DOI:10.11918/202308059

智能体引导的视频重定位网络

郭阿欣,周 圆,霍树伟,李硕士

(天津大学 电气自动化与信息工程学院,天津 300072)

摘 要:视频重定位的目标是在未经剪辑的参考视频中定位与给定查询视频语义相关的片段。这项任务不仅满足用户的实际浏览需求,而且在多种应用场景中发挥着重要作用。由于视频相较于图像、文本等其他数据类型包含更丰富的信息,因此在长视频中准确识别目标片段并确定其时间边界具有较大挑战。将视频重定位任务视为一个序贯决策过程,应用强化学习实现高效且准确的定位。具体而言,提出智能体引导的定位网络(AGLN),通过训练智能体基于学习到的策略逐步执行动作,细化定位片段的时间边界,从而找到与查询视频最相关的片段。此外,AGLN 融合强化学习与监督学习,构建多任务学习框架,助力智能体更有效地探索环境并学习最优策略。在 ActivityNet-VRL 数据集上的实验结果表明,AGLN 在视频重定位任务上的表现优于现有方法,其检索平均准确率达到了 25.9%,相较于目前最佳方法提高了 0.2 个百分点。

关键词:视频重定位;强化学习;智能体;监督学习;多任务学习

中图分类号:TP391.4 **文献标志码**:A **文章编号**:0367-6234(2026)03-0120-09

Agent-guided video re-localization network

GUO Axin, ZHOU Yuan, HUO Shuwei, LI Shuoshi

(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: Video re-localization aims to localize a moment that semantically corresponds to a given query video from an untrimmed reference video. This task not only meets the actual browsing needs of users but also plays an important role in various application scenarios. Since videos contain richer information compared to other data forms like images and text, accurately identifying the target moment in a long video and determining its temporal boundaries are significantly challenging. This paper regarded the video re-localization task as a sequential decision-making process and applied reinforcement learning to achieve efficient and accurate localization. Specifically, this paper proposed an agent-guided localization network (AGLN), which trained an agent to progressively refine temporal boundaries of the localized moment based on the learned policy, thereby finding the most relevant moment to the query video. Additionally, AGLN combined reinforcement learning with supervised learning in a multi-task learning framework, aiding the agent in more effectively exploring the environment and learning the optimal policy. Experimental results on the ActivityNet-VRL dataset demonstrate that AGLN outperforms existing methods in the video re-localization task. The average retrieval accuracy of AGLN is 25.9%, which is 0.2 percentage points higher than the current optimal method.

Keywords: video re-localization; reinforcement learning; agent; supervised learning; multi-task learning

随着计算机技术和智能手机的飞速发展,视频数据量呈爆炸式增长。为帮助用户快速而准确地找到感兴趣的内容,视频检索[1-3]已经成为重要的研究领域,其目标是从大规模视频集合中检索出与查询内容相关的视频。然而,由于多数视频未经剪辑,包含大量与查询无关的内容,增加了用户的浏览负担。为此,研究者提出了如自然语言视频定位(natural language video localization, NLVL)[4-6]等任务,其以文本描述为查询,在长视频中定位相关片段。尽管文本查询简单直观,但其信息表达能力限制。因此,为提供更丰富的查询信息,视频重定位(video re-localization, VRL)[7]任务被提出,以优化检索过程。

VRL 任务旨在根据给定的查询视频,在参考视频中定位与其语义相关的目标片段。查询视频通常是用户感兴趣的短视频剪辑,而参考视频则是未经剪辑的长视频。该任务不仅有助于用户在长视频中迅速而准确地定位感兴趣内容,还在视频监控[8]、

收稿日期:2023-08-17;**录用日期**:2023-11-17;**网络首发日期**:2024-06-25
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20240624.1554.012
基金项目:国家重点研发计划(2020YFC1523204);国家自然科学基金(62171320,U2006211)
作者简介:郭阿欣(1999—),女,硕士研究生;周 圆(1983—),女,教授,博士生导师
通信作者:周 圆,zhouyuan@tju.edu.cn

人员重识别[9]等领域具有实际应用价值。视频重定位的主要挑战在于准确地识别目标片段并确定其时间边界。这源于视频视觉表现形式的多样化,即使表达相似语义信息的视频,也可能因环境、视角等因素在视觉上呈现显著差异。此外,视频由连续图像帧构成,相邻帧之间通常相似度较高,因此精确界定特定片段的时间边界具有一定难度。

现有的视频重定位方法多基于深度学习模型实现,主要分为两种定位策略。其一为基于候选片段的方法[10-11],通常利用滑动窗口技术或者候选生成网络从参考视频中提取候选片段,通过深度学习模型对候选片段和查询视频进行特征提取与匹配,从而选取最佳匹配的片段作为定位结果。但是,为确保定位的高准确性,这类方法常需大量候选片段进行匹配与排序,导致计算冗余、效率低下。其二为基于边界预测的方法[12-15],利用深度学习模型对查询视频和参考视频进行特征提取,通过特征融合和交互预测参考视频中每帧属于目标片段的概率,并使用阈值化和非极大值抑制等策略选取片段作为定位结果。然而,这类方法的预测结果通常存在偏差,可能导致边界定位不准确。

为应对上述挑战,并受强化学习在视频分析领域成功应用[16-20]的启发,提出智能体引导定位网络(agent-guided localization network, AGLN)。AGLN 通过智能体的序贯决策定位目标片段,无需处理大量候选片段,也避免了边界预测不准确带来的定位偏差,从而实现了更高效和准确的定位。AGLN 的核心在于使用一个可学习的智能体,该智能体可根据状态信息自适应地确定边界定位策略,并逐步细化时间边界,以精确锁定目标片段。此外,AGLN 采用多任务学习框架,通过基于策略优化的强化学习训练智能体以学习更优策略,并引入基于位置回归的监督学习以生成更具代表性的状态信息。这有助于智能体更有效地进行探索和学习,进而更好地完成视频重定位任务。

