典型概念驱动的模态缺失深度跨模态检索

夏鑫雨[1)], 朱磊[1)*], 聂秀山[2)], 董国华[3)], 张化祥[1)]

[1)] (山东师范大学信息科学与工程学院 济南 250358)
[2)] (山东建筑大学计算机科学与技术学院 济南 250101)
[3)] (军事认知与脑科学研究所 北京 100850)
(leizhu0608@sdnu.edu.cn)

摘 要: 跨模态检索使用一种模态的数据作为查询条件, 在另一种模态中检索语义相关的数据. 绝大多数的跨模态检索方法仅适用于模态完备条件下的跨模态检索场景, 它们对缺失模态数据的处理能力仍有待提升, 为此, 提出一种典型概念驱动的模态缺失深度跨模态检索模型. 首先提出一个融合多模态预训练网络的多模态 Transformer 模型, 能在模态缺失的情况下充分地进行多模态细粒度语义交互, 提取多模态融合语义并构造跨模态子空间, 同时引导学习生成多模态典型概念; 然后使用典型概念作为跨注意力的键和值来驱动模态映射网络的训练, 使模态映射网络可以自适应地感知查询模态数据中隐含的多模态语义概念, 生成跨模态检索特征, 充分地保留训练提取的多模态融合语义. 在 Wikipedia, Pascal-Sentence, NUS-WIDE 和 XmediaNet 这 4 个基准跨模态检索数据集上的实验结果表明, 所提模型比文中对比模型的平均准确率均值分别提高了 1.7%, 5.1%, 1.6%和 5.4%. 该模型的源代码可在 https://gitee.com/MrSummer123/CPCMR 网站获得.

关键词: 深度跨模态检索; 缺失模态; 多模态 Transformer; 典型概念; 模态映射网络
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00248

Typical Concept-Driven Modality-Missing Deep Cross-Modal Retrieval

Xia Xinyu[1)], Zhu Lei[1)*], Nie Xiushan[2)], Dong Guohua[3)], and Zhang Huaxiang[1)]

[1)] (*School of Information Science and Engineering, Shandong Normal University, Jinan* 250358)
[2)] (*School of Computer Science and Technology, Shandong Jianzhu University, Jinan* 250101)
[3)] (*Institute of Military Cognition and Brain Sciences, Beijing* 100850)

Abstract: Cross-modal retrieval takes one modality data as a query and retrieves semantically relevant data in another modality. Most existing cross-modal retrieval methods are designed for scenarios with complete modality data. However, in real-world applications, incomplete modality data often exists, which these methods struggle to handle effectively. In this paper, we propose a typical concept-driven modality-missing deep cross-modal retrieval model. Specifically, we first propose a multi-modal Transformer integrated with multi-modal pretraining networks, which can fully capture the multi-modal fine-grained semantic interaction in the incomplete modality data, extract multi-modal fusion semantics and construct cross-modal subspace, and at the same time supervise the learning process to generate typical concepts. In addition, the typical concepts are used as the cross-attention key and value to drive the training of the modal mapping network, so

收稿日期: 2023-06-01; 修回日期: 2024-11-20. **基金项目**: 国家自然科学基金(62172263); 山东省自然科学基金(ZR2020YQ47, ZR2019QF002); 山东省高等学校青年创新团队基金(2019KJN040). **夏鑫雨**(1997—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为跨模态检索; **朱磊**(1987—), 男, 博士, 教授, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为大规模多媒体内容分析与检索; **聂秀山**(1981—), 男, 博士, 教授, 博士生导师, 主要研究方向为机器学习与应用、多媒体分析; **董国华**(1991—), 女, 博士, 副研究员, 主要研究方向为人工智能与类脑智能算法研发、生物信息学、信息检索; **张化祥**(1966—), 男, 博士, 教授, 博士生导师, 主要研究方向为机器学习、模式识别、遗传算法、多模态检索、图像识别和稀疏表示、大数据处理技术和行人重识别.

that it can adaptively preserve the implicit multi-modal semantic concepts of the query modality data, generate cross-modal retrieval features, and fully preserve the pre-extracted multi-modal fusion semantics. Experimental results on four benchmark cross-modal retrieval datasets—Wikipedia, Pascal-Sentence, NUS-WIDE, and XmediaNet—show that our proposed method outperforms the existing baseline models in the paper, with average precision improvements of 1.7%, 5.1%, 1.6%, and 5.4%, respectively. The source code of our method is available at: https://gitee.com/MrSummer123/CPCMR.

Key words: deep cross-modal retrieval; incomplete modality; multi-modal Transformer; typical concepts; modality mapping network

跨模态检索使用一种模态的数据作为查询条件, 在另一种模态中检索语义相关的数据, 旨在通过度量不同模态数据的相似性来检索相关多媒体实例[1-2]. 随着社交媒体平台上多模态数据呈爆炸式增长, 跨模态检索已成为人们的重要需求.

不同模态的特征通常具有语义鸿沟, 跨模态检索的主要挑战之一是弥合模态间的异构差异, 即关联异构模态的特征并度量多模态数据的语义相似性. 为了解决上述问题, 目前大多数跨模态检索工作是将多模态数据映射到共享子空间, 再进行语义相似性度量. 根据映射到共享子空间的数据形式, 可以将现有工作大致分为 2 类: 一类是基于连续实值的跨模态检索[3-6], 能够更精确地描述多模态语义信息; 另一类是基于离散二值哈希的跨模态检索[7-11], 能够更高效地进行数据存储与相似度计算.

现有的绝大多数工作中, 都理想地假设用于模型训练的每个多模态实例都是模态完整的, 即一个模态有一个样本, 其他模态存在着与之对应的样本. 然而, 由于不同模态数据在实际收集过程中的难度、成本、用户内容偏好以及面临的隐私问题不同, 容易造成模态缺失问题. 因此模态缺失环境下的跨模态检索数据是不完整的, 这不仅给多模态语义的建模带来重要挑战, 也使得模态间的异质性差距不能够很好地弥合, 限制了跨模态检索的应用.

早期的模态缺失跨模态检索工作主要在浅层学习框架上进行. Guo 等[12]提出一种协同亲和性学习方法(collective affinity learning method, CALM), 构建了模态特定的二部图, 通过推导出的概率模型计算每个模态的完整数据与锚点之间的关系, 分析证明该二部图具有恢复缺失邻接信息的能力; 针对模态数据缺失的问题, Yang 等[13]提出一种基于缺失模态的半监督多模态学习方法(semi-supervised multi-modal learning with incomplete modalities, SLIM), 利用来自无标签数据中的额外信息, 将内在一致性和外在无标注信息集成在一个统一的学习框架中; Wang 等[14]提出一个统一的哈希学习框架(learning to hash on partial multi-modal data, PM^{2}H)处理缺失模态下的跨模态检索问题, 通过图拉普拉斯算子在同一模态下保持模态相似性, 并通过潜在的公共子空间学习, 保持不同模态之间的数据一致性.

近期, 一些学者开始关注在深度网络框架内解决缺失模态跨模态检索问题. PAN[15]通过设计锚点不变损失和辨别损失来共同学习不同模态下的共同表征, 并利用传播策略重建每个模态的均衡表示, 同时保持模态的异质性和语义的一致性. 虽然在处理模态缺失问题上取得了一些进展, 但是仍然存在着 2 个重要问题: (1) 模态内局部结构与模态间数据结构的交互不够, 多模态语义建模不充分; (2) 大部分方法强制将不同模态特征映射到共享子空间, 可能导致模态内语义的缺失.

为了解决上述问题, 本文提出一个典型概念驱动的模态缺失深度跨模态检索模型, 能鲁棒地处理模态缺失问题, 进行细粒度的语义深度交互建模, 充分地保持模态映射网络语义表示的完整性. 一方面, 训练一个多模态融合 Transformer 来学习多模态融合表示. 此过程增加了模态间和模态内的交互过程, 分别在特征分布和语义级别进行对齐, 以缓解多模态语义鸿沟; 为了处理模态缺失问题, 更好地利用缺失模态数据, 提出掩码策略引导缺失模态的信息补全, 使得模型具有较好的鲁棒性; 另一方面, 基于多模态融合表示学习具有代表性的典型概念. 典型概念能够代表一类对象的核心属性或者特征, 在不同模态数据中概括共同表达的语义信息. 典型概念驱动模态映射网络学习, 增强其生成的跨模态检索特征表示语义的

完整性. 在 4 个基准跨模态检索数据集上进行实验, 验证本文模型优于文中对比的完备模态下的跨模态检索方法; 通过鲁棒性实验, 证明该模型在模态缺失跨模态检索任务上的有效性.

