软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(2):985−1006 [doi: 10.13328/j.cnki.jos.006707]

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

视频片段检索研究综述*

王 妍[1], 詹雨薇[1], 罗 昕[1], 刘 萌[2], 许信顺[1]

[1](山东大学 软件学院, 山东 济南 250101)

[2](山东建筑大学 计算机科学与技术学院, 山东 济南 250101)

通信作者: 罗昕, E-mail: luoxin@sdu.edu.cn

摘 要: 视频片段检索旨在利用用户给出的自然语言查询语句, 在一个长视频中找到最符合语句描述的目标视频片段. 视频中包含丰富的视觉、文本、语音信息, 如何理解视频中提供的信息, 以及查询语句提供的文本信息, 并进行跨模态信息的对齐与交互, 是视频片段检索任务的核心问题. 系统梳理了当前视频片段检索领域中的相关工作, 将它们分为两大类: 基于排序的方法和基于定位的方法. 其中, 基于排序的方法又可细分为预设候选片段的方法和有指导地生成候选片段的方法, 而基于定位的方法则可分为一次定位的方法和迭代定位的方法. 同时对该领域的数据集和评价指标进行了介绍, 并对一些模型在多个常用数据集上的性能进行了总结与整理. 此外, 介绍了该任务的延伸工作, 如大规模视频片段检索工作等. 最后, 对视频片段检索未来的发展方向进行了展望.

关键词: 视频片段检索; 自然语言时序定位视频片段; 视频理解; 深度学习; 人工智能

中图法分类号: TP391

中文引用格式: 王妍, 詹雨薇, 罗昕, 刘萌, 许信顺. 视频片段检索研究综述. 软件学报, 2023, 34(2): 985–1006. http://www.jos.org.cn/1000-9825/6707.htm

英文引用格式: Wang Y, Zhan YW, Luo X, Liu M, Xu XS. Survey on Video Moment Retrieval. Ruan Jian Xue Bao/Journal of Software, 2023, 34(2): 985−1006 (in Chinese). http://www.jos.org.cn/1000-9825/6707.htm

Survey on Video Moment Retrieval

WANG Yan[1], ZHAN Yu-Wei[1], LUO Xin[1], LIU Meng[2], XU Xin-Shun[1]

[1](School of Software, Shandong University, Jinan 250101, China)

[2](School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China)

Abstract: Given a natural language sentence as the query, the task of video moment retrieval aims to localize the most relevant video moment in a long untrimmed video. Based on the rich visual, text, and audio information contained in the video, how to fully understand the visual information provided in the video and utilize the text information provided by the query sentence to enhance the generalization and robustness of model, and how to align and interact cross-modal information are crucial challenges of the video moment retrieval. This study systematically sorts out the work in the field of video moment retrieval, and divides them into ranking-based methods and localization-based methods. Thereinto, the ranking-based methods can be further divided into the methods of presetting candidate clips, and the methods of generating candidate clips with guidance; the localization-based methods can be divided into one-time localization methods and iterative localization ones. The datasets and evaluation metrics of this fieldf are also summarized and the latest advances are reviewed. Finally, the related extension task is introduced, e.g., moment localization from video corpus, and the survey is concluded with a discussion on promising trends.

Key words: video moment retrieval; temporal activity localization via language; video understanding; deep learning; artificial intelligence

当下, 社交网络与在线视频平台的兴起, 致使各种各样的未剪辑视频呈爆炸式增长. 对于视频的分析[1]

* 基金项目: 国家自然科学基金(61991411, 61872428, 62006142, 62172256); 山东省重点研发项目(2019JZZY010127); 山东省自然科学基金(ZR2019ZD06, ZR2020QF036)

收稿时间: 2021-04-29; 修改时间: 2021-09-28, 2022-02-17; 采用时间: 2022-05-20; jos 在线出版时间: 2022-07-22

与研究[2]也逐渐成为热点问题. 为满足人们对于搜寻长视频中具有特定语义含义片段的需求, 视频片段检索任务应运而生. 视频片段检索任务需要根据查询语句, 从一个长视频中检索到最符合语句描述的视频片段. 具体来说, 数据集中, 每个被标注的视频片段都与一组注释相关: $\{q,t_s^g,t_e^g\}$. 在检索时, 给定一个查询语句 q, 需要在给定的视频 v 中找到与查询语句 q 最匹配的片段, 并返回片段的起止时间点 t_s^g、t_e^g. 本文中出现的符号及含义见表 1. 在图 1 的示例中, 给定一个完整的视频 v 和一条“a person is eating a sandwich (一个人正在吃三明治)”的查询语句 q, 视频片段检索模型需要在视频 v 中找到与 q 最匹配的视频片段, 并同时预测该片段的开始点和结束点 t_s^g、t_e^g.

表 1 视频片段检索符号

符号	含义	符号	含义
q	查询语句	–	–
v	视频	c	候选片段
t_s	片段的开始时间点	τ_s	候选片段的开始点
t_e	片段的结束时间点	τ_e	候选片段的结束点
g	真实标签上标	s	候选片段的输出分数
p	预测标签上标	M	生成的候选片段的个数
N	批量大小	o	候选片段与真实标签片段的 IoU 值
P	时序点对应的概率值	y	缩放的 IoU 值

图 1 视频片段检索任务的应用示例

一些真实场景中的视频, 如机器人导航[3]、自动驾驶[4]以及监控中的视频[5]等, 包含太多无意义的片段, 如在监控视频中, 异常视频片段出现的时间和频率远远少于正常片段. 使用视频片段检索则可以从长时间的视频中找出异常片段, 从而达到提升效率的目的. 这看起来是一项有挑战性的任务, 因为我们不仅需要理解视频的内容、查询语句的语义信息, 还需要将不同模态的信息进行精确的匹配, 从而达到我们的目的.

视频片段检索任务与动作时序定位任务一脉相承, 区别在于动作时序定位没有办法满足对于包含对象的具体事件的查询. 定位空间语句[6]也是视频片段检索任务的相关任务之一, 其可以视为视频片段检索任务的前期探索. 定位空间语句将视频类别限制为监控视频, 查询语句限制为位置描述语句. Regneri 等人[7]在 MPII Composites 数据集[8]的基础上构建了 TACoS 数据集, 这个数据集在视频片段检索领域得到了广泛的应用. 2017 年, Hendricks 等人[9]和 Gao 等人[10]不约而同地提出了视频片段检索的模型, 他们均采取了将视频划分为候选片段, 然后从中挑选出最匹配片段的方式. 他们不仅简化了问题的复杂度, 而且各自贡献了用于视频片段检索任务的数据集 DiDeMo 及 Charades-STA. 特别地, Gao 等人[10]提出的模型框架更成为了基础框架之一, 其思想被之后的研究工作广泛借鉴, 例如: 通过引入注意力机制、更细致的跨模态交互等技术模块, 使检索性能得到提升. 2019 年, Escorcia 等人[11]将该任务由单个视频的检索扩展到面向视频库的检索, 使视频片段检索任务具有更加广泛的应用可能性, 我们将其称为大规模视频片段检索任务, 并在第 4 节进行简要介绍.

本文前言简短地对视频片段检索任务进行介绍. 根据解决问题角度的不同, 我们在本文中将现有的视频片段检索方法大致分为基于排序的方法和基于定位的方法两类, 并分别在第 1 节、第 2 节进行介绍. 第 3 节对当前视频片段检索任务的常用数据集进行介绍, 并对已有模型在一些数据集上的实验结果进行分析与比

较. 第 4 节中, 我们对模型的类别、提出时间及优缺点进行总结. 第 5 节介绍与该任务有关的探索工作, 并对该领域面临的挑战和发展趋势进行展望.

1 基于排序的方法

本节首先介绍视频检索片段中的基于排序的方法, 这类方法的核心在于对候选片段进行排序. 这种解决方案由于实施简单, 易于解释和理解, 成为了视频片段检索领域的主流方案之一. 具体来说, 根据产生候选片段的过程不同, 可进一步将基于排序的方法细分为预设候选片段的方法和有指导地生成候选片段的方法: 前者是人为地、穷举地切分视频为候选片段, 然后按照与查询语句的相关程度对它们进行排序; 后者则首先利用模型排除掉大多数无关的候选片段, 然后再对生成的候选片段排序. 从模型的输入来看, 预设候选片段的方法会直接将预先切分好的视频片段送入模型, 而有指导地生成候选片段的方法则以视频为模型的输入. 图 2 展示了两者在思路上的区别.

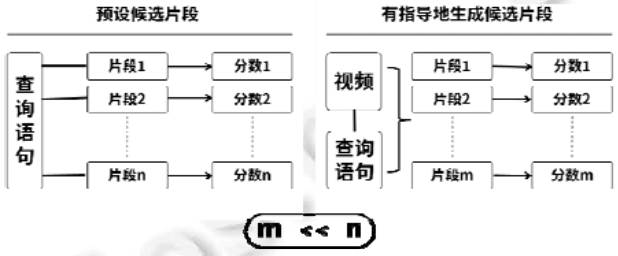

图 2 预设候选片段的方法与有指导地生成候选片段的方法的区别

1.1 预设候选片段的方法

预设候选片段的方法需要在无查询语句信息的情况下对视频预先地进行划分, 生成可能的候选片段集合. 这种方法借鉴了多示例学习的思想, 在训练阶段, 每个候选片段可以被看作是一个带有标签的示例. 划分预设的候选视频片段的方法可以分为以 Gao 等人[10]提出的模型为代表的方法和以 Hendricks 等人[9]提出的模型为代表的方法: 第 1 类方法以跨模态时序回归定位器(cross-modal temporal regression localizer, CTRL)[10]为代表, 以有 80%重叠部分的不同尺度滑动窗口为基准, 对视频进行划分; 第 2 类方法以时刻上下文网络(moment context network, MCN)[9]为代表, 对视频直接进行相同尺度的切分. 这两类方法均采取了先提取多个候选片段再多中选优的思路. 具体来说, 这两类方法最主要的区别是: 相对于 MCN 模型, CTRL 模型构建的滑动窗口片段是层级化的, 且片段之间有大面积重叠, 如图 2 所示.

在预设候选片段的方法中, 第 1 类方法由于更易获得高的准确率而得到更广泛的应用. 以 CTRL 模型为例, 它的整体思路比较简练. 具体地, 它首先分别对候选视频片段和查询语句进行图像和语义理解, 得到相应的特征表示; 然后对多模态表示进行融合, 以进行候选片段的定位与评价. CTRL 模型基本框架如图 3 所示, 它使用 4 个尺度([64 帧、128 帧、256 帧、512 帧])对视频进行片段划分, 其中, 同尺度相邻候选片段间有 80%的重叠. 将得到的视频片段分别与查询语句进行相关性计算, 进而得到相关性由高到低的片段序列.

