中国科学: 信息科学 2022 年 第 52 卷 第 8 期: 1417–1446
SCIENTIA SINICA Informationis
论文

基于精细化多模态关联的自然语言句子在视频中的时序定位方法

袁艺天, 王鑫, 朱文武*

清华大学计算机系, 北京 100084
* 通信作者. E-mail: wwzhu@tsinghua.edu.cn

收稿日期: 2021–04–22; 修回日期: 2021–06–21; 接受日期: 2021–07–22; 网络出版日期: 2022–08–05

科技创新 2030 “新一代人工智能” 重大项目 (批准号: 2020AAA0106300) 和国家自然科学基金委原创探索计划项目 (批准号: 62050110) 资助

摘要 通信技术和移动互联网的发展使多媒体数据逐渐渗透人们的生活, 而视频作为其中一种最具表现力的内容表达方式, 近年来受到了工业界和学术界的广泛关注. 针对视频数据中背景信息较为冗余, 所需分析浏览时间长的特点, 本文介绍了自然语言句子在视频中的时序定位任务, 即在视频中定位与给定自然语言句子语义相关的视频片段, 这样人们可以通过提供明确简洁的文本描述在视频中迅速找寻所关注的特定内容, 从而提高用户的视频浏览体验和搜索效率. 传统方法往往以多模态匹配的框架来解决句子在视频中的时序定位问题, 忽略了自然语言句子中的关键定位线索, 更忽视了自然语言句子对于关联视频内部相关内容的重要指导作用, 因而其时序定位准确率十分有限. 为解决上述难题, 本文提出了多模态共同注意力机制挖掘自然语言句子中与时序定位相关的重要语义细节, 精细地构建句子中各单词和视频内容之间的语义关系. 在此基础上, 我们还提出了语义条件动态归一化机制, 指导视频中与句子语义相关的局部视频内容紧密耦合, 形成明确的视频片段边界, 最后辅以细粒度的边界调整模块, 进而获得更为精准和灵活的时序定位结果. 在公开数据集上的实验验证了本文所提出的机制和方法的有效性. 最后, 本文还从引入视频中的音频信号、考虑弱监督环境下的时序定位问题, 以及构建无偏见时序定位数据集这 3 个方面对自然语言句子在视频中的时序定位问题进行了未来研究方向的展望.

关键词 时序定位, 语义关联, 多模态共同注意力机制, 时序卷积网络, 语义条件动态归一化机制

1 引言

随着通信技术、互联网和社交媒体的快速发展, 视频数据逐渐充斥着人类的日常生活. 与传统的文字、图片等形式的信息媒介有所不同, 视频以图片流的形式, 辅以听觉信息, 可以更为直观、生动地

引用格式: 袁艺天, 王鑫, 朱文武. 基于精细化多模态关联的自然语言句子在视频中的时序定位方法. 中国科学: 信息科学, 2022, 52: 1417–1446, doi: 10.1360/SSI-2021-0138
Yuan Y T, Wang X, Zhu W W. Temporal sentence grounding in videos with fine-grained multimodal correlation (in Chinese). Sci Sin Inform, 2022, 52: 1417–1446, doi: 10.1360/SSI-2021-0138

图 1 (网络版彩图) 网络视频周围往往存在着大量的文本信息, 例如视频标题、视频描述、视频评论等

Figure 1 (Color online) Web videos are surrounded by a lot of natural language sentences, such as video titles, video descriptions, and video comments

展示所发生的事件、所处的情境, 让人们得到更为深刻的感知. 但与此同时, 人们观看浏览视频内容的时间也比阅读文字和查看图片的时间更长, 视频所需的存储空间, 也远大于图片和文字. 另外, 我们还可以观察到, 视频中真正被人们关注的内容往往是十分稀疏的. 例如, 在监控视频中人们更关注一些异常的情境而不是冗长的正常生活场景, 在直播视频中人们更关注对重点商品的推介而不是主播串场时的闲谈等. 因此, 如何在视频中快速地定位抓取值得人们关注的、有价值的视频内容, 对视频的管理和存储以及提升人们搜索观看视频的体验有着重要的意义.

为解决上述问题, 近年来研究者们在视频关键动作检测 (action detection) 这一任务上进行了大量的探索 [1~9]. 具体而言, 视频关键动作检测旨在视频中发掘关键事件和人类的行为动作, 将其进行分类, 并定位其在视频中的时间位置. 然而, 在视频关键动作检测这一问题中, 能够被归类到检测目标中的事件和动作种类有限, 无法涵盖视频中大量种类丰富的内容. 因此, 人们开始寻求以更加灵活的方式来定位识别视频中的内容. 事实上, 视频内容在网络空间中不是独立存在的, 如图 1 所示, 网络视频周遭往往存在着大量的自然语言文本信息, 例如视频标题、视频描述、视频评论等. 这些文本信息, 能够以简洁的语言描述视频中用户所关注的信息, 因而可以为视频中关键内容的定位提供指导. 相比于关键动作检测问题中被检测到的事件总是局限于特定类别集合中, 自然语言能够表达更为丰富的含义且没有类别限制, 基于上述思考, 人们开始探究一项新的任务 [10, 11], 如图 2 所示, 给定一个视频和一句描述该视频内容的自然语言句子, 需要预测该句子所描述的目标视频片段在视频中的时间位置, 该任务也被称为自然语言句子在视频中的时序定位问题 (temporal sentence grounding/localization in videos).

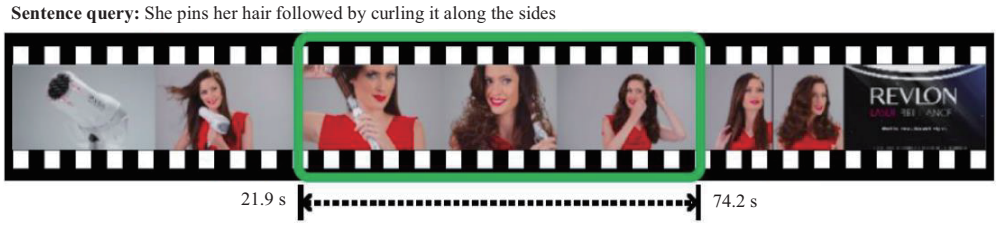

图 2 (网络版彩图) 自然语言句子在视频中的时序定位问题示意图

Figure 2 (Color online) The illustration of temporal sentence grounding in videos