1　相关工作

现有的连续实值和二值哈希这 2 类跨模态检索方法中, 基本流程是将图像模态和文本模态经过线性或非线性映射函数投影到潜在的公共子空间, 通过比较它们在公共子空间的特征表示度量异构模态语义相似性 , 按照实现方式, 它们又可以分为统计学习方法和深度学习方法. 统计学习方法中, 典型相关分析(canonical correlation analysis, CCA)[16]通过典型相关分析学习图像和相关文本的关联语义表示; 在此基础上, Wang 等[17]提出核典型相关分析(kernel canonical correlation analysis, KCCA), 利用核函数方法将原始特征数据映射到更高维的核函数特征空间. 由于 KCCA 具有复杂的函数空间, 可以通过足够的数据表示更高的相关性, 以提高分类器的性能. Rasiwasia 等[18]提出聚类典型相关分析(cluster canonical correlation analysis, Cluster-CCA), 采用聚类典型相关分析方法对多模态数据进行联合降维, 学习具有辨别性的低维特征表示, 在最大化类内相关性的同时, 在子空间内使得类间差异最大; 并提出聚类-核典型相关分析(cluster-kernel canonical correlation analysis, Cluster-KCCA), 扩展了 Cluster-CCA, 以考虑非线性关系. 鉴于基于深度学习的跨模态检索方法其强大的学习能力受到广泛关注. 对应自编码器(correspondence autoencoder, Corr-AE)[19], 通过自编码器学习 2 个模态的嵌入表示, 并以最小化每个模态表示学习误差和 2 个模态嵌入表示之间的相关学习误差训练模型; Zhen 等[20]提出深度有监督跨模态检索方法(deep supervised cross-modal retrieval, DSCMR), 在标签空间和公共表示空间中最小化判别损失, 同时最小化模态不变性损失, 以学习一个公共表示空间, 使得不同模态的样本可以在此空间内直接进行比较; Peng 等[21]提出跨模态对抗生成网络(cross-modal generative adversarial networks, CM-GANs), 通过 GAN 模拟不同模态数据的联合分布, 在生成模型和辨别模型中可以同时约束模态间和模态内的语义相关性, 使得异构模态数据可以有效地关联; Huang 等[22]提出选择性多模态长短期记忆网络(selective multimodal LSTM, sm-LSTM), 设计了一个多模态上下文调节注意力方案, 可以选择性地关注图像和句子实例对, 并通过多模态 LSTM 网络进行局部相似性度量和聚合.

处理大规模多模态数据时, 使用实值特征进行跨模态检索是非常耗时的, 并且需要占用大量的存储空间. 哈希学习生成的哈希码存储要求低且检索速度快, 应对大规模数据具有很高的效率[23]. 深度跨模态哈希(deep cross-modal hashing, DCMH)[24]将哈希技术与深度学习进行结合, 基于端到端的深度学习框架, 将图像和文本模态进行特征提取并生成哈希码; 图卷积网络哈希(graph convolutional network hashing, GCH)[25]通过构建语义编码器指导特征编码器的训练, 并且通过图卷积网络挖掘数据之间的相似性; 语义深度跨模态哈希(semantic deep cross-modal hashing, SDCH)[26]利用语义标签改进特征学习部分, 既保留特征表示的语义信息, 又保持跨模态数据的不变性, 并通过哈希学习保持了汉明空间中不同模态之间哈希码的一致性; 深度联合语义重构哈希(deep joint-semantics reconstructing hashing, DJSRH)[27]构建一种新的联合语义亲和矩阵, 通过不同模态的原始邻域信息捕捉多模态数据内在的语义关联.

现有的大多数跨模态检索工作都是建立在多模态数据完备的理想状态下, 即 2 个模态间的实例是成对存在的, 在文本模态中有图像模态所对应的语义关联实例, 图像模态中有文本模态所对应的语义关联实例, 并且 2 个模态实例之间一一对应. 然而, 在实际的跨模态检索应用中, 模态间的数据并非都是一一对应的, 存在一部分图像缺少对应的文本模态描述信息, 或一部分文本缺少对应的图像模态关联的情况. 解决模态缺失环境下的跨模态检索问题时, 最直接的策略是移除不成对样本. 然而, 简单地移除部分样本会直接减少宝贵的训练样本, 降低检索性能. CALM[12]通过构建模态特定图结构进行缺失邻接信息的恢复; Wang 等[14]提出 PM^2H 引入图引导的哈希学习框架来对抗模态缺失带来的干扰; Jing 等[3]提出双对齐变分自编码器(dual-aligned variational autoencoders, DAVAE), 训练多个自编码器学习不同模态之间的潜在关系, 缓解不同模态的异构差异, 增强嵌入表示的可辨别性. 对于模态缺失实例, 利用变分自编码器来合成子空间表示特征. PAN[15]引入原型一致性建模和传播机制, 实现多模态补全. 虽然这些方法取得了一些进展, 但是它们没有充分考虑模态

内局部结构和模态间数据结构的交互，使得多模态语义建模不充分. 另外，这些方法大部分强制将不同模态特征映射到共同子空间，极易导致模态内语义的缺失.

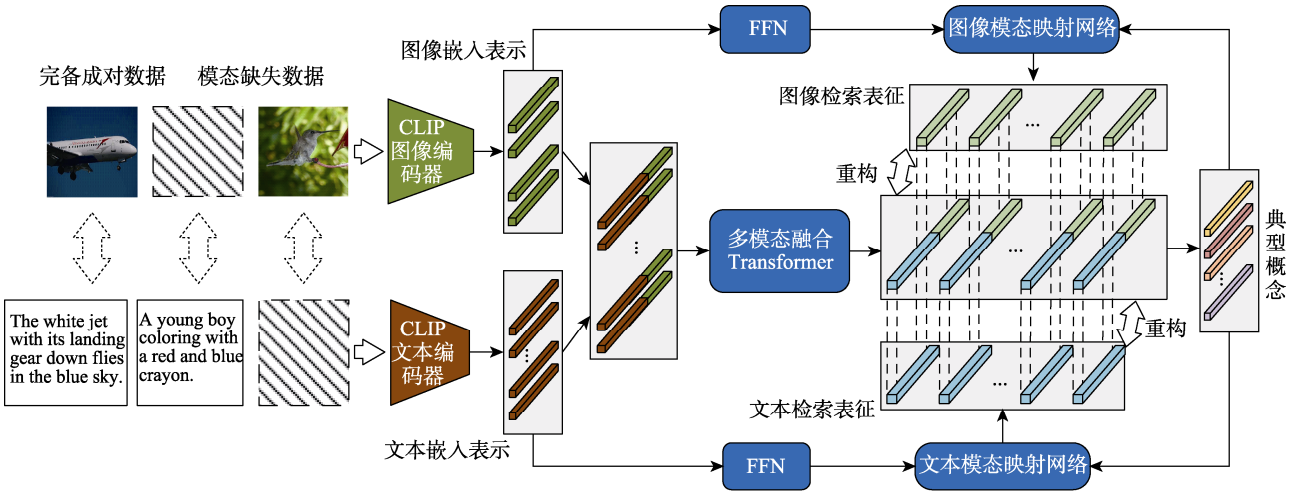
2 本文模型

本文关注缺失模态下的跨模态检索问题，重点探究如何利用不完备多模态数据提升跨模态检索模型的性能. 针对现有模态缺失跨模态检索方法的不足，主要解决 2 个问题: (1) 针对多模态数据语义建模粗糙的问题，通过实现细粒度的深度多模态语义建模和交互，增强跨模态共享空间的多模态语义表征能力; (2) 针对模态特有语义丢失的问题，通过典型概念驱动的模态映射网络学习，捕捉多模态数据的语义分布，进一步提高跨模态检索性能.

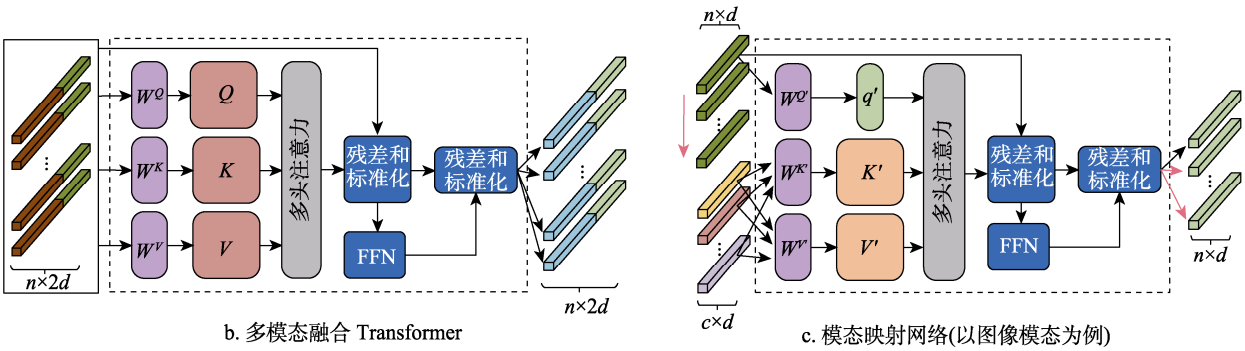
本文模型分为离线模型训练和在线检索 2 个阶段，其框架如图 1 所示. 在离线学习阶段，首先以 CLIP 模型[28]作为多模态数据的特征编码器，分别得到图像和文本的嵌入表示. 针对模态缺失的问题，提出一种掩码补全策略来生成具有模态语义一致的模态缺失样本; 同时，设计多模态融合 Transformer 网络，通过自注意力机制实现多模态数据的细粒度语义交互，学习跨模态共享空间，并通过典型概念学习增强多模态数据表征的语义判别性. 为了学习模态检索表征，设计一种基于跨注意力机制的多模态映射学习模块，利用典型概念与原始数据特征的交互，生成具有强判别性的多模态数据表征. 在线检索阶段中，典型概念驱动的模态映射网络通过前馈神经网络(feed-forward network，FFN)自适应地感知查询样本的模态特征; 然后利用预先提取的多模态融合语义，得到最终的跨模态检索表征.