如图所示, 整个模型基本分为 4 个部分: (1) 对视频滑动窗口片段编码, 使用 C3D 网络[12]直接对片段进行处理, 得到视觉特征表示; (2) 对查询语句编码, 使用 Skip-thought 模型[13]直接获得查询语句的文本特征表示; (3) 融合视觉和文本特征表示, 从而得到跨模态特征表示, 这个过程的关键在于尽可能地保留两个模态的信息以及探索模态间的对应关系; (4) 利用设计的损失函数, 在跨模态特征表示的基础上进行目标片段定位与评价, 将得到的结果作为反馈, 从而进行整个网络的参数更新. 损失函数的设计在训练阶段十分重要, CTRL 模型采用对齐损失函数和定位回归损失函数, 分别如下:

$$L_{align}=\frac{1}{N}\sum_{i=0}^{N}\left[a_c\log(1+\exp(-s_{i,i}))+\sum_{j=0,j\neq i}^{N}a_w\log(1+\exp(s_{i,j}))\right] \tag{1}$$

其中, N 表示批量大小(batch size), $s_{i,j}$ 表示句子 q_j 和候选片段 c_i 之间的对齐分数, a_c 和 a_w 分别表示控制正负片段-查询对的权重. 该对齐损失是为了增大对齐的片段-查询对的对齐分数, 缩小未对齐的片段-查询对的对齐分数.

$$L_{reg}=\frac{1}{N}\sum_{i=0}^{N}[R(t_{s,i}^{*}-t_{s,i})+R(t_{e,i}^{*}-t_{e,i})] \tag{2}$$

$$t_s^{*}=t_s^{p}-\tau_s,\ t_e^{*}=t_e^{p}-\tau_e \tag{3}$$

$$t_s=t_s^{g}-\tau_s,\ t_e=t_e^{g}-\tau_e \tag{4}$$

其中, τ_s 和 τ_e 分别表示候选片段 c 对应的开始点和结束点, R 表示 L_1 正则函数. 该定位回归损失是为了令当前对齐的候选片段接近真实标签. 许多模型在 CTRL 的基础上进行了有意义的创新. 后期基于这种思路的文章, 一部分以模块为单位, 对模块进行改造与提升; 一部分以处理流为单位, 添加有特定目标的处理支线, 极大地提高了模型性能.

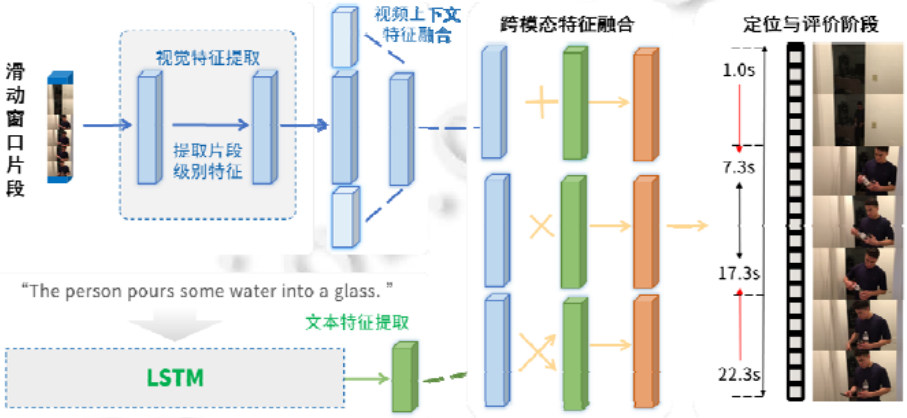

图 3 CTRL 模型基本框架

后期工作对于模型第 1 部分, 即视频编码模块的创新, 主要集中于融入片段的上下文特征. 如专注的跨模态检索网络(attentive cross-modal retrieval network, ACRN)[14]使用注意力机制融合了上下文片段, 还有一些模型尝试利用空间信息来辅助视频片段的时序定位. 如空间和语言时序注意(spatial and language-temporal attention, SLTA)方法[15]增加了物体的局部特征, 使用 Faster R-CNN[16]在视频片段中侦测物体, 结合查询语句中提取出的具体对象的特征, 将物体的视觉信息和文本信息一起融入到跨模态融合特征中, 为最后的回归定位提供了帮助. 跨模态交互网络(cross-modal interaction networks, CMIN)[17]将多头自注意力机制应用于视频片段编码阶段, 同时探索邻近帧和遥远帧对当前帧的影响, 进一步拓展了模型对上下文信息的利用能力. 跨模态和自模态图注意力网络(cross- and self-modal graph attention network, CSMGAN)[18]同样将自注意力机制融入视频片段的编码, 更大程度上为后续阶段保留了候选片段的信息.

模型的第 2 部分是对查询语句的编码. 查询语句作为对视频片段的描述, 包含了丰富的可供挖掘的信息, 在模型训练过程中起到了参考的作用. 时序组合模块化网络(temporal compositional modular network, TCMN)[19]充分利用了自然语言信息, 使用树注意力网络从查询语句中分别提取出主要事件、上下文事件和时序信号这 3 个信息, 再利用两个网络分别计算候选片段和语句的相关分数. Zhang 等人认为, 这种模型能够更好地对复杂长句进行检索. 跨模态交互网络(cross-modal interaction network, CMIN)利用句法图卷积网络构建了查询语句的句法结构树, 以获得更加细致的查询语句表示. 跨模态注意力(cross-modality attention, CMA)模型[20]使用语义词组提取网络, 提取出查询语句中的主要动作, 与查询语句的整体表示一起作为输入, 放入跨模态融合网络中. CSMGAN 模型对查询语句使用了层级化的编码, 构造了单词级别、词组级别及语句级别的特征, 不仅有助于强调语句中的重要信息, 同时可以更好地探索文本的上下文关系. 广义模态内和模态间的

多线性池化模型(generalized intra- and intermodal multilinear pooling model, GIIM)[21]使用词性标注方法提取出了查询语句中的动词, 并经过了标签平滑, 辅助定位任务的完成.

模型第3部分, 模态间交互模块是影响这一任务性能的关键模块. 在这一步中, 需要将两个模态的特征映射到同一维度空间. 目前, 很多模型在探索如何使用注意力机制使两部分的交互更加细致, 包括基于记忆的注意力、共同注意力、自注意力等. CTRL 模型对于跨模态交互部分的处理, 是将视觉特征向量和文本特征向量进行相乘、相加、级联等操作. 在此基础上, 跨模态时刻定位网络(cross-modal moment localization network, ROLE)[22]和 ACRN 两个模型分别使用注意力机制实现了每一帧要重点听取哪个单词以及查询语句要重点观察哪一帧的效果. CMA 模型应用多头自注意力机制去探索文本与视觉表示间更加细致的交互. 细粒迭代注意力网络(fine-grained iterative attention network, FIAN)[23]设计了两个对称的迭代注意力模块, 两个模块分别生成注意视频的文本特征以及注意语句的视觉特征, 然后进行模态间的融合, 最后引入预设候选视频片段的概念, 对各个片段进行打分和重定位. CSMGAN 模型在跨模态特征融合阶段融入了自模态模块, 构造图网络以更全面地计算不同模态间、相同模态内不同主体间的注意力权重. GIIM 模型在原有多模态分解双线性池化(multimodal factorized bilinear pooling, MFB)模型[24]上融入了模态内交互的概念, 得到了细致的跨模态融合表示. 潜在图的共注意力网络(latent graph co-attention network, LoGAN)[25]构造了共同注意力矩阵来指示帧与单词间的注意力分数, 同时设计了信息传递过程用于迭代地更新注意力, 从而引入上下文帧的影响. VAL 模型[26]在CTRL模型提供的跨模态交互前, 在视觉表示的基础上分别应用了通道注意力机制和体元注意力机制, 从而在查询语句的指导下, 保存视频的时序维度和空间维度的信息. 跨模态关系对齐的图卷积框架(CrossGraphAlign)[27]利用文本关系图与视觉关系图来建模查询文本与视频片段中的语义关系, 再通过跨模态对齐的图卷积网络来评估文本关系与视觉关系的相似度, 构建了更加精确的片段检索系统.

这些经典的模态间交互模块深入挖掘了两个模态间的关系, 但繁杂的交互过程也是限制检索速度的瓶颈. 快速视频片段检索(fast video moment retrieval, FVMR)模型[28]率先从检索速度的角度出发, 提出将跨模态交互模块替换为公共子空间, 在测试阶段直接对映射到公共子空间的特征表示做点乘, 大大提升了检索效率.

模型的第 4 部分是对目标片段的定位与评价. 这一部分需要考虑对于损失函数的设计. 在预设候选片段的任务设定中, 损失函数一般包括分数指导的对齐损失函数以及时间节点指导的定位回归损失函数这两部分. GIIM 模型在此基础上, 利用从查询语句中经过词性标注方法提取出的动词, 结合视频提取出的动作进行动作损失函数的构建, 帮助任务的完成. 还有一类模型使用弱监督设定对损失函数部分进行改进, LoGAN 模型是其中的代表之一. 这类模型由于在训练阶段只使用视频和视频片段对应的查询语句而不给出查询语句对应片段的时序边界, 因此仅利用正负视频查询语句对构造损失函数, 而非利用正负片段查询语句对.

以原有 CTRL 模型的思路为主线进行整体流程创新的文章也有很多. 如基于活动概念的定位器(activity concepts based localizer, ACL)[29]在原有处理流上加入了一条活动概念挖掘流, 从动作时序定位的角度, 使视频片段检索任务聚焦活动概念的挖掘, 使得定位更加准确. 语言指导的网络(language guided network, LGN)[30]将文本表示应用到模型的每个过程中, 用于指导每一步表示的生成. Chen 等人[31]将CTRL模型使用的思路应用于粗定位阶段, 在挑选出最相似的滑动窗口片段后, 又将该片段的起止时间点进行扩展, 重新生成一系列滑动窗口, 进行更加细致的定位. 多模态关系图(multi-modal relational graph, MMRG)模型[32]为视频的片段和查询语句分别构建了关系图, 利用多任务的训练来增强关系图的表示, 最后用图匹配和边界回归两个优化目标来完成检索任务. MMRG 模型将视频和查询语句的信息具象化, 剔除掉了干扰性信息, 为检索过程创造了新的载体.

在使用CTRL模型思路的方法中, 弱监督语言定位网络(weakly supervised language localization network, WSLLN)[33]是使用弱监督方式进行学习的模型, 其在定位打分模块设计了对齐分支和侦测分支来生成两个分数, 分别用于指示候选片段与查询语句的相关程度和增强候选片段之间的竞争, 最终结合两个分数得到每个候选片段的最终得分. 值得一提的是: 由于弱监督视频片段检索任务中, 真实标签中仅提供了视频-查询语句对, 因此 WSLLN 的损失函数是通过将得分最高的片段设为伪正例, 而将其他得分低的片段设为反例进行构

造的.

在预设候选片段的方法中，第 2 类模型以 MCN 模型为代表. MCN 模型的基本框架如图 4 所示，其使用的思路与 CTRL 模型类似，但更加简练，包括视频的编码、文本的编码以及相似度计算这 3 个模块. 其中缺少细致的跨模态交互过程，因此在相似度计算时可能会丢失掉部分重要信息.