为了解决自然语言句子在视频中的时序定位问题, 早期的方法主要采用的是两阶段的 "先扫描再调整 (匹配)" 的多模态匹配框架. 这一类方法 [10~12] 首先利用滑动窗口在视频中密集地扫描采样出各种时间跨度不同的候选视频片段, 然后再将自然语言句子与这些候选片段一一进行多模态匹配, 最终挑选匹配度最高的候选片段作为时序定位的结果. 另一类方法舍弃了前述的两阶段框架, 它们或利用长短时记忆循环神经网络处理视频并直接预测目标视频片段的位置 [13], 或将视频视作一个二维时间图, 在此时间图中对视频内容的依赖关系进行建模从而预测目标片段的位置 [14]. 还有一类方法结合了强化学习的思想 [15], 将一个位置任意初始化的片段通过位置跳跃和边界调整的方式逐步地向目标视频片段的位置拉近. 尽管上述方法都取得了不错的时序定位性能, 但是仍然存在如下缺陷:

- 上述基于多模态匹配框架的模型需要在视频中利用滑动窗口密集地采样候选片段, 这带来了大量处理候选片段的额外开销, 其时间效率并不高. 此外, 将视频局部片段和自然语言句子进行整体编码后再匹配, 一则割裂了视频的上下文关系, 二则忽略了自然语言句子中的重要细节 (例如为时序定位提供关键线索的词语、短语等), 从而影响了时序定位的准确性.
- 上述方法认为视频内容与自然语言句子只存在简单的多模态语义匹配关系, 而忽略了自然语言句子对视频内部内容的耦合起到的重要指导作用. 例如原本在视觉层面可以割裂来看的两个视频事件, 自然语言句子却将二者关联起来作为一个完整的复杂事件呈现, 在此情况下, 这两个事件就应该耦合到一起被整体定位. 简单的多模态语义匹配关系无法建立起视频内部内容之间的关联, 因此具有一定的局限性.
- 上述方法总是以在视频中采样的候选片段为基础, 逐步调整这些片段的位置来确定最终的目标片段位置. 因此, 我们可以认为这是一种 "片段级别" 的粗粒度预测. 然而目前选取候选片段的方式往往被预先定义, 导致候选片段的位置、长度都不够灵活多样, 因此所预测的目标片段的边界位置也不够准确, 需要进行更为精细的调整.

为解决以上问题, 本文提出视频和自然语言句子之间的精细化语义关联方法, 不仅考虑视频片段和句子细节的语义匹配关系, 还着重利用自然语言信息来充分耦合视频中的相关内容, 从而更精准高效地预测出自然语言句子所描述的目标片段的位置. 具体而言, 本文的工作主要从以下 3 个方面展开:

- 我们提出了一种基于注意力回归的时序定位方法, 它打破了传统两阶段的 "先扫描再调整" 的框架, 利用一种多模态共同注意力机制生成自然语言句子和视频的注意力权重. 句子注意力权重能够增强句子中对时序定位起关键指导作用的重要词语的影响; 视频注意力权重可以反映视频各部分内容与句子的语义匹配关系, 进而整合出全局视频结构, 并能够由此直接高效地回归出目标片段的位置坐标.
- 上述基于注意力回归的时序定位方法能够根据视频内容动态地衡量句子中与时序定位相关的关键词语, 并生成注意力加权的句子表征; 在此基础上, 我们提出了一种语义条件动态归一化 (semantic conditioned dynamic normalization, SCDN) 机制, 它作用于多层时序卷积网络, 能够基于不同位置的

视频内容动态地引导产生注意力加权的句子表征, 并以此指导时序卷积网络的特征归一化过程, 调整卷积特征图的数据分布, 进而促进与句子语义相关联的视频内容紧密耦合, 为目标片段的预测提供更为清晰的特征分布边界.

- 基于上述语义条件动态归一化机制指导下的时序卷积网络, 我们进一步引入细粒度时间边界调整模块, 它作用于卷积网络底层, 输出单位时间粒度上的视频片段和句子的语义匹配关系, 从而为高层卷积网络所预测的粗粒度的候选视频片段提供更为精细的时间边界调整, 提高时序定位的准确度.

下面将具体回顾自然语言句子在视频中的时序定位任务的相关方法, 然后逐步介绍我们所提出的 3 个方面的工作, 并进行实验分析.

2 相关工作介绍

早期的自然语言句子在视频中的时序定位问题主要拘泥于特定的视觉场景 (例如厨房、实验室等), 并要将多个连续且相互关联的句子同时在视频中进行时序定位 [16~19]. 近年来, Gao 等 [11] 和 Hendricks 等 [10] 将这一问题进一步泛化, 推广到更加多元的视觉场景, 且更强调将单个自然语言句子在视频中进行独立的时序定位. Hendricks 等 [10] 提出了片段上下文网络 (moment contextual network, MCN), 以一种多模态匹配的框架来解决自然语言句子在视频中的时序定位问题. 首先, 他们利用滑动窗口在视频中密集地扫描采样出各种时间跨度不同的候选视频片段, 然后将这些视频片段和自然语言句子映射到同一隐特征空间, 再比较自然语言句子与这些候选片段在隐空间中的距离, 最终选取最靠近自然语言句子的片段作为时序定位的结果. Gao 等 [11] 和 Liu 等 [12] 将目标检测方法的思想应用到时序定位问题中, 并分别提出了跨模态时序回归定位器 (cross-modal temporal regression localizer, CTRL) 和注意力的跨模态检索网络 (attentive cross-modal retrieval network, ACRN). 在这一类方法中, 模型也需要在视频中扫描采样具有多种时间尺度的视频候选片段, 不同于 MCN 的是, CTRL 和 ACRN 将视频候选片段和自然语言句子的特征融合, 并设计了一个位置回归网络从融合特征中预测出候选片段与目标视频片段之间的位置偏移和位置重合度, 这样便能由候选片段推断出目标视频片段的位置. Ge 等 [20] 则进一步改进了上述位置回归网络, 并提出了基于事件活动概念检测的定位器 (activity concepts based localizer, ACL), 它能够从视频和自然语言句子中提取挖掘出重要的事件活动的概念来进一步提升时序定位的准确度. Wu 等 [21] 则在 CTRL 的基础上引入了新的多模态循环融合网络 (multimodal circulant fusion, MCF) 来强化视频和自然语言句子特征的融合, 进而提升时序定位精度. 由于上述所有方法都需要利用滑动窗口从视频中采样大量的候选片段, 因此这些方法会带来大量的计算冗余, 导致模型时间复杂度较高.

