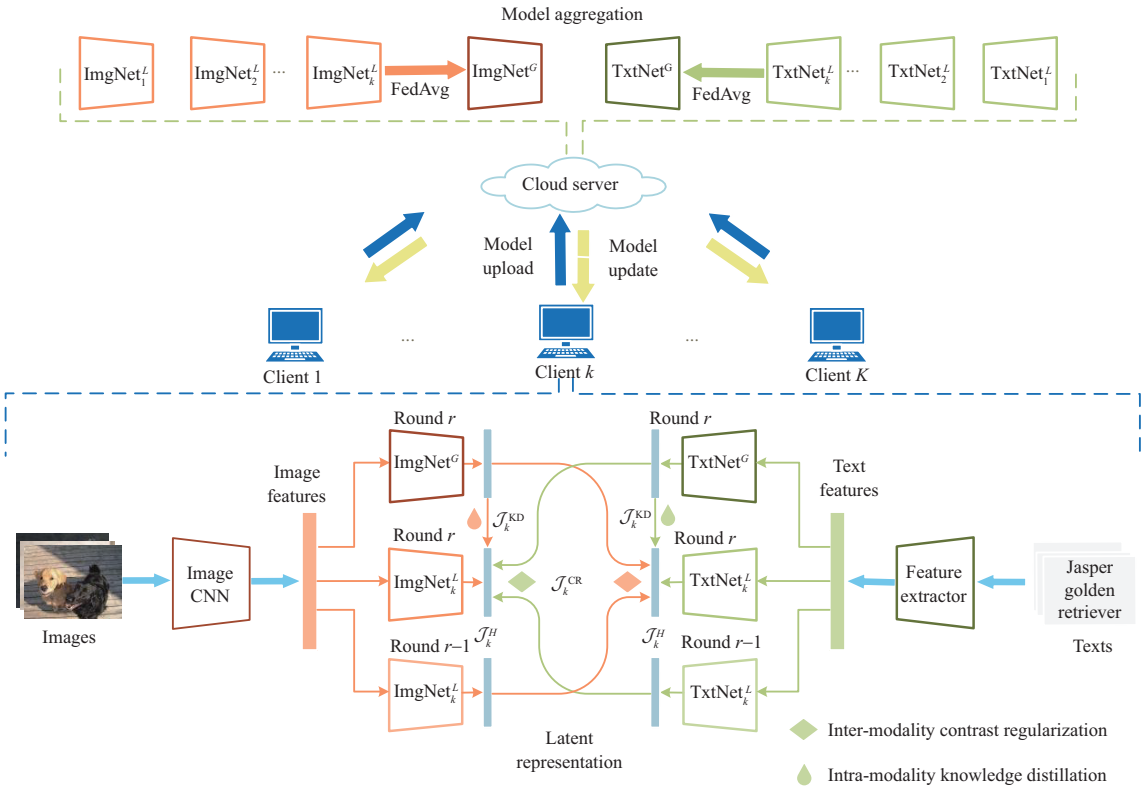

图 1　(网络版彩图) FedUCH 模型结构

Figure 1　(Color online) Model structure of FedUCH

上述联邦优化过程中, 假设联邦跨模态哈希学习过程中有 K 个客户端参与, 则 FedUCH 的总体目标函数被表示为

$$\min_{\Theta^*} \sum_{k=1}^{K} p_k \mathbb{E}_{x^* \sim D^k}[\mathcal{J}_k(\Theta^*; x^*)], \quad \text{s.t.} \quad * \in \{V, T\}, \tag{1}$$

其中 p_k 表示第 k 个客户端参与联邦优化过程时自身模型对全局模型的贡献大小, p_k 一般设置为与客户端本地数据的大小成比例, $\sum_{k=1}^{K} p_k = 1$; $\mathbb{E}_{x^* \sim D^k}[\mathcal{J}_k(\Theta^*; x^*)]$ 表示第 k 个客户端的经验损失, x^* 表示从客户端数据集 D^k 中获取的样本. 对于联邦优化过程, 本文利用基于客户端本地数据的局部目标函数替代总体目标函数, 从而减少数据传输的通信开销并保护客户端本地的数据隐私 [7]. 即在上述过程中, 对于第 r 轮次通信, 客户端接收云服务器加密分发的全局模型参数 Θ_G^V 和 Θ_G^T, 经解密后作为客户端的本地模型参数 Θ_k^V 和 Θ_k^T 并执行 E $(E \geqslant 1)$ 轮梯度下降算法进行优化 [4]:

$$\Theta_k^{*,r+(i)} = \Theta_k^{*,r+(i-1)} - \eta \bigtriangledown \mathcal{J}_k(\Theta_k^{*,r+(i-1)}), \tag{2}$$

其中 η 表示客户端更新本地模型的学习率, $i = 1, 2, \ldots, E$.

在上述客户端模型更新过程中, 客户端通过挖掘本地多模态数据的语义相似性来学习哈希码. 然而, 客户端的本地数据难以刻画整体多模态数据分布, 因此哈希模型在客户端本地训练时受到数据分布不平衡的影响, 最终降低了哈希模型的跨模态检索性能. 为了解决此问题, 我们在客户端中提出了全局 – 局部模态间对比正则化, 使客户端的哈希模型能够充分捕获整体多模态数据间的语义相似性, 并

利用整体多模态数据间的语义相似性加强对客户端本地哈希学习过程的引导. 同时, 我们引入全局 – 局部模态内知识蒸馏进一步获取整体数据中各模态的模态特有知识. 如算法 1 所示, 我们总结了联邦无监督跨模态哈希的整个过程. 下文分别从云服务器和客户端方面来详细介绍本文所提模型.

Algorithm 1 FedUCH

Input: The number of communication rounds R; the number of local epochs E; the set of clients $C = \{C_1, C_2, \ldots, C_K\}$; the set of dataset $D = \{D^1, D^2, \ldots, D^K\}$; the hyper-parameters μ, φ.

Output: Global image Hashing model ImgNet^G, global text Hashing model TxtNet^G.

Server executes:
Initialize $\text{ImgNet}(\cdot, \Theta_G^V)$, $\text{TxtNet}(\cdot, \Theta_G^T)$;
For each round $r = 1, 2, \ldots, R$ **do**
  For each client $k = 1, 2, \ldots, K$ **in parallel do**
    Send the global model $\text{ImgNet}(\cdot, \Theta_G^{V,r})$, $\text{TxtNet}(\cdot, \Theta_G^{T,r})$ to C_k;
    $\Theta_k^{V,r+1}$, $\Theta_k^{T,r+1} \Leftarrow \text{ClientUpdate}(k, \Theta_G^{V,r}, \Theta_G^{T,r})$;
  $\Theta_G^{V,r+1}$, $\Theta_G^{T,r+1} \Leftarrow$ Models aggregation with Eq. (3);
Return $\Theta_G^{V,R}$, $\Theta_G^{T,R}$;
ClientUpdate$(k, \Theta^{V,r}, \Theta^{T,r})$**:**
$\Theta_k^{V,r} \Leftarrow \Theta^{V,r}$, $\Theta_k^{T,r} \Leftarrow \Theta^{T,r}$;
For each local epoch $i = 1, 2, \ldots, E$ **do**
  For each batch b **of** D^k **do**
    Calculates the local corss-modal Hashing loss $\mathcal{J}_k^H$ with Eq. (4);
    Calculates the global-local inter-modality constrastive regularization loss $\mathcal{J}_k^{\text{CR}}$ with Eq. (8);
    Calculates the global-local intra-modality knowledge distillation loss $\mathcal{J}_k^{\text{KD}}$ with Eq. (9);
    $\mathcal{J}_k \Leftarrow \mathcal{J}_k^H + \mu\mathcal{J}_k^{\text{CR}} + \varphi\mathcal{J}_k^{\text{KD}}$;
    Update $\Theta_k^{V,r+1}$, $\Theta_k^{T,r+1}$ using $\mathcal{J}_k$;
Return $\Theta_k^{V,r+1}$ and $\Theta_k^{T,r+1}$ to server.