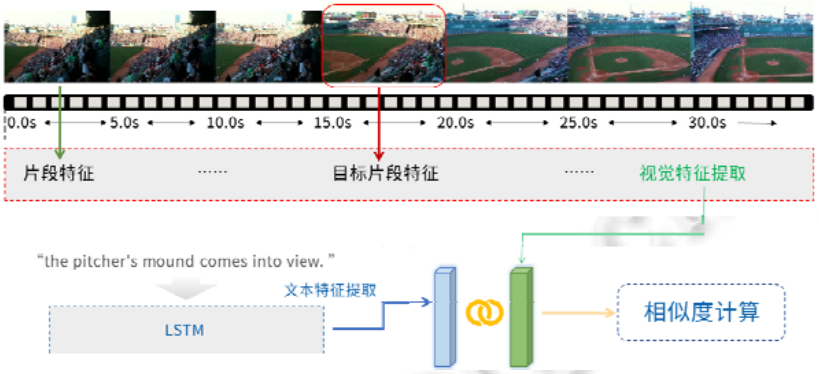

图 4 MCN 模型基本框架

MCN 模型将长度相当的视频统一切分为 5 s 的视频片段，然后将片段的视觉表示与查询语句的文本表示进行相关性计算. 时序模块网络(temporal modular network, TMN)[34]首先对视频进行同一较小尺度的切分，经过模块化时序网络的处理，得到每个小尺度片段与查询语句的相关性，并在最后将 n 个相邻连续的片段进行组合，则可得到 $\sum_{i=1}^{n} i$ 个结果.

视频片段检索任务中，对数据集的标注包括查询语句及其匹配片段的起止时间点两部分，工作量较大. 一直以来，研究者们也在积极探索弱监督和无监督的方法. Gao 等人[35]在预设候选片段的框架下，探索了一种无监督的视频片段检索方法，其利用视频条件下的句子生成器，自主地生成对视频的描述；又引入了图像句子嵌入空间，将生成的描述与分割好的片段进行相关分数的计算；最后结合片段本身的表示，找出得分最高的候选片段作为检索结果. 总的来说，无监督方法有两类思路：第 1 类是利用仅有的视频信息生成查询语句和时序点的伪标签，再应用现有模型的框架完成视频片段检索任务；第 2 类是将两个阶段融合到一起，同时完成伪标签的生成和片段检索两个任务，目前还未出现这个思路的工作. 并且，目前无监督的方法在性能上与有监督的方法存在较大差距，需要研究者们进一步加以探索.

1.2 有指导地生成候选片段的方法

有指导地生成候选片段，即以查询语句或视频本身为指导，探索视频中的哪个片段更有可能与查询语句相关，从而确定候选片段. 该类文献大多是先对整个视频进行处理，得到视频级别的视觉特征，在与查询语句进行模态间融合之后，根据结果产生候选片段，最后对每个候选片段计算分数. 由于是提取视频级别的特征，这类方式能够避免对重复的帧进行多次处理和计算. 根据生成候选片段的阶段不同，该类模型可以分为 3 类：第 1 类是在没有查询语句作为指导的情况下，直接利用视觉信息生成候选片段；第 2 类是在对查询语句进行编码之后，利用文本信息给出的指导生成候选片段；第 3 类则是在对视觉信息和文本信息进行跨模态融合之后，生成候选片段. 一般来讲，候选片段的生成时间越晚，计算资源的浪费就越少，但也更易丢失有用信息. 此外，在有指导地生成候选片段的方法中，弱监督学习和强化学习多次被使用.

第 1 类方法中，Xu 等人[36]在未使用查询语句作为指导的情况下，通过用于动作定位的 R-C3D 网络[37]来生成候选片段. 虽然这种方式生成的候选片段没有考虑与查询语句的相关程度，但是由于使用了预测离中心位置的相对偏移和一系列预设片段长度的方式，获得了长度可变的候选片段. Xu 等人使用视频描述作为辅助任务，通过构造重新生成的查询语句与原本查询语句间的重构损失，来更好地提升模型性能，查询指导的片段提出网络(query-guided segment proposal network, QSPN)[38]和弱监督的语义完成网络(weakly-supervised

semantic completion network, SCN)[39]也应用这种重构损失构造了损失函数.

2D 时序邻接网络(2D temporal adjacent network, 2D-TAN)模型[40]不属于严格意义上的有指导生成候选片段的方法, 它提出了使用 2D 时序邻接网络筛选候选视频片段的方式. 在该筛选过程中, 没有查询语句的指导和参与, 但其利用起始点时间和结束点时间定义两个维度, 在这两个维度构建的坐标系上均匀取样视频片段, 很好地筛选掉大部分无用的候选视频片段, 达到了生成候选片段的效果. 同时, 文献[40]提出的损失函数对模型起到了很好的优化效果, 被之后的文献广泛采纳, 损失函数公式是:

$$L=\frac{1}{M}\sum_{i=1}^{C}y_i\log s_i+(1-y_i)\log(1-s_i) \tag{5}$$

其中, s_i 表示的是片段的输出分数, M 表示设置的生成片段的个数, y_i 由以下公式计算得出:

$$y_i=\begin{cases}0, & o_i\leqslant t_{\min}\\ \dfrac{o_i-t_{\min}}{t_{\max}-t_{\min}}, & t_{\min}<o_i<t_{\max}\\ 1, & o_i>t_{\max}\end{cases} \tag{6}$$

其中, o_i 表示当前候选片段与真实标签片段之间的 IoU 值, 结合两个阈值 $t_{\min}$ 和 $t_{\max}$ 得到缩放的 IoU 值 y_i. 该损失可以将与真实标签之间的 IoU 值作为监督信息, 来不断优化目标片段的输出分数. 基于残差网络的语义调整(semantic modulation based residual network, SMRN)模型[41]沿用了 2D 时序网络的思路来生成候选片段, 并在此基础上引入了残差网络, 通过构造残差块融合了视频片段的上下文信息, 使得不同残差块能够识别跨越不同尺度的活动片段. 多尺度 2D 时序邻接网络(multi-scale temporal adjacent network, MS-2D-TAN)[42]亦沿用了 2D 时序邻接网络的思路, 使用门控卷积网络的方式构造了多尺度的时序网路, 使其能够挖掘不同尺度候选片段的特征.

第 2 类方法则在对查询语句进行编码之后, 生成候选片段. 在生成候选视频片段时, 该类方法大多借助注意力机制来探索视觉特征和文本特征之间的关系. QSPN 模型使用注意力机制将文本特征与经过 C3D 网络编码的视频片段特征相结合, 从而得到与查询语句最相关的片段, 并为这些片段生成更细致的视觉特征表示, 又使用两层 LSTM[43]网络将视觉特征表示与查询语句嵌入相结合, 从而得到模态融合表示, 最后利用视频描述任务作为辅助任务, 构建重构损失函数用于网络的训练. 语义活动提议(semantic activity proposal, SAP)模型[44]先生成视频中取样帧对应的视觉语义向量, 然后利用向量中的相关性信息, 组合相关性高的帧为候选片段, 再进行片段评价并调整时序边界. 时刻对齐网络(moment alignment network, MAN)[45]将文本信息作为动态卷积滤波器对视觉信息进行处理, 筛选出与查询语句相关的候选视频片段, 并结合图卷积网络[46]的思想得到每个片段与查询语句的相关性分数. SCN 模型同样利用查询语句提供的指导进行候选片段的生成, 将候选片段作为输入放入其后的语义完成模块, 构造重构损失函数. 视频-语言对齐网络(video-language alignment network, VLANet)[47]首先预设了候选视频片段, 然后在对视频和语句进行编码之后, 利用相似度信息筛选出候选片段, 同时在损失函数中引入对比学习的概念, 使得训练更加充分. 正则化的双流提议网络(regularized two-branch proposal network, RTBPN)[48]引入多实例学习的思想, 使用文本特征作为过滤器, 将视觉特征分别放入加强流和压缩流, 其中, 加强流用于生成正候选片段, 而压缩流用于生成负候选片段, 利用正负候选片段的差距进行损失函数的构建. 语义条件下的动态调整(semantic conditioned dynamic modulation, SCDM)机制[49]是有指导生成候选片段方法的典型方法之一, 其利用层级化的时序卷积网络生成多尺度特征, 自然地得到了与查询语句相关的不同持续时间的视频片段特征, 充分利用了查询语句中蕴含的指导信息.

第 3 类方法则是在对两个模态的编码及跨模态融合之后生成候选片段, 其中, 时序定位网络(temporal ground net, TGN)模型[50]就使用了这种思路. 具体实现方法是: 在跨模态融合后, 模型会产生一个 $T\times K$ 的矩阵, 指示了 $T\times K$ 个视频片段与查询语句的相关性分数. 视频片段的生成方式是: 以 T 个时间点分别作为结束时间点, 每个时间点向前以σ为步长推移 K 个等差数列, 得到了 K 个视频片段. 这种方式使得前期过程中不会出现对视觉信息的冗余计算. 上下文边界预测(contextual boundary-aware prediction, CBP)模型[51]应用了类似的思

路, 但在最后生成候选片段的时候, 创造性地将视频划分为等长的短候选片段, 再使用时序定位的方法决定需要组合的候选片段的个数. 这两种方式均能产生可变长度的候选片段.

在有指导地生成候选片段的模型中, 也出现了弱监督的设定. 但是由于弱监督的方法在训练阶段只是用视频和视频片段对应的查询语句, 而不给出查询语句对应片段的时序边界, 因此使用基于排序的方法就会面临难以利用片段查询对构建损失函数的问题. SCN 模型应对的方式是: 先遮盖查询语句的重要词汇, 再利用文本解码器重新生成重要词汇, 通过将生成的词汇和原始词汇对比来构造重构损失函数, 对候选片段生成网络进行训练; 测试阶段则直接利用候选片段生成网络, 生成最匹配的片段. RTBPN 模型则是通过构造加强流和压缩流, 自主构造正负对, 用于网络的训练.

语义匹配强化学习(semantic matching reinforcement learning, SM-RL)模型[52]是视频片段检索领域使用强化学习进行探索的早期论文, 但其思路不同于其他对视频的帧进行评分的生成候选片段的论文, 首先使用强化学习的思路, 界定当前帧是否与查询语句相关, 并计算相关概率; 然后定位下一个要观察的帧; 最后将相关性持续大于阈值的连续帧组合为片段, 便是要查询的目标视频片段.

2 基于定位的方法

不同于第 1 节介绍的基于排序的一类方法, 基于定位的这类方法不以候选视频片段为处理单位, 而是以整个视频为处理单位, 直接以片段时间点作为预测目标. 因此, 相对于有指导地生成候选片段的排序方法而言, 基于定位的方法能够较为明显地减少计算成本, 几乎完全消除对于视频的重复处理和计算. 考虑到定位的方式又根据是否经过迭代, 本节将要介绍的基于定位的方法可以进一步划分, 分别是一次定位的方法和迭代定位的方法. 前者直接输出目标预测节点, 后者则会在生成预测节点后对节点进行迭代地调整. 两者在思路上的区别展示在图 5 中.

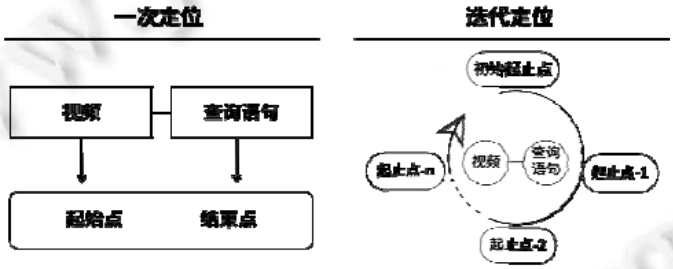

图 5 一次定位的方法与迭代定位的方法的区别

2.1 一次定位的方法

一次定位, 即在经历前期的处理之后, 在模型的最后部分直接产生定位到的起止节点. 这类方法包括基于概率的方法和基于回归的方法: 基于概率的方法分别计算每个时间点可能是目标起止时间点的概率, 将概率最大的时间点分别确定为起止时间点, 并保证结束时间点在开始时间点的后面; 基于回归的方法则通过回归网络直接定位到开始点和结束点的位置.