为了跳出上述利用滑动窗口采样候选视频片段的思路, Xu 等 [22] 提出先利用自然语言句子的语义在视频中预判一些片段提案, 然后再从这些片段提案出发重新生成自然语言句子, 通过计算重新生成的句子与原自然语言查询句子的相似度来对片段提案进行排序, 排序最高的提案则作为最终定位结果. 值得注意的是, 与滑动窗口采样产生的候选视频片段相比, 预判的视频片段提案的数目大大减少, 且在句子语义的指导下也与最终目标片段更具关联性, 因此可以提升模型的时间效率, 这一方法也被称为基于句子重构的视频片段定位方法. 类似地, Chen 等 [23] 提出了基于视觉概念挖掘的定位方法 (semantic proposal for activity localization, SAP), 这一方法也将自然语言句子的语义信息融入到视频片段提案的生成过程中, 并通过检测视频中的视觉概念产生视频视觉概念特征向量, 且自然语言句子和视频视觉概念特征的融合表征将被用来进一步评估和细化所产生的视频片段提案的边界信息. 上述两种方法仍然将视频片段提案的产生和排序分割开来, 因此整体的时序定位模型也无法联合优化.

除此之外, 还有一类方法不需要提前采样和生成任何候选片段和提案, 而是以一种端到端的方式来完成自然语言句子在视频中的时序定位. Chen 等 [13] 提出了时序定位网络 (temporal ground-net, TGN), 它首先将视频均等分割成若干视频单元, 然后利用一个长短时记忆循环神经网络 (long short-term memory, LSTM) 定位器来动态地匹配各视频单元和自然语言句子. 在每一视频单元处, LSTM 定位器都将考虑终止点位于此视频单元内的若干具有不同时间尺度的视频片段, 计算它们作为目标片段的概率, 并以此来对这些视频片段进行排序, 获得最终的时序定位结果. Zhang 等 [24] 则提出了片段对齐网络 (moment alignment network, MAN), 他们利用时序卷积网络中时间卷积层的特征单元都能对应于视频中特定片段的特性, 自然高效地建立了视频卷积特征和候选视频片段之间的映射, 同时还利用图卷积网络 [25] (graph convolutional network, GCN) 来刻画各视频片段在自然语言句子语义指导下的内部关联关系, 从而加强时间边界预测的准确性, 最终通过一次网络前馈计算便能够得到目标视频片段的位置. Rodriguez 等 [26] 提出了一种基于引导注意力机制的 "去提案化" 时序定位方法 ProFree, 此方法利用一个动态滤波器将自然语言句子的语义信息迁移到视频视觉域. 一项新的损失函数也被引入来引导模型将注意力集中在与自然语言句子语义最相关的视频片段区域. Hahn 等 [15] 提出了使用门控注意力机制进行时序定位的端到端框架 TripNet, 它利用强化学习的思想通过学习如何智能地跳过视频区域来有效地在长视频中定位与句子语义相关的片段, 在此网络中利用门控注意力机制加强视频和自然语言句子之间的语义对齐关系. 上述这些方法都只对自然语言句子和细粒度单元视频片段之间的多模态交互进行建模, 而时序定位的预测则是在不同尺度的视频片段上进行的. 因此, 上述方法忽略了自然语言句子与不同粒度或尺度的视频内容之间复杂的匹配关系, 这将影响模型的时序定位精度.

3 基于注意力回归的时序定位

为了解决自然语言句子在视频中的时序定位问题, 我们不妨先从人类面对这一问题时可能采取的策略展开思考. 从视频本身出发, 人们往往习惯于从头到尾完整地浏览视频, 然后再决定给定的自然语言句子究竟是对应于视频中的哪部分片段, 而不是像前述的基于两阶段框架的方法那样, 逐段地去评估句子与各视频片段之间的语义关联性. 在完整浏览视频再作决策的情况下, 视频的上下文信息是完整的, 而且避免了密集地扫描视频、切割视频片段造成的计算量过大的问题, 这种做法使得对于视频的理解更加全面和高效. 从句子的角度出发, 人们往往会更加注意自然语言句子中的一些关键的单词或者短语, 它们可以清晰地提供一些线索来识别目标视频片段. 通过关注与这些关键词相关的局部视频内容, 人们可以逐渐推理扩大区域, 最终定位出目标片段的位置. 因此, 在时序定位问题中, 我们应该更加关注这些句子的细节, 以便于更加准确地预测目标视频片段的位置.

基于以上考虑, 我们提出了一种端到端的基于注意力回归的时序定位模型 (attention based location regression, ABLR). ABLR 模型超越了现有的多模态匹配方法, 可以在一次性扫描视频序列之后直接输出所定位的视频片段的时间坐标. 具体而言, ABLR 首先利用两个双向长短时记忆循环神经网络 LSTM [27] 分别对视频片段和单词序列进行编码, 其中 LSTM 网络的每个输出单元特征都将蕴含序列中的上下文信息. 基于编码后的输出单元特征, 我们设计了一种多模态共同注意力机制来学习视频和句子之间的注意力关系. 其中视频注意力包含了不同视频片段和给定的自然语言句子之间的语义关联, 因此它可以反映全局视频结构. 句子注意力突出显示了单词序列中的关键细节, 因此它可以为后续的时间位置预测提供清晰的指导. 最后, 我们提出了一种基于注意力的位置预测网络, 它能够从前述所预测的注意力输出中回归得到目标视频片段的时间坐标. 联合优化保持上下文的特征编码网络、多模

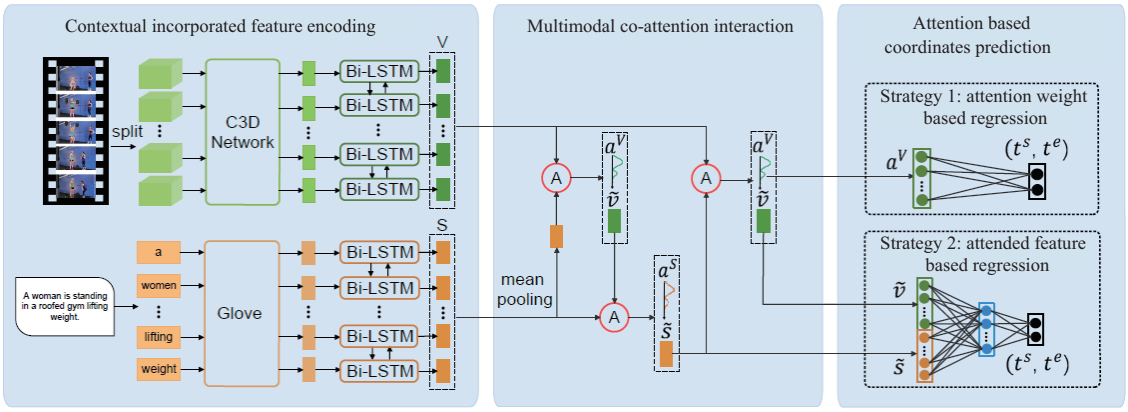

图 3　(网络版彩图) 基于注意力回归的时序定位模型 ABLR 框架图, ABLR 共包含 3 个部分: (1) 保持上下文的特征编码网络; (2) 多模态共同注意力交互网络; (3) 基于注意力的位置预测网络. 此外, 我们设计了两种位置回归策略来预测最终的目标视频片段位置, 即基于注意力权重的回归和基于注意力加权特征的回归.