3.2 服务器端设计

在每一轮次的联邦学习中, 客户端自主选择是否参与该轮联邦优化过程, 云服务器负责统筹协调各客户端上传的局部模型. 为使本文描述更加简洁, 我们假设 FedUCH 中所有客户端均参与每一轮联邦优化过程, 而云服务器则通过聚合所有局部模型生成新的全局模型指导下一轮联邦优化过程. 需要注意的是, FedUCH 同样适用于联邦优化过程中客户端数量随机变动的场景. 由于构建客户端局部模型的深度神经网络往往架构繁琐且参数量庞大, 因此 FedUCH 在模型聚合阶段将多个模态的公共子空间部分在客户端之间进行模型参数共享 [5], 即仅对跨模态哈希函数模块进行聚合, 而不再考虑局部模型中结构复杂的特征映射模块, 从而降低联邦优化过程中的通信开销和模型聚合消耗. 在第 r 轮联邦优化结束后, 所有客户端将本地的局部哈希模型参数加密上传到云服务器, 之后云服务器利用 FedAvg 算法 [4] 聚合所有局部模型, 从而将聚合后的全局哈希模型加密分发给每个客户端, 其中服务器端的模型聚合过程被定义为

$$\begin{cases} \Theta_G^{V,r+1} = \displaystyle\sum_{k=1}^{K} \frac{n_k}{n} \Theta_k^{V,r}, \\ \Theta_G^{T,r+1} = \displaystyle\sum_{k=1}^{K} \frac{n_k}{n} \Theta_k^{T,r}, \end{cases} \tag{3}$$

其中 $n = \sum_{k=1}^{K} n_k$ 表示整体多模态数据的数量, $\Theta_G^{V,r+1}$, $\Theta_G^{T,r+1}$ 分别表示通过模型聚合得到的全局哈希模型. 通过聚合所有客户端的局部模型, 全局哈希模型能够捕获所有客户端中多模态数据间的公共

语义. 同时, 我们能够利用全局哈希模型来引导各个客户端上局部模型的跨模态哈希学习过程, 从而使局部模型能够学习到整体多模态数据间的语义相似性.

3.3 客户端设计

我们将以 DGCPN [19] 为例, 说明本地执行无监督跨模态哈希的过程, 并介绍针对联邦无监督场景下跨模态哈希所提出的全局 – 局部模态间对比正则化与模态内知识蒸馏两个模块.

3.3.1 无监督跨模态哈希模型训练

DGCPN [19] 源于图模型, 通过数据与其邻居之间的关联信息来构建模态间相似性矩阵, 在无监督跨模态哈希检索中取得了显著的效果. 假设在第 k 个客户端中存在一个批次训练数据的松弛实值矩阵: H_k^V =ImgNet$_k^L(F^V,\Theta_k^V)$, H_k^T =TxtNet$_k^L(F^T,\Theta_k^T)$, F^* 表示图像和文本的特征矩阵. 由此可以得到两个同构模态数据的相似矩阵: $C(H_k^V,H_k^V)$ 和 $C(H_k^T,H_k^T)$, 以及两个异构模态数据的相似矩阵: $C(H_k^V,H_k^T)$ 和 $C(H_k^T,H_k^V)$, 其中 $C(\cdot,\cdot)$ 表示余弦相似度. DGCPN [19] 的目标函数如下:

$$\mathcal{J}_k^H = \mathcal{Q}_c + \lambda_1 \mathcal{Q}_g + \lambda_2 \mathcal{Q}_i, \tag{4}$$

其中 λ_1 和 λ_2 为调整损失贡献的超参数;

$$\mathcal{Q}_c = \left\| \mathrm{Tr}(C(H_k^V,H_k^T) - 1.5I) \right\|_2, \tag{5}$$

为共存相似度保持损失, $\mathrm{Tr}(\cdot)$ 表示矩阵的迹, I 表示与 $C(H_k^V,H_k^T)$ 大小相同的单位矩阵, $\|\cdot\|_2$ 表示 L2 范数;

$$\begin{aligned}\mathcal{Q}_g = &\left\| C(H_k^V,H_k^V) - S_{\mathrm{gc}}(H_k^V,H_k^T) \right\|_{\mathrm{F}} \\ &+ \left\| C(H_k^T,H_k^T) - S_{\mathrm{gc}}(H_k^V,H_k^T) \right\|_{\mathrm{F}} \\ &+ \left\| C(H_k^V,H_k^T) - S_{\mathrm{gc}}(H_k^V,H_k^T) \right\|_{\mathrm{F}}\end{aligned} \tag{6}$$

表示图近邻一致性保留损失, 它以 H_k^V, H_k^T 的图近邻一致性矩阵 $S_{\mathrm{gc}}(H_k^V,H_k^T)$ 反映两个数据的相似性, $\|\cdot\|_{\mathrm{F}}$ 表示 Frobenius 范数;

$$\begin{aligned}\mathcal{Q}_i = &\left\| C(H_k^V,H_k^T) - C(H_k^V,H_k^V) \right\|_{\mathrm{F}} \\ &+ \left\| C(H_k^V,H_k^T) - C(H_k^T,H_k^T) \right\|_{\mathrm{F}} \\ &+ \left\| C(H_k^V,H_k^T) - [C(H_k^V,H_k^T)]^{\mathrm{T}} \right\|_{\mathrm{F}} \\ &+ \left\| C(H_k^V,H_k^V) - C(H_k^T,H_k^T) \right\|_{\mathrm{F}}\end{aligned} \tag{7}$$

为模态内与模态间一致性损失, 其中 $[C(H_k^V,H_k^T)]^{\mathrm{T}}$ 表示相似矩阵 $C(H_k^V,H_k^T)$ 的转置矩阵.

3.3.2 全局 – 局部模态间对比正则化

在无监督学习框架下, 客户端通过挖掘局部数据的模态间相似性矩阵作为监督信息来学习哈希码. 但是, 联邦客户端之间数据分布的不平衡性会导致客户端本地数据难以刻画整体的多模态数据分布, 仅依靠局部信息不足以让跨模态哈希模型学习到整体数据上的相似性知识, 进而会影响客户端模型的训练, 使得局部模型出现偏差, 导致检索性能下降. 一般来说, 聚集所有客户端的多模态数据来训练跨模态哈希模型, 所获得的整体模态间相似性语义能够优于单个客户端上的模态间相似性语义 [6]. 全局

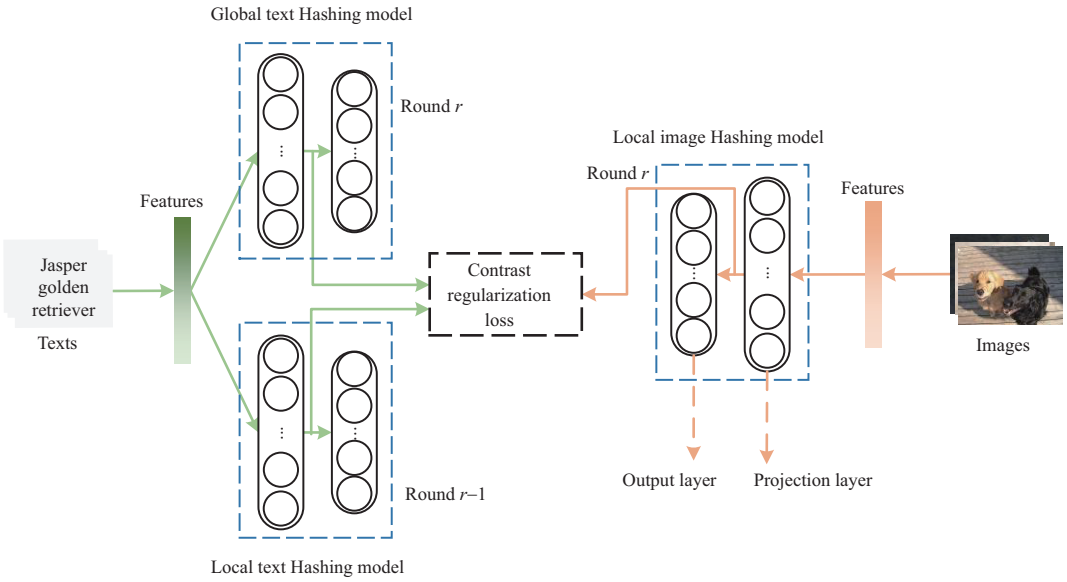

图 2 (网络版彩图) 全局 – 局部模态间对比正则化

Figure 2 (Color online) Global-local inter-modal contrast regularization

哈希模型能够捕获整体多模态数据上的模态间相似性语义, 得益于它聚合了所有客户端本地哈希模型的知识, 从而实现了更优越的性能.

