软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(5):2120−2132 [doi: 10.13328/j.cnki.jos.007021]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向跨模态检索的查询感知双重对比学习网络*

尹梦冉[1,2], 梁美玉[1,2], 于 洋[1,2], 曹晓雯[1,2], 杜军平[1,2], 薛 哲[1,2]

[1](北京邮电大学 计算机学院 (国家示范性软件学院), 北京 100876)

[2](智能通信软件与多媒体北京市重点实验室 (北京邮电大学), 北京 100876)

通信作者: 梁美玉, E-mail: meiyu1210@bupt.edu.cn

摘 要: 近期, 跨模态视频语料库时刻检索 (VCMR) 这一新任务被提出, 它的目标是从未分段的视频语料库中检索出与查询语句相对应的一小段视频片段. 现有的跨模态视频文本检索工作的关键点在于不同模态特征的对齐和融合, 然而, 简单地执行跨模态对齐和融合不能确保来自相同模态且语义相似的数据在联合特征空间下保持接近, 也未考虑查询语句的语义. 为了解决上述问题, 提出一种面向多模态视频片段检索的查询感知跨模态双重对比学习网络 (QACLN), 该网络通过结合模态间和模态内的双重对比学习来获取不同模态数据的统一语义表示. 具体地, 提出一种查询感知的跨模态语义融合策略, 根据感知到的查询语义自适应地融合视频的视觉模态特征和字幕模态特征等多模态特征, 获得视频的查询感知多模态联合表示. 此外, 提出一种面向视频和查询语句的模态间及模态内双重对比学习机制, 以增强不同模态的语义对齐和融合, 从而提高不同模态数据表示的可分辨性和语义一致性. 最后, 采用一维卷积边界回归和跨模态语义相似度计算来完成时刻定位和视频检索. 大量实验验证表明, 所提出的 QACLN 优于基准方法.

关键词: 跨模态语义融合; 跨模态检索; 视频时刻定位; 对比学习

中图法分类号: TP18

中文引用格式: 尹梦冉, 梁美玉, 于洋, 曹晓雯, 杜军平, 薛哲. 面向跨模态检索的查询感知双重对比学习网络. 软件学报, 2024, 35(5): 2120–2132. http://www.jos.org.cn/1000-9825/7021.htm

英文引用格式: Yin MR, Liang MY, Yu Y, Cao XW, Du JP, Xue Z. Query Aware Dual Contrastive Learning Network for Cross-modal Retrieval. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2120–2132 (in Chinese). http://www.jos.org.cn/1000-9825/7021.htm

Query Aware Dual Contrastive Learning Network for Cross-modal Retrieval

YIN Meng-Ran[1,2], LIANG Mei-Yu[1,2], YU Yang[1,2], CAO Xiao-Wen[1,2], DU Jun-Ping[1,2], XUE Zhe[1,2]

[1](School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China)

[2](Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia (Beijing University of Posts and Telecommunications), Beijing 100876, China)

Abstract: Recently, a new task named cross-modal video corpus moment retrieval (VCMR) has been proposed, which aims to retrieve a small video segment corresponding to a query statement from an unsegmented video corpus. The key point of the existing cross-modal video text retrieval work is the alignment and fusion of different modal features. However, simply performing cross-modal alignment and fusion cannot ensure that semantically similar data from the same modal remain close under the joint feature space, and the semantics of query statements are not considered. To solve the above problems, this study proposes a query-aware cross-modal dual contrastive learning network for multi-modal video moment retrieval (QACLN), which achieves the unified semantic representation of different modal data by

* 基金项目: 国家自然科学基金 (62192784, U22B2038, 62172056, 62272058); 中国人工智能学会-华为 MindSpore 学术奖励基金 (CAAIXSJLJJ-2021-007B)
本文由“多模态协同感知与融合技术”专题特约编辑孙立峰教授、宋新航副研究员、蒋树强教授、王莉莉教授、申恒涛教授推荐.
收稿时间: 2023-03-26; 修改时间: 2023-06-08, 2023-08-16; 采用时间: 2023-08-23; jos 在线出版时间: 2023-09-11
CNKI 网络首发时间: 2024-02-09

combining cross-modal and intra-modal contrastive learning. First, the study proposes a query-aware cross-modal semantic fusion strategy, obtaining the query-aware multi-modal joint representation of the video by adaptively fusing multi-modal features such as visual modal features and caption modality features of the video according to the aware query semantics. Then, a cross-modal and intra-modal dual contrastive learning mechanism for video and text query is proposed to enhance the semantic alignment and fusion of different modalities, which can improve the discriminability and semantic consistency of data representations of different modalities. Finally, the 1D convolution boundary regression and cross-modal semantic similarity calculation are employed to perform moment localization and video retrieval. Extensive experiments demonstrate that the proposed QACLN outperforms the benchmark methods.

Key words: cross-modal semantic fusion; cross-modal retrieval; video moment localization; contrastive learning

随着互联网上多媒体数据的爆炸增长, 数据模态也变得多样化, 人们的检索需求也越来越精细化, 呈现出从单模态检索到跨模态检索的趋势. 与传统的单模态检索方法相比, 不同特征空间的数据存在语义理解的差距, 各个模态的数据具有多样性. 因此, 如何获取不同模态数据的统一语义表示成为跨模态检索的关键问题. 为了解决这一问题, 研究人员提出各种方法[1], 通过跨模态语义关联学习与融合, 减小不同模态数据间的语义差距并保留数据的可判别性和语义一致性. 近年来, 视频检索 (video retrieval, VR) 任务和单视频时刻检索 (single video moment retrieval, SVMR)[2–4]任务已取得显著进展. 视频检索任务旨在从大规模视频数据库中根据用户查询语句找到相关的视频. 用户输入一段查询文本, 而输出则是与查询相关的整段视频, 如图 1 所示. 单视频时刻检索是视频检索领域的一个子任务, 其目标是根据用户提供的查询文本, 从单个视频中找到与查询相关的视频时刻片段. 其输入包括用户的查询文本和单个视频, 而输出则是与查询相关的视频时刻或片段. 在 SVMR 任务的基础上, Escorcia 等人[5]提出了一种视频语料库时刻检索任务 (video corpus moment retrieval, VCMR), 它的目标是从未分段的视频语料库中检索出与查询语句相对应的一小段视频片段. 与 SVMR 任务从单视频中检索相关片段不同的是, VCMR 视频检索涵盖了更广泛的范围. 这 3 个任务共同面临的挑战是如何有效地将用户查询文本与视频内容进行匹配, 以找到最相关的视频片段. 它们在输入和输出上的差异主要体现在针对不同视频数据规模和查询需求的设计, 能够满足不同粒度的检索需求. 视频检索是一个广泛的领域, 包括了更多的视频信息, 针对整个视频内容进行的检索. 而单视频时刻检索和视频语料库时刻检索则是视频检索的两个子领域, 更注重时间和时刻的准确性, 能够满足用户更精准需求.