在基于概率的方法中, 根据模型关注的目标, 可以分为两类模型: 第 1 类模型通过直接计算每个点成为起止时间点的概率值, 分别确定下起止时间点的位置; 第 2 类模型仍然是以起止时间点为定位对象, 但并非直接计算概率值, 而是构造了起止时间点的概率分布函数.

在第 1 类模型中, 定位网络(localizing network, L-NET)[53]是典型代表之一, 其在保证结束时间点在起始时间点之后的情况下, 直接挑选概率最大的时间点作为预测结果. 并且, Chen 等人[53]深入探索了两个模态间的交互问题, 令两种模态分别去强调对方模态中的相关信息, 弱化无关信息, 同时考虑了上下文时刻对每个时刻的影响.

目前, 大多数模型仅仅对数据集的偏差进行了建模, 而忽略了对视频内容的学习. 基于这一动机, Yang 等

人[54]提出了引入因果介入的视频片段检索模型. 他们将因果关系理论引入到视频片段检索任务中, 通过分析视频的时序位置与视频、查询语句以及预测结果之间的因果联系, 在模型的拟合过程中降低对数据分布特征的关注, 为视频片段检索任务打开了一条新的研究道路.

相对于直接计算每个点可能是目标起止节点的概率值, 第 2 类模型则同时构造真实节点时序位置和预测节点时序位置的概率分布函数, 利用两个分布的对比, 获得了更大的灵活性和容错性. 跨模态时刻定位(cross-modal moment localization, XML)模型[55]在大规模视频片段检索任务上, 尝试使用基于概率分布函数的定位方法, 性能得到了有效提高. 这是因为该方法不仅对视频内容进行编码, 还同时引入了影视作品中的字幕, 充分利用了视频提供的视觉、文本信息, 得到了更加有用的定位依据. 在 XML 模型提供的思路的基础上, 多语言的跨模态时刻定位模型(multilingual cross-modal moment localization, MXML)模型[56]扩展了中文标注, 通过参数共享和邻域限制的方式, 促进中英文两种语言的共同学习. 使用对比学习的检索和定位网络(retrieval and localization network with contrastive learning, ReLoCLNet)[57]分别在视频级别和帧级别通过对比学习增强视频中多模态信息和查询语句文本信息的特征表示能力. 对抗双向交互网络(adversarial bi-directional interaction network, ABIN)[58]引入对抗学习的思想, 利用生成器生成预测到的节点概率分布函数, 同时引入辅助的对抗学习任务, 将预测的概率分布函数与真实标签对应的分布函数对比, 从而提高模型性能. Rodriguez-Opazo 等人[59]充分利用了查询语句给出的参考信息, 将文本表示作为动态滤波器对视觉特征进行增强, 并利用得到的注意力数值构建注意力损失函数, 同时利用预测时间点和真实标签时间点的分布而非数值构建损失函数. 视频跨度定位网络(video span localizing network, VSLNet)[60]则将视觉和文本编码模块、跨模态融合模块和定位模块组合之后作为基础框架, 方便实现扩展和迁移.

还有一些方法可以看作是对密集回归的改进, 属于基于回归的方法. 如密集回归网络(dense regression network, DRN)[61]通过计算视频所有时序节点到起止节点的距离, 并构造语义匹配头和 IoU 回归头, 从两个维度对每个节点的预测结果进行评价. Zeng 等人[61]认为, 这种方法可以更好地利用数据集给出的真实标签, 虽然数据集中每一个文本查询与视频片段对只能提供一个起止时间点标签, 但是处于起止时间点之中的所有时间节点都可以作为中间节点用来训练. 密集的自底向上定位(dense bottom-up grounding, DEBUG)模型[62]通过对中间节点进行定位, 同时构造了两条处理跨模态间特征的支线, 既可以判定该节点是否属于真实标签提供的时序边界, 又可以评价定位相关度的置信分数. 密集预测的图 FPN (graph-FPN with dense predictions, GDP)模型[63]在 DEBUG 模型的基础上, 通过在跨模态间特征融合和定位模块之间增加图卷积网络和编码模块, 可以生成融合上下文特征的不同尺度特征, 使得定位更加细致、准确. 专注的跨模态相关度匹配(attentive cross-modal relevance matching, ACRM)模型[64]使用相同的处理流, 分别对每个节点属于目标起、止、中间节点的概率大小进行预测.

直接使用回归网络定位起止时间点是比较直观的回归模型. 例如, 基于注意力的定位回归(attention based location regression, ABLR)模型[65]使用了共同注意力机制对两个模态的表示进行了细致的探索, 并在最后使用回归网络获得起止时间点的定位. Mun 等人[66]探索了文本特征的分层表示, 分别获得了单词级别、词组级别以及句子级别的编码, 将得到的特征与视觉信息进行更细致的交互, 最终利用多层感知机进行起止时间点的定位. 分层的视觉-文本图(hierarchical visual-textual graph, HVTG)模型[67]在模态间特征交互阶段依次构造了物体-语句子图、物体-物体子图以及语句引导的节点集成图, 生成了结构化的、有重点的视觉特征表示, 然后利用 ReLU 函数对视觉表示转化的分数进行回归, 从而定位起止节点. 双通道交互网络(dual path interaction network, DPIN)[68]将编码好的视觉和文本特征表示分别送入自底向上对视频进行定位的支线以及自顶向下对片段进行打分的支线, 同时设计一个交互模块进一步增强两条支线之间的关联. 发现物体关系(discovering object relationships, DORi)模型[69]首先使用空间子图探索关键帧中物体间的空间位置关系, 再将利用图模型获得的帧视觉表示放入时序子图中得到概率分布, 最后经过两层全连接层获得起止节点的预测结果. 提取片段的定位(extractive clip localization, ExCL)模型[70]同样使用了回归的思路, 通过将跨模态融合部分和节点预测部分结合起来, 并利用多层感知机、单层 LSTM 网络、两层 LSTM 网络这 3 种方式对起止时间点的预测结果

进行拟合. 使用回归网络的方式对起止时间点进行定位的优势是可以直接获得结果, 无需进行额外的操作, 缺点是缺乏可解释性. 使用双对比学习的介入性视频定位(interventional video grounding with dual contrastive learning, IVG-DCL)模型[71]通过引入因果关系和对比学习的思想, 可以解决数据集偏差的问题, 提取更加有效的特征.

大部分弱监督模型采用一次定位, 这是因为弱监督方法仅利用视频和视频片段对应的查询语句作为训练依据, 而不给出查询语句对应片段的时序边界. 因此, 弱监督模型无法像排序方法一样自然地利用候选片段标注与真实标签标注的偏差作为损失函数. 比如, 文本指导的注意力(text-guided attention, TGA)模型[72]首先将视频切分为时间片段, 然后以文本描述语句为指导, 找到相关的时间片段, 在训练阶段, 利用视频与查询语句的正对和负对构造损失函数, 测试阶段则直接定位到相关时序边界.

在一次定位的模型中, 基于概率的方法可以以概率为基础构建损失函数:

$$L=-\frac{1}{N}\sum_{i}^{N}\log(P(t_s^{g,i}))+\log(P(t_e^{g,i})) \tag{7}$$

其中, P 表示时序点对应的概率值. 该损失函数可以使真实标签对应的时序帧获得更大的概率值.

基于回归的方法则将预测时间点与真实标签时间点之间的差分函数作为损失函数:

$$L=-\frac{1}{N}\sum_{i}^{N}F(t_s^{p,i}-t_s^{g,i})+F(t_e^{p,i}-t_e^{g,i}) \tag{8}$$

其中, F 可以是 L_1 损失函数、L_2 损失函数或均方误差函数等.

2.2 迭代定位的方法

迭代定位的方法首先随机或粗略定位起止时间点, 再以当前片段与查询语句的相关性为指导, 对起止时间点进行调整, 不断向最优目标靠近. 这类方法通常使用强化学习的思路, 将视频片段检索任务看作一个序贯决策问题, 通过设定状态、动作、奖赏等迭代调整起止时间点的位置, 直至达到满意结果或迭代结束.

He 等人[73]提出了一种基于强化学习的方法, 通过将经过 Skip-thought[13]处理的文本表示以及经过 C3D 网络[12]处理的视觉表示看作环境, 将初始定位节点视为状态, 设定了 7 个动作(以δ的幅度向前移动开始/结束时间点、以δ的幅度向后移动开始/结束时间点、以δ的幅度向前/后同时移动开始和结束时间点以及停止), 在网络训练过程中, 将监督信息融入损失函数中. 对抗的视频片段检索(adversarial video moment retrieval, AVMR)模型[74]将强化学习模块作为对抗网络中的生成器, 来生成和调整目标片段的起止时间, 同时利用贝叶斯个性化排序作为对抗网络中的判别器, 对生成器的定位效果进行评价, 从而提供强化学习定位过程中的奖赏依据. 空间时序强化学习(spatio-temporal ReinfOrcementlearniNG, STRONG)模型[75]沿用了 AVMR 模型的思路, 不同的是, 既构造了处理时序维度的强化学习模块, 也设计了处理空间维度的强化学习模块, 可以对关键帧的活动主体进行定位, 从而提高时序定位的性能. 敏捷网络(tripping network, TripNet)[76]也使用了与 AVMR 模型类似的思路, 通过将动作空间由固定动作扩展为可指定移动帧数的动作(例如, 向前移动开始时间点调整为: 将开始时间点向前移动 h 帧、j 帧、1 s 帧数等), 使得模型的迭代过程更具灵活性. 作者通过实验验证了该模型只需浏览 32%–41%的视频即可完成定位, 这证明迭代定位的方式可以在保证检索效果的情况下大幅降低检索成本. 基于树结构策略的稳步强化学习(tree-structured policy based progressive reinforcement learning, TSP-PRL)模型[77]用树结构构造了分层的动作空间, 用于指示不同细致程度的动作策略, 如在根节点下设置了代表缩小、向左微调、向右大调等策略的中间节点, 并在这些节点下设置了叶子节点, 用于决定调整的幅度大小.

3 数据集与实验

3.1 数据集

可用于视频片段检索领域的数据集包括 Regneri 等人[7]提出的 TACoS、Hendricks 等人[9]提出的 DiDeMo、Gao 等人[10]在 Charades 数据集[78]的基础上提出的 Charades-STA、Krishna 等人[79]提出的 ActivityNetCaptions、

Hendricks 等人[80]在 DiDeMo 数据集的基础上进行调整得到的 TEMPO-TL 及 TEMPO-HL 两个数据集, 以及 Lei 等人[55]提出的 TVR 数据集. 这些数据集的基本信息见表 2.