Figure 3　(Color online) The framework of the attention based location regression (ABLR) model, which consists of three parts: (1) contextual incorporated feature encoding; (2) multimodal co-attention interaction; (3) attention based coordinates prediction. Besides, we design two different regression strategies to predict the target segment locations, i.e., attention weight based regression, and attended feature based regression.

态共同注意力交互网络和基于注意力的位置预测网络, 我们所提出的 ABLR 模型能够高效准确地定位自然语言句子在视频中的位置.

本节首先定义研究问题, 接着介绍我们提出的基于注意力回归的时序定位模型 ABLR 以及它的学习过程.

3.1　问题定义

假设视频 V 与一组自然语言句子在语义 $\{(S,\tau^s,\tau^e)\}$ 相关联, 即任意一个自然语言句子 S 都在视频 V 中能够找到一个视频片段, 其起始和结束时间点分别为 τ^s 和 τ^e, 它所对应的视频内容正是 S 所描述的. 那么给定视频 V 和自然语言句子 S, 我们的任务是预测 S 所描述的视频片段对应的时间坐标 (τ^s,τ^e).

3.2　ABLR 模型

如图 3 所示, 基于注意力回归的时序定位模型 ABLR 具有 3 个模块, 即保持上下文的特征编码网络, 多模态共同注意力交互网络和基于注意力的位置预测网络. 从输入的视频和自然语言句子到输出的时间坐标, 我们所提出的 ABLR 模型是可以通过端到端的方式联合优化的. 下面将对 ABLR 的 3 个模块进行具体的介绍.

保持上下文的特征编码网络. 由于时序定位旨在全局视频中确定一个片段以匹配自然语言句子提供的语义信息, 因此特定的视频内容和全局视频上下文信息都是不可忽视的关键元素. 一些现有方法声称它们已将上下文信息整合到视频片段中. 但是, 它们通过一些硬编码方式执行此操作: 即融合全局视频的平均特征 [10] 或以预定义的尺度 [11,12] 在一定程度上扩展候选视频片段的边界. 实际上, 含混地融合全局视频特征将会对时序定位的过程带来干扰, 以预定义尺度限制视频片段的扩展将无法维持视频中的长序关系. 为了克服这些问题, 我们采用双向长短时记忆循环神经网络 (bi-directional LSTM) 进行视频编码.

对于每个视频 V, 首先按时间顺序将其平均分成 M 个单位视频片段 $\{v_1,\ldots,v_j,\ldots,v_M\}$. 然后, 我们应用广泛使用的三维卷积网络对这些视频片段进行编码:

$$\boldsymbol{x}_j = \mathrm{3DCNN}(v_j), \tag{1}$$

其中 $\boldsymbol{x}_j$ 是视频片段 v_j 的三维卷积特征. 然后, 我们使用双向 LSTM 生成包含上下文信息的视频片段特征. 形式化的定义如下:

$$\boldsymbol{h}_j^{\mathrm{f}}, \boldsymbol{c}_j^{\mathrm{f}} = \mathrm{LSTM}^{\mathrm{f}}(\boldsymbol{x}_j, \boldsymbol{h}_{j-1}^{\mathrm{f}}, \boldsymbol{c}_{j-1}^{\mathrm{f}}), \quad \boldsymbol{h}_j^{\mathrm{b}}, \boldsymbol{c}_j^{\mathrm{b}} = \mathrm{LSTM}^{\mathrm{b}}(\boldsymbol{x}_j, \boldsymbol{h}_{j+1}^{\mathrm{b}}, \boldsymbol{c}_{j+1}^{\mathrm{b}}). \tag{2}$$

双向 LSTM 由两个独立的子网络组成, 其中前向 $\mathrm{LSTM}^{\mathrm{f}}$ 从视频的开头向结尾移动, 而后向 $\mathrm{LSTM}^{\mathrm{b}}$ 从视频的结尾向开头移动. 视频片段 v_j 的最终特征 $\boldsymbol{v}_j$ 是在位置 j 处通过级联前向和后向 LSTM 的输出计算的:

$$\boldsymbol{v}_j = f\big(\boldsymbol{W}_v(\boldsymbol{h}_j^{\mathrm{f}} \| \boldsymbol{h}_j^{\mathrm{b}}) + \boldsymbol{b}_v\big). \tag{3}$$

$f(\cdot)$ 在本文中表示激活函数, 即 ReLU. 依据长短时记忆循环神经网络 LSTM 的特性, 相邻的视频片段将相互影响, 因此, 每个视频片段的表征都将会包含视频的上下文信息. 最终我们将视频表示为$\boldsymbol{V} = [\boldsymbol{v}_1,\ldots,\boldsymbol{v}_j,\ldots,\boldsymbol{v}_M] \in \mathbb{R}^{h_v \times M}$, 其中任意一列都表示某一视频片段的 h_v 维特征.

为了探索自然语言句子中的细节信息, 我们也使用双向 LSTM 来编码句子, 就像文献 [28] 一样. 与一般的 LSTM 直接将句子整体编码为一个特征向量不同, 双向 LSTM 将句子 S 中的原始单词特征序列作为输入, 针对任一单词 s_j, 它都将输出一个包含句子上下文信息的单词特征向量 $\boldsymbol{s}_j$. 编码句子特征的双向 LSTM 其精确定义类似于式 (2), 其中输入的原始单词特征是 300 维的 Glove [29] 词嵌入向量. 最终, 自然语言句子的特征形式表示为 $\boldsymbol{S} = [\boldsymbol{s}_1,\ldots,\boldsymbol{s}_j,\ldots,\boldsymbol{s}_N] \in \mathbb{R}^{h_s \times N}$. 下文将句子和视频双向 LSTM 的输出维度统一为 h, 即 $h_v = h_s = h$.

多模态共同注意力交互网络. 在以前的方法中, 视觉注意力机制主要集中在识别不同视觉任务中的 "看哪里" 的问题. 这样, 它也能够应用于自然语言句子在视频中时序定位问题, 因为这一任务恰好是要在句子语义的指导下准确定位视频中需要注意的片段. 此外, 明确 "听谁的" 这一问题, 即找到句子中为时序定位提供关键线索的词语或短语也十分重要, 因为突出这些关键词将为时序定位提供更为明确的目标.