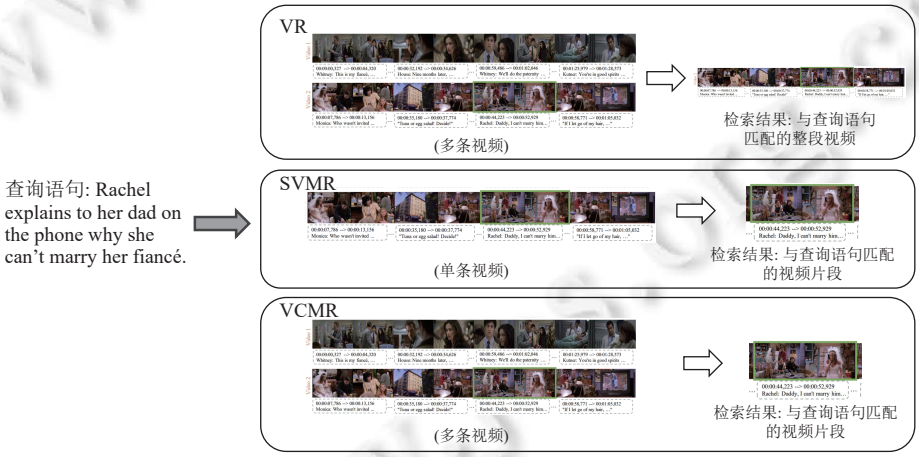

图 1　视频检索任务、单视频时刻检索任务和视频语料库时刻检索任务的区别

对于视频语料库时刻检索任务, 目前有两种主流方法. 第 1 种方法是分别从视频和查询中提取特征, 并在后期通过添加线性变换和注意力机制来完成特征对齐与融合, 如基于后期融合的跨模态时刻定位模型 (XML)[6]、基于对比学习的检索和定位网络 (ReLoCLNet)[7]. 该类方法的优点在于可以对视觉特征进行预编码和存储, 只需要对查询语句进行编码, 因此具有快速推理的特点. 第 2 种方法是通过跨模态注意力机制在早期阶段进行跨模态交互融合, 以便跨模态信息可以作为推理过程中的参考, 以获得更准确的多模态特征, 如面向视频跨模态检索的分层多模

态编码方法 (HAMMER)[8]和情境感知排序方法 (CONQUER)[9]等. 然而, 这类方法需要更多的计算资源和时间, 并且可能会存在过拟合的风险.

由于对比学习的特性, 即减小相匹配的样例之间的语义差距, 对比学习常用于消除多模态数据的模态差异, 利用表示同一语义的异构数据构建不同模态间的对应关系. 在跨模态视频文本检索工作中, 关键点在于不同模态特征的对齐和融合. 然而, 简单地执行模态间对齐和融合只能确保不同模态之间的相似数据保持接近, 从而忽略了联合特征空间中模态内语义相似数据. 因此, 本文基于跨模态对比学习机制建立跨模态语义关联, 同时利用模态间和模态内双重对比学习方法, 从而获得更准确的跨模态语义关联.

与此同时, 视频中蕴含的信息是多模态的, 且在语义层面上是互补的. 然而, 当前大多数方法在视频和文本跨模态语义表示学习中通常只采用了视频中视觉模态数据, 而忽略了视频中的字幕和语音等多模态信息. 此外, 现有的跨模态视频检索方法没有考虑查询语句的语义. 换句话说, 现有的方法在进行跨模态视频检索时没有充分考虑查询语句的语义信息与视频数据中不同模态数据之间的关联性. 针对以上问题, 本文提出了一种查询感知的跨模态融合 (QCF) 策略, 根据查询语句的语义, 自适应地学习视频中视觉模态和字幕模态的融合权重, 从而对视频数据中不同模态的特征进行加权和融合, 以获得视频的多模态联合语义表示.

综上所述, 本文提出了一个面向多模态视频片段检索的查询感知双重对比学习网络 (QACLN), 通过建立跨模态对比学习机制进行跨模态语义关联学习, 自适应地融合不同模态的语义特征, 实现不同模态数据的语义对齐和交互, 最终提高跨模态统一语义表示的可判别性和语义一致性.

主要贡献包括 3 个方面.

(1) 构建了面向多模态视频时刻定位和检索的查询感知的跨模态双重对比学习网络, 通过将查询语义感知和模态间及模态内双重对比学习集成在一个统一的框架中进行联合优化, 增强了不同模态数据的统一语义表示和跨模态检索性能.

(2) 提出了一种基于查询感知的跨模态语义融合策略, 根据查询语义自适应地融合视频中的视觉模态和字幕模态等多模态特征, 获取视频的多模态联合语义表示.

(3) 提出了一种面向视频-文本跨模态检索的跨模态双重对比学习机制, 联合模态间和模态内双重对比学习, 增强不同模态的语义对齐和融合, 从而提高不同模态数据表示的判别性和语义一致性.

本文第 1 节介绍跨模态视频文本检索工作的相关工作. 第 2 节介绍本文构建的面向多模态视频片段检索的查询感知跨模态双重对比学习网络 (QACLN). 第 3 节通过对比实验验证了所提方法的有效性. 最后总结全文.

1 相关工作

1.1 跨模态视频检索

近年来, 跨模态视频/视频片段检索研究取得了显著进展, 涌现了许多令人印象深刻的方法. 由于不同模态之间存在特征异构和语义鸿沟的问题, 跨模态视频/视频片段检索的核心挑战是通过跨模态语义关联学习与融合, 获取不同模态数据的统一语义表征. 跨模态表征学习的目标是将来自不同模态 (如文本、图像、视频、语音等) 的信息映射到同一个语义空间中, 进行联合建模和表示. 在跨模态表征学习中, 一些经典的方法包括基于跨模态哈希的模型[10]和基于无监督细粒度的模型[11]. 其中, 基于哈希映射的方法, 通过将不同模态的数据映射为同一空间中更紧凑的哈希码, 可以提升跨模态检索的效率; 而基于无监督细粒度的网络的模型可以减少来源不同的模态的数据之间的模态差异, 不依赖标签完成跨模态任务.

在跨模态语义融合方面, 根据特征融合的方式, 这些方法主要可以分为后期融合的方法和前期融合的方法, 前者通过从视频和查询中提取特征, 并在后期通过添加线性变换和注意力机制来完成特征对齐; 而后者则是通过跨模态注意力机制在早期编码阶段进行跨模态交互融合. Lei 等人[6]提出了一种跨模态时刻定位 (cross-modal moment localization, XML) 模型, 该模型基于视频-文本特征关系进行视频检索与时刻定位, 并构建了一个新的跨模态视频检索数据集, 用于 VCMR 任务, 称为 TVR 数据集. 该数据集不仅包含视频信息, 还包含相应的字幕文本. 为了获得

更准确的多粒度特征, ReLoCLNet[7]在 XML 模型的基础上添加了帧级和视频级的对比学习方法, 以实现更好的特征学习. HERO[12]通过分层方法逐渐融合粗粒度、细粒度和多模态信息. 这 3 个模型属于后期融合方法. 与上述方法不同的是, HAMMER[8]、CONQUER[9]和使用跨模态编码器在前期编码阶段来完成查询文本和视频的融合. 跨模态编码器可以根据查询内容更好地生成融合特征, 从而获得更细粒度和更全面的信息. Zhang 等人[8]提出了一种分层多模态编码器 (hierarchical multi-modal encoder, HAMMER), 通过查询和视频之间的细粒度跨模态交互学习来联合训练视频检索和时刻定位模型. 然而, 上述方法无法对特征进行预编码和存储, 导致计算成本增加和检索效率降低. FLAT 模型[8]的架构与 HAMMER 类似, 只是省略了分层注意力机制, 只使用了一个平铺的多模态记忆编码器来对视频和问题进行编码. 但由于它没有对视频和文本进行多层次的建模, 因此在处理较长的视频时可能会受到一定影响.