受现有工作的启发 [4,6,34], 我们提出全局 – 局部模态间对比正则化. 具体来说, 让一种模态的全局哈希模型产生的表示来对本地另一种模态的哈希模型进行正则化, 通过模型之间的正则化对本地跨模态学习过程加以约束, 让局部哈希模型能够感知到整体数据上的模态间相似性语义, 加强对客户端本地哈希学习过程的引导. 以本地图像模态为例, 让局部的图像哈希模型尽可能地学习到全局文本哈希模型中的语义相似性知识, 这样能够增加全局模型与局部模型中跨模态信息的交互, 同时降低先前局部文本哈希模型中语义相似性知识的影响, 因为局部文本哈希模型中含有局部训练所带来的模型偏差. 以客户端的图像模态为例, 如图 2 所示, 详细说明全局 – 局部模态间对比正则化的实现方式.

对于图像和文本两个模态来说, 哈希映射模块负责将不同模态的特征映射到统一的特征空间中, 最后经过输出层映射为固定长度的哈希码. 为了便于表示, 如图 2, 采用 $\mathcal{P}^*(\cdot,\omega^*)$ 表示投影层表示的映射函数, 其中 ω^* 表示文本或图像的网络参数, $*\in\{V,T\}$. 对于第 k 个客户端中文本特征 F^T 而言, 经过第 r 轮全局文本哈希模型中投影层得到的特征可以表示为

$$Z_G^{T,r}=\mathcal{P}_G^{T,r}(F^T,\omega_G^T). \tag{8}$$

同理, 文本模态特征经过第 $r-1$ 轮局部文本哈希模型中投影层得到的特征可以表示为

$$Z_k^{T,r-1}=\mathcal{P}_k^{T,r-1}(F^T,\omega_k^T). \tag{9}$$

图像模态特征经过第 r 轮局部图像哈希模型中投影层得到的特征可以表示为

$$Z_k^{V,r}=\mathcal{P}_k^{V,r}(F^V,\omega_k^V). \tag{10}$$

我们的目标是将不同模态特征经过特征投影层映射到公共的特征空间中, 在公共空间中对不同模态的特征表示进行对比学习, 减小 $Z_k^{V,r}$ 和 $Z_G^{T,r}$ 之间的距离, 增大 $Z_k^{V,r}$ 和 $Z_k^{T,r-1}$ 之间的距离. 因此,

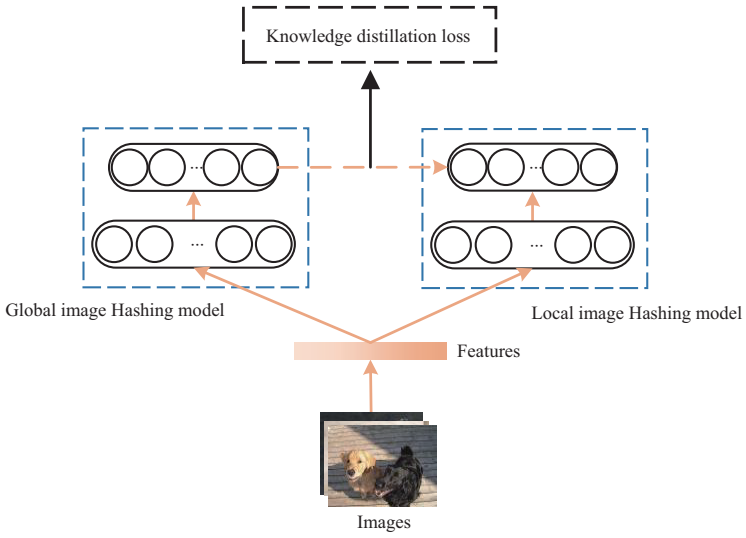

图 3　(网络版彩图) 全局 – 局部模态内知识蒸馏

Figure 3　(Color online) Global-local intra-modal knowledge distillation

与以往工作中损失函数的形式保持一致 [34], 我们将全局 – 局部模态间对比正则化损失 $\mathcal{J}_k^{\mathrm{CR}}$ 定义为

$$\mathcal{J}_k^{\mathrm{CR}} = -\log\frac{\exp(C(Z_k^V, Z_G^T)/\tau)}{\exp(C(Z_k^V, Z_G^T)/\tau) + \exp(C(Z_k^V, Z_k^T)/\tau)}, \tag{11}$$

其中 $\exp(\cdot)$ 表示以自然常 e 为底的指数函数, $C(\cdot,\cdot)$ 表示测量不同模态特征之间的余弦相似性, τ 表示温度系数.

全局 – 局部模态间对比正则化利用共享的全局跨模态哈希模型让局部哈希模型能够充分感知整体数据的相似性语义, 加强对本地跨模态哈希学习过程的引导, 对局部模型中的模态间相似性知识起到了一定的补充作用, 防止了因数据不平衡性而导致的模型偏移. 但是, 对于一种模态的当前轮次模型更新, 局部模型仅学习到了局部数据和全局模型中的跨模态知识, 并没有学习到模态内的全局知识. 这样训练出的模型可能会导致模态内知识缺失, 从而导致模型聚合的过程变得困难 [9]. 因此, 我们又引入了全局 – 局部模态内知识蒸馏进一步地获取模态内的全局知识.

3.3.3　知识蒸馏

受到领域自适应 [35] 中学习域公共知识的启发, 联邦学习可以看作跨客户端进行知识集成的过程 [35]. 通过学习每个客户端数据集上的知识, 可以找到泛化性能更好的模型. 从全局来看, 全局模型来源于所有客户端模型的聚合, 可以视作源域模型. 从局部来看, 客户端自身所拥有的数据并不完整, 需要获得额外的知识补充, 可以看作目标域模型. 从源域自适应到目标域, 一种非常有效的方法就是进行知识蒸馏 [36,37]. 因此, 为了进一步利用各模态全局模型中的知识, 将全局知识转移到局部模型训练中去, 我们采用知识蒸馏策略在全局与局部之间做模态内的知识传递. 如图 3 是以图像模态为例进行知识蒸馏的示意图.

具体而言, 对于第 k 个客户端中的图像和文本两个模态对来说, 我们将其分别输入到全局图像哈希模型和全局文本哈希模型中, 得到的松弛实值可以表示为 H_G^V 和 H_G^T. 同理, 将其分别输入到局部图像哈希模型和局部文本哈希模型中, 得到的松弛实值可以表示为 H_k^V 和 H_k^T. 我们在模态间对比正则化损失的基础之上添加一个模态内的蒸馏项, 缩小局部模型产生的松弛实表示与每轮全局模型产生

松弛表示之间的距离, 在同一模态内让全局哈希模型对局部哈希模型进行指导:

$$\mathcal{J}_k^{\mathrm{KD}} = \mathcal{D}(H_G^*, H_k^*), \tag{12}$$

其中 $\mathcal{D}(\cdot,\cdot)$ 用来衡量同一模态内全局模型与局部模型输出之间的差异, 这种差异可以通过 KL (Kullback-Leibler) 散度来衡量:

$$\min_{\Theta_k^*} \mathbb{E}_{x^* \sim D^k}[\mathcal{F}(x^*;\Theta_G^*)||\mathcal{F}(x^*;\Theta_k^*)], \tag{13}$$

其中 $\mathcal{F}$ 表示图像或文本哈希函数.

通过全局 – 局部模态间对比正则化和模态内知识蒸馏两种策略, 我们在模态间与模态内分别施加约束, 做到了在模态内与模态间共同对局部训练进行正则化. 最终的客户端损失形式如下所示:

$$\mathcal{J}_k = \mathcal{J}_k^{\mathrm{H}} + \mu\mathcal{J}_k^{\mathrm{CR}} + \varphi\mathcal{J}_k^{\mathrm{KD}}, \tag{14}$$

其中 μ, φ 为平衡损失的超参数.

4 实验

4.1 数据集

在实验中, Wikipedia [38], MIRFlickr-25K [39], IAPR TC-12 [40], MS-COCO [41] 和 NUS-WIDE [42] 5 个常用的跨模态检索数据集被用来测试本文所提出的联邦无监督跨模态哈希检索方法. 上述数据集中, Wikipedia 是唯一的单标签跨模态数据集, 共包含属于 10 个类别的 2866 个图像 – 文本对. IAPR TC-12 包含 255 个类别的 20000 个多标签图像 – 文本对. MIRFlickr-25K 由属于 24 个类别的 25000 个多标签图像 – 文本对构成, 本文遵循 DCMH [3] 中的实验设置, 共选择 20015 个样本进行实验. MS-COCO 包含属于 80 个类别的 123287 个多标签图像 – 文本对. NUS-WIDE 数据集包含 269648 幅带有文本标签的网络图像, 其中每个图像 – 文本对由一个或多个标签注释, 我们选择了 21 个最常见类别的 195834 对样本进行实验.