1　智能体引导的定位网络

AGLN 的整体结构如图 1 所示,其通过智能体迭代地细化定位片段的时间边界,实现准确的定位结果。

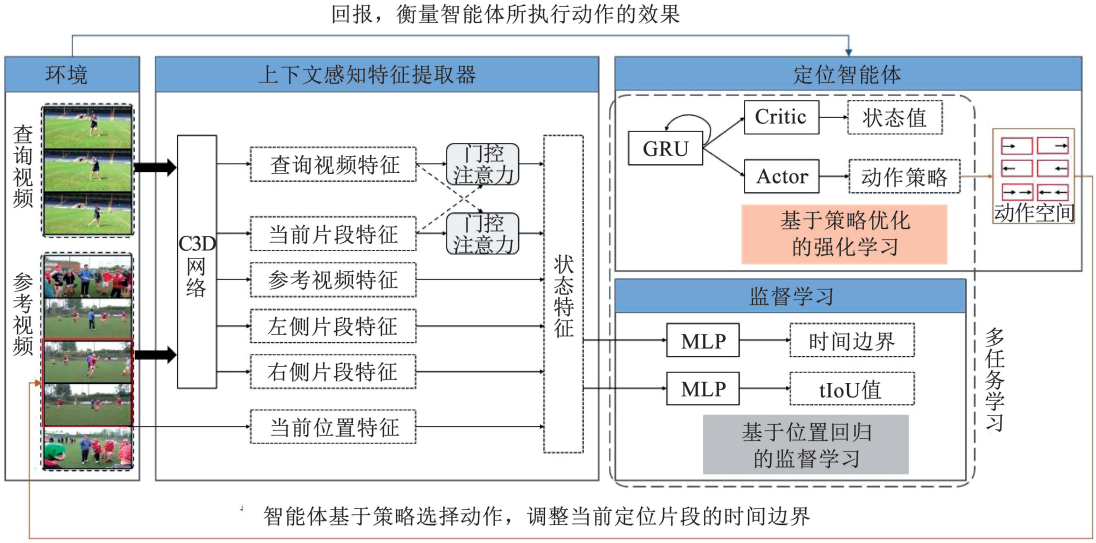

图 1　智能体引导的定位网络(AGLN)的整体结构

Fig. 1　Overall architecture of AGLN

具体而言,对于给定的查询视频和参考视频,首先在参考视频中确定一个初始定位片段,其时间边界表示为$[L_s^0, L_e^0]$,其中 L_s^0 和 L_e^0 分别表示片段的起始边界和终止边界。在每一轮迭代中,上下文感知特征提取器将包含查询视频、参考视频以及当前定位片段的环境信息编码作为状态特征。定位智能体根据状态特征生成动作策略,该策略定义了所有可能动作的概率分布,智能体基于这一策略选择相应动作以调整定位片段的时间边界。随后,环境根据动作结果更新状态,并反馈给智能体一个回报,反映所执行动作的效果。通过不断迭代上述过程,智能体逐步细化定位片段的时间边界,直至找到最佳定位结果或达到预设的最大迭代步数 $T_{\max}$。

1.1　上下文感知特征提取器

上下文感知特征提取器对视频进行特征提取和编码,充分利用视频的上下文信息生成反映当前环境状态的特征。此处环境由查询视频、参考视频以及当前定位片段组成。

首先使用预训练的 C3D 网络[21]对视频进行处理,以提取片段级别的特征。C3D 网络的输入是由连续的 16 帧视频组成的片段,且相邻片段之间没有重叠;接着通过主成分分析法(principal component

analysis,PCA)将 C3D 网络输出的 4 096 维特征投影至 500 维,以此作为视频的特征表示。对于查询视频和参考视频,可以得到相应的视频特征序列,表示为

$$F_{\mathrm{Q}} = \{f_{\mathrm{q}}^{i}\}_{i=1}^{n_{\mathrm{q}}}, F_{\mathrm{R}} = \{f_{\mathrm{r}}^{j}\}_{j=1}^{n_{\mathrm{r}}} \tag{1}$$

式中:f_{q}^{i} 和 f_{r}^{j} 分别表示查询视频和参考视频的片段级别特征;n_{q} 和 n_{r} 分别表示两个视频特征序列的长度。

在每个时间步,参考视频被划分为 3 部分:左侧片段、当前片段和右侧片段。从参考视频的特征序列中选取相应的部分,可得到 3 个片段的特征序列,分别表示为 F_{l}^{t-1}、F_{c}^{t-1} 和 F_{r}^{t-1}。为了获取固定长度的视频特征,对查询视频、参考视频、左侧片段、当前片段和右侧片段的特征序列进行平均池化,分别得到特征向量 f_{q}、f_{r}、f_{l}^{t-1}、f_{c}^{t-1} 和 f_{r}^{t-1}。这些特征都将作为状态特征编码的输入。其中,查询视频和参考视频特征反映了整个视频的全局信息,当前片段特征反映了定位片段的局部信息,而左侧片段和右侧片段特征则提供了视频的上下文信息。此外,为了明确表示当前定位片段的状态,将其归一化的时间边界 $l^{t-1}=[L_{\mathrm{s}}^{t-1}/N, L_{\mathrm{e}}^{t-1}/N]$(位置特征)也作为状态特征编码的一部分。其中 L_{s}^{t-1} 和 L_{e}^{t-1} 分别表示当前定位片段的起始边界和终止边界,而 N 为参考视频的长度。