基于以上考虑, 我们通过引入多模态共同注意力机制 [30] 建立视频与句子之间的关联关系. 在这种注意力机制中, 我们依次在生成视频和句子的注意力向量之间进行交替. 该机制的计算过程包含 3 个步骤: (1) 根据初始句子特征指导生成视频注意力向量; (2) 根据由注意力加权后的视频特征指导生成句子注意力向量; (3) 根据由注意力加权后的句子特征指导生成最终的视频注意力向量. 具体来说, 注意力函数 $\tilde{\boldsymbol{z}} = A(\boldsymbol{Z};\boldsymbol{g})$ 以视频 (或句子) 特征 $\boldsymbol{Z}$ 和指导向量 $\boldsymbol{g}$ 作为输入, 输出视频 (或句子) 的注意力权重 $\boldsymbol{a}^z$ 以及由注意力加权后的特征 $\tilde{\boldsymbol{z}}$. 具体定义如下:

$$\boldsymbol{a}^z = \mathrm{softmax}\Big(\boldsymbol{u}_a^{\mathrm{T}} \tanh\big(\boldsymbol{U}_z\boldsymbol{Z} + (\boldsymbol{U}_g\boldsymbol{g})\boldsymbol{1}^{\mathrm{T}} + \boldsymbol{b}_a\boldsymbol{1}^{\mathrm{T}}\big)\Big), \quad \tilde{\boldsymbol{z}} = \sum a_j^z \boldsymbol{z}_j, \tag{4}$$

其中 $\boldsymbol{U}_g, \boldsymbol{U}_z \in \mathbb{R}^{k\times h}$, $\boldsymbol{b}_a, \boldsymbol{u}_a \in \mathbb{R}^k$ 为注意力函数的参数, $\boldsymbol{1}$ 是所有元素均为 1 的向量. 如图 3 的中间部分所示, 在交替生成注意力的第 1 步中, $\boldsymbol{Z}$ 以 $\boldsymbol{V}$ 赋值, $\boldsymbol{g}$ 是通过平均所有单词特征获得的全局句子特征. 在第 2 步中, $\boldsymbol{Z}$ 以 $\boldsymbol{S}$ 赋值, $\boldsymbol{g}$ 是由第 1 步输出的注意力权重加权后的视频特征. 在最后一步中, $\boldsymbol{Z}$ 以 $\boldsymbol{S}$ 赋值, $\boldsymbol{g}$ 是由第 2 步输出的注意力权重加权后的句子特征.

通过以上过程, 可以将视频注意力权重 $\boldsymbol{a}^V$ 视为时间维度的一种特征, 其中单个元素 a_j^V 表示第 j 个视频片段和自然语言句子之间的语义关联. 因此, 整个视频注意力权重将反映视频的全局时间结

构, 同时通过注意力权重加权求和得到全局视频特征中, 与句子语义相关的特定视频内容的影响将被增大. 同时, 我们还基于视频内容来计算句子中各词语的注意力权重, 因此句子中具有较大注意力权重的关键词和短语将为定位过程提供更强有力的指导.

基于注意力的位置预测网络. 给定视频注意力权重向量 $\boldsymbol{a}^V$, 定位目标视频片段位置的一种可能的方法是选择并合并一些注意力权重值更高的视频片段 [5, 31]. 但是, 这种做法属于一种后处理策略, 而且将位置预测模块与前述网络模块割离开来, 将会造成整体网络陷入次优解. 为了避免这个问题, 我们提出了一种新颖的基于注意力的位置预测网络, 该网络直接探索了注意力权重向量与目标视频片段位置之间的相关性. 具体而言, 基于注意力的位置预测网络将视频注意力权重向量或注意力加权后的特征作为输入, 由此回归出目标视频片段的时间坐标. 另外, 我们还根据此种思路设计了两种位置回归策略: 一种是基于注意力权重的回归, 另一种是基于注意力加权特征的回归.

基于注意力权重的回归将视频注意力权重 $\boldsymbol{a}^V$ 作为时间维度的一种特征, 并通过回归网络得到目标视频片段的时间坐标:

$$\boldsymbol{t} = (t^s, t^e) = f\big(\boldsymbol{W}_{aw}(\boldsymbol{a}^V)^{\mathrm{T}} + \boldsymbol{b}_{aw}\big), \tag{5}$$

其中 $\boldsymbol{W}_{aw} \in \mathbb{R}^{2\times M}$ 和 $\boldsymbol{b}_{aw} \in \mathbb{R}^2$ 是回归参数, f 是激活函数 ReLu. (t^s, t^e) 分别是所预测视频片段的开始时间和结束时间, 值域在 0∼1 之间, 它们是由视频片段的绝对位置除以全局视频长度后得到的归一化时间.

基于注意力加权特征的回归首先将注意力加权后的视频特征 $\tilde{\boldsymbol{v}}$ 和句子特征 $\tilde{\boldsymbol{s}}$ 融合为多模态特征向量:

$$\boldsymbol{f} = f\big(\boldsymbol{W}_f(\tilde{\boldsymbol{v}} \| \tilde{\boldsymbol{s}}) + \boldsymbol{b}_f\big), \tag{6}$$

其中 $\boldsymbol{W}_f \in \mathbb{R}^{h\times 2h}$ 和 $\boldsymbol{b}_f \in \mathbb{R}^h$ 是用于特征融合操作的参数. 然后, 我们将 $\boldsymbol{f}$ 作为回归网络的输入预测目标片段的时间位置坐标:

$$\boldsymbol{t} = (t^s, t^e) = f(\boldsymbol{W}_{af}\boldsymbol{f} + \boldsymbol{b}_{af}), \tag{7}$$

其中 $\boldsymbol{W}_{af} \in \mathbb{R}^{2\times h}$ 和 $\boldsymbol{b}_{af} \in \mathbb{R}^2$ 是基于注意力加权特征的回归网络的参数. 如式 (5) 和 (7) 所示, 我们以单层完全连接操作的形式定义了时间坐标回归网络. 实际上, 在我们的实验设置中, 输入数据和输出时间坐标之间有两个完全连接层.

上述两种回归策略都将全局视频内容视为时序定位的基础, 但分别适用于不同的视频场景. 我们将在实验部分更为具体地讨论它们的影响.