由于上述数据集文本模态构成方式不同, 因此每个数据集的文本特征表示方式也不相同. 其中, Wikipedia 数据集采用了 1024 维的 LDA (latent Dirichlet allocation) 特征表示, MIRFlickr-25K 和 NUS-WIDE 数据集分别采用 1386 维和 1000 维的 Tag 特征表示; 对于 IAPR TC-12 和 MS-COCO 数据集, 采用 Bert 模型 [43] 提取了 1024 维的深度文本特征表示. 而对于图像模态, 均采用在 ImageNet [44] 数据集上预训练的 CNN-F [45] 网络获取 4096 维的特征表示.

我们将每个数据集划分为检索集和测试集, 并选取部分检索集样本构成训练集, 因此检索集与测试集中的数据不重叠, 具体的数据集划分情况如表 1 所示. 同时, 为模拟联邦环境中数据分布的不平衡性, 我们采用先前工作 [21, 46] 中的设置, 利用狄利克雷 (Dirichlet) 分布对每个数据集进行非独立同分布划分, 并通过无量纲分布参数 α 控制多模态数据的非独立同分布程度, 图 4 展示了 $\alpha = 0.5$ 时 Wikipedia 数据集的多模态数据划分情况.

4.2 实现细节

本文 FedUCH 中的客户端哈希模型可以是任意无监督跨模态哈希方法, 因此采用了 3 种无监督跨模态哈希方法 [17~19] 验证 FedUCH 的普适性. 由于 3 种无监督跨模态哈希方法的网络结构不同, 实验中均依据原始工作中的网络设置构建图像哈希模型与文本哈希模型, 同时所有基线方法的参数设置

表 1　5 个基准数据集的划分

Table 1　Patition of the five benchmark datasets

Dataset	Retrieval set	Training set	Query set
Wikipedia	2173	2173	462
IAPR TC-12	18000	5000	2000
MIRFlickr-25K	16015	5000	2000
MS-COCO	121287	5000	2000
NUS-WIDE	193834	5000	2000

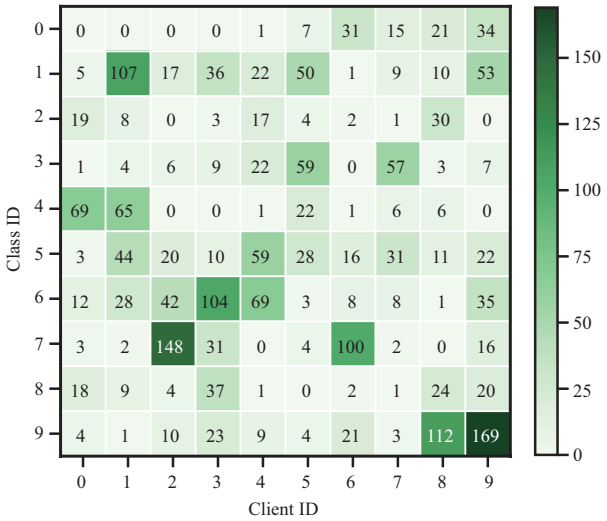

图 4　(网络版彩图) Wikipedia 数据集上 $\alpha = 0.5$ 时的标签不平衡划分结果

Figure 4　(Color online) Label imbalance partition result for $\alpha = 0.5$ on Wikipedia dataset

也遵循原始文献中的说明. 针对每个无监督跨模态哈希方法, 计算了联邦框架中各个客户端上检索性能的平均值作为该哈希方法联邦学习的最终结果. 本文代码已公开至网站[1)].

在联邦优化过程中, 我们为每个数据集设定了不同的加密通信轮次 R, 其中 Wikipedia 上的通信轮次 $R = 25$, MIRFlickr-25K 和 IAPR TC-12 的通信轮次 $R = 30$, 而 MS-COCO 和 NUS-WIDE 的通信轮次 $R = 35$. 在每轮次的加密通信中, 各个客户端均利用本地数据训练 5 轮局部哈希模型, 温度系数 τ 设定为 1.0. 针对损失函数的超参数 μ 和 φ, 分别从 0.2 调整到 0.8 进行超参数组合测试联邦检索的最终性能, 各个数据集的最优超参数组合如下: Wikipedia 和 MIRFlickr-25K: $\mu = 0.6$, $\varphi = 0.4$; IAPR TC-12: $\mu = 0.5$, $\varphi = 0.5$; MS-COCO: $\mu = 0.4$, $\varphi = 0.7$; NUS-WIDE: $\mu = 0.3$, $\varphi = 0.7$.

4.3　评价指标

与以往的跨模态检索工作相同, 本文使用 mAP (mean average precision) [47] 评估所有跨模态哈希方法的检索性能. mAP 同时度量了跨模态检索模型的相似性搜索精度和返回结果排序情况, 检索模型

1) https://gitee.com/JZL629/feduch.

表 2 Wikipedia 数据集上的检索性能对比

Table 2 The retrieval performance comparison on Wikipedia dataset

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
Standalone [19]	0.239	0.282	0.333	0.344	0.203	0.220	0.229	0.233
FedAvg [4]	0.323	0.382	0.417	0.435	0.266	0.269	0.294	0.313
FedProx [7]	0.311	0.362	0.418	0.422	0.242	0.271	0.272	0.260
MOON [6]	0.346	0.379	0.442	0.450	0.281	0.282	0.297	0.295
FedUCH	**0.364**	**0.390**	**0.454**	**0.460**	**0.311**	**0.321**	**0.336**	**0.347**

搜索出来的相关样本越多且排名越靠前, 其 mAP 值就会越高, 反之越低. 因此, mAP 被定义如下:

$$\mathrm{mAP} = \frac{1}{q}\sum_{i=1}^{q}\mathrm{AP}_i, \tag{15}$$

其中 q 表示查询样本的数量, AP_i 表示第 i 个查询样本的平均检索精度, 假设给定第 i 个查询样本的前 k 个返回结果, 其 AP 被定义为

$$\mathrm{AP}_i = \frac{\sum_{j=1}^{k}(P(j)\times R(j))}{\sum_{j=1}^{k}R(j)}, \tag{16}$$

其中 $P(j)$ 表示前 j 个返回结果中相关样本的检索精度, $R(j)$ 表示第 j 个检索结果是否与查询样本相关, 如果相关则 $R(j)=1$, 反之 $R(j)=0$. 本文基于每个查询样本返回的前 50 个检索结果计算 mAP 值来度量跨模态哈希模型的检索性能, 即 mAP@50.

4.4 对比实验

为验证本文所提方法的有效性, 我们在 5 个被广泛使用的跨模态检索数据集上进行了大量的对比实验. 遵循以往的联邦跨模态检索工作 [5], 首先在对比实验中设置了 Standalone [5] 方法, 其表示在无联邦学习框架下仅利用客户端本地数据训练跨模态哈希模型得到的检索性能. 同时, 3 种联邦学习方法被引入用来验证本文所提 FedUCH 的有效性, 包括主流的联邦平均算法 FedAvg [4], 两种先进的联邦学习框架 FedProx [7] 和 MOON [6]. 为使对比实验结果更加准确可靠, Standalone 和所有联邦学习方法均采用 DGCPN [19] 作为基础的客户端模型. 需要注意的是, DGCPN 可以被替换为任意的无监督跨模态哈希方法, 例如 DJSRH [17], JDSH [18] 等, 我们将在消融实验中详细分析不同无监督跨模态哈希方法对联邦跨模态检索性能的影响. 本小节根据数据集的类型和规模分别探讨了本文所提 FedUCH 和对比方法的检索性能.