将所有的视频特征以及位置特征经过全连接层处理后,进一步使用门控注意力机制[22]增强查询视频和当前定位片段之间的特征交互,表示为

$$\tilde{f}_{\mathrm{q}} = \sigma(f_{\mathrm{c}}^{t-1}) \odot f_{\mathrm{q}}, \tilde{f}_{\mathrm{c}}^{t-1} = \sigma(f_{\mathrm{q}}) \odot f_{\mathrm{c}}^{t-1} \tag{2}$$

式中,σ 表示 Sigmoid 激活函数,$\odot$ 表示哈达玛积。最后将所有特征进行拼接并通过全连接层 ϕ 处理,得到最终的状态特征,表示为

$$s^{t} = \phi(\tilde{f}_{\mathrm{q}}, \tilde{f}_{\mathrm{c}}^{t-1}, f_{\mathrm{r}}, f_{\mathrm{l}}^{t-1}, f_{\mathrm{r}}^{t-1}, l^{t-1}) \tag{3}$$

1.2 定位智能体

定位智能体根据状态特征生成动作策略,并基于策略从动作空间中选择动作来调整当前定位片段的时间边界。在本文的任务中,动作空间共定义了 6 种动作:1)起始边界向前移动 δ;2)终止边界向前移动 δ;3)起始边界向后移动 δ;4)终止边界向后移动 δ;5)起始边界和终止边界同时向前移动 δ;6)起始边界和终止边界同时向后移动 δ。其中 δ 表示移动步长。

定位智能体采用演员—评论家(actor-critic,AC)[23]算法来生成状态值和动作策略。在每个时间步,状态特征 s^{t} 首先通过 GRU[24]层,然后输入至演员分支和评论家分支。演员分支利用一个全连接层和一个 Softmax 层生成动作策略 $\pi(a_{i}^{t}|s^{t},\theta_{\pi})$,表示动作空间中各动作的概率分布;评论家分支则利用一个全连接层生成状态值 $V(s^{t},\theta_{v})$,表示对智能体在当前状态所获回报的估计,用来评估所执行动作的优劣。其中,$a_{i}^{t}(i=0,1,\cdots,5)$ 表示动作空间中的动作,θ_{π} 和 θ_{v} 分别为演员分支和评论家分支的网络参数。

智能体执行动作后,环境会相应更新,并反馈给智能体一个回报函数。该函数用于衡量不同动作的执行效果,以引导智能体学习更优策略。视频重定位任务的目标是精准定位目标片段,环境反馈的回报函数应鼓励智能体逐步细化定位片段的时间边界,使其更接近于目标片段。因此,本文使用时间交并比(temporal Intersection over Union, tIoU)来设计回报函数,tIoU 衡量了定位片段 $L^{t}=[L_{\mathrm{s}}^{t}, L_{\mathrm{e}}^{t}]$ 和目标片段 $G=[g_{\mathrm{s}}, g_{\mathrm{e}}]$ 之间的接近程度,其计算公式为

$$\mathrm{tIoU}^{t} = \frac{\min(g_{\mathrm{e}}, L_{\mathrm{e}}^{t}) - \max(g_{\mathrm{s}}, L_{\mathrm{s}}^{t})}{\max(g_{\mathrm{e}}, L_{\mathrm{e}}^{t}) - \min(g_{\mathrm{s}}, L_{\mathrm{s}}^{t})} \tag{4}$$

如果智能体的动作导致定位片段与目标片段之间的 tIoU 值提高,则将获得相应的奖励,特别当 $\mathrm{tIoU}\geqslant 0.5$ 时,奖励程度更大。反之,如果 tIoU 保持不变或降低,智能体将受到相应的惩罚。此外,若智能体的动作导致定位片段的起始边界出现在终止边界之后,将会受到更大程度的惩罚。综合上述情况,环境在第 t 时间步反馈给智能体的回报函数定义为:

$$r_{t} = \begin{cases} \tau + \mathrm{tIoU}^{t}, & \mathrm{tIoU}^{t} > \mathrm{tIoU}^{t-1} \cap \mathrm{tIoU}^{t} \geqslant 0.5 \\ \tau, & \mathrm{tIoU}^{t} > \mathrm{tIoU}^{t-1} \cap \mathrm{tIoU}^{t} < 0.5 \\ -0.1 * \tau, & \mathrm{tIoU}^{t-1} \geqslant \mathrm{tIoU}^{t} \geqslant 0 \\ -\tau, & \text{else} \end{cases} \tag{5}$$

式中:tIoU^{t-1} 和 tIoU^{t} 分别表示智能体在执行动作前后的 tIoU 值;τ 用于控制奖励或惩罚的程度。

在整个序贯决策过程中,智能体通过逐步执行动作实现最终任务目标。由于当前时间步的动作可能影响未来的决策,因此需要计算第 t 时间步的累积回报,以综合考虑当前时间步的立即回报和未来时间步的潜在回报。累积回报的定义为

$$R_{t} = \begin{cases} r_{t} + \gamma * V(s^{t},\theta_{v}), & t = T_{\max} \\ r_{t} + \gamma * R_{t+1}, & t = 1,\cdots,T_{\max}-1 \end{cases} \tag{6}$$

式中 γ 为折扣因子,用于控制未来回报对累积回报的影响。

1.3 多任务学习

为了学习更具代表性的状态特征,使其能够准确地反映环境信息,从而帮助智能体更有效地进行探索和学习,本文采用结合强化学习与监督学习的多任务学习框架。其中,基于策略优化的强化学习