1.2 对比学习与互信息

对比学习[13]作为无监督学习的一个重要技术被广泛应用于计算机视觉、自然语言处理等领域. 它的目标是使匹配的正样例在同一语义空间中尽可能接近, 而使负样例彼此远离. 对比学习和跨模态交互融合可以相互促进, 例如 CrossCLR[14]利用对比学习机制来学习不同模态数据之间的语义关联, 以实现更好的跨模态匹配和融合. 互信息 (mutual information, MI)[15]是衡量两个变量是否相关的重要标准, 也被广泛用于无监督特征学习. 然而, 由于计算高维连续随机变量的互信息较为困难, 如何最大化互信息一直是一个难题. 随机变量互信息的估计方法 (mutual information neural estimation, MINE)[16]是一种使用神经网络有效地计算互信息的方法. 本文借鉴了深度互信息最大化 (deep infomax, DIM) [17]的思路, 通过最大化视频-文本对的互信息来指导编码器更好地学习视频文本特征. DIM[17]提出了两种最大化互信息下限的方法: 基于 Jensen Shannon 散度估计[17]和基于 InfoNCE 估计[18]. 现有的互信息最大化工作主要是针对模态间特征融合提出的, 本文提出模态间和模态内双重对比学习机制, 指导编码器更好地学习视频文本特征, 同时最大化匹配视频-文本对的相互信息.

2 QACLN 模型设计

本文提出了一个面向多模态视频片段检索的查询感知跨模态双重对比学习网络, 旨在提高视频片段检索的准确性和效率. 该网络的总体框架如图 2 所示, 主要包括 5 个组件, 分别是用于学习查询特征的文本编码器、用于学习视频多模态特征的视频编码器、查询感知的跨模态语义融合组件、模态间与模态内双重对比学习组件以及视频片段定位和视频检索组件. 在编码器组件中, 利用基于多层自注意力机制的 Transformer 模型来学习文本和视频的语义特征表示, 该组件由两部分构成: 一个用于学习查询语句特征的查询编码器, 另一个用于学习视频多模态特征的视频编码器. 为了更好地对齐视频特征和查询文本特征, 构建模态间和模态内双重对比学习组件, 该模块包括帧级和视频级对比学习 VideoCL 和 FrameCL, 以及视频模态内对比学习 VVCL, 通过最大化互信息来学习更丰富的视觉表示. 此外, 本文提出了一种查询感知的跨模态融合 (QCF) 策略. 通过从查询语句中学习融合权重, 从而对两种不同模态的特征进行自适应加权和融合, 以获得视频的多模态联合语义表征. 对于视频检索组件, 通过计算语义相似度来计算查询和视频之间的语义匹配度; 对于视频时刻定位组件, 基于两个卷积滤波器检测 1D 相似信号中的起始端点边, 从而实现视频片段开始时刻和结束时刻的定位和检索.

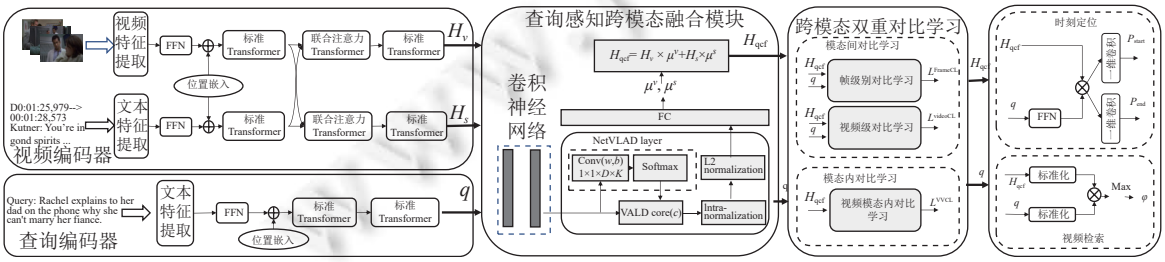

图 2 面向多模态视频片段检索的查询感知双重对比学习网络

2.1 问题描述与定义

定义视频语料库 $\mathcal{V}=\{V^1,V^2,\ldots,V^M\}$, 其中 M 表示视频的数量, $V^k=[f_i]_{i=0}^{T-1}$ 表示第 k 个视频有 T 帧. 对于查询语句 $\mathbf{Q}=[f_i]_{i=0}^{T-1}$, 本文提出方法的目标是从视频语料库 $\mathcal{V}$ 中检索与之相关的目标时刻片段, 目标时刻片段的起止时间分别用 i^s 和 i^e 表示. 因此本文方法主要包括两个步骤: 1) 视频检索: 从视频语料库 $\mathcal{V}$ 中找到视频 V^*, 这里 V^* 表示包含目标片段的视频; 2) 时刻定位: 在 V^* 中定位目标时刻. 目标时刻 m^* 表示为 $m^*=\{v_i|i=i^s,\ldots,i^e\}$, 其中 $0\leqslant i^s\leqslant i^e\leqslant n_v-1$.

对于查询语句 $\mathbf{Q}$ 中的单词, 从预先训练好的词嵌入模型或语言模型中获得初始编码 $\mathbf{Q}=[q_i]_{i=0}^{n_q-1}\in\mathbb{R}^{d_\omega\times n_q}$, 其中 d_ω 表示文本特征的维度. 对于视频 $V\in\mathcal{V}$, 将其划分为 n_v 个视频片段, 用预训练好的特征编码器将其编码为视觉特征 $\mathbf{V}=[v_i]_{i=0}^{n_v-1}\in\mathbb{R}^{d_v\times n_v}$, 其中 d_v 表示视觉特征的维度. 对于视频而言, 除了视觉特征外, 它还包括多模态信息. 以 TVR 数据集[6]为例, 其视频有字幕. 用 S 表示视频的字幕, 用 $\mathbf{S}\in\mathbb{R}^{d_\omega\times n_v}$ 表示从字幕中提取的文本特征.

2.2 基于多层自注意力机制的视频-文本编码器

对于文本编码器, 首先基于文本特征提取器将查询语句中的单词转换为相应的特征 $\mathbf{Q}$. 然后, 通过前馈神经网络将获得的特征 $\mathbf{Q}$ 投影到维度为 d 的特征空间中. 其次, 使用两个基于多层注意力机制的 Transformer[19]来更好地捕获查询语句的上下文表示, 查询语句经过查询编码器后得到查询特征向量 $q\in\mathbb{R}^d$.

对于视频编码器, 给定带有字幕的视频, 首先使用视觉特征提取器和文本特征提取器分别获得它们的单模态特征 $\mathbf{V}\in\mathbb{R}^{d_v\times n_v}$ 和 $\mathbf{S}\in\mathbb{R}^{d_\omega\times n_v}$. 与查询编码器类似, 本文基于两个前馈神经网络将 $\mathbf{V}$ 和 $\mathbf{S}$ 投影到维度为 d 的特征空间中, 然后基于 Transformer 进行上下文语义推理和特征表示学习. 与查询文本编码器不同, 不仅基于两个多层注意力机制的 Transformer, 而是在第 1 个 Transformer 之后, 基于联合注意力 Transformer[20]以便更好地捕捉视频中视频和字幕的多模态联合表示. 最后, 基于标准 Transformer 细化编码, 输出视频视觉向量与视频字幕向量 $H_v\in\mathbb{R}^{d\times d_v}$ 和 $H_s\in\mathbb{R}^{d\times d_v}$.