4.4.1 在 Wikipedia 数据集上的对比实验

相比于其他数据集, Wikipedia 是一个小规模单标签数据集, 因此所有对比方法在该数据集上的跨模态哈希学习难度较大. 表 2 记录了在 Wikipedia 数据集上文本检索图像和图像检索文本的实验结果, 从中可以发现每个客户端独立训练 (即 Standalone 方法) 的检索性能较差, 所有联邦跨模态检索方法在两种任务上均显著优于 Standalone 方法. 此外, FedUCH 的检索性能要优于其余 3 种联邦跨模态检索方法, 其中在文本检索图像任务上 FedUCH 的检索性能超过 1%∼4%, 在图像检索文本任务上超过 3%∼5%. 因此, 实验结果表明, 在各个客户端拥有数据量较少的情况下, 联邦学习框架能够在保

表 3 MIRFlickr-25K 数据集上的检索性能对比

Table 3 The retrieval performance comparison on MIRFlickr-25K dataset

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
Standalone [19]	0.679	0.697	0.702	0.711	0.693	0.717	0.733	0.746
FedAvg [4]	0.729	0.753	0.769	0.777	0.742	0.763	0.781	0.790
FedProx [7]	0.727	0.747	0.758	0.769	0.743	0.764	0.779	0.791
MOON [6]	0.735	0.756	0.775	0.785	0.752	0.772	0.794	0.802
FedUCH	**0.748**	**0.765**	**0.785**	**0.793**	**0.761**	**0.786**	**0.805**	**0.814**

表 4 IAPR TC-12 数据集上的检索性能对比

Table 4 The retrieval performance comparison on IAPR TC-12 dataset

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
Standalone [19]	0.521	0.587	0.603	0.612	0.513	0.554	0.590	0.594
FedAvg [4]	0.573	0.623	0.644	0.650	0.564	0.586	0.633	0.634
FedProx [7]	0.585	0.623	0.649	0.653	0.573	0.614	0.629	0.640
MOON [6]	0.572	0.628	0.641	0.643	0.567	0.605	0.638	0.628
FedUCH	**0.602**	**0.641**	**0.652**	**0.661**	**0.591**	**0.630**	**0.645**	**0.650**

护客户端数据隐私的同时构建全局跨模态哈希模型, 从而提高各个客户端在本地数据上的跨模态检索性能. 同时, FedUCH 通过全局跨模态哈希模型对局部模型进行引导训练和正则化约束, 从而明显改善了联邦学习过程中客户端局部模型的学习能力和检索性能.

4.4.2 在 MIRFlickr-25K, IAPR TC-12 数据集上的对比实验

MIRFlickr-25K 和 IAPR TC-12 数据集均为样本数量 2 万左右的多标签数据集, 其实验结果如表 3 和 4 所示, 由于数据规模的影响, MIRFlickr-25K 和 IAPR TC-12 数据集划分到每个客户端的跨模态数据量明显大于 Wikipedia 数据集, 因此每个客户端独立训练 (即 Standalone 方法) 也可以获得较好的跨模态检索性能.

在两种检索任务上, 联邦跨模态检索方法相对于每个客户端独立训练仍然可以取得 3%∼5% 的检索性能提升, 这进一步证明了联邦跨模态学习的有效性. 同时, 本文提出的 FedUCH 仍然能够在目前联邦学习方法的基础上取得 1%∼3% 的性能提升, 这说明在客户端局部数据较多的情况下, FedUCH 仍然是显著有效的.

4.4.3 在 MS-COCO, NUS-WIDE 数据集上的对比实验

MS-COCO 和 NUS-WIDE 数据集均为样本数量超过 10 万的多标签数据集, 表 5 和 6 记录了两个数据集上所有方法的检索性能. 根据表 5 和 6 的实验结果, 每个客户端独立训练的跨模态哈希模型在检索性能上已经接近甚至超过了 FedAvg 方法 [4], 这主要是由于 FedAvg 算法依靠模型聚合来缓解客户端之间的数据不平衡性, 而当客户端本地数据规模较大时, 局部模型难以在联邦过程中能够获取有指导意义的样本相似性知识, 并且模型平均可能会导致客户端本地知识的丢失从而造成跨模态检索性能的略微下降. 相对于 FedAvg 方法, 其他联邦学习方法通过不同方式缓解客户端之间的数据不平衡

表 5 MS-COCO 数据集上的检索性能对比

Table 5 The retrieval performance comparison on MS-COCO dataset

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
Standalone [19]	0.557	0.623	0.675	0.707	0.494	0.583	0.608	0.616
FedAvg [4]	0.602	0.683	0.714	0.710	0.591	0.643	0.663	0.659
FedProx [7]	0.614	0.680	0.722	0.739	0.586	0.646	0.665	0.669
MOON [6]	0.615	0.679	0.738	0.726	0.597	0.649	0.678	0.683
FedUCH	**0.635**	**0.715**	**0.753**	**0.760**	**0.621**	**0.665**	**0.705**	**0.710**

表 6 NUS-WIDE 数据集上的检索性能对比

Table 6 The retrieval performance comparison on NUS-WIDE dataset

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
Standalone [19]	0.663	0.724	0.739	0.754	0.668	0.731	0.760	0.765
FedAvg [4]	0.647	0.724	0.744	0.757	0.648	0.725	0.764	0.767
FedProx [7]	0.675	0.727	0.745	0.759	0.684	0.738	0.761	0.761
MOON [6]	0.664	0.722	0.747	0.751	0.681	0.722	0.759	0.768
FedUCH	**0.682**	**0.737**	**0.756**	**0.766**	**0.690**	**0.745**	**0.775**	**0.781**

性, 因此能够更好地平衡全局模型与局部模型中包含的样本相似性知识. 然而, 目前的联邦学习方法并没有针对无监督跨模态检索任务的特定改进, 本文所提出的 FedUCH 得益于跨模态对比约束和知识蒸馏框架, 因此相较于其他联邦学习方法仍然有一定的检索性能提升. 具体而言, 在 MS-COCO 数据集上, FedUCH 相对于表现最好的联邦跨模态检索方法能够提升 1%~3% 的检索性能, 而在 NUS-WIDE 数据集上, FedUCH 能够获得 1% 左右的检索性能提升.

4.5 消融实验

4.5.1 模块消融实验

为了验证本文所提方法中各模块的有效性, 设计了如下变体方法来分别探究 FedUCH 中不同模块的对最终联邦跨模态检索性能的贡献:

- FedUCH-1. 移除 FedUCH 客户端中的对比正则化模块和知识蒸馏模块, 该变体等同于将无监督跨模态哈希方法与 FedAvg 框架相结合;
- FedUCH-2. 移除 FedUCH 客户端中的知识蒸馏模块, 仅保留对比正则化模块;
- FedUCH-3. 移除 FedUCH 客户端中的对比正则化模块, 仅保留知识蒸馏模块.

表 7~11 分别记录了在 5 个数据集上 32 和 64 bit 哈希码的消融实验结果. 根据表 7~11 的实验结果, 当移除对比正则化模块和知识蒸馏模块后, FedUCH 的跨模态检索性能效果相似于 FedAvg 方法. 当增加对比正则化模块或知识蒸馏模块后, FedUCH 均获得不同程度的跨模态检索性能提升, 这说明全局模型在联邦跨模态哈希学习中的有效性和重要性, 并且对比正则化模块和知识蒸馏模块均有益于客户端模型的联邦学习过程. 此外, FedUCH-2 在大多数情况下的跨模态检索性能优于 FedUCH-3, 这说明全局模型与局部模型的模态间语义交互要更优于局部模态内语义交互.

表 7 Wikipedia 上的消融结果

Table 7 Ablation results on Wikipedia

Method	Text to image		Image to text	
	32 bit	64 bit	32 bit	64 bit
FedUCH-1	0.382	0.417	0.269	0.294
FedUCH-2	0.386	0.421	0.311	0.321
FedUCH-3	0.385	0.432	0.312	0.329
FedUCH	**0.390**	**0.454**	**0.321**	**0.336**

表 8 MIRFlickr-25K 上的消融结果

Table 8 Ablation results on MIRFlickr-25K

Method	Text to image		Image to text	
	32 bit	64 bit	32 bit	64 bit
FedUCH-1	0.753	0.769	0.763	0.781
FedUCH-2	0.761	0.781	0.772	0.792
FedUCH-3	0.757	0.777	0.769	0.785
FedUCH	**0.765**	**0.785**	**0.786**	**0.805**

表 9 IAPR TC-12 上的消融结果

Table 9 Ablation results on IAPR TC-12

Method	Text to image		Image to text	
	32 bit	64 bit	32 bit	64 bit
FedUCH-1	0.623	0.644	0.586	0.633
FedUCH-2	0.638	0.646	0.621	0.640
FedUCH-3	0.630	0.648	0.615	0.642
FedUCH	**0.641**	**0.652**	**0.630**	**0.645**

表 10 MS-COCO 上的消融结果

Table 10 Ablation results on MS-COCO

Method	Text to image		Image to text	
	32 bit	64 bit	32 bit	64 bit
FedUCH-1	0.683	0.714	0.643	0.663
FedUCH-2	0.705	0.748	0.660	0.677
FedUCH-3	0.701	0.734	0.658	0.670
FedUCH	**0.715**	**0.753**	**0.665**	**0.705**