表 2 数据集的基本信息

数据集名称	#视频数/#片段数	#句子数	视频来源	视频领域
TACoS	127/7 206	18 226	厨房	烹饪
DiDeMo	10 464/26 892	40 543	Flicker	开放
Charades-STA	6 672/11 772	16 124	活动	室内活动
ActivityNet captions	19 209/–	71 942	活动	开放
TVR	21 793/–	108 965	影视剧	开放

下面对几个数据集的内容进行简单介绍.

(1) TACoS 数据集[7]是在 MPII Composites[8]的基础上, 通过众包的方式采集的视频片段描述语句, 官方网站提供了视频数据和查询语句语料库. 但其视频内容限制在烹饪领域, 视觉语义变化较小, 限制了下游任务模型的性能. 在数据集的划分上, 本领域的大多数模型沿用了 Gao 等人[10]提出的思路, 将数据集以 2:1:1 的比例划分为训练集、验证集和测试集进行实验;

(2) DiDeMo 数据集[9]收集了 10 464 个 25–30 s 的视频, 在公开发布的数据集中, 提供了将每个视频切分为 5–6 个 5 s 长度的片段后, 每个片段经过 VGG 网络[81]处理得到的 4 096 维视觉特征, 同时给出了每个单词经过 GloVe[82]处理之后得到的 300 维词嵌入. 视频内容多来自日常生活, 场景丰富. 在数据集的划分上, 大部分文献沿用 Hendricks 等人[9]提出的思路, 将视频进行如下切分: 8 395 个用于训练, 1 065 个用于验证, 1 004 个用于测试;

(3) Charades-STA 数据集[10]是在 Charades 数据集的基础上进行构建的, Charades 数据集的官方网站提供了视频的特征、对视频的长描述以及视频片段的动作标记. Gao 等人[10]则将视频的长描述和动作标记扩展为片段的自然语言描述, 发布在公开的代码中, 但未提供自然语言描述的词嵌入或查询语句特征, 需自行进行文本特征的提取工作. 视频内容主要为室内活动. 在数据集的划分上, 大多数模型沿用了 Gao 等人提出的思路, 将数据集中的 13 898 个片段描述对作为训练集, 4 233 个片段描述对作为测试集;

(4) ActivityNet Captions 数据集[79]最早是被提出应用于视频描述任务中的, 因其中包括了时序片段注释以及相应的自然语言描述, 非常适合用于视频片段检索任务. 数据集的发布网站中提供了视频的 C3D[12]特征以及视频片段的自然语言描述, 文本特征仍需要自行提取. 该数据集提供了多达 19 209 个视频, 且场景多样化, 每个视频至少对两个片段进行了标注. 大多数模型将数据集以 2:1:1 的比例划分为训练集、验证集和测试集;

(5) TEMPO-TL 及 TEMPO-HL 数据集[80]是在 DiDeMo 数据集的基础上分别采取机械和人工手段, 利用时序词(如 before、after)将短句构造为复杂长句得到的. 长句标注能够更好地为模型提供查询语句时序信息;

(6) TVR 数据集[55]: Lei 等人从无到有, 构建了一个针对视频片段检索任务的数据集. 不同于其他数据集, 其构建过程不仅依据视频内容, 且利用影视作品中的字幕, 更加准确且细致地对片段进行定位, 达到了用更长的句子描述更短片段的效果. 数据集从 6 个影视剧中挑选出 21 793 个视频, 构造了 108 965 个查询语句及对应的视频片段时序位置. 数据集的官方网站提供了查询语句的自然语言描述、视频字幕及经过 ResNet[83]提取的视觉特征. Lei 等人还使用该数据集对大规模视频片段检索任务进行了探索, 获得了性能的突破. 其视频内容来源于美国情景剧, 场景丰富, 包含很多针对较短视频片段的查询语句描述.

3.2 评价度量

视频片段检索任务中一些常见的评价指标如下, 通常来说, 这些指标的数值越大, 代表方法的性能越好:

(1) $R(n,m)$, 也可被写为“$R@n,IoU=m$”, 表示在返回的前 n 个结果中, 交并比指标(IoU)大于 m ($\in(0,1]$)的结果(至少 1 个)占总体 n 个返回结果的比例. 例如: 在某次测试中, 共有 1–8 号共 8 个样本, 其对应返回的 IoU 值分别为 0.1–0.8, 若将 m 设置为 0.7, 则满足条件的为 7–8 共 2 个样本, 那么 $R(n,m)$即 $R(8,0.7)$的结果应为 2/8=0.25. 因为大部分基于定位的方法对于每个测试样例仅会返回一个结果, 所以只会出现 $R(1,m)$的情形;

(2) mIoU, 也称为平均 IoU, 表示的是所有测试样例的第 1 个返回结果对应 IoU 的平均值;

(3) Rank@n, 也写作 Rank-n accuracy, 表示的是最匹配结果出现在前 n 个结果中的百分比, 一般 n 取 1 或 5. 等同于 $R(n,1)$指标;

(4) Acc@0.5, 适用于定位任务, 表示的是模型产生的返回结果(一个查询语句仅返回一个结果)与真实标签间的 IoU 高于 0.5 的比例, 等同于 $R(1,0.5)$指标.

3.3 性能比较

本文对一些模型在 Charades-STA、TACoS、DiDeMo 以及 ActivityNet Captions 这 4 个数据集上的结果进行了汇总, 由于不同文章中对相同模型的测试可能得到不同的结果, 我们以原文提供的结果为准, 分别在表 3–表 6 中进行展示(其中, “(weak)”表示弱监督的方法, “–”表示原文未给出相应指标的实验结果, 加粗的数字表示在使用相同视觉特征提取网络和同一个评价指标的情况下最好的实验结果).

表 3 在 Charades-STA 数据集上的实验结果

模型简称	类别	*R*@1 *IoU*=0.7	*R*@1 *IoU*=0.5	*R*@1 *IoU*=0.3	*R*@5 *IoU*=0.7	*R*@5 *IoU*=0.5	*R*@5 *IoU*=0.3	视觉特征
SM-RL[52]	排序	11.17	24.36	–	32.08	61.25	–	VGG
SAP[44]	排序	13.36	27.42	–	38.15	66.37	–	VGG
2D-TAN[40]	排序	**23.25**	**39.81**	–	**52.15**	**79.33**	–	VGG
ROLE[22]	排序	–	12.12	25.26	–	40.59	70.13	C3D
TGA (weak)[72]	定位	8.84	19.94	32.14	33.51	65.52	86.58	C3D
CTRL[10]	排序	8.89	23.63	–	29.52	58.92	–	C3D
EFRC[36]	排序	15.00	33.80	53.00	43.90	77.30	94.60	C3D
QSPN[38]	排序	15.80	35.60	54.70	**45.40**	**79.40**	**95.60**	C3D
R-W-M[73]	定位	–	36.70	–	–	–	–	C3D
DEBUG[62]	定位	**17.69**	**37.39**	**54.95**	–	–	–	C3D
MAN[45]	排序	22.72	46.53	–	53.72	86.23	–	TAN
ExCL[70]	定位	22.40	44.10	61.50	–	–	–	I3D
SCDM[49]	排序	**33.43**	**54.44**	–	58.08	74.43	–	I3D
DRN[61]	定位	31.75	53.09	–	**60.05**	**89.06**	–	I3D
STRONG[75]	定位	19.30	50.14	**78.10**	–	–	–	ResNet+ConvLSTM
AVMR[74]	定位	–	**54.59**	77.72	–	72.78	88.92	ResNet+ConvLSTM

表 4 在 TACoS 数据集上的实验结果

模型简称	类别	*R*@1 *IoU*=0.5	*R*@1 *IoU*=0.3	*R*@1 *IoU*=0.1	*R*@5 *IoU*=0.5	*R*@5 *IoU*=0.3	*R*@5 *IoU*=0.1	视觉特征
SM-RL[52]	排序	15.95	20.25	26.51	27.84	38.47	50.01	VGG
SAP[44]	排序	**18.24**	–	**31.15**	**28.11**	–	**53.51**	VGG
ABLR[65]	定位	9.30	18.90	31.60	–	–	–	Bi-LSTM
CTRL[10]	排序	13.30	18.32	24.32	25.42	36.69	48.73	C3D
ACRN[14]	排序	14.62	19.52	24.22	24.88	34.97	47.42	C3D
CMIN[17]	排序	18.05	24.64	32.48	27.02	38.46	62.13	C3D
TGN[50]	排序	18.90	21.77	41.87	31.02	39.06	53.40	C3D
SCDM[49]	排序	21.17	26.11	–	32.18	40.16	–	C3D
2D-TAN[40]	排序	**25.32**	**37.29**	**47.59**	**45.04**	**57.81**	**70.31**	C3D
ExCL[70]	定位	28.00	45.50	–	–	–	–	I3D
AVMR[74]	定位	49.13	**72.16**	89.77	64.40	83.37	94.26	ResNet+ConvLSTM
STRONG[75]	定位	**49.73**	72.14	**90.85**	–	–	–	ResNet+ConvLSTM

表 5 在 DiDeMo 数据集上的实验结果

模型简称	类别	Rank@1	Rank@5	mIoU	视觉特征
TGA (weak)[72]	定位	12.19	39.74	24.92	C3D
EFRC[36]	排序	**13.23**	**46.98**	**27.57**	C3D
MAN[45]	排序	27.02	81.70	41.16	TAN
WSLLN(weak)[33]	排序	18.40	54.40	27.40	VGG
TMN[34]	排序	22.92	76.08	35.17	VGG
MCN[9]	排序	28.10	78.21	41.08	VGG
TGN[50]	排序	28.23	79.26	42.97	VGG
TCMN[19]	排序	28.90	79.00	41.03	VGG
SM-RL[52]	排序	**31.06**	**80.45**	**43.94**	VGG

表 6 在 ActivityNet Captions 数据集上的实验结果

模型简称	类别	*R*@1 *IoU*=0.7	*R*@1 *IoU*=0.5	*R*@1 *IoU*=0.3	*R*@5 *IoU*=0.7	*R*@5 *IoU*=0.5	*R*@5 *IoU*=0.3	视觉特征
ABLR[65]	定位	–	36.79	55.67	–	–	–	Bi-LSTM
ExCL[70]	定位	24.10	42.70	62.30	–	–	–	I3D
QSPN[38]	排序	13.60	27.70	45.30	38.30	59.20	75.70	C3D
SCDM[49]	排序	19.86	36.75	54.80	41.53	64.99	77.29	C3D
R-W-M[73]	定位	–	36.90	–	–	–	–	C3D
WSLLN (weak)[33]	排序	22.70	42.80	**75.40**	–	–	–	C3D
CMIN[17]	排序	23.88	43.40	63.61	50.73	67.95	80.54	C3D
2D-TAN[40]	排序	**27.38**	44.05	58.75	**62.26**	76.65	**85.65**	C3D
DRN[61]	定位	24.36	**45.45**	–	50.30	**77.97**	–	C3D

在模型性能的对比中, 有如下几点因素会对模型的结果造成影响.