2.3 基于查询感知的视频多模态联合表征学习

为了更好地指导视频中不同模态的特征进行自适应融合, 通过对查询语义的感知, 为视频中视觉模态特征和字幕模态特征分配不同的权重进行自适应地融合, 以获得视频的多模态联合特征表示. 具体而言, 一些查询语句更多地描述视频中的语言交流信息, 而这些描述预计不会出现在视觉画面内容中, 例如“瑞秋向父亲解释为什么她不能嫁给未婚夫”. 相反, 有些查询更注重视觉特征描述, 因此应该赋予它们更高的视觉权重, 例如“Cady 从房子前面的桌子上拾取一些文档”.

具体地, 本文将视觉特征和字幕特征编码为固定长度的向量表示 $N_{v/s}$. 这个向量表示具有比原始的视觉特征更佳的表达能力和可区分性, 能够更好地反映视频中的语义信息, 从而为后续的查询感知跨模态融合提供更好的语义特征.

$$N_{v/s}=NetVLAD(H_{v/s})\in\mathbb{R}^{d_{v/w}\times n_v} \tag{1}$$

由于 Softmax 层能够将输入的向量归一化为概率分布, 因此本文构建一个具有 Softmax 层的全连接网络来获得查询感知的视觉特征和字幕特征的自适应融合权重 μ^v 和 μ^s, 计算方法分别如公式 (2) 所示:

$$\mu^s=\frac{\mathrm{e}^{W_sq}}{\mathrm{e}^{W_vq}+\mathrm{e}^{W_sq}},\ \mu^v=\frac{\mathrm{e}^{W_vq}}{\mathrm{e}^{W_vq}+\mathrm{e}^{W_sq}} \tag{2}$$

其中, q 表示查询向量, W_v 和 W_s 是全连接层的权重矩阵.

最终, 通过对视觉模态特征和字幕模态特征的自适应加权融合, 得到查询感知的视频多模态联合表征 H_{qcf}, 如公式 (3) 所示:

$$H_{\mathrm{qcf}}=H_v\times\mu^v+H_s\times\mu^s \tag{3}$$

2.4 跨模态双重对比学习机制

为了充分学习模态内和模态间的语义相似性约束来进一步增强不同模态数据的语义关联学习能力, 进一步缩小跨模态语义鸿沟问题, 本节提出了面向视频-文本跨模态检索的跨模态双重对比学习机制, 联合模态间和模态内

对比学习技术进行模态内表征优化和模态间语义融合, 从而更好地保持模态内和模态间的语义相似性, 以增强不同模态数据的语义对齐和融合, 从而提高不同模态数据表示的判别性和语义一致性.

(1) 基于对比学习的模态内表征优化

简单地执行模态间语义对齐忽略了模态内数据的语义一致性, 不能确保来自同一模态的相似语义的数据表示保持接近. 如果当前数据有噪声, 问题将变得更为严重. 因此, 为了优化跨模态表征学习的表征质量, 本文基于对比学习进行模态内表征优化, 通过最大化开始时刻/结束时刻与目标片段之间的互信息, 完成视频模态内特征的对齐, 如图 3 所示.

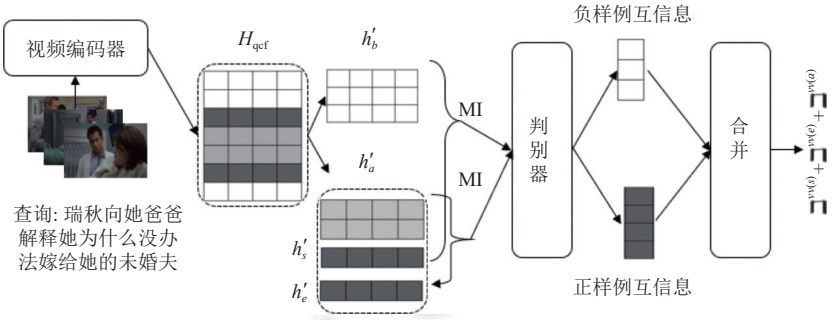

图 3 模态内对比学习机制

基于模态内对比学习机制对 QCF 组件生成的多模态联合表征向量 H_{qcf} 进行模态内表征优化. $h'_a=\{h'_i|i=i^s,\ldots,i^e\}\in\mathbb{R}^{d\times n_v}$ 作为正样例, 表示属于目标片段内的特征, 并且 $h'_b=\{h'_i\big|i=0,\ldots,i^{s-1},i^{e+1},\ldots,n_{v-1}\}\in\mathbb{R}^{d\times n_v-n_t}$ 作为表示位于目标片段外部的负样例. 将 $\beth^{vv(s)}$ 和 $\beth^{vv(e)}$ 表示为目标片段开始时刻和结束时刻与其他视频片段之间的互信息, 将 $\beth^{vv(a)}$ 作为正样例和其他剪辑之间的互信息, h'_s 和 h'_e 表示目标时刻的开始和结束边界的视频表示, h'_a 表示正样例. $\beth^{vv(s)}$ 、$\beth^{vv(e)}$ 与 $\beth^{vv(a)}$ 如下所示:

$$\beth^{vv(a)}=\mathbb{E}_{h'_a}[-sp(-C(h'_a,h'_a))]-\mathbb{E}_{h'_b}[sp(C(h'_a,h'_b))] \tag{4}$$

$$\beth^{vv(s)}=\mathbb{E}_{h'_a}[-sp(-C(h'_s,h'_a))]-\mathbb{E}_{h'_b}[sp(C(h'_s,h'_b))] \tag{5}$$

$$\beth^{vv(e)}=\mathbb{E}_{h'_a}[-sp(-C(h'_e,h'_a))]-\mathbb{E}_{h'_b}[sp(C(h'_e,h'_b))] \tag{6}$$

其中, C 表示 $d\times d$ 维的判别器, $sp(z)=\log(1+\mathrm{e}^z)$ 是 softplus 激活函数.

模态内对比学习损失函数为:

$$\mathcal{L}^{\text{VVCL}}=-\frac{1}{3}(\beth^{vv(s)}+\beth^{vv(e)}+\beth^{vv(a)}) \tag{7}$$

(2) 基于对比学习的模态间语义融合

除了基于对比学习的模态内表征优化之外, 本文还考虑了基于对比学习的模态间语义融合. 模态间对比损失函数由两部分组成, 分别为视频级模态间对比学习 VideoCL 和帧级模态间对比学习 FrameCL.

对于视频级的模态间对比学习 VideoCL, 将语义相关的视频和查询视为正样例对 $P=(c,q)$, 将不相关的视频与查询视作为负样例对 $N=(c',q')$. 其目的是减少在统一语义空间下正样例之间的距离, 增加负样例之间的间距. 换句话说, 在跨模态统一表示空间中, 语义相关的视频和查询文本彼此更接近.

经过编码器和 QCF 模块后, 查询文本被编码为 d 维的特征表示 $q\in\mathbb{R}^d$, 视频被编码为多模态融合特征表示 $H_{\text{qcf}}=\{h^0_{\text{qcf}},h^0_{\text{qcf}},\ldots,h^{n_v-1}_{\text{qcf}}\}\in\mathbb{R}^{d\times n_v}$, 其中 h^i_{qcf} 表示每个视频片段对应的特征向量.