3.3 ABLR 模型的学习

首先, 将 ABLR 的训练集表示为 $\{(V_i, \tau_i, S_i, \tau_i^{\mathrm{s}}, \tau_i^{\mathrm{e}})\}_{i=1}^K$, V_i 是长度为 τ_i 的视频, S_i 是描述特定视频片段的自然语言句子, 该片段在视频 V_i 中的位置由起点和终点 $(\tau_i^{\mathrm{s}}, \tau_i^{\mathrm{e}})$ 标识. 请注意, 一个视频可以对应于多句自然语言句子, 分别描述其不同的视频片段, 因此不同的训练样本可能对应于同一视频.

我们将句子 S_i 所对应的视频片段的开始和结束时间位置归一化为 $\tilde{\boldsymbol{t}}_i = (\tilde{t}_i^s, \tilde{t}_i^e) = (\tau_i^s/\tau_i, \tau_i^e/\tau_i)$, 这被视为位置预测的目标. 根据真实的和预测的时间坐标对 $(\tilde{\boldsymbol{t}}_i, \boldsymbol{t}_i)$, 我们设计了一个注意力回归损失函数以优化时间坐标预测, 它以平滑 L1 函数 $R(x)$ [32] 的形式定义:

$$L_{\mathrm{reg}} = \sum_{i=1}^{K}[R(\tilde{t}_i^s - t_i^s) + R(\tilde{t}_i^e - t_i^e)]. \tag{8}$$

通常, 在其他计算机视觉任务中, 注意力的学习过程都没有明确的注意力监督信息作为指导. 但是在 ABLR 中, 我们需要直接从视频注意力权重中回归时间坐标. 因此, ABLR 所学习到的视频注意

力权重的准确性将对随后的位置回归产生很大的影响. 基于上述考虑, 我们设计了一个注意力校准损失函数, 它指导多模态共同注意力交互网络生成与真实视频片段时间坐标相一致的视频注意力权重:

$$L_{\text{cal}} = -\sum_{i=1}^{K} \frac{\sum_{j=1}^{M} m_{i,j} \log(a_j^{V_i})}{\sum_{j=1}^{M} m_{i,j}}, \tag{9}$$

其中 $m_{i,j}$ 表示视频 V_i 中的第 j 个视频片段是否在句子 S_i 所对应的目标视频片段的真实时间窗口 $(\tau_i^{\text{s}}, \tau_i^{\text{e}})$ 之内, 若是, 则 $m_{i,j} = 1$, 否则 $m_{i,j} = 0$. 显然, 注意力校准损失会促使真实窗口内的视频片段具有更高的注意力权重. 对于句子注意力权重, 由于缺少监督信息, 我们只能像一般情况那样从总体模型中隐含地学习它.

ABLR 定位系统的总体损失函数包括上述注意力回归和注意力校准损失:

$$L = \alpha L_{\text{reg}} + \beta L_{\text{cal}}, \tag{10}$$

其中, α 和 β 是控制两个损失项之间权重的超参, 它们的值由网格搜索确定.

基于上述总体损失函数, 我们所提出的 ABLR 模型就可以从特征编码网络到坐标预测网络进行端到端的训练. 在测试阶段, 我们将视频和自然语言句子输入到 ABLR 模型中, 然后输出句子所对应的视频片段的归一化时间坐标, 将其乘以视频时间长度来获得其绝对位置.

4 基于语义条件动态归一化的时序定位

针对自然语言句子在视频中的时序定位问题, 目前已有方法主要侧重于在语义上匹配句子和单个视频片段, 而忽略了句子的在关联理解视频各部分内容中的重要指导作用. 例如, 图 4 中显示的目标视频片段主要包含两个不同的活动: “女人走过房间” 和 “女人在沙发上读书”. 如果不参考句子, 这两个不同的活动就很难作为一个整体关联起来. 但是, 给定的自然语言句子清楚地表明 “女人拿着书穿过房间去沙发上看书”. 理解了句子的语义含义, 人们便可以轻松地将两个活动关联在一起, 从而精准地确定句子所描述的整体活动在视频中的位置. 因此, 如何利用句子的语义信息来指导相关视频内容随着时间的关联和耦合, 对句子在视频中的时序定位这一任务至关重要. 其次, 视频中包含的活动通常具有不同的视觉特征和不同的时间尺度. 因此, 自然语言句子在指导关联视频内部内容时也应该考虑到不同的视觉特征和尺度信息, 并能够随着视频内容的变化而动态发展.

基于上述考虑, 我们提出了一种新颖的 SCDN 机制, 该机制利用句子语义信息来指导微调多层时序卷积网络中的特征归一化过程. 具体而言, SCDN 通过参考句子语义表征控制产生卷积网络中特征归一化操作的缩放和平移参数, 从而控制视频在各时序卷积层中的特征分布. 这样, 每一次特征调整都将促进下一层时序卷积更好地将句子语义所指示的视频内容在时间上互相关联和耦合. 此外, 参照前述基于注意力回归的时序定位模型, SCDN 在处理各层卷积特征图的不同位置时, 还将利用所对应的卷积特征指导产生强调不同句子细节 (词语) 的注意力加权句子表征, 使得基于句子语义的归一化过程能够动态地随着时序视频流变化发展, 进而让不同尺度、不同位置下的视频内容更好地与自然语言句子的语义对齐. 将 SCDN 应用到多层时序卷积网络中, 我们提出的模型可以自然地建模句子和视频之间的交互行为, 从而为自然语言句子在视频中的时序定位任务提供了一种新颖而有效的架构.

4.1 语义条件动态归一化机制指导下的时序卷积网络模型

给定一个视频 V 和一个自然语言句子 S, 自然语言句子在视频中的时序定位任务旨在视频中确定一个片段, 定位其开始和结束位置, 使得该片段所示的视频内容与句子的语义含义相对应. 为了

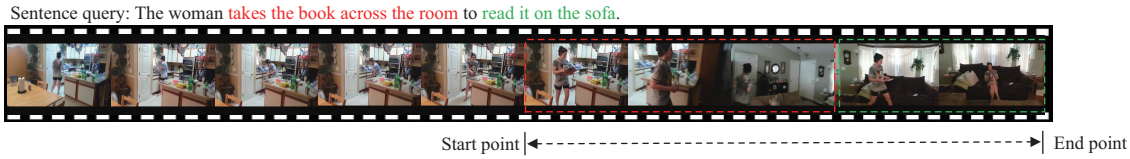

图 4 (网络版彩图) 本图展示了自然语言句子在视频中的时序定位问题. 我们所提出的语义条件动态归一化机制利用句子语义信息来指导时序卷积网络中的特征归一化过程, 因而可以促进句子语义所指示的视频内容 (例如图中用红色和绿色框所标注的视频片段) 在时间上互相关联和耦合.