4.5.2 方法普适性实验

为验证本文所提 FedUCH 的普适性, 我们分别利用 DJSRH [17] 和 JDSH [18] 作为客户端无监督跨模态哈希方法测试 FedUCH 的最终检索性能. 表 12~15 分别记录了 DJSRH, JDSH 两种方法在 MIRFlickr-25K 和 NUS-WIDE 数据集上的实验结果. 根据实验结果可以看出, 本文所提 FedUCH 的跨模态检索性能均优于其他联邦跨模态检索方法, 这证明对任意无监督跨模态哈希检索方法在联邦环境

表 11 NUS-WIDE 上的消融结果

Table 11 Ablation results on NUS-WIDE

Method	Text to image		Image to text	
	32 bit	64 bit	32 bit	64 bit
FedUCH-1	0.724	0.744	0.725	0.764
FedUCH-2	0.732	0.751	0.736	0.772
FedUCH-3	0.729	0.748	0.730	0.769
FedUCH	**0.737**	**0.756**	**0.745**	**0.775**

表 12 本地方法为 DJSRH 时, 在 NUS-WIDE 数据集上的实验结果

Table 12 Experimental results on the NUS-WIDE dataset when the local method is DJSRH

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
FedAvg [4]	0.573	0.620	0.660	0.654	0.525	0.616	0.649	0.701
FedProx [7]	0.552	0.610	0.644	0.629	0.561	0.621	0.642	0.673
MOON [6]	0.569	0.616	0.652	0.664	0.582	0.639	0.673	0.695
FedUCH	**0.581**	**0.624**	**0.661**	**0.677**	**0.596**	**0.644**	**0.679**	**0.718**

表 13 本地方法为 JDSH 时, 在 NUS-WIDE 数据集上的实验结果

Table 13 Experimental results on the NUS-WIDE dataset when the local method is JDSH

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
FedAvg [4]	0.661	0.694	0.717	0.722	0.700	0.723	0.751	0.760
FedProx [7]	0.663	0.705	0.715	0.733	0.702	0.724	0.731	0.763
MOON [6]	0.678	0.699	0.721	0.735	0.705	0.728	0.749	0.758
FedUCH	**0.689**	**0.717**	**0.728**	**0.743**	**0.710**	**0.737**	**0.759**	**0.773**

表 14 本地方法为 DJSRH 时, 在 MIRFlickr-25K 数据集上的实验结果

Table 14 Experimental results on MIRFlickr-25K dataset when the local method is DJSRH

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
FedAvg [4]	0.667	0.684	0.695	0.684	0.667	0.677	0.687	0.698
FedProx [7]	0.666	0.677	0.682	0.672	0.678	0.683	0.679	0.683
MOON [6]	0.673	0.681	0.690	0.705	0.687	0.710	0.726	0.736
FedUCH	**0.681**	**0.689**	**0.711**	**0.725**	**0.693**	**0.720**	**0.735**	**0.741**

中的性能均有提升, 这证明了本文方法具有一定的通用性.

4.6 参数敏感性分析

本小节在 5 个数据集的 64 bit 哈希码上进行了参数敏感性实验, 从而探究超参数 μ 和 φ 对 FedUCH 最终检索性能的影响. 我们取 $\varphi = [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]$, 并将其与 $\mu = [0.2, 0.3, 0.4, 0.5,$

表 15　本地方法为 JDSH 时, 在 MIRFlickr-25K 数据集上的实验结果

Table 15　Experimental results on MIRFlickr-25K dataset when the local method is JDSH

Method	Text to image				Image to text			
	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
FedAvg [4]	0.738	0.762	0.772	0.775	0.757	0.789	0.777	0.795
FedProx [7]	0.733	0.743	0.763	0.760	0.748	0.755	0.770	0.792
MOON [6]	0.742	0.755	0.766	0.781	0.765	0.772	0.782	0.800
FedUCH	**0.749**	**0.769**	**0.784**	**0.790**	**0.776**	**0.795**	**0.801**	**0.812**

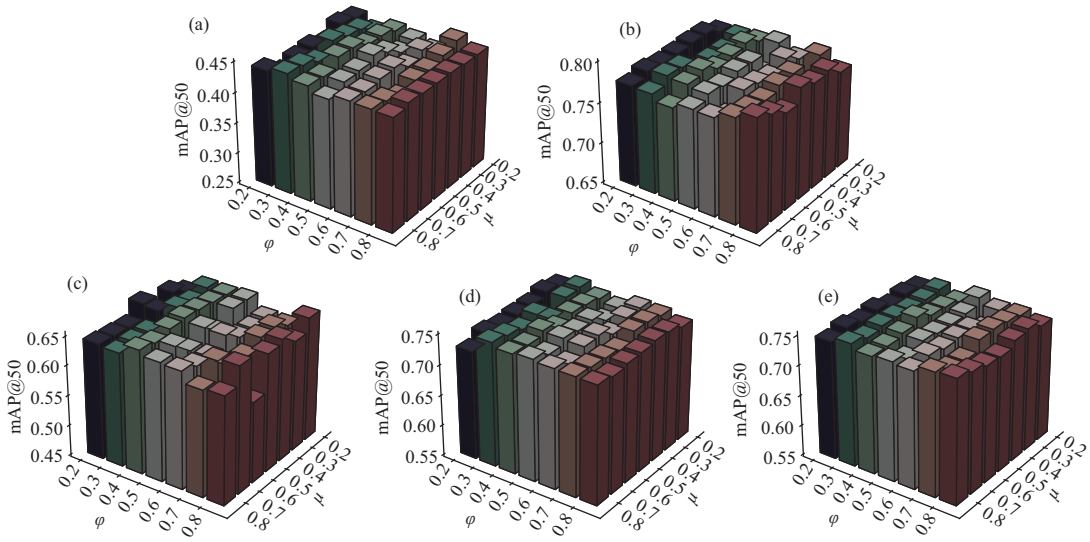

图 5　(网络版彩图) 5 个基准数据集文本检索图像任务的参数敏感性实验结果

Figure 5　(Color online) Experimental results of parameter sensitivity of text-to-image retrieval task on five benchmark datasets. (a) Wikipedia; (b) MIRFlickr-25K; (c) IAPR TC-12; (d) MS-COCO; (e) NUS-WIDE

$0.6, 0.7, 0.8]$ 两两组合进行参数敏感性实验. 每一组 μ 和 φ 超参数都会获得若干客户端的 mAP 值, 我们将该组超参数设定下所有客户端的 mAP 值取平均来分析超参数影响. 图 5 和 6 分别记录了 5 个数据集上不同超参数组合对于跨模态检索性能的影响. 综合图中的实验结果分析, 不同超参数组合对于图像检索文本任务的检索性能影响较大, 而对于文本检索图像任务影响较小. 具体来说, 在大多数情况下, 当固定超参数 μ 时, φ 的不同取值对于两种检索任务的性能影响较小; 当固定超参数 φ 时, μ 的不同取值对于图片检索文本任务的性能影响较大, 对于文本检索任务的性能影响较小. 需要注意的是, 虽然 μ 的不同取值对于图片检索文本性能影响较大, 但 FedUCH 仍然具有较高的检索精度, 因此整体来看 FedUCH 对超参数变化具有一定的鲁棒性.

4.7　模型收敛性分析

在实际联邦框架应用中, 云服务器和客户端的通信轮次往往限制了最终联邦学习的性能. 为验证本文所提 FedUCH 在实际应用中的可行性, 我们测试了联邦框架中客户端上无监督跨模态哈希模型的收敛性情况. 实验结果如图 7 和 8 所示, FedUCH 中客户端模型在不同数据集上的模型收敛性不同. 为实现 FedUCH 中客户端模型的完全收敛, Wikipedia 数据集上需要进行 25 个联邦通信轮次,