由于噪声对比估计 (noise-contrastive estimation, NCE)[21]是一种用于学习模型参数的损失函数, 通过构造一个负样例集合, 将正样例与负样例进行对比, 并最小化两者之间的交叉熵损失, 从而实现模型参数的优化. 本文基于噪声对比损失来计算视频级对比学习 VideoCL:

$$\beth^{\mathrm{Vcl}}=\log\frac{\sum_{c,q\in\mathcal{P}}\mathrm{e}^{f(c)^{\top}\cdot g(q)}}{\sum_{c,q\in\mathcal{P}}\mathrm{e}^{f(c)^{\top}\cdot g(q)}+\sum_{c',q'\in\mathcal{P}}\mathrm{e}^{f(c')^{\top}\cdot g(q')}} \tag{8}$$

$$\alpha=\mathrm{Softmax}(W_{\alpha}\cdot H_{\mathrm{qcf}})\in\mathbb{R}^{n_v},\ \ c=\sum_{i=1}^{n_v-1}\alpha_i\times h_{\mathrm{qcf}}^{i}\in\mathbb{R}^{d} \tag{9}$$

其中, $f(\cdot)$ 和 $g(\cdot)$ 表示参数映射, 将视频和查询特征投影到相同的语义特征空间中. $\mathrm{e}^{f(c)^{\top}\cdot g(q)}$ 表示向量 c 和 q 的互信息. c 是由 H_{qcf} 转化得来的视频特征. 最终, 视频级对比学习损失函数表示为:

$$\mathcal{L}^{\mathrm{VideoCL}}=-\beth^{\mathrm{Vcl}} \tag{10}$$

对于帧级对比学习 FrameCL, 旨在引导两个编码器更好地区分与查询文本匹配的片段和非匹配片段, 如图 4 所示. 本文基于互信息最大化 (MI) 的判别方法来计算正负样例之间的对比损失. FrameCL, 即最大化查询 q 和目标时刻 h'_a 之间的互信息, 如公式 (11) 所示:

$$\beth^{\mathrm{Fcl}}=\mathbb{E}_{h'_a}\left[-sp(-C(q,h'_a))-\mathbb{E}_{h'_b}\left[sp(C(q,h'_b))\right]\right] \tag{11}$$

最终, 帧级对比学习损失函数表示为:

$$\mathcal{L}^{\mathrm{FrameCL}}=-\beth^{\mathrm{Fcl}} \tag{12}$$

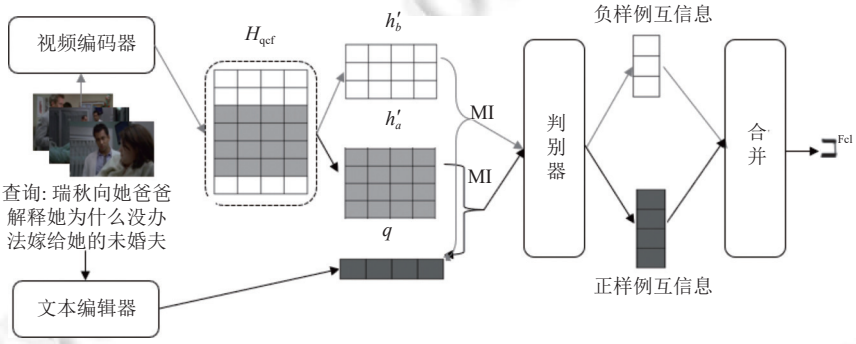

图 4 模态间对比学习机制

2.5 基于语义匹配的跨模态视频检索与时刻定位

首先, 计算每个视频和查询之间的余弦相似度, 用 φ 表示, 并选择相似度最高的视频和查询文本对作为正样例.

$$\varphi=\max\left(\frac{H_{\mathrm{qcf}}}{\|H_{\mathrm{qcf}}\|}\frac{q}{\|q\|}\right)\in\mathbb{R}^{n_v} \tag{13}$$

对于每一对正样例 (q_i,v_i) 而言, 选择两对负样例 (q_i,v_j) 和 (q_z,v_i), 假设两个负样例的相似度分别为 φ' 和 φ''. 采用铰链损失 (hinge loss) 计算视频检索模块的损失函数, 如公式 (14) 所示:

$$\mathcal{L}^{\mathrm{VR}}=\frac{1}{n}\sum_{i}\left[\max\left(0,\Delta\frac{1}{N}\varphi'-\varphi\right)+\max\left(0,\Delta\frac{1}{N}\varphi''-\varphi\right)\right] \tag{14}$$

受图像处理中边缘检测器[22]的启发, 使用卷积开始-结束检测器 (convolutional start-end detector, ConvSE) 和两个一维卷积滤波器来学习检测分数曲线中目标片段的开始边缘和结束边缘. 目标片段与查询语句在语义空间中彼此接近, 并且它们的相似度高于目标时刻之外彼此远离的查询片段, 自然地在目标时刻周围形成可检测的边缘. 具体来说, 本文使用两个滤波器来生成开始时刻预测分数 S_{start} 和结束时刻预测分数 S_{end}:

$$S_{\mathrm{start/end}}=Conv1D_{\mathrm{start/end}}(S_{\mathrm{video\text{-}query}})\in\mathbb{R}^{n_v} \tag{15}$$

$$S_{\mathrm{video\text{-}query}}=H_{\mathrm{qcf}}^{\top}\cdot q \tag{16}$$

其中, $S_{\mathrm{video\text{-}query}}$ 表示视频和查询语句的相似性分数.

基于最大化联合概率的方法来预测与查询相匹配的目标片段的开始和结束时刻. 对于预测片段 $(\hat{i}^s,\hat{i}^e)$ 的置信

分数 p^{se} 计算方法如下所示:

$$(\hat{i}^s,\hat{i}^e)=\arg\max_{a^s,a^e}P_{\text{start}}(a^s)\times P_{\text{end}}(a^e) \tag{17}$$

$$p^{se}=P_{\text{start}}(\hat{i}^s)\times P_{\text{end}}(\hat{i}^e) \tag{18}$$

其中, $P_{\text{start/end}}$ 是 $S_{\text{start/end}}$ 经过 Softmax 函数标准化后得到的表示概率的参数, $\arg\max()$ 指的是函数输出最大值时的输入或参数值, a^s 、 a^e 表示以 $\hat{i}^s$ 和 $\hat{i}^e$ 为自变量的函数, $0\leqslant\hat{i}^s\leqslant\hat{i}^s\leqslant n_v-1$.

时刻定位模块损失函数为:

$$\mathcal{L}^{\text{ML}}=\frac{1}{2}[f(P_{\text{start}},Y_{\text{start}})+f(P_{\text{end}},Y_{\text{end}})] \tag{19}$$

其中, f 表示交叉熵损失函数, Y_{start} 和 Y_{end} 是 i^s 和 i^e 的独热编码.

综上, 可以得到 QACLN 模型最终的损失函数, 如公式 (20) 所示:

$$\mathcal{L}=\lambda_1\times\mathcal{L}^{\text{VR}}+\lambda_2\times\mathcal{L}^{\text{ML}}+\lambda_3\times\mathcal{L}^{\text{VideoCL}}+\lambda_4\times\mathcal{L}^{\text{FrameCL}}+\lambda_5\times\mathcal{L}^{\text{VVCL}} \tag{20}$$

其中, 超参数 λ_1 、 λ_2 、 λ_3 、 λ_4 和 λ_5 分别对应于视频检索模块、时刻定位模块、模态间视频级对比学习 VideoCL 、模态间帧级 FrameCL 和模态内对比学习 VVCL 的权重系数.