引导智能体学习最优策略,逐步细化边界定位结果以接近目标片段;同时,基于位置回归的监督学习使网络能够预测定位片段的目标位置,并评估当前定位片段与目标片段的匹配程度。

1)基于策略优化的强化学习

AC 算法通过演员和评论家的协作优化智能体策略。演员的目标是学习策略(动作选择函数),使智能体在给定状态下选择最优动作。演员分支采用策略梯度方法更新参数,并引入优势函数以更有效地更新策略。其损失函数为

$$L'_{\mathrm{A}}(\theta_\pi) = -\sum_t (\log \pi(a_i^t \mid s^t, \theta_\pi)) * A_t,$$
$$A_t = R_t - V(s^t, \theta_v) \tag{7}$$

式中,A_t 为优势函数,衡量在给定状态下执行某个动作相对于平均情况的优劣程度。为了增加动作的多样性,进一步引入动作策略的熵[25],因此最终的损失函数为

$$L_{\mathrm{A}}(\theta_\pi) = L'_{\mathrm{A}}(\theta_\pi) - \lambda_0 \sum_t H(\pi(a_i^t \mid s^t, \theta_\pi) \tag{8}$$

式中,$H(\cdot)$ 表示熵,λ_0 为权重因子,用于控制熵项在整个损失中的权重。

评论家的目标是评估演员所执行动作的价值,即对当前状态所获回报进行估计。通过评估演员的动作,评论家提供策略改进的反馈。评论家分支通过最小化值函数误差更新参数,通常使用均方误差作为损失函数,定义为

$$L_{\mathrm{C}}(\theta_v) = \sum_t (R_t - V(s^t, \theta_v))^2 \tag{9}$$

对于基于策略优化的强化学习,总损失为演员分支和评论家分支的损失函数之和,表示为

$$\mathrm{loss}_{\mathrm{r}} = L_{\mathrm{A}}(\theta_\pi) + \lambda_1 * L_{\mathrm{C}}(\theta_v) \tag{10}$$

式中 λ_1 为权重因子。

2)基于位置回归的监督学习

监督学习部分包括时间边界回归和 tIoU 回归,输入均为状态特征 s^t。时间边界回归通过多层感知机 (multilayer perceptron, MLP)预测目标片段的时间边界,MLP 包含两个全连接层,其损失函数定义为:

$$L_{\mathrm{loc}} = \sum_t \left[y^t(\mid g_{\mathrm{s}} - P_{\mathrm{s}}^t \mid + \mid g_{\mathrm{e}} - P_{\mathrm{e}}^t \mid) \right] / 2,$$
$$y^t = \begin{cases} 1, & \mathrm{tIoU}^{t-1} \geqslant 0.5 \\ 0, & \mathrm{tIoU}^{t-1} < 0.5 \end{cases} \tag{11}$$

式中:$[P_{\mathrm{s}}^t, P_{\mathrm{e}}^t]$ 表示预测的时间边界;y^t 为一个指示值,表示仅在 tIoU≥0.5 时考虑时间边界回归。

tIoU 回归同样利用包含两个全连接层的 MLP 预测当前定位片段与目标片段之间的 tIoU 值,其损失函数为

$$L_{\mathrm{tIoU}} = \sum_t \mid \mathrm{tIoU}^{t-1} - P_{\mathrm{tIoU}}^t \mid \tag{12}$$

式中,P_{tIoU}^t 表示预测的 tIoU 值。在测试阶段,使用 tIoU 预测值作为定位智能体的停止信号。具体而言,定位智能体与环境交互 $T_{\max}$ 步,获得一系列 tIoU 预测值,这些值反映了定位片段与目标片段之间的匹配程度。因此,从这些预测值中选择 tIoU 最大的时间步,将其对应的定位片段作为最终结果。

基于位置回归的监督学习总损失为时间边界回归和 tIoU 回归的损失之和,表示为

$$\mathrm{loss}_{\mathrm{s}} = L_{\mathrm{loc}} + \lambda_2 * L_{\mathrm{tIoU}} \tag{13}$$

式中 λ_2 为权重因子。

AGLN 通过最小化损失$\mathrm{loss}_{\mathrm{r}} + \lambda_3\ \mathrm{loss}_{\mathrm{s}}$ 进行端到端的训练,其中 λ_3 为权重因子。通过同时优化这两部分损失,AGLN 能够学习更具代表性的状态特征。这有助于智能体学习更优的动作策略,从而实现更准确的目标片段定位。

2　实验与结果分析

本节首先介绍了实验设置,包括数据集、性能评估指标、训练细节以及用于对比的方法;然后对本文提出方法与对比方法进行定量比较,并展示了所提方法进行视频重定位的可视化示例;最后进行算法分析,验证了所提方法中超参数的影响以及主要组件的有效性,并对其运行效率进行比较与分析。

2.1　实验设置

1)数据集

ActivityNet 是用于动作定位的大规模数据集,包含来自 200 个动作类别的视频,每个视频中的动作片段使用相应的动作类别进行标注。Feng 等[26]通过对 ActivityNet 中的视频序列进行重新组合,构建了适用于视频重定位任务的新数据集 ActivityNet-VRL。ActivityNet-VRL 是目前唯一公开用于视频重定位任务的数据集,该数据集包含 9 530 个未剪辑的视频,每个视频仅包含 1 个动作片段。数据集按动作类别进行划分,其中训练集包含 160 个动作类别,而验证集和测试集则分别包含 20 个动作类别。这种划分方式确保了用于验证和测试的动作类别在训练期间是不可见的,从而使得模型能够定位未知的动作类别。在视频重定位任务中,每组数据都由成对的查询视频和参考视频组成。每个视频中的动作片段都可以作为查询视频,与之配对的参考视频则可以从具有相同动作类别的未剪辑视频中选择。在训练阶段,查询视频和参考视频是随机匹配的,而在验证和测试阶段,查询视频和参考视频的配对是固定的。