Figure 4 (Color online) The illustration of temporal sentence grounding in videos task. Our proposed SCDN relies on the sentence to guide the temporal convolution operations, which can thereby temporally correlate and compose the various sentence-related activities (highlighted in red and green) for more accurate grounding results.

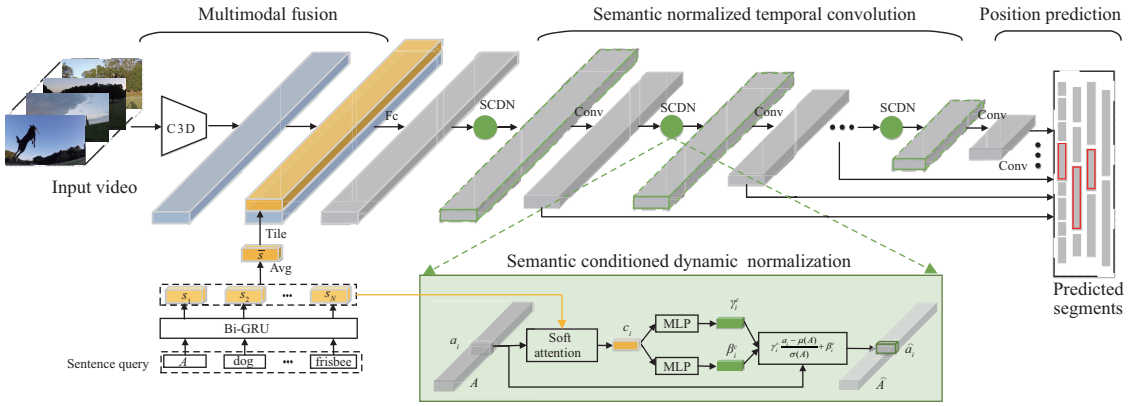

图 5 (网络版彩图) 语义条件动态归一化机制指导下的时序卷积网络模型框架图. 此模型一共包含 3 个子模块: 多模态融合网络、基于语义条件动态归一化的时序卷积网络和位置预测网络. 这 3 个子模块紧密耦合, 可以联合优化.

Figure 5 (Color online) An overview of the proposed temporal convolution architecture with semantic conditioned dynamic normalization, which includes three fully-coupled modules that can be globally optimized: multimodal fusion, semantic normalized temporal convolution, and position prediction.

解决这一任务, 首先我们将视频表示为一些单位视频片段 (video clip, 时长为 1 s 的小视频) 的序列 $\boldsymbol{V}=\{\boldsymbol{v}_t\}_{t=1}^{T}$, 句子则表示为一些单词的序列 $\boldsymbol{S}=\{\boldsymbol{s}_n\}_{n=1}^{N}$.

我们提出了一种新颖的模型来处理自然语言句子在视频中的时序定位任务, 模型的框架图如图 5 所示. 具体而言, 我们所提出的模型由 3 部分组成, 即多模态融合网络, 基于语义条件动态归一化的时序卷积网络和位置预测网络. 这 3 个部分是完全耦合在一起的, 因此整体模型可以以端到端的方式进行训练.

4.1.1 多模态融合网络

自然语言句子在视频中的时序定位任务需要充分理解句子和视频的含义, 并建立二者的语义关联. 因此, 我们首先让每个单位视频片段都与句子的语义信息进行交互融合:

$$\boldsymbol{f}_t=\mathrm{ReLU}\left(\boldsymbol{W}^{\mathrm{f}}\left(\boldsymbol{v}_t\|\bar{\boldsymbol{s}}\right)+\boldsymbol{b}^{\mathrm{f}}\right), \tag{11}$$

其中, $\boldsymbol{W}^{\mathrm{f}}$ 和 $\boldsymbol{b}^{\mathrm{f}}$ 是可学习的参数. $\bar{\boldsymbol{s}}$ 表示全局的句子特征编码, 它可以通过简单地平均 $\boldsymbol{S}$ 中的各单词特征获得. 通过这种多模态融合产生的特征 $\boldsymbol{F}=\{\boldsymbol{f}_t\}_{t=1}^{T}\in\mathcal{R}^{T\times d_f}$ 能够比较细粒度地捕获句子和视频片段之间的语义交互. 下面即将介绍的基于语义条件动态归一化的时序卷积网络将把这些多模态融合

特征在时序上相互关联, 从而把与句子语义相关的视频内容耦合到一起, 为时序定位提供较为清晰的时间边界信息.

4.1.2 基于语义条件动态归一化的时序卷积网络

如前所述, 不同句子所描述的视频内容可以具有多种时间尺度, 占据不同的时间位置. 因此, 我们应该利用多模态融合表征 $\boldsymbol{F}$ 在多种时间尺度下全面刻画视频中活动内容的多样性. 受到高效的视频动作检测网络的启发 [2,33], 我们将一种多层时序卷积网络作用于多模态融合表征上, 从而层序地产生具有多种时间尺度的特征图, 以涵盖视频中持续时长多种多样的视觉动作. 此外, 为了充分利用句子的语义指导作用, 我们提出了 SCDN 机制, 该机制依靠句子语义来指导时序卷积网络中的特征归一化过程, 以便随着时间的推移更好地关联和耦合与句子相关的视频内容. 下文首先回顾基础的时序卷积网络, 然后将详细介绍所提出的 SCDN 机制.

时序卷积网络. 本文的标准时序卷积运算表示为 $\mathrm{Conv}(\theta_k, \theta_s, d_h)$, 其中 θ_k, θ_s 和 d_h 分别表示卷积核的大小、步幅大小和滤波器的数量. 同时, 非线性激活函数 (例如 ReLU) 将级联在每一个卷积运算之后, 以构成基本的时序卷积层. 通过分别将 θ_k 设置为 3 并将 θ_s 设置为 2, 每个卷积层能够将输入特征图的时间维度减半, 同时扩展特征图内特征单元的视觉感受野范围. 在多模态融合表征 $\boldsymbol{F}$ 上堆叠多个时序卷积层, 我们便能构建起一个具有层级结构的时序卷积网络, 在此网络中任意一层特征图中的单个特征单元都将对应于原始视频中的一个特定视频片段. 为了简化后续定义, 我们将第 k 个时序卷积层的输出特征图表示为 $\boldsymbol{A}_k = \{\boldsymbol{a}_{k,i}\}_{i=1}^{T_k} \in \mathbb{R}^{T_k \times d_h}$, 其中 $T_k = T_{k-1}/2$ 是时间维度, 而 $\boldsymbol{a}_{k,i} \in \mathbb{R}^{d_h}$ 表示第 k 层特征图中的第 i 个特征单元.