3 实验结果与分析

3.1 数据集与性能评价指标

本文采用 TVR 数据集[6]验证本文提出方法的性能, 并进行实验对比. 该数据集包含 10.9 万个查询和 21.8 万个视频, 这些视频收集于 6 个不同类型的电视节目中, 其中每个查询语句都与一个视频紧密关联. 视频中包含字幕, 平均长度为 76.2 s, 查询语句平均包含 13.4 个单词. 以 1.5 s 为单位将每个视频划分为若干视频片段. 由于查询文本的描述可能与连续视频片段相关, 所以视频与查询文本的匹配不是一对一的.

本文对以下 3 个跨模态视频检索任务进行了实验: 视频语料库时刻检索 (VCMR)、单视频时刻检索 (SVMR) 和视频检索 (VR). 性能评价评估指标是召回率 Recall@K, 指前 top$K$ 结果中检索出的相关结果数和库中所有的相关结果数的比率, 衡量的是检索系统的查全率. 设置 $K\in\{1,10,100\}$. 为 VCMR 和 SVMR 任务设置了附加的 IoU 阈值 Recall@K, IoU = μ . 规定如下:若预测片段与目标片段的重叠度高于设定阈值, 则认定预测结果为正确. 这里, 使用交并比 (IoU) 度量预测片段与目标片段的重叠度.

3.2 对比方法

为了验证本文提出的 QACLN 的性能, 将其与目前几种优秀的跨模态视频检索算法进行实验对比. MCN[23]方法是基于局部和全局特征的跨模态时刻定位网络, 是领域内开山之作. CAL[5]是第 1 个提出视频语料库时刻检索 (VCMR) 任务的工作. MEE[24]主要关注如何处理大规模注释视频字幕数据集缺乏的问题. ExCL[25]专门为单视频时刻检索任务设计, 是一种细粒度的跨模态交互学习方法. XML[6] 用于多模态时刻检索任务, 使用了一种新颖的卷积起始-结束检测器 (ConvSE), 为未来的工作提供了一个强有力的起点. ReLoCLNet[7]在 XML 的基础上, 增加了模态间的对比学习方法. HERO[12]通过分层方法逐渐融合粗粒度、细粒度和多模态信息. HAMMER[8]通过查询文本和视频之间的细粒度跨模态交互学习来联合训练视频检索和时刻定位模型. FLAT 是 HAMMER 的一个变体, 区别在于 FLAT 省略了分层注意力机制. 其中, MCN[23]、CAL[5]、MEE[24]、XML[6]、ReLoCLNet[7]和 HERO[12]属于后期融合的方法, FLAT[8]和 HAMMER[8]属于前期跨模态交互融合的学习方法.

3.3 实验设置

使用 ResNet-152[26]、I3D[27]来提取视频的时空特征, 使用 12 层 RoBERTa[28]提取文本特征. 整个模型训练时长大约为 4 h 15 min. 学习率设置为 1E–4. 超参数 λ_1 、 λ_2 、 λ_3 、 λ_4 和 λ_5 分别对应于视频检索模块、时刻定位模块、模态间视频级对比学习 VideoCL、模态间帧级 FrameCL 和模态内对比学习 VVCL 的权重系数, 参数的取值在第 3.4.4 节中详细讨论.

3.4 实验结果与分析

3.4.1 模型性能对比实验结果与分析

表 1 报告了 VCMR 任务下不同方法在 TVR 数据集上的跨模态检索的 Recall 指标值. 观察表 1 的实验结果可以发现, QACLN 在 Recall@10 和 Recall@100 两个指标上优于所有基准方法. HERO 是在大型数据集 (How-100M) 上预先训练的, 具有强大的性能, 然而, 它仍然受限于未使用查询感知来自适应融合多模态视频特征. 与 HAMMER 和 FLAT 相比, QACLN 的性能在 Recall@10 和 Recall@100 表现更佳, 这是因为 HAMMER 和 FLAT 更多地关注于正样例之间距离的减小, 并未考虑对负样例的处理. QACLN 不仅关注于最小化正样例之间的距离, 还加大负样例之间距离. 由于 ReLoCLNet 只考虑了模态间对比学习方法, 而忽略了保持同模态数据的语义一致性, 故 QACLN 的性能略好于 ReLoCLNet.

表 1 VCMR 任务下不同方法 (Recall@K) 的性能比较 (%)

方法	IoU=0.5			IoU=0.7		
	K=1	K=10	K=100	K=1	K=10	K=100
XML[6]	—	—	—	2.62	9.05	22.47
HERO[12]	—	—	—	2.98	10.65	18.25
ReLoCLNet[7]	7.42	20.01	43.34	3.74	13.21	31.02
FLAT[8]	8.45	21.14	30．75	4.61	11.29	16.24
HAMMER[8]	**9.19**	21.28	31.25	**5.13**	11.38	16.71
QACLN	7.46	**21.75**	**44.23**	3.97	**13.35**	**31.83**

表 2 和表 3 列出了针对 VR 和 SVMR 任务提出的 QACLN 模型在 TVR 数据集上与各基准方法对比的结果. QACLN 远远超过了 VR 子任务上不同 Recall@K 的所有基准方法. 由于没有考虑使用对比学习来指导语义空间中样例对之间的距离, XML 表现出比 QACNL 更低的性能. ReLoCLNet 在 XML 模型的基础上引入对比学习, 且获得了令人满意的结果. 然而, 与 QACLN 相比, 由于忽略了模态内数据的语义关联, 导致性能受到限制. 从表 3 中还可以发现 QACLN 优于 SVMR 子任务上的所有基线. 这是由于提出的模型具有如下几点先进性. 首先, QACLN 模型在训练时采用了联合训练的方法, 将文本模态和视觉模态的信息融合在一起进行训练, 使得模型能够更好地学习到两种模态之间的关联性; 其次, QACLN 模型采用了双流注意力机制, 分别对文本和视频特征进行注意力加权, 提高了模型对关键信息的关注度从而更准确地进行视频时刻检索; 最后 QACLN 模型采用了层次结构编码器, 能够更好地对视频进行多层次的表示学习, 从而更好地将视频特征与文本特征进行对齐.

表 2 VR 任务下不同方法的性能比较 (%)

方法	VR, Recall@K		
	K=1	K=10	K=100
MCN[23]	0.05	0.66	3.59
CAL[5]	0.28	1.68	8.55
MEE[24]	7.56	29.88	73.07
ReLoCLNet[4,7]	21.1	55.5	89.88
QACLN	**21.19**	**57.18**	**90.49**

表 3 SVMR 任务下不同方法的性能比较 (%)

方法	SVMR, Recall@1	
	IoU=0.5	IoU=0.7
MCN[23]	13.08	5.06
CAL[5]	12.07	4.68
ExCL[25]	31.34	14.19
ReLoCLNet[4,7]	30.53	14.23
QACLN	**32.16**	**14.35**

图 5 显示了当 IoU 阈值设定为不同值时, VCMR 任务在 Recall@1 和 Recall@10 两个评价指标下的不同值. 选择 ReLoNet 作为对比方法, 它与 ReLoCLNet 的结构相似, 但 ReLoNet 采用对比学习机制. 由图 5 结果可见, QACLN 总是优于 ReLoNet. 这说明提出的跨模态双重对比学习机制在视频时刻检索任务中起着不可或缺的作用. 图 6 还显示了 QACLN 和 ReLoNet 在 TVR 数据集上在 K 取不同值时 VR 任务的结果. 同样, 当 μ 取从 1–10 不同值时, QACLN 仍超过了基准方法. 并且, 在更严格的指标下, 即 μ 数值越大时, 相对性能改善率更高.