(1) 对模型性能影响较大的是提取视觉特征使用的网络, 视频片段检索领域较常使用的网络有 VGG-16、C3D、I3D[84]这 3 种, 根据 Gao 等人[10]给出的实验数据, 我们得知: C3D 相对于 VGG-16 可取得更好的效果; 同时, 根据 Zeng 等人[61]给出的实验数据, I3D 相对于 C3D 也能取得更好的实验效果; 另外, AVMR 模型[74]和 STRONG 模型[75]均使用了 ResNet+ConvLSTM[85]的方式对视觉特征进行提取, 也获得了很好的效果;

(2) 第 2 个影响因素是文本特征的提取方式, 较为常见的有两种方式: 第 1 种是首先用例如 GloVe 或 Skip-gram 的 Word2Vec[86]方法获得词嵌入向量, 然后使用 LSTM[43]网络将词嵌入向量转化为以句子为单位的文本特征; 第 2 种是直接使用 Skip-thought[13]获得句子的文本特征. 根据 Gao 等人[10]给出的实验结果可以发现, 使用第 2 种方式会比第 1 种方式获得更优的效果;

(3) 一些研究表明[73]: 使用视频的 RGB 流和光流作为共同的输入, 也能够小幅提升性能;

(4) 对基于排序的方法中使用预设滑动窗口方式的模型来说, 使用更加细致且更多尺度的滑动窗口也有助于性能的提升[17].

在这 4 个因素中, 文本特征提取方式对于实验结果的影响较小, 最后两个因素又未得到广泛应用, 因此为了公平且有效地对模型进行比较, 我们只将视觉特征的提取方式作为比较模型性能的考虑因素.

可以发现: 模型在 Charades-STA 上取得的效果普遍更好, 其次是 ActivityNet Captions, 在 TACoS 上取得的效果最差. 相较于 TACoS 来说, Charades-STA 和 ActivityNet Captions 这两个数据集提供的视频数量多、视频持续时间长、场景更加复杂多样, 而 TACoS 的视频都是相同场景下不同主角进行的不同烹饪活动, 这导致不同视频间仅有细微的差异, 从而增加了模型学习的难度.

在对数据集上的模型实验结果进行分析后, 有如下的发现.

(1) 无论在哪种类别的方法中, 注意力机制都是简单且有效的. 如在基于排序的方法中, 相比于 CTRL 模型, ACRN 模型和 CMIN 模型分别在视觉模块使用注意力机制和多头注意力机制融合了上下文片段, 获得了性能的提升;

(2) 在基于排序的方法中, 有指导地生成候选片段的模型相对于预设候选片段的模型, 能够有效提升模

型性能. 例如仅依据视觉特征生成候选片段的 EFRC 模型以及在查询语句的指导下生成候选片段的 QSPN 模型, 它们在 Charades-STA 数据集的表现上相对于 CTRL 模型提升了 5%–10%. 其中, QSPN 模型之所以获得了更优的性能, 是因其不仅利用查询语句给出的指导, 而且利用视频描述任务重构损失函数, 获得了更多的监督信息;

(3) 在基于定位的方法中, DRN 模型获得最优的效果. 究其原因, 是其充分利用了监督信息, 为模型的拟合过程提供了足够的信息;

(4) 当前的视频片段检索领域存在一类弱监督方法, 这类方法不给出查询语句的匹配片段对应的时序边界, 而只应用视频整体及视频片段的查询语句描述. 弱监督方法取得的实验结果普遍低于使用全监督的方法, 如 TGA 模型采取了与 CTRL 模型基本类似的思路, 不同的是采取了弱监督的设定, 于是导致了相对较弱的实验结果;

(5) 强化学习的方法能够实现较好的性能. 如使用排序思路的 SM-RL 模型, 在 DiDeMo 数据集的使用 VGG 网络的方法中获得了最优的性能. 又比如在 TACoS 数据集上, 使用定位思路的 AVMR 模型和 STRONG 模型因为引入了强化学习以及不同的视觉特征提取模式, 所以获得了较大的性能提升;

(6) 在 Charades-STA 数据集的使用 VGG 网络的方法中、在 TACoS 数据集的使用 C3D 网络的方法中以及在 ActivityNet Captions 数据集的使用 C3D 网络的方法中, 2D-TAN 模型均取得了最优的效果. 其在没有查询语句指导的情况下, 使用了对视频片段均匀挑选来生成候选片段的方式, 获得了非常好的效果, 证明了这种方式的简单、有效.

3.4 检索结果的可视化

我们分别从预设候选片段的方法、有指导地生成候选片段的方法、一次定位的方法和迭代定位的方法挑选了 CTRL 模型、2D-TAN 模型、DRN 模型和 AVMR 模型这 4 个典型模型, 并随机挑选了一个示例, 来展示各个模型的检索结果, 如图 6 所示. 其中, 绿色箭头标注指示的是真实标签, 黑色箭头标注则是这 4 个模型的检索结果, IoU 表示当前示例下, 各个模型的检索结果与真实标签间的交并比. 个例不足以说明整个模型的优劣, 但可以在一定程度上反映不同方法在检索过程中的特点. 通过分析这些结果, 我们可以得到以下发现.

(1) AVMR 模型属于迭代定位的方法. AVMR 模型基于强化学习来进行迭代定位, 可以达到对个例的微观级别的修正, 因此在本示例中取得了较为优秀的结果;

(2) CTRL 模型属于预设候选片段的方法. 该方法通过人为设定的方式提供了大量的候选视频片段, 并通过多中选优的策略给出最终的检索结果, 整体思路简单、有效, 在本示例中也取得了较好的结果;

(3) 2D-TAN 模型是一种有指导地生成候选片段的方法. 其在筛选候选片段时采用的是均匀取样策略, 这一策略可能会导致更优的候选视频片段被遗漏. 尽管该方法在本示例中的效果不理想, 但不可否认, 2D-TAN 模型具有很好的效率和性能;

(4) DRN 模型是一种一次定位的方法, 这类方法对模型的建模能力提出了较高的要求. DRN 模型可以充分地对数据集进行标注, 从匹配分数和 IoU 值两个维度对回归结果进行评价, 从而达到同时兼顾相关性和精确的目的.

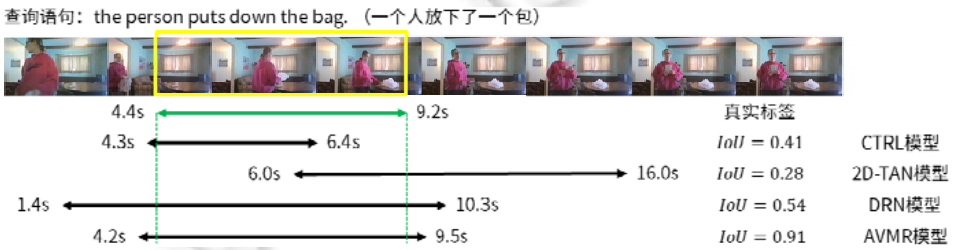

图 6 检索结果示例

4 模型总结

在这一节, 我们将对前述模型进行简要总结, 并提出当前方法可能的改进方向及发展趋势.

表 7 将前述模型的所属类别及提出时间进行了汇总(模型简称首先以原文中提出的为准; 若原文未提出, 则选用其他文章对其模型的简称, 那么可能出现多个). 表 8 对各类方法的主要思想及特点进行了对比总结.

表 7 模型汇总

类别	子分类	子分类	模型简称				
			2017	2018	2019	2020	2021
基于排序	预设候选片段	基于CTRL模型	CTRL[10]	ROLE[22], ACRN[14], VAL[26]	SLTA[15], CMIN[17], TCMN[19], LoGAN[25], ACL[29]	CSMGAN[18], CMA[20] GIIM[21], FIAN[23], LGN[30] CTF[31], WSLLN[33]	U-VMR[35], MMRG[32], FVMR[28]
		基于MCN模型	MCN[9]	TMN[34]	–	–	–
	有指导地生成候选片段	–	–	TGN[50]	SAP[44], MAN[45], SCDM[49] SM-RL[52]	SCN[39], 2D-TAN[40], SMRN[41], MS-2D-TAN[42], EFRC[36], QSPN/T-to-C/MLVI[38], VLANet[47], RTBPN[48], CBP[51]	–
基于定位	一次定位	基于概率	–	–	L-NET[53]	XML[55], ABIN[58], PfTML-GA/TMLGA[59], VSLNet[60], DRN[61]	DCM[54], MXML[56], ReLoCLNet[57]
		基于回归	–	–	ABLR[65], ExCL[70], TGA[72]	DEBUG[62], GDP[63], ACRM[64], LGI/LGVTI[66] HVTG[67], DPIN[68]DORi[69]	IVG-DCL[71]
	迭代定位	–	–	–	R-W-M/READ[73]	AVMR[74], STRONG[75]; TripNet[76], TSP-PRL[77]	–

表 8 各类方法的对比

类别	子分类	主要思想	处理单位	冗余计算	对模型要求
基于排序	预设候选片段	提取足够多候选片段, 多中选优	片段	过多	低
	有指导地生成候选片段	以查询语句或视频本身为指导, 确定候选片段	视频	较少	较低
基于定位	一次定位	直接产生定位得到的起止节点	视频	无	较高
	迭代定位	随机或粗略定位起止时间点, 迭代调整	视频	无	高

按照时间顺序梳理, 2017 年, CTRL 模型和 MCN 模型同时提出视频片段检索任务, 并给出了有效的解决方案, 让研究者认识到这个任务的价值和挑战. 2018 年, 研究者开始关注细粒度的视频内容和文本含义的理解, 尝试使用注意力机制解决问题[14,22]. 2019 年, 解决问题的思路开始多元化, 大家开始尝试跳出预设候选片段的思路, 使用有指导地生成候选片段的思路、一次定位的思路和迭代定位的思路. 2020 年, 研究者对这一任务的研究热情高涨, 出现了大量研究工作[40,61,74]. 2021 年, 受相关领域的影响, 研究者开始探索如何更好地理解视频和文本表达的意义[32,54].

结合表 7 和表 8, 我们可以得到以下发现.

(1) 从出现的时间上可以看到: 领域内先是提出了预设候选片段的思路, 然后给出有指导地生成候选片段的思路, 再到后来的基于定位的思路. 对模型效率的需求在不断提升, 为了达到这一目的, 越来越多的方法选择消除冗余计算. 这也会是未来领域内研究的一个重要方向;

(2) 基于排序的方法中, 预设候选片段方法的发展已经趋于成熟. 我们发现: 基于这种方法的创新主要集中于加入注意力机制(我们将图模型看作是建模一种更为复杂的注意力关系)、挖掘辅助概念表示和从模型优化的角度构造损失函数这 3 类, 其中又以加入注意力机制的创新居多. 因为预设候选片段的方法与其说是提供了一种思路, 不如说是提出了一个框架, 在一个固定的框架上, 可以做的创新终究是有限的;

(3) 相对于预设候选片段的方法, 有指导地生成候选片段的方法和基于定位的方法需要额外克服一个弊端, 就是在拟合过程中, 模型会慢慢偏向于学习数据集偏差而非视频内容. 这一现象由 Yang 等

人[54]发现并进行了一定的探讨，为今后的研究者提供了一个新的思考方向. 而大部分现有方法忽视了数据本身偏差的影响，进而导致模型更多地学到数据集中时序动作位置的偏差，从而影响了模型的训练和预测. 预设候选片段的方法由于会均匀地选取候选片段，因此更小概率会被数据分布干扰;

(4) 近期的文章中，不论是对任务中因果关系的讨论，还是使用无监督学习进行视频片段检索任务，研究重点开始更多关注研究对视频内容和查询语句真正的语义理解. 之前的模型大多也是从将本任务看作是认知任务的角度出发，但模型的拟合过多依赖于数据集本身，且模型的泛化性和鲁棒性较差，因此更像是将视频片段检索任务看作是一个感知任务来完成的. 但该任务的本意则应该是多模态语义理解的认知任务. 这些文章的出现，也为研究者们提供了新的研究思路和方向.