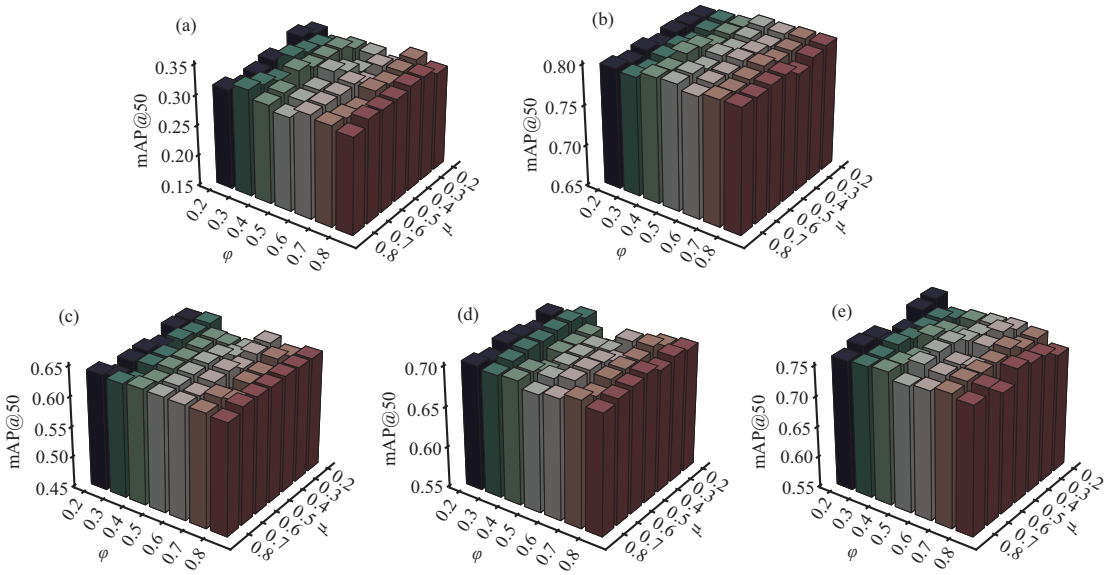

图 6 (网络版彩图) 5 个基准数据集图像检索文本任务的参数敏感性实验结果

Figure 6 (Color online) Experimental results on parameter sensitivity of image-to-text retrieval task with five benchmark datasets. (a) Wikipedia; (b) MIRFlickr-25K; (c) IAPR TC-12; (d) MS-COCO; (e) NUS-WIDE

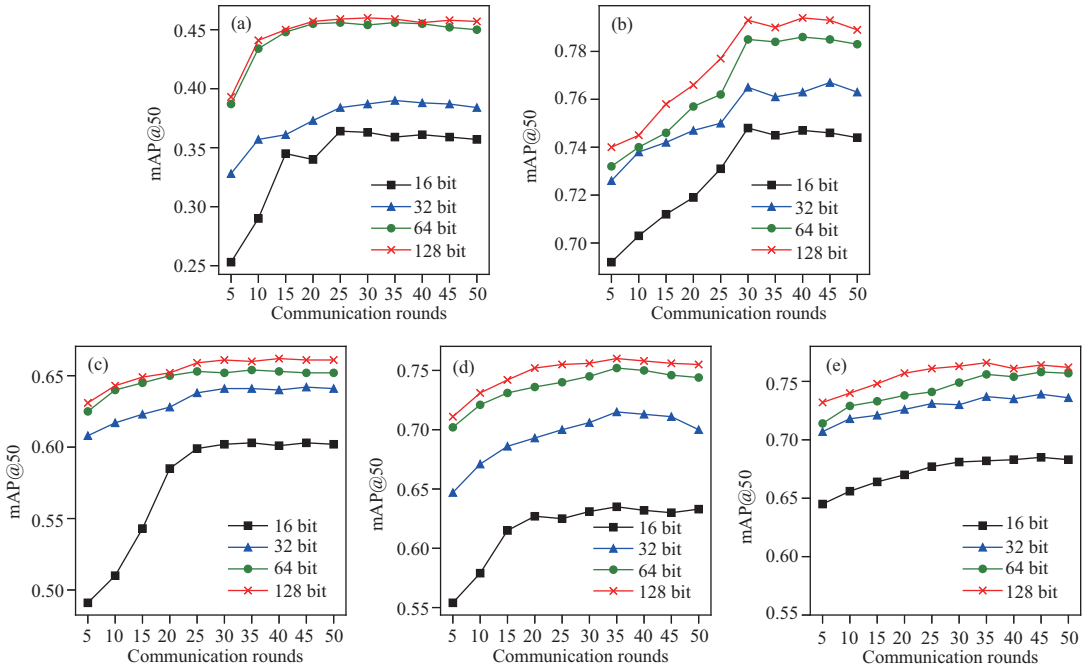

图 7 (网络版彩图) 5 个基准数据集文本检索图像任务的模型收敛性实验结果

Figure 7 (Color online) Experimental results of model convergence for text-to-image retrieval task with five benchmark datasets. (a) Wikipedia; (b) MIRFlickr-25K; (c) IAPR TC-12; (d) MS-COCO; (e) NUS-WIDE

MIRFlickr-25K 和 IAPR TC-12 数据集上需要进行 30 个联邦通信轮次, MS-COCO 和 NUS-WIDE 数据集上需要进行 35 个联邦通信轮次. 以上实验结果说明客户端用户仅需要与云服务器进行少量的模

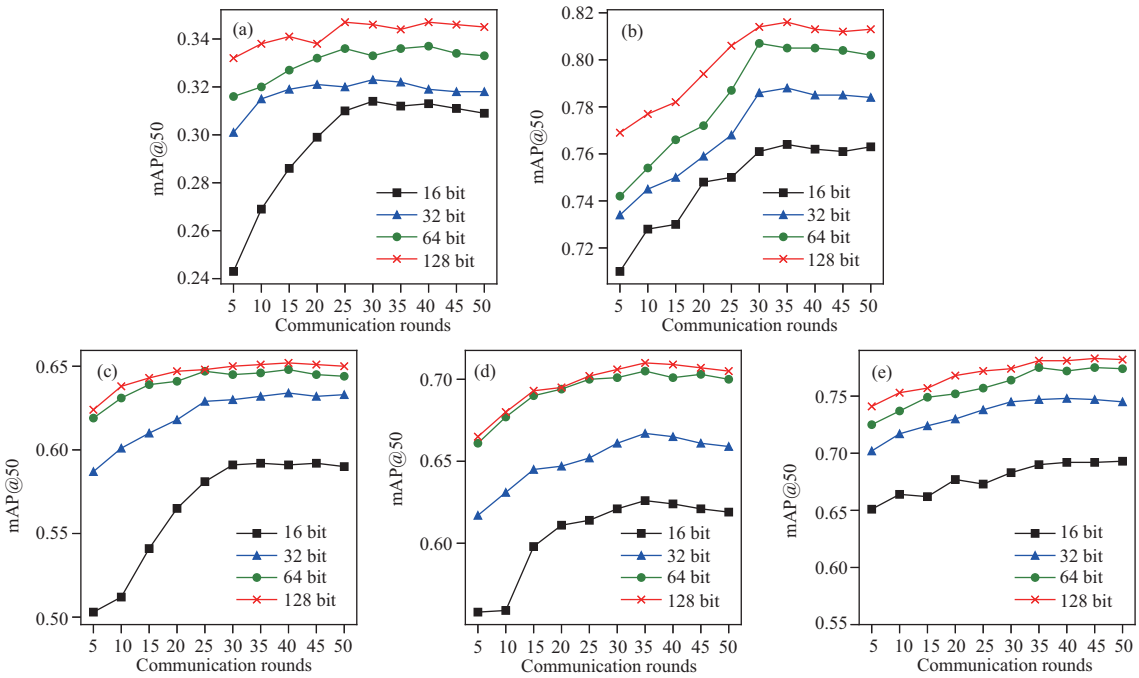

图 8　(网络版彩图) 5 个基准数据集图像检索文本任务的模型收敛性实验结果

Figure 8　(Color online) Experimental results of model convergence for image-to-text retrieval task with five benchmark datasets. (a) Wikipedia; (b) MIRFlickr-25K; (c) IAPR TC-12; (d) MS-COCO; (e) NUS-WIDE

型加密传输即可得到最终的跨模态哈希检索模型, 这实现了高效的联邦训练并充分保护了客户端的本地数据隐私.

5　结论

本文提出了一种新的联邦无监督跨模态哈希检索模型 FedUCH, 在保护客户端隐私的情况下学习不依赖语义标注的跨模态检索模型. 考虑到联邦环境中多模态数据分布的不平衡性, 本文利用全局模型产生的表示在不同模态间对本地模型进行正则化训练引导, 并在同一模态内进行知识蒸馏感知全局语义. 5 个基准跨模态检索数据集上的大量的实验研究表明, FedUCH 有效地提高了联邦无监督学习环境下的跨模态检索性能. 对于未来工作, 我们首先将探究如何依照每个客户端的数据分布特点, 对客户端模型采取 “个性化” 措施, 进一步提升局部模型的效果. 其次, 我们希望进一步探究客户端中模态不平衡, 数据不完整等情况下的联邦跨模态检索问题.