2)性能评估指标

在视频重定位任务中,通过预测定位片段与目标片段之间的匹配程度评估模型性能,常用的评估指标是平均准确率(mean average precision, mAP)。对于一个查询视频,首先计算模型预测定位片段与目标片段之间的 tIoU。如果 tIoU 值不低于设定的阈值 m,则判定为定位正确;否则视为定位错误。mAP 表示所有查询视频中定位正确的视频所占的比例,可用以下公式表示:

$$\text{mAP} = \frac{1}{N_q}\sum_{i=1}^{N_q}\text{acc}(q_i,m),$$

$$\text{acc}(q_i,m) = \begin{cases}1, & \text{tIoU} \geqslant m\\ 0, & \text{tIoU} < m\end{cases} \tag{14}$$

式中:q_i 代表查询视频;m 是设定的 tIoU 阈值;N_q 表示查询视频的总数。在本文实验中,将 m 设置为 $\{0.5, 0.6, 0.7, 0.8, 0.9\}$,分别计算不同阈值下的 mAP,然后计算这些 mAP 的平均值。

3)训练细节

本文使用 2016 年 ActivityNet 挑战赛发布的视频特征作为模型输入,这些特征由预训练的 C3D 网络(时间分辨率为 16 帧,空间分辨率为 112 × 112)提取得到。在网络训练阶段,使用学习率为 1×10^{-3}的 Adam 优化器[27],将批次设置为 32;移动步长 δ 为 $N/10$;最大步数 $T_{\max}$为 10;回报函数中的参数 τ 为 1;累积回报中的折扣因子 γ 为 0.3;损失函数中的权重因子 λ_0 为 0.1;λ_1、λ_2 和 λ_3 则均被设置为 1。初始定位片段基于 C3D 网络得到的视频特征序列确定。具体来说,首先对查询视频特征序列进行平均池化,计算其与参考视频特征序列中每个特征向量之间的相似度,然后选择相似度最高的片段作为初始定位片段。

4)对比方法

为了验证本文所提方法的优越性,将其与其他多个方法进行了性能对比。由于目前针对视频重定位任务设计的方法比较少,为了进行充分比较,本文将类似任务设计的方法应用于视频重定位任务,并将其性能与本文方法进行对比。

随机(Random)方法:从参考视频中随机选择一个片段作为定位片段,要求选取片段的起始边界小于终止边界。

统计先验(Statistical prior)方法:对训练集中所有视频的长度进行归一化处理,统计视频中目标片段的平均起始边界和终止边界,分别为 0.326 5 和 0.650 9,之后利用该先验知识在测试集上进行预测。

帧级别(Frame-level)方法:基于回溯表和对角块[28]的思想。首先对查询视频和参考视频的特征序列进行归一化处理,计算所有特征向量之间的 L2 距离得到一个距离表;再通过动态规划搜索平均距离最小的对角块,将其对应的片段作为最终的定位片段。

视频级别(Video-level)方法:通过滑动窗口技术生成多个候选片段,同时采用 LSTM 对视频进行特征编码。模型训练过程中使用三元组损失[29],以区分匹配和不匹配的片段;测试时则选择匹配程度最高的候选片段作为定位结果。

SST[30]:是一种用于动作定位的方法。对于给定的视频,SST 会检测其中所有可能的动作片段,并为每个片段分配一个置信度分数。在本文中,将 SST 应用于视频重定位任务。使用训练集中的视频训练模型,在测试阶段,从生成的动作片段中选择置信度分数最高的片段作为定位片段。

CGBM:利用交叉门双线性匹配来捕获视频之间的复杂交互,将参考视频中的每个帧级特征与经过注意力加权的查询视频特征进行匹配,并将时间边界预测问题转换为一个分类问题,即通过判断参考视频中每一帧是否与查询视频相关来找到定位片段。

RWM:是一种用于自然语言视频定位的方法。基于强化学习算法,训练一个智能体迭代地读取文本描述,观看视频以及定位片段,然后移动定位边界,以找到最佳匹配的片段。在本文中,将 RWM 中的自然语言特征提取器 Skip-thought[31]替换为视频特征提取器 C3D,将其应用于视频重定位任务。

2D-TAN[32]:是一种用于自然语言视频定位的方法。利用二维时序图建模片段之间的时间关系,不仅能够有效地感知视频的上下文信息,还能够学习到具有判别性的特征,用以区分语义复杂的片段。在本文中,将 2D-TAN 中的自然语言特征提取器 Glove[33]替换为视频特征提取器 C3D,并将其应用于视频重定位任务。

MABAN[34]:是一种用于自然语言视频定位的方法。采用多智能体强化学习框架,训练两个智能体,分别在不同策略下定位片段的起始边界和终止边界。本文在 MABAN 的基础上使用一个 C3D 网络对查询视频进行特征提取,将其应用于视频重定位任务。

URL[35]:是一种用于自然语言视频定位的方法。采用一种对抗学习框架,利用强化学习作为生成器,产生一系列候选片段,同时使用多任务学习模块作为鉴别器,评估片段和查询文本之间的相关性。生成器和鉴别器在对抗学习的过程中相互促进,实现了对视频片段排名和定位性能的共同优化。在本