2.1 符号定义与描述

给定具有 n 个实例的多模态数据集 O，其包含 n_{p} 个成对样本实例 $\{o_i\}_{i=1}^{n_{\mathrm{p}}}$ 和 n_{q} 个模态缺失实例 $\{\bar{o}_i\}_{i=1}^{n_{\mathrm{q}}}$，其中，$n_{\mathrm{q}}$ 个模态缺失实例由 n_{v} 个文本缺失实例 $\{\hat{o}_i\}_{i=1}^{n_{\mathrm{v}}}$ 和 n_{t} 个图像缺失实例 $\{\tilde{o}_i\}_{i=1}^{n_{\mathrm{t}}}$ 组成，即



a. 本文模型



b. 多模态融合 Transformer　　c. 模态映射网络(以图像模态为例)

图 1　本文模型框架

$n=n_{\mathrm{p}}+n_{\mathrm{q}}=n_{\mathrm{p}}+n_{\mathrm{v}}+n_{\mathrm{t}}$. 模态缺失跨模态检索的任务是, 综合利用完备模态数据 $\{o_i\}_{i=1}^{n_{\mathrm{p}}}$ 和不完备模态数据 $\{\overline{o}_i\}_{i=1}^{n_{\mathrm{q}}}$ 训练一个跨模态检索模型. 具体的符号定义与描述如表 1 所示.

表 1　符号定义与描述

符号	描述
v_i,t_i,l_i	成对样本实例 o_i 的图像、文本和标签
$\hat{v}_i,\hat{l}_i$	文本缺失实例的图像和标签
$\check{t}_i,\check{l}_i$	图像缺失实例的文本和标签
$\hat{t}_i,\check{v}_i$	实例的重构文本和重构图像
$f_i^{\mathrm{v}},f_i^{\mathrm{t}}$	实例的图像特征和文本特征
$\hat{f}_i^{\mathrm{v}},\hat{f}_i^{\mathrm{t}}$	实例的图像特征和占位文本特征
$\check{f}_i^{\mathrm{t}},\check{f}_i^{\mathrm{v}}$	实例的文本特征和占位图像特征
c	多模态数据集的类别数量
X-Y-Z	模态数据缺失实例, $X:Y:Z=n_{\mathrm{p}}:n_{\mathrm{v}}:n_{\mathrm{t}}$

2.2　多模态融合 Transformer

传统的多模态 Transformer 以成对的图像-文本为输入, 经过拼接或叠加等操作后生成成对数据的共享语义表示. 在实际的跨模态检索应用中, 模态缺失是经常存在的. 为了解决模态缺失条件下的多模态语义建模问题, 受到文献[29]的启发, 本文利用特征级的掩码策略生成缺失模态样本的占位特征, 该过程定义为

$$\begin{cases}\hat{\boldsymbol{f}}_i^{\mathrm{t}}=\dfrac{\lambda}{p+m^{\mathrm{t}}}\left(\displaystyle\sum_{j=1}^{s}\boldsymbol{f}_j^{\mathrm{t}}+\sum_{k=1}^{m^{\mathrm{t}}}\check{\boldsymbol{f}}_k^{\mathrm{t}}\right)\\ \check{\boldsymbol{f}}_i^{\mathrm{v}}=\dfrac{\lambda}{p+m^{\mathrm{v}}}\left(\displaystyle\sum_{j=1}^{s}\boldsymbol{f}_j^{\mathrm{v}}+\sum_{q=1}^{m^{\mathrm{v}}}\hat{\boldsymbol{f}}_q^{\mathrm{v}}\right)\end{cases}\tag{1}$$

其中, $\hat{\boldsymbol{f}}_i^{\mathrm{t}}$ 和 $\check{\boldsymbol{f}}_i^{\mathrm{v}}$ 分别表示实例 $\hat{o}_i$ 和 $\check{o}_i$ 的占位文本特征和占位图像特征; s 表示对应批次下成对实例 o_i 的样本个数; m^{t} 和 m^{v} 分别表示对应批次下图像缺失实例 $\check{o}_i$ 和文本缺失实例 $\hat{o}_i$ 的样本个数; λ 表示缩放因子. 因此, 通过掩码占位策略, 数据集 O 能够被扩展为标准完备多模态数据集,

$$O'=\{v_i,t_i,l_i\}_{i=1}^{n_{\mathrm{p}}}\bigcup\{\hat{v}_i,\hat{t}_i,\hat{l}_i\}_{i=1}^{n_{\mathrm{v}}}\bigcup\{\check{v}_i,\check{t}_i,\check{l}_i\}_{i=1}^{n_{\mathrm{t}}}.$$

基于标准的 Transformer 架构, 设计多模态融合 Transformer, 建模图像-文本对的跨模态对应关系以及成对样本实例间的细粒度语义关联. 在拼接特征编码器的输出 $\boldsymbol{f}^{\mathrm{v}},\boldsymbol{f}^{\mathrm{t}}\in\mathbb{R}^D$ 后, 多模态融合 Transformer 以多模态特征 $\boldsymbol{f}\in\mathbb{R}^D$ 为输入, 并利用自注意力机制生成多模态数据表征. 在自注意力机制中, 多模态特征 $\boldsymbol{f}$ 被用来构造查询、键和值, 公式为

$$\begin{cases}\boldsymbol{q}_i=\boldsymbol{f}_i\boldsymbol{W}^{\mathrm{Q}}\\ \boldsymbol{k}_j=\boldsymbol{f}_j\boldsymbol{W}^{\mathrm{K}}\\ \boldsymbol{v}_j=\boldsymbol{f}_j\boldsymbol{W}^{\mathrm{V}}\end{cases}.$$

其中, $\boldsymbol{W}^{\mathrm{Q}}\in\mathbb{R}^{D\times D_{\mathrm{k}}},\boldsymbol{W}^{\mathrm{K}}\in\mathbb{R}^{D\times D_{\mathrm{k}}},\boldsymbol{W}^{\mathrm{V}}\in\mathbb{R}^{D\times D_{\mathrm{v}}}$ 分别表示自注意力机制中查询、键和值的参数矩阵; D_{k}, D_{v} 分别表示键和值的维度; $\boldsymbol{f}_i$ 表示查询模态数据的特征; $\boldsymbol{f}_j$ 表示键、值模态数据的特征. 对于任一图像-文本对, 多模态融合 Transformer 能够生成表征能力更强的多模态特征, 公式为

$$\boldsymbol{f}_i'=A_{\mathrm{self}}(\boldsymbol{q}_i,\boldsymbol{K},\boldsymbol{V})=\sum_{j=1}^{m}\mathrm{Softmax}\left(\boldsymbol{q}_i\boldsymbol{k}_j^{\mathrm{T}}\Big/\sqrt{D_k}\right)\boldsymbol{v}_j.$$

其中, m 表示对应批次下的样本个数; $A_{\mathrm{self}}(\cdot)$ 表示自注意力机制; $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别表示键矩阵和值矩阵. 进一步, 利用 FFN 和直连结构生成数据集 O' 的最终多模态表征, 公式为

$$\boldsymbol{z}_i=R_2\left(R_1\left(\boldsymbol{f}_i'\right)\right).$$

其中, $\boldsymbol{z}_i=\left\{\boldsymbol{z}_i^{\mathrm{v}},\boldsymbol{z}_i^{\mathrm{t}}\right\}\in\mathbb{R}^{2D}$, 表示多模态融合 Transformer 最终输出的多模态特征;

$$R_1(\cdot)=N_{\mathrm{norm}}\left(D_{\mathrm{drop}}(\cdot)+\boldsymbol{f}_i\right);$$

$$R_2(\cdot)=N_{\mathrm{norm}}\left(D_{\mathrm{drop}}\left(\mathrm{FFN}(\cdot)\right)+(\cdot)\right);$$

$N_{\mathrm{norm}}(\cdot)$ 表示层正则化函数; $D_{\mathrm{drop}}(\cdot)$ 表示 Dropout 函数.

为了确保同一模态内相同类别的数据表征包含一致的类别语义, 将 $\left\{\boldsymbol{z}_i^{\mathrm{v}},\boldsymbol{z}_i^{\mathrm{t}}\right\}$ 进行线性映射, 再利用交叉熵损失分别监督图像模态和文本模态的一致性表征学习过程, 公式分别为

$$L_{\mathrm{recon}}^{\mathrm{v}}=-\frac{1}{m}\sum_{i=1}^{m}\ln\left(\mathrm{e}^{\boldsymbol{w}_{l_i}^{\mathrm{T}}\boldsymbol{z}_i^{\mathrm{v}}+\boldsymbol{b}_{l_i}}\Big/\sum_{j=1}^{c}\mathrm{e}^{\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z}_i^{\mathrm{v}}+\boldsymbol{b}_j}\right),$$

$$L_{\mathrm{recon}}^{\mathrm{t}}=-\frac{1}{m}\sum_{i=1}^{m}\ln\left(\mathrm{e}^{\boldsymbol{w}_{l_i}^{\mathrm{T}}\boldsymbol{z}_i^{\mathrm{t}}+\boldsymbol{b}_{l_i}}\Big/\sum_{j=1}^{c}\mathrm{e}^{\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z}_i^{\mathrm{t}}+\boldsymbol{b}_j}\right).$$