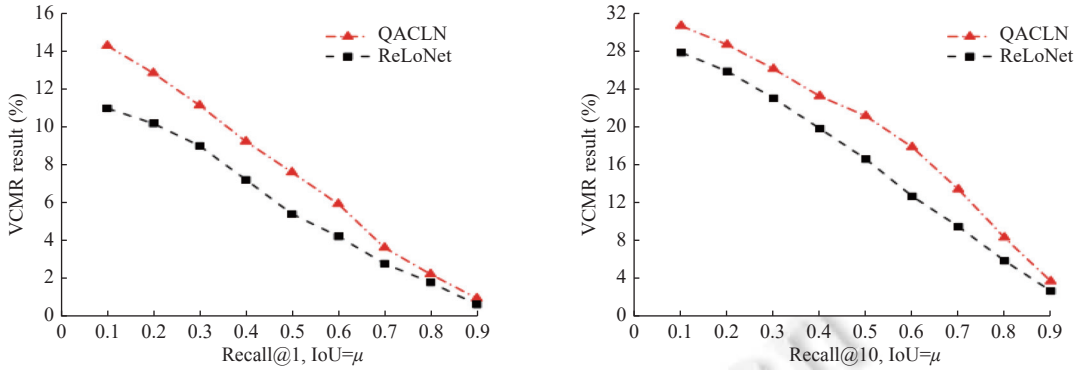

图 5 IoU 为不同值时, VCMR 在 Recall@1 和 Recall@10 下的性能分析

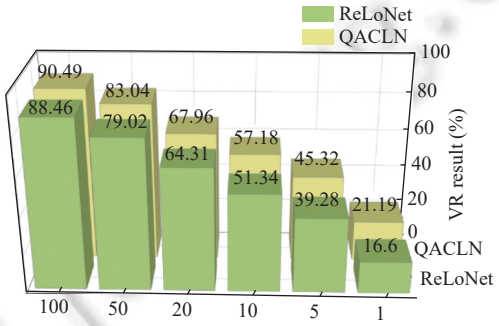

图 6 VR 在不同 Recall@K 下的性能分析

3.4.2 模型时间效率对比实验结果与分析

如表 4, 本文对 QACLN 模型的检索效率也进行了验证. 可以观察到, QACLN 的检索效率远高于 HAMMER. 根据表 1 可知, 在 Recall@100 指标下 QACLN 的检索精度与 HAMMER 相近, 尽管 HAMMER 在更严格的度量 (例如 Recall@1, IoU=0.5) 上效果略好于 QACLN, 但是提出的模型在检索效率方面比 HAMMER 快 41 倍. 由于 HAMMER 基于跨模态注意力机制来融合文本和视觉模态之间的信息, 属于前期融合编码方法, 从而导致高计算成本. 总之, 本文提出的 QACLN 模型既保证了检索效率, 又保证了检索精度. 总体而言, 提出的 QACLN 模型具有更好的性能.

表 4 QACLN 与 HAMMER 检索效率对比

方法	总时间 (s)	平均时间 (ms)
HAMMER[8]	2378.67	218.33
QACLN	58.01	5.32

3.4.3 消融实验

为了验证本文提出的 QACLN 模型中各个组件的有效性, 对提出的模型的 3 种不同的变种 QACLN w/o inter, QACLN w/o intra 和 QACLN w/o qcf 进行对比分析. QACLN w/o inter 表示在 QACLN 的基础上去掉模态间对比学习模块. QACLN w/o intra 表示在 QACLN 的基础上去掉模态内对比学习模块. QACLN w/o qcf 表示在 QACLN 的基础上去掉 QCF 模块.

由表 5 可知, 与 QACLN w/o inter 相比, QACLN 取得了更好的性能, 这主要是因为模态间对比学习可以进一步增强不同模态数据的语义表示一致性. 模态间对比学习包括两个对比对象: VideoCL 和 FrameCL. VideoCL 使用噪声对比损失来增强与正样例之间的相似度, 并降低与负样例之间的相似度, 这与跨模态视频文本检索的目标一

致, 因此, 它有利于跨模态视频检索任务结果的提升. FrameCL 旨在区分视频中的目标时刻和非目标时刻, 使其更加准确地区分目标时刻开始与结束边界. FrameCL 通过更好地定位与查询语句匹配的视频片段起止时刻, 来帮助视频的细粒度检索任务取得更好的结果. 与 QACLN w/o intra 相比, QACLN 获得了更好的性能. 因为简单地执行模态间对齐忽略了每个模态中的数据语义相似性, 从而导致学习到的特征表示降级. 模态内对比学习的目标是通过比较同一模态内的不同样例的相似性来学习更好的模态内表示, 从而提供模态内数据的互补优势, 并确保来自同一模态的相似输入的语义距离更加接近. 与 QACLN w/o qcf 相比, QACLN 的性能更优, 这表明通过查询感知来学习视频的多模态融合特征有助于提升视频的时刻定位和视频检索性能.

表 5 QACLN 消融实验结果 (Recall@K)(%)

模型	VCMR						SVMR					
	IoU=0.5			IoU=0.7			IoU=0.5			IoU=0.7		
	K=1	K=10	K=100	K=1	K=10	K=100	K=1	K=10	K=100	K=1	K=10	K=100
QACLN w/o inter	6.57	16.1	34.02	3.25	11.14	26.09	29.47	61.74	84.05	13.71	43.37	70.86
QACLN w/o intra	7.42	20.01	43.34	3.74	13.21	31.02	30.53	62.43	85.19	14.23	44.39	71.13
QACLN w/o qcf	7.57	21.29	43.36	3.95	**13.69**	31.6	30.78	62.46	85.87	14.07	44.63	71.07
QACLN	**7.64**	**21.75**	**44.23**	**3.97**	13.35	**31.83**	**32.16**	**63.31**	**86.57**	**14.47**	**45.03**	**71.68**

3.4.4 参数敏感性分析实验

在本文中设定了不同的模块参数, 以评估其对跨模态表示学习与搜索的影响. 主要关注模态间对比学习 (λ_3, λ_4) 和模态内对比学习 (λ_5) 的贡献程度, 因此固定 λ_1 =1、λ_2 =0.01. 通过调整损失权重 λ_3 、λ_4 和 λ_5 的取值控制模态间对比学习 (λ_3, λ_4) 和模态内对比学习 (λ_5) 的贡献程度. 设置了 4 组参数: (1) $\{\lambda_3=0.01,\lambda_4=0.01,\lambda_5=0.001\}$; (2) $\{\lambda_3=0.05,\lambda_4=0.05,\lambda_5=0.001\}$; (3) $\{\lambda_3=0.03,\lambda_4=0.03,\lambda_5=0.001\}$; (4) $\{\lambda_3=0.03,\lambda_4=0.03,\lambda_5=0.01\}$. 由表 6 中实验结果可知, 损失权重取值为参数组 (3) 时得到最佳结果, 因此本文在所有实验中都使用这组参数值. 值得注意的是, 当损失权重取值为参数组 (4) 时, 模型性能大大降低, 这证明模态间对比学习比模态内对比学习在跨模态语义表示学习过程中发挥更重要的作用.