文中,将 URL 应用于视频重定位任务,将其中的查询文本替换为查询视频。

2.2　与对比方法的定量比较结果

表 1 展示了本文提出的 AGLN 和其他对比方法的性能评估结果,包括不同 tIoU 阈值{0.5, 0.6, 0.7, 0.8, 0.9}下的 mAP 以及平均值。从表中的结果可以看出,相较于现有方法中的最佳结果,AGLN 在所有 tIoU 阈值下的平均 mAP 从 25.7% 提升至 25.9%,这表明 AGLN 能够获得更加准确的定位结果。

表 1　不同方法的性能评估结果

Tab. 1　Performance evaluation results of different methods

方法	mAP/%						方法	mAP/%					
	0.5	0.6	0.7	0.8	0.9	Avg.		0.5	0.6	0.7	0.8	0.9	Avg.
Random	16.2	11.0	5.4	2.9	1.2	7.3	RWM	41.7	31.0	21.2	11.9	4.2	22.0
Statistical prior	25.4	16.5	2.3	2.3	1.2	10.7	2D-TAN	39.6	33.9	26.0	18.5	6.0	24.8
Frame-level	18.8	13.9	9.6	5.0	2.3	9.9	MABAN	37.5	28.9	20.2	12.0	4.5	20.6
Video-level	24.3	17.4	12.0	5.9	2.2	12.4	URL	42.3	32.3	21.7	13.5	5.6	23.1
SST	33.2	24.7	17.3	7.8	2.7	17.1	**AGLN**	**45.6**	**34.8**	**25.4**	**15.8**	**7.8**	**25.9**
CGBM	43.5	35.1	27.3	16.2	6.5	25.7							

2.3　可视化示例

为了全面验证所提出的 AGLN 的有效性,列示模型预测的几个定位片段示例,如图 2 所示。尽管查询视频和参考视频中目标片段表达的是相似的语义信息,但在持续时间和外观表示上均存在显著差异,而且目标片段和参考视频中的其余部分在视觉上是很相似的。因此,在参考视频中精准地定位目标片段是有难度的。然而,AGLN 能够有效地应对上述挑战,在参考视频中准确地定位到与目标片段非常接近的位置。值得注意的是,这些示例中的动作类别并未包含在训练集中,这进一步证明了 AGLN 具有定位未知动作类别的能力。

图 2　AGLN 预测的定位片段示例

Fig. 2　Example of localized moment predicted by AGLN

2.4　算法分析

1) 累积回报中折扣因子的影响

在序贯决策过程中,需要将未来时间步的回报追溯到当前时间步,以计算累积回报,从而在动作序列内部建立联系。为了探究累积回报中的折扣因子 γ 对模型性能的影响,在实验中将 γ 的取值范围设置为 0.1 ~ 0.9,步长为 0.2。使用不同值的 γ 对模型进行训练,并比较在 tIoU 阈值为 0.5 下的 mAP,

结果如图 3 所示。当 γ 取 0.3 时,mAP 值最高;当 γ 取值过小时,累积回报几乎不受未来回报的影响,从而忽略了连续动作之间的序列依赖性;相反,当 γ 取值过大时,累积回报过度依赖于未来回报的影响,即使当前时间步的动作是失败的尝试,也可能因为未来时间步的积极回报而获得累积回报。这两种情况均不是合理的奖励机制。因此,将折扣因子 γ 设置为 0.3。

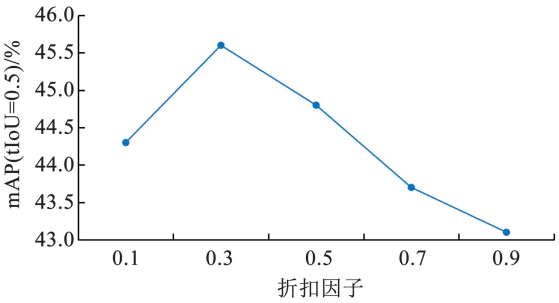

图 3 累积回报中折扣因子的影响

Fig. 3 Influence of discount factor in accumulated reward

表 2 AGLN 不同变体的性能评估结果

Tab. 2 Performance evaluation results of different variants of AGLN

方法	mAP/%					
	0.5	0.6	0.7	0.8	0.9	Avg.
w/o context	42.6	33.1	23.6	14.2	7.1	24.1
w/max step	41.5	29.3	19.7	11.4	4.0	21.2
w/o Loc-R	39.5	29.3	19.4	10.2	3.5	20.4
w/o tIoU-R	40.6	28.0	20.6	11.6	3.4	20.8
w/o Reg	37.5	26.6	19.0	10.3	2.4	19.1
w/stop action	45.0	33.8	23.8	12.9	5.6	24.2
w/reward-BE	41.9	29.6	19.2	12.1	4.7	21.5
AGLN	**45.6**	**34.8**	**25.4**	**15.8**	**7.8**	**25.9**

2)利用视频上下文信息的有效性

为了验证 AGLN 的上下文感知特征提取器在利用视频上下文信息方面的有效性,在生成状态特征时去除了左侧片段和右侧片段的特征,并将相应的模型变体记为“w/o context”。从表 2 中的结果可以看出,与 AGLN 相比,这一变体在所有 tIoU 阈值下的平均 mAP 从 25.9% 降低至 24.1%。这表明,充分利用视频的上下文信息能够更全面且准确地反映环境状态,对模型性能具有显著积极影响。