其中, $\boldsymbol{w}_j$ 和 $\boldsymbol{b}_j$ 分别表示第 j 类样本的映射权重和偏置向量. 因此, 最终的多模态交叉熵损失被定义为

$$L_{\mathrm{class}}=L_{\mathrm{class}}^{\mathrm{v}}+L_{\mathrm{class}}^{\mathrm{t}}=-\frac{1}{m}\sum_{i=1}^{m}\left(\ln\frac{\mathrm{e}^{\boldsymbol{w}_{l_i}^{\mathrm{T}}\boldsymbol{z}_i^{\mathrm{v}}+\boldsymbol{b}_{l_i}}}{\sum_{j=1}^{c}\mathrm{e}^{\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z}_i^{\mathrm{v}}+\boldsymbol{b}_j}}+\ln\frac{\mathrm{e}^{\boldsymbol{w}_{l_i}^{\mathrm{T}}\boldsymbol{z}_i^{\mathrm{t}}+\boldsymbol{b}_{l_i}}}{\sum_{j=1}^{c}\mathrm{e}^{\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z}_i^{\mathrm{t}}+\boldsymbol{b}_j}}\right)\tag{2}$$

2.3　典型概念学习

由于成对图像-文本中的细粒度语义并非一一

对应, 并且掩码策略引导生成的模态缺失样本携带着同一模态内的多类别语义信息, 因此, 成对的多模态样本往往存在着大量的非对称语义和噪声信息. 为了获得多模态数据的去噪特征表示并增强多模态语义判别性, 设计典型概念学习, 进一步增强多模态数据特征的语义表征能力. 典型概念指在模态数据中普遍存在且具有代表性的特征或属性, 这些特征或属性在不同模态间具有高度的一致性和相关性.

首先定义多模态数据集 O' 的典型概念为 $\boldsymbol{P}=\left\{\boldsymbol{p}_i \middle| \boldsymbol{p}_i \in \mathbb{R}^D, i=1,2,\cdots,c\right\}$, 随机初始化 c 个典型概念 $\{\boldsymbol{p}_i\}_{i=1}^c$. 在训练阶段, 通过协同优化多模态融合 Transformer 的输出特征 $\{\boldsymbol{z}_i\}_{i=1}^m$ 和典型概念 $\{\boldsymbol{p}_i\}_{i=1}^c$ 来实现 2 个目标: (1) 优化多模态特征 $\{\boldsymbol{z}_i\}_{i=1}^m$, 使其分布在对应的典型概念附近, 从而使该特征的类间差异最大且类内差异最小; (2) 优化典型概念 $\{\boldsymbol{p}_i\}_{i=1}^c$, 使其保留成对多模态数据的公共语义信息, 去除典型概念中的多模态噪声信息. 为了实现多模态特征 $\{\boldsymbol{z}_i\}_{i=1}^m$ 和典型概念 $\{\boldsymbol{p}_i\}_{i=1}^c$ 的协同优化, 多模态均方误差损失被定义为

$$L_{\text{center}}=\frac{1}{m}\sum_{i=1}^{c}\sum_{j}\left[\left(\boldsymbol{z}_{i,j}^{\text{v}}-\boldsymbol{p}_i\right)^2+\left(\boldsymbol{z}_{i,j}^{\text{t}}-\boldsymbol{p}_i\right)^2\right] \quad (3)$$

其中, p_i 表示第 i 个类别的典型概念特征; $\boldsymbol{z}_{i,j}^{\text{v}},\boldsymbol{z}_{i,j}^{\text{t}}$ 分别表示属于第 i 个类别中第 j 个实例的图像和文本. CLIP4CMR[4]通过 2 个独立的模态编码器学习原型, 约束多模态特征的生成. 与 CLIP4CMR 不同, 本文通过典型概念学习多模态类表征, 它们的主要区别在于: (1) 学习方式不同. CLIP4CMR 利用多个模态语义单独约束类别原型学习, 而本文的典型概念通过多模态融合 Transformer 学习得到且具有多模态融合语义; (2) 学习目的不同. CLIP4CMR 将待检索模态表征映射到相应类别的原型附近, 而本文的典型概念作为跨注意力机制的键和值辅助模态映射网络生成模态检索特征.

2.4 模态映射网络

在跨模态在线检索过程中, 模态映射网络用于生成查询模态数据的检索特征(然后直接用于相似度计算). 在学习模态映射网络时, 大部分跨模态检索方法, 强制将模态映射特征与跨模态子空间特征进行对齐. 研究表明[10], 这种强制性的语义对齐容易造成语义丢失. 与已有方法不同, 本文基于学习到的典型概念, 提出基于跨注意力机制的模态映射网络学习模态检索表征. 与自注意力机制不同, 本文提出的跨注意力机制通过查询模态数据提取检索特征, 同时利用学习到的典型概念构造跨注意力机制的键和值. 该过程的公式为

$$\begin{cases}\boldsymbol{q}_i'=\text{FFN}\left(\boldsymbol{f}_j^*\right)\boldsymbol{W}^{\text{Q}'}\\ \boldsymbol{k}_j'=\boldsymbol{p}_i\boldsymbol{W}^{\text{K}'}\\ \boldsymbol{v}_j'=\boldsymbol{p}_i\boldsymbol{W}^{\text{V}'}\end{cases}.$$

其中, $\boldsymbol{W}^{\text{Q}'},\boldsymbol{W}^{\text{K}'},\boldsymbol{W}^{\text{V}'}$ 分别表示跨注意力机制中查询、键和值的参数矩阵; $\boldsymbol{f}_j^*\in\left\{\boldsymbol{f}_j^{\text{v}},\boldsymbol{f}_j^{\text{t}}\right\}$ 表示查询模态数据的特征. 借助跨注意力机制和典型概念, 模态映射网络根据

$$\boldsymbol{h}_i^*=C_{\text{att}}\left(\boldsymbol{q}_i',\boldsymbol{K}',\boldsymbol{V}'\right)=\sum_{j=1}^{c}\text{Softmax}\left(\boldsymbol{q}_i'\boldsymbol{k}_j'^{\text{T}}\Big/\sqrt{d_{k'}}\right)\boldsymbol{v}_j' \quad (4)$$

生成模态检索特征. 其中, C_{att} 表示跨注意力机制, $\boldsymbol{K}'$ 和 $\boldsymbol{V}'$ 分别表示键矩阵和值矩阵. 通过计算每个查询模态数据和典型概念之间的关联, 并根据关联关系自适应地组合典型概念, 生成模态检索特征. 由于典型概念本身包含典型的多模态语义, 式(4)能保存预先提取的多模态融合语义, 当某一模态的数据缺失时, 通过强化其他模态中与“典型概念”相关的特征弥补信息的不足, 保持检索的准确性和鲁棒性. 通过交叉熵保持多模态特征 $\{\boldsymbol{h}_i^{\text{v}},\boldsymbol{h}_i^{\text{t}}\}_{i=1}^{\text{m}}$ 与多模态融合语义特征 $\{\boldsymbol{z}_i^{\text{v}},\boldsymbol{z}_i^{\text{t}}\}_{i=1}^{\text{m}}$ 的一致性, 基于交叉熵损失的多模态映射约束被定义为

$$L_{\text{recon}}=L_{\text{recon}}^{\text{v}}+L_{\text{recon}}^{\text{t}}=\\ \text{KL}\left[p\left(\boldsymbol{z}^{\text{v}}\right)\|q\left(\boldsymbol{h}^{\text{v}}\right)\right]+\text{KL}\left[p\left(\boldsymbol{z}^{\text{t}}\right)\|q\left(\boldsymbol{h}^{\text{t}}\right)\right] \quad (5)$$

2.5 在线跨模态检索

在线检索过程中, 查询文本或图像首先利用预训练模型 CLIP 作为特征编码器, 得到对应的嵌入表示. 在 FFN 进一步提取嵌入表示后, 模态映射网络自适应地感知嵌入表示和预先提取的多模态融合语义, 最终得到查询模态表征 $\boldsymbol{h}_i^{\text{q},*}$. 该过程定义为

$$\begin{cases}\boldsymbol{q}_i^{\text{q},*}=\text{FFN}\left(\boldsymbol{f}_j^{\text{q},*}\right)\boldsymbol{W}^{\text{Q}*}\\ \boldsymbol{k}_j'=\boldsymbol{p}_i\boldsymbol{W}^{\text{K}'}\\ \boldsymbol{v}_j'=\boldsymbol{p}_i\boldsymbol{W}^{\text{V}'}\\ \boldsymbol{h}_i^{\text{q},*}=\sum_{j=1}^{C}\text{Softmax}\left(\boldsymbol{q}_i^{\text{q},*}\boldsymbol{k}_j'^{\text{T}}\Big/\sqrt{d_{k'}}\right)\boldsymbol{v}_j'\end{cases}.$$

其中, $\boldsymbol{f}_j^{\text{q},*}$ 表示查询模态数据, 包括查询文本特征 $\boldsymbol{f}_j^{\text{q,t}}$ 和查询图像特征 $\boldsymbol{f}_j^{\text{q,v}}$.

2.6　优化目标和算法

基于典型概念学习的多模态 Transformer 包含多模态融合 Transformer、典型概念学习和模态映射网络 3 个模块, 因此, 整体多模态 Transformer 框架可以进行端到端优化, 其总体目标函数表示为

$$L=\alpha L_{\text{class}}+\beta L_{\text{center}}+\gamma L_{\text{recon}} \tag{6}$$

其中, α, β, γ 表示平衡各损失项的超参数.

算法 1. 本文算法的整体框架.