参考文献

1　Chen Z, Du H, Chen Y F, et al. Cross-modal video moment retrieval based on visual-textual relationship alignment. Sci Sin Inform, 2020, 50: 862–876 [陈卓, 杜昊, 吴雨菲, 等. 基于视觉 – 文本关系对齐的跨模态视频片段检索. 中国科学: 信息科学, 2020, 50: 862–876]

2　Li Z X, Ling F, Tang Z J, et al. Unsupervised cross-media Hashing retrieval based on multi-head attention network. Sci Sin Inform, 2021, 51: 2053–2068 [李志欣, 凌锋, 唐振军, 等. 基于多头注意力网络的无监督跨媒体哈希检索. 中国科学: 信息科学, 2021, 51: 2053–2068]

3 Jiang Q Y, Li W J. Deep cross-modal Hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3232–3240
4 McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017. 1273–1282
5 Zong L L, Xie Q J, Zhou J H, et al. FedCMR: federated cross-modal retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021. 1672–1676
6 Li Q B, He B S, Song D. Model-contrastive federated learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021. 10713–10722
7 Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems, 2020. 429–450
8 Wang J Y, Liu Q H, Liang H, et al. Tackling the objective inconsistency problem in heterogeneous federated optimization. In: Proceedings of Advances in Neural Information Processing, 2020. 7611–7623
9 Yao D Z, Pan W N, Dai Y T, et al. Local-global knowledge distillation in heterogeneous federated learning with non-IID data. 2021. ArXiv:2107.00051
10 Kaur P, Pannu H S, Malhi A K. Comparative analysis on cross-modal information retrieval: a review. Comput Sci Rev, 2021, 39: 100336
11 Lin Z J, Ding G G, Hu M Q, et al. Semantics-preserving Hashing for cross-view retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3864–3872
12 Tang J, Wang K, Shao L. Supervised matrix factorization Hashing for cross-modal retrieval. IEEE Trans Image Process, 2016, 25: 3157–3166
13 Cao Y, Long M S, Wang J M, et al. Deep visual-semantic Hashing for cross-modal retrieval. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 1445–1454
14 Ding G G, Guo Y C, Zhou J L. Collective matrix factorization Hashing for multimodal data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. 2083–2090
15 Liu H, Ji R R, Wu Y J, et al. Cross-modality binary code learning via fusion similarity Hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 6345–6353
16 Wu G S, Lin Z J, Han J G, et al. Unsupervised deep Hashing via binary latent factor models for large-scale cross-modal retrieval. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018. 2854–2860
17 Su S P, Zhong Z S, Zhang C. Deep joint-semantics reconstructing Hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 3027–3035
18 Liu S, Qian S S, Guan Y, et al. Joint-modal distribution-based similarity Hashing for large-scale unsupervised deep cross-modal retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020. 1379–1388
19 Yu J, Zhou H, Zhan Y B, et al. Deep graph-neighbor coherence preserving network for unsupervised cross-modal Hashing. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021. 4626–4634
20 Yang Q, Liu Y, Chen T, et al. Federated machine learning. ACM Trans Intell Syst Technol, 2019, 10: 1–19
21 Zhao Y, Li M, Lai L Z, et al. Federated learning with non-IID data. 2018. ArXiv:1806.00582
22 Li X, Huang K X, Yang W H, et al. On the convergence of FedAvg on non-IID data. In: Proceedings of the 8th International Conference on Learning Representations, 2020. 1–26
23 Fallah A, Mokhtari A, Ozdaglar A E. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. In: Proceedings of Advances in Neural Information Processing, 2020. 3557–3568
24 Smith V, Chiang C K, Sanjabi M, et al. Federated multi-task learning. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 4424–4434
25 Zhang F D, Kuang K, You Z Y, et al. Federated unsupervised representation learning. 2020. ArXiv:2010.08982
26 Zhuang W M, Gan X, Wen Y G. Collaborative unsupervised visual representation learning from decentralized data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 4892–4901
27 Liu F L, Wu X, Ge S, et al. Federated learning for vision-and-language grounding problems. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020. 11572–11579
28 Xiong B, Yang X, Qi F, et al. A unified framework for multi-modal federated learning. Neurocomputing, 2022, 480: 110–118

29 Zhao Y C, Barnaghi P M, Haddadi H. Multimodal federated learning on IoT data. In: Proceedings of the 7th IEEE/ACM International Conference on Internet-of-Things Design and Implementation, 2022. 43–54

30 Chen J Y, Zhang A D. FedMSplit: correlation-adaptive federated multi-task learning across multimodal split networks. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022. 87–96

31 Chen S J, Li B C. Towards optimal multi-modal federated learning on non-IID data with hierarchical gradient blending. In: Proceedings of the IEEE Conference on Computer Communications, 2022. 1469–1478

32 Xu J, Xu Z X, Walker P B, et al. Federated patient Hashing. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020. 6486–6493

33 Yang M L, Xu J, Liu Y, et al. FedHAP: federated Hashing with global prototypes for cross-silo retrieval. 2022. ArXiv:2207.05525

34 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, 2020. 1597–1607

35 Peng X C, Huang Z J, Zhu Y Z, et al. Federated adversarial domain adaptation. In: Proceedings of the 8th International Conference on Learning Representations, 2020. 1–19

36 Zhu Z D, Hong J Y, Zhou J Y. Data-free knowledge distillation for heterogeneous federated learning. In: Proceedings of the 38th International Conference on Machine Learning, 2021. 12878–12889

37 Lin T, Kong L J, Stich S U, et al. Ensemble distillation for robust model fusion in federated learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 2351–2363

38 Rasiwasia N, Costa P J, Coviello E, et al. A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th International Conference on Multimedia, 2010. 251–260

39 Huiskes M J, Lew M S. The MIR flickr retrieval evaluation. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008. 39–43

40 Escalante H J, Hernández C A, Gonzalez J A, et al. The segmented and annotated IAPR TC-12 benchmark. Comput Vision Image Understand, 2010, 114: 419–428

41 Lin T Y, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context. In: Proceedings of European Conference on Computer Vision, 2014. 740–755

42 Chua T S, Tang J H, Hong R C, et al. NUS-WIDE: a real-world web image database from National University of Singapore. In: Proceedings of the 8th ACM International Conference on Image and Video Retrieval, 2009. 1–9

43 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186

44 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis, 2015, 115: 211–252

45 Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British Machine Vision Conference, 2014. 1–12

46 Li Q B, Diao Y Q, Chen Q, et al. Federated learning on non-IID data silos: an experimental study. In: Proceedings of the 38th IEEE International Conference on Data Engineering, 2022. 965–978

47 Liu W, Mu C, Kumar S, et al. Discrete graph Hashing. In: Proceedings of Advances in Neural Information Processing, 2014. 3419–3427

Federated unsupervised cross-modal Hashing

Lei ZHU[1,3*], Jingzhi LI[1], Tianshi WANG[1], Jingjing LI[2] & Huaxiang ZHANG[1*]

1. *School of Information Science and Engineering, Shandong Normal University, Jinan* 250358*, China*;
2. *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu* 611731*, China*;
3. *Peng Cheng Laboratory, Shenzhen* 518055*, China*
* Corresponding author. E-mail: leizhu0608@gmail.com, huaxzhang@sdnu.edu.cn

Abstract Federated cross-modal retrieval uses decentralized clients to learn a shared cross-modal retrieval model to reduce the high maintenance cost associated with centralized multimodal training data and solve the data privacy problem in cross-modal retrieval in distributed data storage scenarios. However, most existing federated cross-modal retrieval methods rely on many semantic annotations, limiting the scalability of the retrieval model in large-scale applications. In this paper, an unsupervised federated cross-modal Hashing retrieval model is proposed to learn a cross-modal Hashing retrieval model not dependent on semantic annotations under the premise of protecting the privacy of client data. Because of the unbalanced distribution of multimodal data in a federated learning environment, local information is insufficient for the model to learn the inter-modal similarity of the overall data, which affects the retrieval performance. To solve this problem, this paper proposes a global and local intra-modal contrastive regularization, which imposes constraints on the local Hashing model of a single modality with a global Hashing model of a different modality. This ensures that the local Hashing model can fully perceive the overall semantic similarity of data and enhance the supervision of the local cross-modal hash learning process. Moreover, this paper introduces a global-local intra-modal knowledge distillation strategy to further obtain specific global knowledge of the intra-modality. Experimental results on five benchmark cross-modal retrieval datasets demonstrate the effectiveness of the proposed method.

Keywords federated learning, multimodal learning, unsupervised learning, cross-modal retrieval, unsupervised cross-modal Hashing