3)基于位置回归的监督学习的有效性

为了验证 AGLN 在多任务学习框架中进行基于位置回归的监督学习的有效性,依次去除监督学习部分的时间边界回归和 tIoU 回归训练模型的多个变体。将去除时间边界回归的模型记为“w/o Loc-R”,去除 tIoU 回归的模型记为“w/o tIoU-R”,同时去除这两部分的模型记为“w/o Reg”。值得注意的是,去除 tIoU 回归后,无法在测试阶段使用 tIoU 预测值作为智能体的停止信号。因此,对于上述变体模型,直接选取最大时间步的定位片段作为最终结果。为了进行比较,还展示了完整模型在测试阶段直接选取最大时间步的定位片段作为最终结果的性能评估情况,记为“w/max step”。从表 2 中的结果可以看出,与“w/max step”相比,所有模型变体的性能均有所下降。去除时间边界回归时,所有 tIoU 阈值下的平均 mAP 值从 21.2% 降低至 20.4%;去除 tIoU 回归时,所有 tIoU 阈值下的平均 mAP 值从 21.2% 降低至 20.8%;而同时去除这两部分时,所有 tIoU 阈值下的平均 mAP 值从 21.2% 降低至 19.1%。这充分表明了基于位置回归的监督学习的有效性。通过利用目标片段的时间边界标注作为监督信息,可以帮助智能体获得更准确和全面的状态特征,进而更有效地探索环境,实现任务目标。

4)测试阶段停止信号的有效性

为了验证 AGLN 在测试阶段使用 tIoU 回归生成的 tIoU 预测值作为智能体停止信号的有效性,在动作空间中添加了一个额外的动作,记为“stop”,并将对应的模型变体记为“w/stop action”。在训练和测试阶段,当智能体选择执行 stop 动作时,该时间步的定位片段便被视为最终的定位结果。从表 2 中的结果可以看出,与 AGLN 相比,引入 stop 动作的模型性能有所下降。这主要是因为在动作空间中引入 stop 动作时,很难为其设计一个合适的回报机制,导致智能体难以有效地终止迭代过程。

5)基于 tIoU 设计回报函数的有效性

为了验证 AGLN 基于 tIoU 设计回报函数的有效性,采用定位片段与目标片段之间的边界误差(boundary error,BE)作为另一种回报函数的设计依据,其计算方式如公式(15)所示,指的是定位片段的起始边界和终止边界与真实值之间的偏差。智能体执行动作后,若边界误差减小,会给予奖励,反之则给予惩罚。将使用这种回报函数设计的模型记为“w/reward-BE”。从表 2 中的结果可以看出,w/reward-BE 的性能明显低于 AGLN,证明了基于 tIoU 设计回报函数的有效性。相比于 BE,tIoU 不仅衡量了边界的准确性,还综合考虑了两个片段的整体重合程度,从而更全面地评估了匹配程度。因此,基于 tIoU 设计的回报函数能够更有效地引导智能体找到与目标片段匹配的定位片段,实现更精确的目标片段定位。

$$\mathrm{BE} = |g_s - L_s^t| + |g_e - L_e^t| \tag{15}$$

6)模型的运行效率

为了全面评估所提出的 AGLN 的有效性,本文在配备 12 GB GPU 内存的 NVIDIA GTX 1080Ti 显卡上统计了其运行时间。模型总共训练 200 轮次,耗时 187.7 分钟。在实际查询阶段,模型的处理速度约为 20.3 帧/秒。进一步地,将 AGLN 与其他现有方法如 CGBM 和 MABAN 等在参数量、计算复杂度和推理时间方面进行了比较。其中,计算复杂度是通过乘积累加运算数(multiply-accumulate operations, MACs)进行衡量的,而推理时间则表示测试阶段处理单个样本所需的平均时间。根据表 3 中的数据,AGLN 在这些方面的表现均更出色,特别是在计算复杂度上。这一结果证明了 AGLN 不仅在性能上表现优异,而且在运行效率方面也具有显著优势。

表 3　模型的运行效率比较

Tab. 3　Comparison of operational efficiency of models

方法	参数量/10^6	MACs /10^6	推理时间/s
CGBM	1.74	190.56	0.051
MABAN	12.19	238.21	0.071
AGLN	1.56	1.56	0.010

3　结　语

本文基于强化学习提出了一个智能体引导的定位网络 AGLN,用于实现视频重定位任务。AGLN 的核心思想在于通过智能体的序贯决策,实现对目标片段的精准定位。智能体基于学习到的策略,逐步细化定位片段的时间边界,以找到与查询视频在语义上最相关的片段。此外,AGLN 将基于策略优化的强化学习与基于位置回归的监督学习融合于多任务学习框架,进一步提升了模型的性能。在 ActivityNet-VRL 数据集上进行实验,充分证明了 AGLN 在视频重定位任务上的优越性和有效性。

参考文献

[1]HU Weiming, XIE Nianhua, LI Li, et al. A survey on visual content-based video indexing and retrieval[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011, 41(6): 797. DOI: 10.1109/TSMCC.2011.2109710

[2]JIANG Chen, HUANG Kaiming, HE Sifeng, et al. Learning segment similarity and alignment in large-scale content based video retrieval[C]//Proceedings of the 29th ACM International Conference on Multimedia. Online: ACM, 2021: 1618. DOI: 10.1145/3474085.3475301

[3]GE Yuying, GE Yixiao, LIU Xihui, et al. Bridging video-text retrieval with multiple choice questions [C]//Proceedings of theIEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 16146. DOI: 10.1109/CVPR52688.2022.01569

[4]GAO Jiyang, SUN Chen, YANG Zhenheng, et al. Tall: Temporal activity localization via language query[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5277. DOI: 10.1109/ICCV.2017.563

[5]XIAO Shaoning, CHEN Long, ZHANG Songyang, et al. Boundary proposal network for two-stage natural language video localization [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Online: AAAI, 2021: 2986. DOI: 10.1609/aaai.v35i4.16406