输入. 多模态数据集 $\{o_i\}_{i=1}^{n}$, 其中, 包含 n_{p} 个成对样本实例, n_{v} 个文本缺失实例和 n_{t} 个图像缺失实例, 多模态融合 Transformer 的学习率 ω_{F}, 概念原型驱动的模态映射网络的学习率 ω_{P}, 超参数 α,β,γ.

输出. 多模态融合 Transformer 参数 Θ_{F}、概念原型驱动的模态映射网络的参数(分别对应于图像模态和文本模态) $\Theta_{\text{P}}^{\text{v}},\Theta_{\text{P}}^{\text{t}}$.

初始化. 初始化多模态融合 Transformer 参数 Θ_{F}、图像模态映射网络参数 $\Theta_{\text{P}}^{\text{v}}$ 和文本模态映射网络参数 $\Theta_{\text{P}}^{\text{t}}$.

WHILE　$t\leqslant$最大迭代次数

Step1. 从数据集 $\{o_i\}_{i=1}^{n}$ 按批次取数据.

Step2. 根据式(1)计算模态缺失数据所在批次中成对样本均值作为缺失样本掩码.

Step3. 通过多模态融合 Transformer 学习特征表示, 根据式(2)计算多模态交叉熵损失 L_{class}.

Step4. 通过概念原型学习, 根据式(3)计算多模态均方误差损失 L_{center}.

Step5. 通过概念原型驱动的多模态映射网络生成对应多模态表征 $\boldsymbol{h}^{\text{v}},\boldsymbol{h}^{\text{t}}$, 根据式(5)计算映射损失 L_{recon}.

Step6. 按照式(6)计算总体损失函数.

Step7. 根据梯度下降法更新网络参数

$$\left(\Theta_{\text{F}}\right)_{t+1}=\left(\Theta_{\text{F}}\right)_{t}-\omega_{\text{F}}\frac{\partial L}{\left(\partial\Theta_{\text{F}}\right)_{t}},$$

$$\left(\Theta_{\text{P}}^{\text{v}}\right)_{t+1}=\left(\Theta_{\text{P}}^{\text{v}}\right)_{t}-\omega_{\text{P}}\frac{\partial L}{\left(\partial\Theta_{\text{P}}^{\text{v}}\right)_{t}},$$

$$\left(\Theta_{\text{P}}^{\text{t}}\right)_{t+1}=\left(\Theta_{\text{P}}^{\text{t}}\right)_{t}-\omega_{\text{P}}\frac{\partial L}{\left(\partial\Theta_{\text{P}}^{\text{t}}\right)_{t}}.$$

END WHILE

3　实验及结果分析

3.1　数据集

本文在 Wikipedia[30], Pascal-Sentence[31], NUS-WIDE-10K[32]和 XmediaNet[33]这 4 个公开的跨模态检索数据集上进行实验, 这些数据集已被广泛用于跨模态检索方法的性能评测[3-5,7-10]. 与这些方法一致, 本文将每个数据集划分为训练集和测试集. 首先基于训练集训练跨模态检索模型, 然后在测试集中测试跨模态检索模型的性能. 为了使实验性能对比更加公平可信, 保持与先前工作[4]相同的数据划分. 数据集的划分如下:

(1) Wikipedia 数据集[30]收集了来自 Wikipedia 网站中 10 个类别的图像-文本对, 每个图像-文本对属于一个类别且由一幅图像和一段文本描述组成. 实验中, 采用 2 157 个图像-文本对作为训练集, 462 个图像-文本对作为测试集.

(2) Pascal-Sentence 数据集[31]收集了来自 2008 PASCAL development kit 中 20 个类别的图像-文本对, 其中, 每幅图像对应 5 个句子描述. 实验中, 采用 800 个图像-文本对作为训练集, 100 个图像-文本对作为测试集.

(3) NUS-WIDE-10K 数据集[32]共包含 10 个类别的图像-文本对. 实验中, 将 NUS-WIDE-10K 划分为 2 部分: 8 000 个样本对作为训练集, 1 000 个样本对作为测试集.

(4) XmediaNet 数据集[33]共包含 200 个类别的图像-文本对, 实验中, 采用 32 000 个图像-文本对作为训练集, 4 000 个图像-文本对作为测试集.

3.2　实验设置

按照标准的 Transformer 编码器架构设计多模态融合 Transformer, 其中, Transformer 中多头自注意力的头部数量为 2, 编码器数量为 2. 初始化典型概念时, 随机生成 1 024 维的特征向量并做归一化处理. 典型概念驱动的模态映射网络中, 利用基于跨注意力机制的 Transformer 模块生成 1 024 维度特征表示, 其中, 模态映射网络中多头跨注意力的头部数量设为 2, 并在网络输出层之后使用归一化生成最终的多模态特征表示.

模型训练参数设置如下: 在训练阶段, 采用 ADAM 优化器并设置 batch size 为 128, 多模态融合 Transformer 的学习率为 0.000 1, 典型概念驱动的 Transformer 学习率为 0.001, 掩码取值缩放因子为−1. 在 NUS-WIDE 数据集上, 训练轮次为 80; 其余数据集上的训练轮次均为 100.

3.3　评价指标

与已有跨模态检索方法[3-5,7-10]一致, 本文采用平均准确率均值(mean average precision, mAP)作为跨模态检索性能的评价指标. 跨模态检索是双

向检索, 包括图像检索文本(image to text, I2T)和文本检索图像(text to image, T2I). 以 I2T 为例, 首先计算每个图像查询的检索结果, 根据与图像查询样本的相似度, 把数据库中的文本数据从高到低进行排序并计算平均准确率(average precision, AP); 然后计算所有图像查询检索结果 AP 的平均值, 得到最终的检索性能指标 mAP. AP 的计算公式为

$$\mathrm{AP}=\frac{1}{P}\sum_{k=1}^{n}\frac{P_k}{k}\times G(k).$$

其中, P 表示测试集中与给定查询样本相似的样本数量; n 表示测试集的样本数量; $p(k)$ 表示前 k 个检索结果中与查询样本相似的样本数量; $G(k)$ 是一个指示函数, 当第 k 个检索样本与查询样本相似时, 其值为 1, 反之则为 0.

3.4 模态完备跨模态检索

为了验证本文模型在完备模态跨模态检索任务上的有效性, 选取 16 种跨模态检索方法作为对比方法, 包括 CCA[16], KCCA[17]和 JRL[5]这 3 种传统统计学习方法, 以及 Corr-AE[19], CMDN[34], JFSSL[35], ACMR[30], JLSLR[36], MCSM[33], CCL[37], CM-GANs[6], DSCMR[20], PAN[15], MCCN[38], C^{3}CMR[39]和 CLIP4CMR[4]这 13 种深度学习方法.

在 4 个基准数据集上, 将本文模型与 16 种跨模态检索方法进行实验, mAP 的对比结果如表 2 所示, 其中, 实验数据均来自文献[4]. 可以看出, 与对比方法相比, 本文模型均获得最优结果; 特别地, 在 Pascal-Sentence 和 XmediaNet 数据集上, 与 MCCN[38]相比, 本文模型的 mAP 平均值分别提高 9.4%和 5.4%, 与最优对比方法 CLIP4CMR 相比, 本文模型的 mAP 平均值分别提高 5.1%和 5.4%; 深度学习方法比传统机器学习方法的检索性能有明显的提升, 说明深度跨模态检索方法借助深度神经网络强大的特征表示能力, 能够在跨模态检索任务上展现优越的性能.

表 2 在 4 个基准数据集上 17 种方法的 mAP 值对比

方法	Wikipedia			Pascal-Sentence			NUS-WIDE			XmediaNet		
	I2T	T2I	平均值	I2T	T2I	平均值	I2T	T2I	平均值	I2T	T2I	平均值
CCA[16]	0.298	0.273	0.286	0.203	0.208	0.206	0.167	0.181	0.174	0.212	0.217	0.215
KCCA[17]	0.438	0.389	0.414	0.488	0.446	0.467	0.351	0.356	0.354	0.252	0.270	0.261
Corr-AE[19]	0.442	0.429	0.436	0.532	0.521	0.527	0.441	0.494	0.468	0.469	0.507	0.488
JRL[5]	0.479	0.428	0.454	0.563	0.505	0.534	0.466	0.499	0.483	0.488	0.405	0.447
CMDN[34]	0.487	0.427	0.457	0.544	0.526	0.535	0.492	0.542	0.517	0.485	0.516	0.501
JFSSL[35]	0.458	0.426	0.442	0.553	0.542	0.548	0.514	0.523	0.519	0.525	0.518	0.521
ACMR[30]	0.468	0.412	0.440	0.538	0.544	0.541	0.519	0.542	0.531	0.536	0.519	0.528
JLSLR[36]	0.473	0.440	0.456	0.568	0.551	0.560	0.536	0.531	0.534	0.544	0.553	0.549
MCSM[33]	0.516	0.458	0.487	0.598	0.598	0.598	0.522	0.546	0.534	0.540	0.550	0.545
CCL[37]	0.505	0.457	0.481	0.576	0.569	0.573	0.506	0.535	0.521	0.537	0.528	0.533
CM-GANs[6]	0.521	0.466	0.494	0.603	0.604	0.604	0.536	0.551	0.543	0.567	0.551	0.559
PAN[15]	0.517	0.462	0.489	0.686	0.689	0.688	0.590	0.571	0.581	0.669	0.660	0.665
DSCMR[20]	0.521	0.478	0.499	0.674	0.682	0.678	0.611	0.615	0.613	0.697	0.693	0.695
MCCN[38]	0.552	0.487	0.520	0.705	0.712	0.709				0.741	0.743	0.742
C^{3}CMR[39]	0.561	0.475	0.518	0.702	0.715	0.708	0.573	0.589	0.581	0.748	0.744	0.746
CLIP4CMR[4]	0.592	0.574	0.583	0.742	0.733	0.738	0.609	0.621	0.615	0.746	0.758	0.742
本文	0.606	0.581	0.593	0.780	0.772	0.776	0.617	0.633	0.625	0.784	0.781	0.782

3.5 模态缺失跨模态检索

3.5.1 数据划分

为了更全面地测试缺失模态下的跨模态检索性能, 在不同模态缺失率下优化跨模态检索模型. 例如, 50%-25%-25%表示训练集样本对的 50%数据包含图像和文本; 25%的训练集样本只包含图像, 没有文本数据; 剩下的 25%只包含文本, 没有图像数据. 在模态缺失的跨模态检索实验中, 共生成 10 种不同缺失率下的数据集, 分别为 50%-50%-0%, 50%-0%-50%, 50%-25%-25%, 30%-70%-0%, 30%-0%-70%, 30%-35%-35%, 10%-90%-0%, 10%-0%-90%, 10%-45%-45%和 100%-0%-0%.