5 探索与展望

虽然视频片段检索任务是近年来刚被提出的研究领域，但该问题受到了较为广泛的关注，有着较好的发展. 本节将对该领域潜在的发展进行一定的预测.

5.1 端到端的模型架构

视频片段检索属于视频理解领域的边界敏感任务，其他类似的边界敏感任务还包括动作时序定位、逐步定位等. Xu 等人[87]认为: 这类任务有一个共同的问题，即视频视觉特征的提取与后续的定位过程是割裂开来的，视觉特征的表示对于定位结果有很大影响，但定位结果却无法作为反馈优化视觉特征的提取过程. 于是，他们将边界信息融入视觉特征的提取过程，使得提取的特征更加适用于视频片段检索这类边界敏感任务. Lei 等人[88]也针对视频片段检索任务尝试使用端到端的设计思路，通过设置有效的预训练和微调规则，使端到端训练变得可行，并达到了很好的效果.

5.2 大规模视频片段检索任务

传统的视频片段检索任务是在单一视频中定位最匹配查询语句的片段，考虑到一些真实应用需要针对一条查询语句从多个视频中进行查找，因此出现了一种新的视频片段检索范式. 该范式被称为大规模视频片段检索任务，其在大规模视频集合中查找与查询语句最相关的视频片段. 近期已有一些研究工作对大规模视频片段检索任务进行了探索.

Escorcia 等人[11]率先提出了解决该任务的方法，其设计的模型沿用了 MCN 模型[9]中的相似性比较思路，将多个视频划分为等长的视频片段并作为输入，此外还引入了重排序检索模块，使得检索结果能够更加细致地匹配查询语句. 分层的时刻对齐网络(hierarchical moment alignment network, HMAN)[89]也使用了这种思路，并在视觉特征的提取阶段使用多层卷积网络，从而产生对应不同长度片段的视觉特征. 这一思路无论在单视频片段检索还是大规模视频片段检索中均能取得很好的效果. 分层的多模态编码器(hierarchical multi-modal encoder, HAMMER)模型[90]将大规模视频检索任务拆分成两个子任务: 首先从大规模视频中检索与查询语句相关的视频，这一步称为视频检索; 然后对得分高的视频进行细致的视频片段检索，这一步称为片段定位. 并且构建了分层的跨模态编码器对视觉信息进行帧级别、片段级别和视频级别这 3 个不同粒度的编码. Lei 等人[55]构建了一个由影视剧组成的数据集 TVR，并设计了一种名为 XML 的模型，该模型将视觉特征和视频字幕作为共同的输入，获得了很好的大规模视频片段检索的效果. XML 模型提供了一种新颖的解决大规模视频片段检索任务的思路，即视频的多模态信息和查询语句分别进行特征生成，通过生成的特征间的交互，共同完成视频检索和片段定位任务. 鉴于该思路的有效性，一些后续研究也沿用了 XML 的思路，如 MXML 模型[56]、ReLoCLNet 模型[57]等.

5.3 认知模型

当前，大部分的视频片段检索模型往往过分依赖于训练数据，通过对数据集特性的学习达到良好的检索效果，而缺少对视觉和文本模态真正的语义认知. Yang 等人[54]的实验结果表明: 尽管当前现有的模型可以获

得有效的检索结果, 但是取得好结果的原因有一部分是由于模型对数据集偏差的拟合, 而非对视频内容的真正理解. Yang 等人通过将因果关系引入视频片段检索任务, 深入挖掘任务中各元素的因果联系, 为模型赋予了更好的鲁棒性和泛化性, 最终让模型具备了举一反三的潜力. 针对这个问题, Yuan 等人[91]也进行了研究, 并对数据集和模型评价指标进行了改进. 具体来说, 针对数据集存在的标记偏置, 其重新对 Charades-STA 数据集和 ActivityNet Captions 数据集进行了切分; 并在评价指标中加入了折扣参数, 作为对某些特定场景下虚高的 IoU 值的惩罚.

6 结束语

对于视觉信息和文本信息的语义理解问题, 是近些年来科研工作者研究的热点. 随着相关技术的不断发展, 视频片段检索领域的研究不断深入, 并展现出了良好的发展和应用前景. 本文对目前视频片段检索的研究进展进行了详细阐述, 同时分析了该领域目前面临的挑战, 最后对该领域进行了展望.

References:

[1] Wang S, Wang WY, Chen SZ, Jin Q. Video memorability prediction based on global and local information. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 1969−1979 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5935.htm [doi: 10.13328/j.cnki.jos.005935]

[2] Yu Q, Gao Y, Huo J, Zhuang YK. Discriminative joint multi-manifold analysis for video-based face recognition. Ruan Jian Xue Bao/Journal of Software, 2015, 26(11): 2897−2911 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4894.htm [doi: 10.13328/j.cnki.jos.004894]

[3] Liu T, Wang SL, Zhan NJ. Safety verification of trajectory planning for multiple robots. Ruan Jian Xue Bao/Journal of Software, 2017, 28(5): 1118−1127 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5207.htm [doi: 10.13328/j.cnki.jos.005207]

[4] Zhu XL, Wang HC, You HM, Zhang WH, Zhang YY, Liu S, Chen JJ, Wang Z, Li KQ. Survey on testing of intelligent systems in autonomous vehicles. Ruan Jian Xue Bao/Journal of Software, 2021, 32(7): 2056−2077 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6266.htm [doi: 10.13328/j.cnki.jos.006266]

[5] Zhang GM, Li QB, Zhang P, Cheng SJ. Defending code reuse attacks based on running characteristics monitoring. Ruan Jian Xue Bao/Journal of Software, 2019, 30(11): 3518−3534 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5539.htm [doi: 10.13328/j.cnki.jos.005539]

[6] Tellex S, Kollar T, Shaw G, Roy N, Roy D. Grounding spatial language for video search. In: Proc. of the 12th Int'l Conf. on Multimodal Interfaces; the 7th Int'l Workshop on Machine Learning for Multimodal Interaction. New York: Association for Computing Machiery, 2010. 31:1−31:8.

[7] Regneri M, Rohrbach M, Wetzel D, Thater S, Schiele B, Pinkal M. Grounding action descriptions in videos. Trans. of the Association for Computational Linguistics, 2013, 1: 25−36.

[8] Rohrbach M, Regneri M, Andriluka M, Amin S, Pinkal M, Schiele B. Script data for attribute-based recognition of composite activities. In: Proc. of the12th European Conf. on Computer Vision. 2012. 144−157.

[9] 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. Piscataway: IEEE Computer Society, 2017. 5804−5813.

[10] Gao J, Sun C, Yang Z, Nevatia R. TALL: Temporal activity localization via language query. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Piscataway: IEEE Computer Society, 2017. 5277−5285.

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

[12] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Piscataway: IEEE Computer Society, 2015. 4489−4497.

[13] Kiros R, Zhu Y, Salakhutdinov R, Zemel RS, Torralba A, Urtasun R, Fidler S. Skip-thought vectors. In: Proc. of the 2015 Advances in Neural Information Processing Systems. Cambridge: MIT, 2015. 3294−3302.

[14] Liu M, Wang X, Nie L, He X, Chen B, Chua TS. Attentive moment retrieval in videos. In: Proc. of the 41st Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. New York: Association for Computing Machiery, 2018. 15–24.

[15] Jiang B, Huang X, Yang C, Yuan J. Cross-modal video moment retrieval with spatial and language-temporal attention. In: Proc. of the 2019 Int'l Conf. on Multimedia Retrieval. New York: Association for Computing Machiery, 2019. 217–225.

[16] Ren S, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149.

[17] Zhang Z, Lin Z, Zhao Z, Xiao Z. Cross-modal interaction networks for query-based moment retrieval in videos. In: Proc. of the 42nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: Association for Computing Machiery, 2019. 655–664.

[18] Liu D, Dong J, Qu X, Zhou P, Liu XY, Xu Z. Jointly cross- and self-modal graph attention network for query-based moment localization. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machiery, 2020. 4070–4078.

[19] Zhang S, Su J, Luo J. Exploiting temporal relationships in video moment localization with natural language. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machiery, 2019. 1230–1238.

[20] Zhang B, Li Y, Yuan C, Xu D, Jiang P, Shan Y. A simple yet effective method for video temporal grounding with cross-modality attention. arXiv:2009.11232, 2020.

[21] Yu Z, Song Y, Yu J, Wang M, Huang Q. Intra- and inter-modal multilinear pooling with multitask learning for video grounding. Neural Processing Letters, 2020, 52(3): 1863–1879.

[22] Liu M, Tian Q, Wang X, Chen B, Nie L, Chua TS. Cross-modal moment localization in videos. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machiery, 2018. 843–851.

[23] Qu X, Tang P, Zou Z, Cheng Y, Dong J, Zhou P, Xu Z. Fine-grained iterative attention network for temporal language localization in videos. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machiery, 2020. 4280–4288.

[24] Yu Z, Yu J, Fan J, Tao D. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Piscataway: IEEE Computer Society, 2017. 1839–1848.

[25] Tan R, Xu H, Saenko K, Plummer BA. LoGAN: Latent graph co-attention network for weakly-supervised video moment retrieval. In: Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. 2021. 2083–2092.

[26] Song X, Han Y. VAL: Visual-attention action localizer. In: Proc. of the 19th Pacific-RIM Conf. on Multimedia. Berlin: Springer, 2018. 340–350.

[27] Chen J, Du H, Wu YF, Xu T, Chen EH. Cross-modal video moment retrieval based on visual-textual relationship alignment. Scientia Sinica Informationis, 2020, 50: 862–876 (in Chinese with English abstract).

[28] Gao J, Xu C. Fast video moment retrieval. In: Proc. of the 2021 IEEE Int'l Conf. on Computer Vision. Piscataway: IEEE Computer Society, 2021. 1523–1532.

[29] Ge R, Gao J, Chen K, Nevatia R. MAC: Mining activity concepts for language-based temporal localization. In Proc. of the 2019 IEEE Winter Conf. on Applications of Computer Vision. 2019. 245–253.

[30] Liu K, Yang X, Chua T seng, Ma H, Gan C. Language guided networks for cross-modal moment retrieval. arXiv:2006.10457, 2020.

[31] Chen Z, Ma L, Luo W, Tang P, Wong KYK. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. arXiv:2001.09308, 2020.

[32] Zeng Y, Cao D, Wei X, Liu M, Zhao Z, Qin Z. Multi-modal relational graph for cross-modal video moment retrieval. In: Proc. of the 2021 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2021. 2215–2224.

[33] Gao M, Davis LS, Socher R, Xiong C. WSLLN: Weakly supervised natural language localization networks. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Stroudsburg: ACL, 2020. 1481–1487.