[6]SUN Xin, GAO Jialin, ZHU Yizhe, et al. Videomoment retrieval via comprehensive relation-aware network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 5281. DOI: 10.1109/TCSVT.2023.3250518

[7]FENG Yang, MA Lin, LIU Wei, et al. Video re-localization[C]//Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 55. DOI: 10.1007/978-3-030-01264-9_4

[8]胡薰尹, 管业鹏. 基于3D-LCRN 视频异常行为识别方法[J]. 哈尔滨工业大学学报, 2019, 51(11): 183
HU Xunyin, GUAN Yepeng. 3D-LCRN based video abnormal behavior recognition[J]. Journal of Harbin Institute of Technology, 2019, 51(11): 183. DOI:10.11918/j.issn.0367-6234.201812005

[9]ZAHRA A, PERWAIZ N, SHAHZAD M, et al. Person re-identification: A retrospective on domain specific open challenges and future trends[J]. Pattern Recognition, 2023, 142: 109669. DOI: 10.1016/j.patcog.2023.109669

[10]TANG Haoyu, ZHU Jihua, GAO Zan, et al. Attention feature matching for weakly-supervised video relocalization [C]//Proceedings of the 2nd ACM International Conference on Multimedia in Asia. Online: ACM, 2021. DOI: 10.1145/3444685.3446317

[11]HUO Shuwei, ZHOU Yuan, WANG Ruolin, et al. Semantic relevance learning for video-query based video moment retrieval [J]. IEEE Transactions on Multimedia, 2023. DOI: 10.1109/TMM.2023.3250088

[12]FENG Yang, MA Lin, LIU Wei, et al. Spatio-temporal video re-localization by Warp lSTM[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1288. DOI: 10.1109/CVPR.2019.00138

[13]HUNG Y H, HSU K J, JENG S K, et al. Weakly-supervised video re-localization with multiscale attention model[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 11077. DOI:10.1609/aaai.v34i07.6763

[14]CHEN Long, LU Chujie, TANG Siliang, et al. Rethinking the bottom-up framework for query-based video localization [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 10551. DOI:10.1609/aaai.v34i07.6627

[15]HUO Shuwei, ZHOU Yuan, XIANG Wei, et al. Weakly-supervised content-based video moment retrieval using low-rank video representation[J]. Knowledge-Based Systems, 2023, 277: 110776. DOI: 10.1016/j.knosys.2023.110776

[16]WU Wenhao, HE Dongliang, TAN Xiao, et al. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6221. DOI: 10.1109/ICCV.2019.00632

[17]XU Wanru, MIAO Zhenjiang, YU Jian, et al. Deep reinforcement learning for weak human activity localization [J]. IEEE Transactions on Image Processing, 2019, 29: 1522. DOI: 10.

1109/TIP. 2019. 2942814

[18] HE Dongliang, ZHAO Xiang, HUANG Jizhou, et al. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 8393. DOI:10.1609/aaai.v33i01.33018393

[19] YOON U N, HONG M D, JO G S. Unsupervised video summarization based on deep reinforcement learning with interpolation[J]. Sensors, 2023, 23(7): 3384. DOI: 10.3390/s23073384

[20] SUN M J, XIAO J M, LIM E G, et al. Unified multi-modality video object segmentation using reinforcement learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(12): 6890. DOI: 10.1109/TCSVT.2023.3284165

[21] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 4489. DOI: 10.1109/ICCV.2015.510

[22] CHAPLOT D S, MYSORE SATHYENDRA K, PASUMARTHI R K, et al. Gated-attention architectures for task-oriented language grounding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 2819. DOI:10.1609/aaai.v32i1.11832

[23] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. 2nd ed. Cambridge: MIT Press, 2018

[24] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014: 1724. DOI: 10.3115/v1/d14-1179

[25] LI Debang, WU Huikai, ZHANG Junge, et al. A2-RL: Aesthetics aware reinforcement learning for image cropping[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8193. DOI: 10.1109/CVPR.2018.00855

[26] HEILBRON F C, ESCORCIA V, GHANEM B, et al. Activitynet: A large-scale video benchmark for human activity understanding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 961. DOI: 10.1109/CVPR.2015.7298698

[27] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015. DOI: 10.48550/arXiv.1412.6980

[28] CHOU C L, CHEN H T, LEE S Y. Pattern-based near-duplicate video retrieval and localization on web-scale videos[J]. IEEE Transactions on Multimedia, 2015, 17(3): 382. DOI: 10.1109/TMM.2015.2391674

[29] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 815. DOI: 10.1109/CVPR.2015.7298682

[30] BUCH S, ESCORCIA V, SHEN Chuanqi, et al. Sst: Single-stream temporal action proposals[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6373. DOI: 10.1109/CVPR.2017.675

[31] KIROS R, ZHU Y K, SALAKHUTDINOV R, et al. Skip-thought vectors[C]//Advances in Neural Information Processing Systems, 2015, 28: 3294. DOI:10.48550/arXiv.1506.06726

[32] ZHANG Songyang, PENG Houwen, FU Jianlong, et al. Learning 2d temporal adjacent networks for moment localization with natural language[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12870. DOI:10.1609/aaai.v34i07.6984

[33] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014: 1532. DOI: 10.3115/v1/d14-1162

[34] SUN Xiaoyang, WANG Hanli, HE Bin. Maban: Multi-agent boundary-aware network for natural language moment retrieval[J]. IEEE Transactions on Image Processing, 2021, 30: 5589. DOI: 10.1109/TIP.2021.3086591

[35] ZENG Yawen, CAO Da, LU Shaofei, et al. Moment is important: Language-based video moment retrieval via adversarial learning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 18(2): 1. DOI: 10.1145/3478025

(编辑 丁晓清)