3.5.2 与不同方法性能对比

为了验证本文模型在缺失跨模态检索任务中

的有效性, 与当前基于 CLIP 多模态预训练模型且针对模态缺失跨模态检索任务的主流方法 CLIP4CMR[4]进行实验, 结果如表 3 所示. 为了使实验结果更加可靠, 沿用 CLIP4CMR 的数据集和实验设置; 每个实验重复 5 次, 取 5 次的均值作为最终结果.

从表 3 可以看出, 本文模型在 4 个基准数据集上的 mAP 均超过对比方法, 详细分析如下:

(1) 在 Wikipedia 数据集上, 随着模态缺失率的不断增高, 2 种方法的 mAP 值都在逐渐降低; 但是, 即使在模态缺失比较严重的实验设置下, 本文模型跨模态检索的 mAP 值最大下降幅度仅为 8.6%, 能够保证一定的跨模态检索性能.

表 3　在 4 个基准数据集上不同缺失率下 2 种方法的 mAP 值对比

缺失率	Wikipedia		Pascal-Sentence		NUS-WIDE		XmediaNet	
	CLIP4CMR[4]	本文	CLIP4CMR[4]	本文	CLIP4CMR[4]	本文	CLIP4CMR[4]	本文
50%-50%-0%	0.575	0.584	0.726	0.753	0.595	0.623	0.735	0.773
50%-0%-50%	0.574	0.586	0.726	0.755	0.604	0.622	0.745	0.773
50%-25%-25%	0.574	0.585	0.709	0.758	0.600	0.625	0.744	0.779
30%-70%-0%	0.561	0.566	0.684	0.741	0.581	0.614	0.713	0.755
30%-0%-70%	0.566	0.580	0.664	0.737	0.595	0.616	0.736	0.765
30%-35%-35%	0.573	0.582	0.717	0.746	0.599	0.626	0.736	0.777
10%-90%-0%	0.539	0.542	0.569	0.637	0.540	0.593	0.615	0.701
10%-0%-90%	0.522	0.559	0.600	0.649	0.572	0.608	0.685	0.728
10%-45%-45%	0.570	0.570	0.685	0.725	0.592	0.623	0.730	0.772
100%-0%-0%	0.580	0.593	0.738	0.776	0.612	0.625	0.755	0.782

(2) 在 Pascal-Sentence 数据集上, 随着模态缺失率的不断增加, mAP 出现较为明显的下降趋势. 在不同模态缺失率下, CLIP4CMR 和本文模型的 mAP 最大差值分别为 0.169 和 0.139, 说明本文模型的性能更为稳定, 模型的鲁棒性更强.

(3) 在 NUS-WIDE 数据集上, 本文模型在不同模态缺失率下表现稳定, 不同模态缺失率下 mAP 值的最大差值仅为 0.032, CLIP4CMR 的最大差值为 0.072, 进一步说明本文模型在不同数据集、不同模态缺失率下的效果比 CLIP4CMR 方法更加稳定.

(4) XmediaNet 是 4 个基准数据集中数据量最大的数据集, 跨模态检索 mAP 值也是 4 个基准数据集中最高的. 随着训练集数据量的增大, 不同模态缺失率下, CLIP4CMR 的 mAP 最大差值为 0.140, 而本文模型的 mAP 最大差值为 0.081, 证明本文模型具有更优越的检索性能且对模态缺失具有更强的鲁棒性.

为了直观地展示本文模型性能的变化趋势, 分别在 Pascal-Sentence 和 NUS-WIDE 数据集上绘制了 Top100-mAP, 实验结果如表 4 所示.

此外, 对传统跨模态检索在模态缺失数据集上不同缺失比率下的性能进行实验, 结果如表 5 所示. 可以看出随着缺失率的增加, 传统的跨模态检索方法性能急剧降低; 而本文模型即使在缺失率较为严重的情况下, 依然能够保持稳定的性能.

表 4　2 种数据集上不同缺失率下本文模型 Top100-mAP

缺失率	Pascal-Sentence		NUS-WIDE	
	I2T	T2I	I2T	T2I
50%-25%-25%	0.752	0.744	0.703	0.651
30%-35%-35%	0.746	0.734	0.710	0.660
10%-45%-45%	0.720	0.732	0.704	0.656
100%-0%-0%	0.778	0.760	0.699	0.653

3.6　消融实验

通过逐个移除本文模型中的关键模块, 验证该模型的有效性.

(1) 移除多模态融合 Transformer 模块. 为了使图像和文本模态的向量映射到公共子空间, 使用 Transformer 中的标准 FFN 结构代替多模态融合 Transformer 模块, 消融实验结果如表 6 所示. 可以看出, 使用多模态融合 Transformer 模块能够有效地获取不同模态之间的信息, 使 2 个模态的表示保持语义一致, 获得更优越的跨模态检索结果.

(2) 典型概念驱动的模态映射网络的消融实验结果如表 7 所示, 其中, FFN 表示 Transformer 中的标准 FFN 结构, Cro-att 表示仅使用跨注意力机制的 Transformer 结构, FFN+Cro-att 表示典型概念驱动的模态映射网络. 可以看出, 典型概念驱动的模态映射网络能够更好地捕获多模态语义, 生成判别力更强的多模态表示特征.

表 5 在不同缺失率下 8 种方法的 mAP 比较

方法	50%-25%-25%			30%-35%-35%			10%-45%-45%		
	I2T	T2I	平均值	I2T	T2I	平均值	I2T	T2I	平均值
JFSSL[35]	0.129	0.133	0.131	0.128	0.129	0.129	0.124	0.132	0.128
FSH[40]	0.152	0.147	0.150	0.146	0.143	0.145	0.145	0.139	0.142
JRL[5]	0.489	0.479	0.484	0.398	0.367	0.382	0.112	0.176	0.144
SCM[41]	0.560	0.614	0.587	0.546	0.578	0.562	0.466	0.485	0.476
MTFH[42]	0.406	0.642	0.524	0.366	0.609	0.487	0.125	0.109	0.117
SDML[43]	0.449	0.441	0.445	0.442	0.444	0.443	0.376	0.354	0.365
DSCMR[20]	0.679	0.695	0.687	0.650	0.670	0.660	0.556	0.565	0.560
本文	0.762	0.754	0.758	0.747	0.745	0.746	0.729	0.720	0.725

表 6 多模态融合 Transformer 模块的消融实验 mAP 结果对比

成对数据率/%	Wikipedia		Pascal-Sentence	
	融合模块	FFN	融合模块	FFN
50	0.585	0.578	0.758	0.747
30	0.582	0.576	0.746	0.731
10	0.570	0.571	0.725	0.722
100	0.593	0.587	0.776	0.757

(3) 针对模态缺失样本补全的消融实验结果如表 8 所示, 其中, Paired+Only 表示模态缺失数据补全后的跨模态检索性能, Paired 表示仅利用完备模态数据进行模型优化后的跨模态检索性能. 可以看出, 使用掩码补全策略的检索方法比仅利用完备模态数据的检索方法具有更加优越的跨模态检索性能, 证明了本文提出的掩码补全策略能够增强跨模态模型的多模态语义表征能力和检索性能.