表 6 参数敏感性分析实验结果 (%)

参数组	VCMR, Recall@10		SVMR, Recall@1		VR, Recall@10
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7	
(1)	20.75	13.06	60.23	41.82	52.81
(2)	20.72	**13.58**	61.56	43.68	56.85
(3)	**21.75**	13.35	**63.31**	**45.03**	**57.15**
(4)	18.3	11.48	62.45	43.95	56.67

4 总 结

本文提出了一种面向多模态视频时刻检索的查询感知跨模态双重对比学习网络 (QACLN), 建立了跨模态双重对比学习机制, 联合模态内与模态间对比学习来增强不同模态数据的语义对齐和融合, 提高不同模态数据表示的可判别性和语义一致性. 此外, 提出了一种基于查询感知的跨模态语义融合策略, 通过自适应地融合视频中视觉模态和字幕模态等多模态的数据来学习获得视频的多模态联合语义表示. 实验结果表明提出的 QACLN 在视频检索、视频时刻定位与检索任务中均具有高效率和高检索精度的优点.

References:

[1] Liang MY, Du JP, Yang CX, Xue Z, Li HS, Kou FF, Geng Y. Cross-media semantic correlation learning based on deep hash network and semantic expansion for social network cross-media search. IEEE Trans. on Neural Networks and Learning Systems, 2020, 31(9): 3634–3648. [doi: 10.1109/TNNLS.2019.2945567]

[2] Liu DZ, Qu XY, Dong JF, Zhou P, Cheng Y, Wu ZC, Xie YL. Context-aware biaffine localizing network for temporal sentence

grounding. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 11230–11239. [doi: 10.1109/CVPR46437.2021.01108]

[3] Wang Z, Chen JJ, Jiang YG. Visual Co-occurrence alignment learning for weakly-supervised video moment retrieval. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 1459–1468. [doi: 10.1145/3474085.3475278]

[4] Yang X, Wang SS, Dong J, Dong JF, Wang M, Chua TS. Video moment retrieval with cross-modal neural architecture search. IEEE Trans. on Image Processing, 2022, 31: 1204–1216. [doi: 10.1109/TIP.2022.3140611]

[5] Escorcia V, Soldan M, Sivic J, Ghanem B, Russell B. Temporal localization of moments in video collections with natural language. arXiv:1907.12763V1, 2019.

[6] Lei J, Yu LC, Berg TL, Bansal M. TVR: A large-scale dataset for video-subtitle moment retrieval. In: Proc. of the 16th European Conf. on Computer Vision (ECCV 2020). Glasgow: Springer, 2020. 447–463. [doi: 10.1007/978-3-030-58589-1_27]

[7] Zhang H, Sun AX, Jing Wei, Nan GS, Zhen LL, Zhou JT, Goh RSM. Video corpus moment retrieval with contrastive learning. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2021. 685–695. [doi: 10.1145/3404835.3462874]

[8] Zhang BW, Hu HX, Lee J, Zhao M, Chammas S, Jain V, Le E, Sha F. A hierarchical multi-modal encoder for moment localization in video corpus. arXiv:2011.09046, 2020.

[9] Hou ZJ, Ngo CW, Chan WK. CONQUER: Contextual query-aware ranking for video corpus moment retrieval. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 3900–3908. [doi: 10.1145/3474085.3475281]

[10] Wang YX, Tian JR, Chen ZD, Luo X, Xu XS. Label enhancement based discrete cross-modal hashing method. Ruan Jian Xue Bao/Journal of Software, 2023, 34(7): 3438–3450 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6536.htm [doi: 10.13328/j.cnki.jos.006536]

[11] He XT, Peng YX. Unsupervised fine-grained video categorization via adaptation learning across domains and modalities. Ruan Jian Xue Bao/Journal of Software, 2021, 32(11): 3482–3495 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6058.htm [doi: 10.13328/j.cnki.jos.006058]

[12] Li LJ, Chen YC, Cheng Y, Gan Z, Yu LC, Liu JJ. HERO: Hierarchical encoder for video+language omni-representation pre-training. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 2046–2065. [doi: 10.18653/v1/2020.emnlp-main.161]

[13] Chen T, Kornblith S, Norouzi M, Hinton GE. A simple framework for contrastive learning of visual representations. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 149.

[14] Zhang BC, Li L, Zha ZJ, Huang QM. Contrastive cross-modal representation learning based active learning for visual question answer. Chinese Journal of Computers, 2022, 45(8): 1730–1745 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2022.01730]

[15] Bell AJ, Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 1995, 7(6): 1129–1159. [doi: 10.1162/neco.1995.7.6.1129]

[16] Belghazi MI, Baratin A, Rajeswar S, Ozair S, Bengio Y, Hjelm RD, Courville AC. Mutual information neural estimation. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 530–539.

[17] Buchwalter W, Hjelm RD, Bachman P. Learning representations by maximizing mutual information across views. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2019. 1392.

[18] van Den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

[19] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[20] Zhu LC, Yang Y. ActBERT: Learning global-local video-text representations. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8743–8752. [doi: 10.1109/CVPR42600.2020.00877]

[21] Miech A, Alayrac JB, Smaira L, Laptev I, Sivic J, Zisserman A. End-to-end learning of visual representations from uncurated instructional videos. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9876–9886. [doi: 10.1109/CVPR42600.2020.00990]

[22] Szeliski R. Computer Vision: Algorithms and Applications. London: Springer, 2011.

[23] Hendricks LA, Wang O, Shechtman E, Sivic J, Darrell T, Russell B. Localizing moments in video with natural language. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5804–5813. [doi: 10.1109/ICCV.2017.618]

[24] Miech A, Laptev I, Sivic J. Learning a text-video embedding from incomplete and heterogeneous data. arXiv:1804.02516, 2018.

[25] Ghosh S, Agarwal A, Parekh Z, *et al*. ExCL: Extractive clip localization using natural language descriptions. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis:

Association for Computational Linguistics, 2019. 1984–1990. [doi: 10.18653/v1/N19-1198]

[26] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[27] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4724–4733. [doi: 10.1109/CVPR.2017.502]

[28] Liu YH, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized bert pretraining approach. arXiv:1907.11692, 2019.

附中文参考文献:

[10] 王永欣, 田洁茹, 陈振铎, 罗昕, 许信顺. 基于标记增强的离散跨模态哈希方法. 软件学报, 2023, 34(7): 3438–3450. http://www.jos.org.cn/1000-9825/6536.htm [doi: 10.13328/j.cnki.jos.006536]

[11] 何相腾, 彭宇新. 跨域和跨模态适应学习的无监督细粒度视频分类. 软件学报, 2021, 32(11): 3482–3495. http://www.jos.org.cn/1000-9825/6058.htm [doi: 10.13328/j.cnki.jos.006058]

[14] 张北辰, 李亮, 查正军, 黄庆明. 基于跨模态对比学习的视觉问答主动学习方法. 计算机学报, 2022, 45(8): 1730–1745. [doi: 10.11897/SP.J.1016.2022.01730]

尹梦冉(1999－), 女, 硕士生, CCF 学生会员, 主要研究领域为跨模态检索, 自然语言处理.

曹晓雯(1998－), 女, 硕士生, 主要研究领域为深度学习, 跨模态搜索.

梁美玉(1985－), 女, 教授, 博士生导师, CCF 专业会员, 主要研究领域为人工智能, 跨模态数据挖掘与搜索, 计算机视觉.

杜军平(1963－), 女, 教授, 博士生导师, CCF 会士, 主要研究领域为人工智能, 机器学习, 模式识别.

于洋(2000－), 男, 硕士生, 主要研究领域为深度学习, 跨模态检索.

薛哲(1987－), 男, 副教授, CCF 专业会员, 主要研究领域为机器学习, 数据挖掘, 多媒体数据分析.