[34] Liu B, Yeung S, Chou E, Huang DA, Fei-Fei L, Niebles JC. Temporal modular networks for retrieving complex compositional activities in videos. In: Proc. of the 15th European Conf. on Computer Vision. 2018. 569–586.

[35] Gao J, Xu CS. Learning video moment retrieval without a single annotated video. IEEE Trans. on Circuits and Systems for Video Technology, 2022, 32(3): 1646−1657.

[36] Xu H, He K, Sigal L, Sclaroff S, Saenko K. Text-to-clip video retrieval with early fusion and re-captioning. arXiv:1804.05113, 2018.

[37] Xu H, Das A, Saenko K. R-C3D: Region convolutional 3d network for temporal activity detection. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Piscataway: IEEE Computer Society, 2017. 5794−5803.

[38] Xu H, He K, Plummer BA, Sigal L, Sclaroff S, Saenko K. Multilevel language and vision integration for text-to-clip retrieval. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2019. 9062−9069.

[39] Lin Z, Zhao Z, Zhang Z, Wang Q, Liu H. Weakly-supervised video moment retrieval via semantic completion network. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2020. 11539−11546.

[40] Zhang S, Peng H, Fu J, Luo J. Learning 2D temporal adjacent networks for moment localization with natural language. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2020. 12870−12877.

[41] Chen C, Gu X. Semantic modulation based residual network for temporal language queries grounding in video. In: Proc. of the 17th Int'l Symp. on Neural Networks. 2020. 119−129.

[42] Zhang S, Peng H, Fu J, Lu Y, Luo J. Multi-scale 2D temporal adjacent networks for moment localization with natural language. arXiv:2012.02646, 2020.

[43] Hochreiter S, UrgenSchmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735−1780.

[44] Chen S, Jiang Y. Semantic proposal for activity localization in videos via sentence query. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2019. 8199−8206.

[45] Zhang D, Dai X, Wang X, Wang YF, Davis LS. MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2019. 1247−1257.

[46] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proc. of 5th Int'l Conf. on Learning Representations. 2017. 1−14.

[47] Ma M, Yoon S, Kim J, Lee Y, Kang S, Yoo CD. VLANet: Video-language alignment network for weakly-supervised video moment retrieval. In: Proc. of the 16th European Conf. on Computer Vision. 2020. 156−171.

[48] Zhang Z, Lin Z, Zhao Z, Zhu J, He X. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machiery, 2020. 4098−4106.

[49] Yuan Y, Ma L, Wang J, Liu W, Zhu W. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Proc. of the 2019 Advances in Neural Information Processing Systems. Cambridge: MIT, 2019. 534−544.

[50] Chen J, Chen X, Ma L, Jie Z, Chua TS. Temporally grounding natural sentence in video. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018. 162−171.

[51] Wang J, Ma L, Jiang W. Temporally grounding language queries in videos by contextual boundary-aware prediction. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2020. 12168−12175.

[52] Wang W, Huang Y, Wang L. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2019. 334−343.

[53] Chen J, Ma L, Chen X, Jie Z, Luo J. Localizing natural language in videos. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2019. 8175−8182.

[54] Yang X, Feng F, Ji W, Wang M, Chua TS. Deconfounded video moment retrieval with causal intervention. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. New York: Association for Computing Machiery, 2021. 1−10.

[55] Lei J, Yu L, Berg TL, Bansal M. TVR: A large-scale dataset for video-subtitle moment retrieval. In: Proc. of the 16th European Conf. on Computer Vision. 2020. 447−463.

[56] Lei J, Berg TL, Bansal M. MTVR: Multilingual moment retrieval in videos. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Stroudsburg: ACL, 2021. 726−734.

[57] Zhang H, Sun A, Jing W, Nan G, Zhen L, Zhou T, Goh R. Video corpus moment retrieval with contrastive learning. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: Association for Computing Machiery, 2021. 685−695.

[58] Zhang Z, Zhao Z, Zhang Z, Lin Z, Wang Q, Hong R. Temporal textual localization in video via adversarial bi-directional interaction networks. IEEE Trans. on Multimedia, 2021, 23: 3306−3317.

[59] Rodriguez-Opazo C, Marrese-Taylor E, Saleh FS, Li H, Gould S. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In: Proc. of the 2020 IEEE Winter Conf. on Applications of Computer Vision. 2020. 2453−2462.

[60] Zhang H, Sun A, Jing W, Zhou JT. Span-based localizing network for natural language video localization. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020. 6543−6554.

[61] Zeng R, Xu H, Huang W, Chen P, Tan M, Gan C. Dense regression network for video grounding. In: Proc. of the 2020 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2020. 10284−10293.

[62] Lu C, Chen L, Tan C, Li X, Xiao J. DebuG: A dense bottom-up grounding approach for natural language video localization. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Stroudsburg: ACL, 2020. 5143−5152.

[63] Chen L, Lu C, Tang S, Xiao J, Zhang D, Tan C, Li X. Rethinking the bottom-up framework for query-based video localization. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2020. 10551−10558.

[64] Tang H, Zhu J, Liu M, Gao Z, Cheng Z. Frame-wise cross-modal match for video moment retrieval. arXiv:2009.10434, 2009.

[65] Yuan Y, Mei T, Zhu W. To find where you talk: Temporal sentence localization in video with attention based location regression. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2019. 9159−9166.

[66] Mun J, Cho M, Han B. Local-global video-text interactions for temporal grounding. In: Proc. of the 2020 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2020. 10807−10816.

[67] Chen S, Jiang YG. Hierarchical visual-textual graph for temporal activity localization via language. In: Proc. of the 16th European Conf. on Computer Vision. 2020. 601−618.

[68] Wang H, Zha ZJ, Chen X, Xiong Z, Luo J. Dual path interaction network for video moment localization. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machiery, 2020. 4116−4124.

[69] Rodriguez C, Marrese-Taylor E, Fernando B, Li H, Gould S. DORi: Discovering object relationship for moment localization of a natural-language query in video. In: Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. 2021. 1078−1087.

[70] Ghosh S, Agarwal A, Parekh Z, Hauptmann A. 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. Stroudsburg: ACL, 2019. 1984−1990.

[71] Nan G, Qiao R, Xiao Y, Liu J, Leng S, Zhang H, Lu W. Interventional video grounding with dual contrastive learning. In: Proc. of the 2021 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2021. 2765−2775.

[72] Mithun NC, Paul S, Roy-Chowdhury AK. Weakly supervised video moment retrieval from text queries. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2019. 11592−11601.

[73] He D, Zhao X, Huang J, Li F, Liu X, Wen S. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2019. 8393−8400.

[74] Cao D, Zeng Y, Wei X, Nie L, Hong R, Qin Z. Adversarial video moment retrieval by jointly modeling ranking and localization. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machiery, 2020. 898−906.

[75] Cao D, Zeng Y, Liu M, He X, Wang M, Qin Z. STRONG: Spatio-temporal reinforcement learning for cross-modal video moment localization. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. New York: Association for Computing Machiery, 2020. 4162−4170.

[76] Hahn M, Kadav A, Rehg JM, Graf HP. Tripping through time: Efficient localization of activities in videos. In: Proc. of the 31st British Machine Vision Conf. 2020.

[77] Wu J, Li G, Liu S, Lin L. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Menlo Park: AAAI, 2020. 12386−12393.

[78] Sigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A. Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Proc. of the 14th European Conf. on Computer Vision. 2016. 510−526.

[79] Krishna R, Hata K, Ren F, Fei-Fei L, Niebles JC. Dense-captioning events in videos. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Piscataway: IEEE Computer Society, 2017. 706−715.

[80] Hendricks LA, Wang O, Shechtman E, Sivic J, Darrell T, Russell B. Localizing moments in video with temporal language. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018. 1380−1390.

[81] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of 3rd Int'l Conf. on Learning Representations. 2015. 1−14.

[82] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014. 1532−1543.

[83] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2016. 770−778.

[84] 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. Piscataway: IEEE Computer Society, 2017. 4724−4733.

[85] Joe Yue-Hei Ng, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: Deep networks for video classification. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2015. 4694−4702.

[86] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proc. of the 1st Int'l Conf. on Learning Representations. 2013.

[87] Xu M, Pérez-Rúa JM, Escorcia V, Martínez B, Zhu X, Zhang L, Ghanem B, Xiang T. Boundary-sensitive pre-training for temporal localization in videos. In: Proc. of the 2021 IEEE Int'l Conf. on Computer Vision. Piscataway: IEEE Computer Society, 2021. 7200−7210.

[88] Lei J, Li L, Zhou L, Gan Z, Berg TL, Bansal M, Liu JJ. Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proc. of the 2021 IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2021. 7331−7341.

[89] Paul S, Mithun NC, Roy-Chowdhury AK. Text-based localization of moments in a video corpus. arXiv:2008.08716, 2020.

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

[91] Yuan Y, Lan X, Wang X, Chen L, Wang Z, Zhu W. A closer look at temporal sentence grounding in videos: Dataset and metric. In: Proc. of the 2nd Int'l Workshop on Human-centric Multimedia Analysis, Virtual Event. New York: Association for Computing Machiery, 2021. 13−21.

附中文参考文献:

[1] 王帅, 王维莹, 陈师哲, 金琴. 基于全局和局部信息的视频记忆度预测. 软件学报, 2020, 31(7): 1969−1979. http://www.jos.org.cn/1000-9825/5935.htm [doi: 10.13328/j.cnki.jos.005935]

[2] 于谦, 高阳, 霍静, 庄韫恺. 视频人脸识别中判别性联合多流形分析. 软件学报, 2015, 26(11): 2897−2911. http://www.jos.org.cn/1000-9825/4894.htm [doi: 10.13328/j.cnki.jos.004894]

[3] 刘涛, 王淑灵, 詹乃军. 多机器人路径规划的安全性验证. 软件学报, 2017, 28(5): 1118−1127. http://www.jos.org.cn/1000-9825/5207.htm [doi: 10.13328/j.cnki.jos.005207]

[4] 朱向雷, 王海弛, 尤翰墨, 张蔚珩, 张颖异, 刘爽, 陈俊洁, 王赞, 李克秋. 自动驾驶关于智能系统研究测试的研究综述. 软件学报, 2021, 32(7): 2056−2077. http://www.jos.org.cn/1000-9825/6266.htm [doi: 10.13328/j.cnki.jos.006266]

[5] 张贵民, 李清宝, 张平, 程三军. 基于运行特征监控的代码复用攻击防御. 软件学报, 2019, 30(11): 3518–3534. http://www.jos.org.cn/1000-9825/5539.htm [doi: 10.13328/j.cnki.jos.005539]

[27] 陈卓, 杜昊, 吴雨菲, 徐童, 陈恩红, 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.

王妍(1997－), 女, 硕士生, 主要研究领域为视频理解, 多媒体分析, 信息检索.

詹雨薇(1997－), 女, 博士生, CCF 学生会员, 主要研究领域为计算机视觉, 机器学习, 信息检索.

罗昕(1992－), 男, 博士, 助理研究员, CCF 专业会员, 主要研究领域为机器学习, 多媒体内容分析与搜索.

刘萌(1991－), 女, 博士, 教授, CCF 专业会员, 主要研究领域为多媒体分析, 信息检索.

许信顺(1975－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为机器学习, 信息检索, 数据挖掘, 图像/视频分析与检索.