表 7 典型概念驱动的模态映射网络的消融实验 mAP 结果对比

缺失率	Wikipedia			Pascal-Sentence			NUS-WIDE			XmediaNet		
	FFN	Cro-att	FFN+Cro	FFN	Cro-att	FFN+Cro	FFN	Cro-att	FFN+Cro	FFN	Cro-att	FFN+Cro
50%-50%-0%	0.568	0.569	0.584	0.719	0.758	0.753	0.613	0.572	0.623	0.712	0.746	0.773
50%-0%-50%	0.573	0.575	0.586	0.729	0.746	0.755	0.616	0.587	0.622	0.704	0.756	0.773
50%-25%-25%	0.571	0.582	0.585	0.731	0.755	0.758	0.619	0.621	0.625	0.714	0.763	0.779
30%-70%-0%	0.557	0.552	0.566	0.696	0.744	0.741	0.602	0.605	0.614	0.694	0.724	0.755
30%-0%-70%	0.560	0.560	0.580	0.704	0.730	0.736	0.609	0.612	0.616	0.690	0.724	0.755
30%-35%-35%	0.567	0.570	0.582	0.713	0.742	0.746	0.618	0.622	0.626	0.710	0.749	0.777
10%-90%-0%	0.516	0.507	0.542	0.578	0.657	0.637	0.583	0.595	0.593	0.633	0.651	0.701
10%-0%-90%	0.531	0.521	0.559	0.598	0.646	0.649	0.595	0.546	0.608	0.636	0.674	0.728
10%-45%-45%	0.562	0.571	0.570	0.690	0.728	0.725	0.614	0.619	0.623	0.713	0.721	0.772
100%-0%-0%	0.584	0.581	0.593	0.746	0.770	0.776	0.621	0.614	0.625	0.715	0.786	0.782

表 8 模态缺失样本补全的消融实验 mAP 结果对比

缺失率	Wikipedia		Pascal-Sentence		NUS-WIDE		XmediaNet	
	Paired+Only	Paired	Paired+Only	Paired	Paired+Only	Paired	Paired+Only	Paired
50%-50%-0%	0.584	0.572	0.753	0.734	0.623	0.611	0.773	0.756
50%-0%-50%	0.586	0.572	0.755	0.734	0.622	0.611	0.773	0.756
50%-25%-25%	0.585	0.572	0.758	0.734	0.625	0.612	0.779	0.754
30%-70%-0%	0.566	0.555	0.741	0.710	0.614	0.602	0.755	0.735
30%-0%-70%	0.580	0.555	0.737	0.710	0.616	0.602	0.765	0.735
30%-35%-35%	0.582	0.558	0.746	0.697	0.626	0.605	0.777	0.734
10%-90%-0%	0.542	0.514	0.637	0.593	0.593	0.574	0.701	0.667
10%-0%-90%	0.559	0.514	0.649	0.593	0.608	0.574	0.728	0.667
10%-45%-45%	0.570	0.521	0.725	0.563	0.623	0.569	0.772	0.665
100%-0%-0%	0.593	0.593	0.776	0.776	0.625	0.625	0.782	0.782

(4) 在 Wikipedia 和 Pascal-Sentence 数据集上, 模态样本缺失率为 50%-50%-0%, 50%-25%-25%, 30%-70%-0%, 30%-35%-35%, 10%-90%-0% 和 10%-45%-45%下, 对不同掩码取值时的跨模态检索模型进行消融实验, 结果如表 9 和表 10 所示. 可以看出, 在 2 个数据集上, 当缩放因子 λ 为 1, −1 和 0 时, 掩码策略均能发挥较优性能; 为了体现采取式(1)策略生成占位特征的优越性, 选取占位掩码固定填充为−1, 1 和 100 的性能效果作为对比; 综合比较这些掩码取值, 当 λ=−1 时, 模态缺失补全的跨模态检索方法能够取得最优的检索性能.

表 9　在 Wikipedia 数据集上掩码占位策略的消融实验 mAP 结果对比

掩码策略	50%-50%-0%	50%-25%-25%	30%-70%-0%	30%-35%-35%	10%-90%-0%	10%-45%-45%
$\lambda=1$	0.580	0.582	0.570	0.579	0.538	0.574
$\lambda=-1$	0.584	0.585	0.566	0.582	0.542	0.570
$\lambda=0$	0.585	0.584	0.571	0.577	0.540	0.570
−1	0.497	0.521	0.420	0.440	0.289	0.290
1	0.479	0.522	0.435	0.424	0.293	0.290
100	0.495	0.458	0.392	0.438	0.314	0.175

表 10　在 Pascal-Sentence 数据集上掩码占位策略的消融实验 mAP 结果对比

掩码策略	50%-50%-0%	50%-25%-25%	30%-70%-0%	30%-35%-35%	10%-90%-0%	10%-45%-45%
$\lambda=1$	0.753	0.758	0.741	0.746	0.636	0.724
$\lambda=-1$	0.753	0.758	0.741	0.746	0.637	0.725
$\lambda=0$	0.753	0.758	0.741	0.746	0.639	0.724
−1	0.601	0.678	0.413	0.564	0.245	0.319
1	0.600	0.664	0.491	0.565	0.256	0.325
100	0.590	0.514	0.467	0.397	0.279	0.193

3.7　参数敏感性分析

基于典型概念学习的多模态 Transformer 包含多模态融合 Transformer、典型概念学习和模态映射网络, 式(6)中包含 α,β,γ 这 3 个超参数. 为了分析 3 个超参数对模型性能的影响, 在 4 个基准数据集上进行参数敏感性实验, 结果如图 2 所示. 本文实验中, 测试一个超参数时, 其他超参数被固定, 逐个测试每个超参数的敏感性.

从图 2 可以看出: (1) 对于 α, 4 个数据集增长趋势类似; 随着超参数的增加, Wikipedia 数据集, 的 mAP 出现轻微上升, 当 α 超过 10 时, mAP 开始下降, 当 α 取值在 10~1 000 时, mAP 急速下降, 所以 α 的最优取值范围为[0.001,10]; 对于 Pascal-Sentence, NUS-WIDE 和 XmediaNet 数据集, α 的最优取值范围也同样为[0.001,10]. (2) 对于超参数 β, 4 个基准数据集上的性能变化趋势基本一致, 在[0.001,0.1]检索性能都呈现急速增长趋势, 之后性能逐渐趋于平稳. 所以, β 最优取值范围为[0.1, 1 000]. (3) 超参数 γ 的性能变化趋势与超参数 α 和 β 类似, 其最优取值范围为[0.1,1 000].

结合超参数 α,β,γ 的敏感性实验可以看出, 在超参数设置发生变化时, 跨模态检索性能具有一定的稳定性, 表明本文提出的模态缺失跨模态检索方法具有一定的鲁棒性和抗干扰能力.

3.8　收敛性分析

为了分析提出的跨模态检索模型的收敛性, 进行了收敛性实验, 结果如图 3 所示. 可以看出, 随着训练轮数的增加, 损失值呈现快速下降的趋势, 之后逐渐趋于平稳; Wikipedia 和 Pascal-Sentence 数据集在第 100 个训练轮次附近开始于稳定; NUS-WIDE 在第 80 个训练轮次后损失值下降渐缓; 随着训练轮数的增加, XmediaNet 在第 200 个训练轮次开始趋向平稳. 实验结果表明, 本文模型在经过有限的训练轮数后快速收敛, 降低了跨模态检索模型的训练消耗.

4　结　语

本文提出一个典型概念驱动的模态缺失跨模态检索模型. 首先设计了一个多模态融合 Transformer

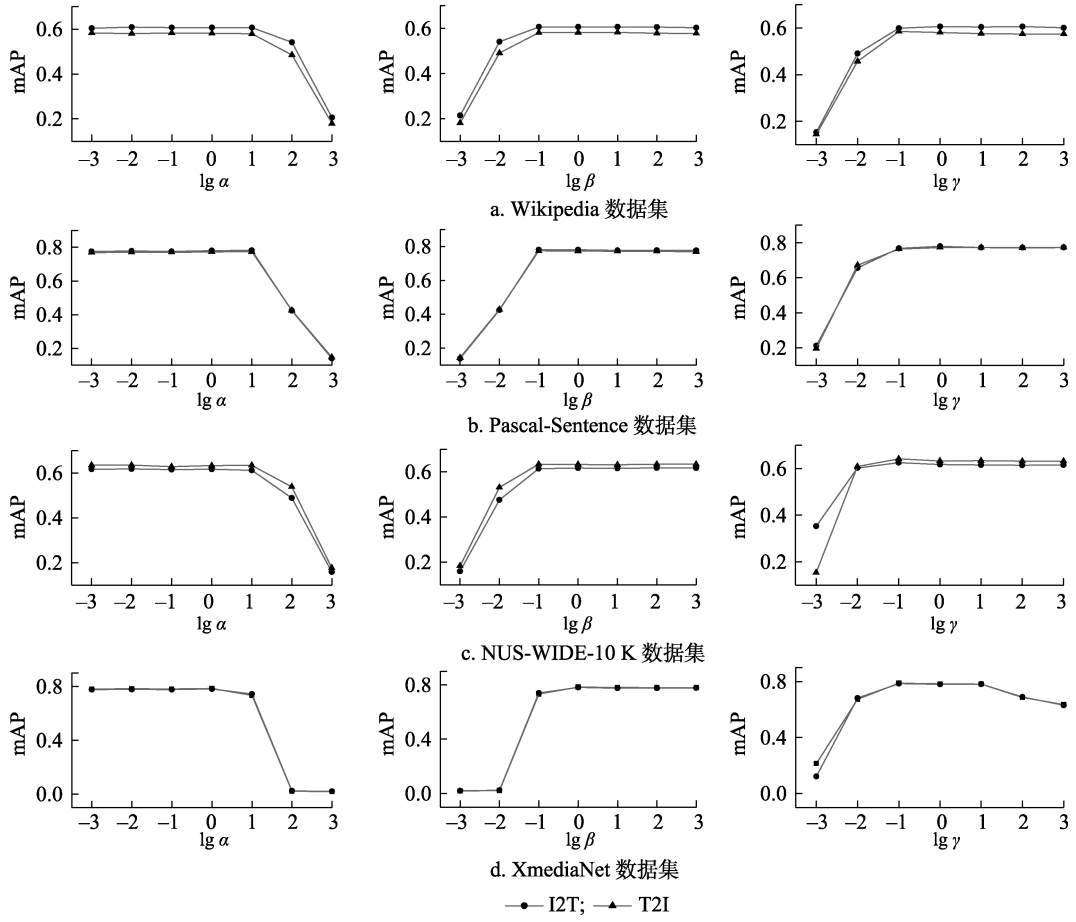


图 2 各数据集上不同超参数值对检索性能的影响

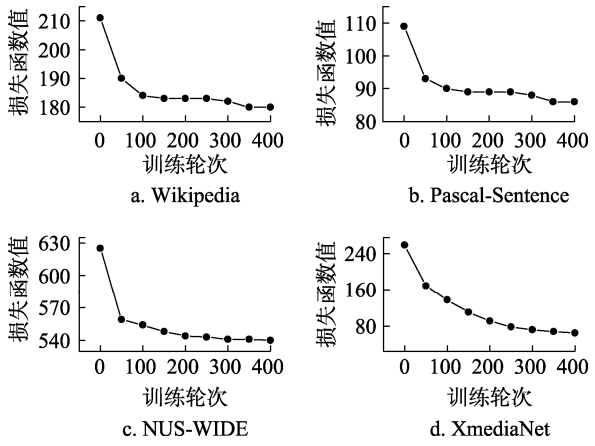


图 3 本文模型在不同数据集上跨模态检索收敛曲线对比

构建跨模态共享空间，并利用掩码策略实现缺失模态的信息补全；然后通过模态内和模态间的细粒度语义交互和典型概念学习，增强跨模态共享空间的多模态语义表示能力；最后通过典型概念驱动的 Transformer 模型学习模态映射网络，增强模态映射网络的语义判别性. 在 4 个基准跨模态检索数据集上的实验结果证明，在模态完备和模态缺失下的跨模态检索任务上，本文模型都取得了优越的性能.

参考文献(References):

[1] Peng Y X, Huang X, Zhao Y Z. An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(9): 2372-2385

[2] Yin Qiyue, Huang Yan, Zhang Junge, *et al*. Survey on deep learning based cross-modal retrieval[J]. Journal of Image and Graphics, 2021, 26(6): 1368-1388(in Chinese)
(尹奇跃, 黄岩, 张俊格, 等. 基于深度学习的跨模态检索综述[J]. 中国图象图形学报, 2021, 26(6): 1368-1388)

[3] Jing M M, Li J J, Zhu L, *et al*. Incomplete cross-modal retrieval with dual-aligned variational autoencoders[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 3283-3291

[4] Zeng Z X, Mao W J. A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval[OL]. [2023-06-01]. https://arxiv.org/abs/2201.02772

[5] Zhai X H, Peng Y X, Xiao J G. Learning cross-media joint representation with sparse and semisupervised regularization[J]. IEEE Transactions on Circuits and Systems for Video Tech-

nology, 2014, 24(6): 965-978

[6] Zhao Peng, Ma Taiyu, Li Yi, *et al*. Cross-modal retrieval based on full-modal autoencoder with generative adversarial mechanism[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(10): 1486-1494(in Chinese)
(赵鹏, 马泰宇, 李毅, 等. 融合全模态自编码器和生成对抗机制的跨模态检索[J]. 计算机辅助设计与图形学学报, 2021, 33(10): 1486-1494)

[7] Tan W T, Zhu L, Guan W L, *et al*. Bit-aware semantic transformer hashing for multi-modal retrieval[C] //Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2022: 982-991

[8] Zhu L, Lu X, Cheng Z Y, *et al*. Flexible multi-modal hashing for scalable multimedia retrieval[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(2): Article No.14

[9] Wang T, Zhu L, Cheng Z Y, *et al*. Unsupervised deep cross-modal hashing with virtual label regression[J]. Neurocomputing, 2020, 386: 84-96

[10] Tan W T, Zhu L, Li J J, *et al*. Teacher-student learning: efficient hierarchical message aggregation hashing for cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2023, 25: 4520-4532

[11] Zeng Chao, Bai Cong, Ma Qing, *et al*. Adversarial projection learning based hashing for cross-modal retrieval[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(6): 904-912(in Chinese)
(曾超, 白琮, 马青, 等. 基于对抗投影学习的跨模态哈希检索[J]. 计算机辅助设计与图形学学报, 2021, 33(6): 904-912)

[12] Guo J, Zhu W W. Collective affinity learning for partial cross-modal hashing[J]. IEEE Transactions on Image Processing, 2020, 29: 1344-1355

[13] Yang Y, Zhan D C, Sheng X R, *et al*. Semi-supervised multi-modal learning with incomplete modalities[C] //Proceedings of the 27th International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 2998-3004

[14] Wang Q F, Si L, Shen B. Learning to hash on partial multi-modal data[C] //Proceedings of the 24th International Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 3904-3910

[15] Zeng Z X, Wang S, Xu N, *et al*. PAN: prototype-based adaptive network for robust cross-modal retrieval[C] //Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2021: 1125-1134

[16] Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639-2664

[17] Wang W R, Livescu K. Large-scale approximate kernel canonical correlation analysis[OL]. [2023-06-01]. https://arxiv.org/abs/1511.04773v2

[18] Rasiwasia N, Mahajan D, Mahadevan V, *et al*. Cluster canonical correlation analysis[C] //Proceedings of the 17th International Conference on Artificial Intelligence and Statistics. New York: PMLR Press, 2014: 823-831

[19] Feng F X, Wang X J, Li R F. Cross-modal retrieval with correspondence autoencoder[C] //Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 7-16

[20] Zhen L L, Hu P, Wang X, *et al*. Deep supervised cross-modal retrieval[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 10386-10395

[21] Peng Y X, Qi J W. CM-GANs: cross-modal generative adversarial networks for common representation learning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2019, 15(1): Article No.22

[22] Huang Y, Wang W, Wang L. Instance-aware image and sentence matching with selective multimodal LSTM[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 7254-7262

[23] Chen Ning, Duan Youxiang, Sun Qifeng. Literature review of cross-modal retrieval research[J]. Journal of Frontiers of Computer Science & Technology, 2021, 15(8): 1390-1404(in Chinese)
(陈宁, 段友祥, 孙歧峰. 跨模态检索研究文献综述[J]. 计算机科学与探索, 2021, 15(8): 1390-1404)

[24] Jiang Q Y, Li W J. Deep cross-modal hashing[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3270-3278

[25] Xu R Q, Li C, Yan J C, *et al*. Graph convolutional network hashing for cross-modal retrieval[C] //Proceedings of the 28th International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 982-988

[26] Lin Q B, Cao W M, He Z H, *et al*. Semantic deep cross-modal hashing[J]. Neurocomputing, 2020, 396: 113-122

[27] Su S P, Zhong Z S, Zhang C. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 3027-3035

[28] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision[C] // Proceedings of the 38th International Conference on Machine Learning. New York: PMLR Press, 2021: 8748-8763

[29] He K M, Chen X L, Xie S N, *et al*. Masked autoencoders are scalable vision learners[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 15979-15988

[30] Rasiwasia N, Pereira J C, Coviello E, *et al*. A new approach to cross-modal multimedia retrieval[C] //Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press, 2010: 251-260

[31] Rashtchian C, Young P, Hodosh M, *et al*. Collecting image annotations using amazon's mechanical turk[C] //Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Stroudsburg: ACL, 2010: 139-147

[32] Chua T S, Tang J H, Hong R C, *et al*. NUS-WIDE: a real-world web image database from national university of Singapore[C] // Proceedings of the ACM International Conference on Image and Video Retrieval. New York: ACM Press, 2009: Article No.48

[33] Peng Y X, Qi J W, Yuan Y X. Modality-specific cross-modal similarity measurement with recurrent attention network[J]. IEEE Transactions on Image Processing, 2018, 27(11): 5585-5599

[34] Peng Y X, Huang X, Qi J W. Cross-media shared representation by hierarchical learning with multiple deep networks[C] // Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 3846-3853

[35] Wang K Y, He R, Wang L, *et al*. Joint feature selection and subspace learning for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2010-2023

[36] Wu J L, Lin Z C, Zha H B. Joint latent subspace learning and regression for cross-modal retrieval[C] //Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2017: 917-920

[37] Peng Y X, Qi J W, Huang X, *et al*. CCL: cross-modal correlation learning with multigrained fusion by hierarchical network[J]. IEEE Transactions on Multimedia, 2018, 20(2): 405-420

[38] Zeng Z X, Sun Y, Mao W J. MCCN: multimodal coordinated clustering network for large-scale cross-modal retrieval[C] // Proceedings of the 29th ACM International Conference on Multimedia. Cambridge: MIT Press, 2021: 5427-5435

[39] Wang J S, Gong T T, Zeng Z X, *et al*. C^3CMR: cross-modality cross-instance contrastive learning for cross-media retrieval[C] // Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 4300-4308

[40] Liu H, Ji R R, Wu Y J, *et al*. Cross-modality binary code learning via fusion similarity hashing[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6345-6353

[41] Zhang D Q, Li W J. Large-scale supervised multimodal hashing with semantic correlation maximization[C] //Proceedings of the 28th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2014: 2177-2183

[42] Liu X, Hu Z K, Ling H B, *et al*. MTFH: a matrix tri-factorization hashing framework for efficient cross-modal retrieval[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(3): 964-981

[43] Hu P, Zhen L L, Peng D Z, *et al*. Scalable deep multimodal learning for cross-modal retrieval[C] //Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2019: 635-644