语义条件动态归一化机制. 在视频中定位特定活动或事件, 除了视频各部分的具体内容外, 它们之间的关联关系也起着十分重要的作用. 针对自然语言句子在视频中的时序定位这一任务, 所提供的输入句子更是为视频内容之间的相关性提供了丰富的语义指示, 它能使邻近的视频内容随时间相互关联耦合, 组合成一个完整的事件. 基于以上考虑, 我们提出了一种新颖的语义条件动态归一化机制 SCDN, 该机制能够依靠句子语义信息动态地控制每个时序卷积层中的特征归一化过程.

具体而言, 如图 6(b) 所示, 给定句子的特征表示 $\boldsymbol{S} = \{\boldsymbol{s}_n\}_{n=1}^{N}$ 和某一时序卷积层输出的特征图 $\boldsymbol{A} = \{\boldsymbol{a}_i\}$ (我们在此省略了层号), 针对此特征图中的每一个特征单元 $\boldsymbol{a}_i$, 我们以它为指导信息, 采用注意力机制生成该特征单元相对于句子中每个词 $\boldsymbol{s}_n$ 的注意力权重 ρ_i^n, 并得到注意力加权的句子表征 $\boldsymbol{c}_i$:

$$\rho_i^n = \mathrm{softmax}\big(\boldsymbol{w}^{\mathrm{T}} \tanh\left(\boldsymbol{W}^{\mathrm{s}} \boldsymbol{s}_n + \boldsymbol{W}^{a} \boldsymbol{a}_i + \boldsymbol{b}\right)\big), \quad \boldsymbol{c}_i = \sum_{n=1}^{N} \rho_i^n \boldsymbol{s}_n, \tag{12}$$

其中 $\boldsymbol{w}$, $\boldsymbol{W}^{\mathrm{s}}$, $\boldsymbol{W}^{a}$ 和 $\boldsymbol{b}$ 是可学习的参数. 然后, 我们使用两个具有 `tanh` 激活函数的全连接 (fully-connected layer, FC) 层来分别生成两个控制向量 $\gamma_i^c \in \mathbb{R}^{d_h}$ 和 $\beta_i^c \in \mathbb{R}^{d_h}$:

$$\gamma_i^c = \tanh(\boldsymbol{W}^{\gamma} \boldsymbol{c}_i + \boldsymbol{b}^{\gamma}), \quad \beta_i^c = \tanh(\boldsymbol{W}^{\beta} \boldsymbol{c}_i + \boldsymbol{b}^{\beta}), \tag{13}$$

其中 $\boldsymbol{W}^{\gamma}$, $\boldsymbol{b}^{\gamma}$, $\boldsymbol{W}^{\beta}$ 和 $\boldsymbol{b}^{\beta}$ 是可学习的参数. 最后, 根据生成的控制向量 γ_i^c 和 β_i^c, 我们将特征单元 $\boldsymbol{a}_i$ 调整为

$$\hat{\boldsymbol{a}}_i = \gamma_i^c \cdot \frac{\boldsymbol{a}_i - \mu(\boldsymbol{A})}{\sigma(\boldsymbol{A})} + \beta_i^c. \tag{14}$$

采用 SCDN 机制, 我们能够基于句子特征精细地控制时序卷积网络中各层特征图的缩放和平移过程. 这样, 每一层特征图的更新都将充分受到句子语义信息的干预和指导, 进一步促使后续的时序卷积

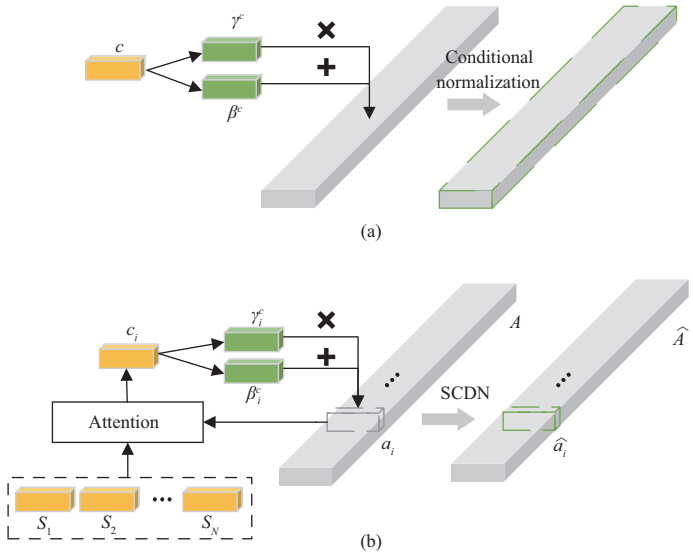

图 6 (网络版彩图) 条件归一化机制和我们所提出的语义条件动态归一化机制比较图

Figure 6 (Color online) The comparison between conditional normalization and our proposed semantic conditioned dynamic normalization. (a) Conditional normalization; (b) semantic conditioned dynamic normalization

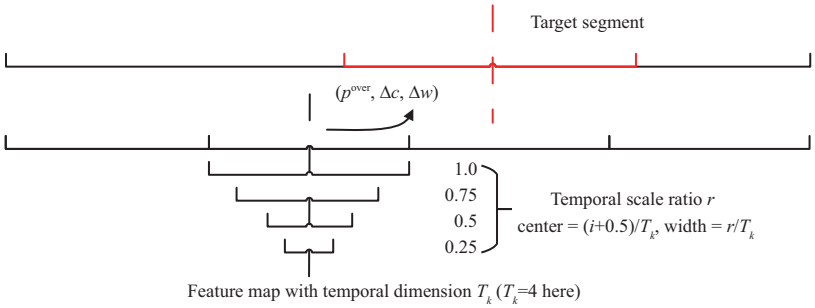

图 7 (网络版彩图) 位置预测网络辅助说明图

Figure 7 (Color online) The supplementary illustration of the position prediction procedure

层随着时间的推移紧密地耦合与句子相关联的视频内容. 如图 5 的中间部分所示, 将所提出的 SCDN 应用于每两个相邻的时序卷积层之间, 便得到了基于语义条件动态归一化的时序卷积网络.

讨论. 如图 6 所示, 我们提出的 SCDN 与现有的条件批/实例规范化 [34, 35] 有所不同. 在条件批/实例规范化中, 相同的 γ^c 和 β^c 参数将在整个批次/实例中应用. 而如等式 (12)~(14) 所示, 我们的 SCDN 通过参考不同的视频内容, 动态地增强句子中不同单词的影响并生成对应的注意力加权的句子语义特征, 从而为每个特征图中的不同特征单元生成了不同的 γ^c 和 β^c 参数. 这种动态归一化使每个特征单元都可以与每个单词进行语义交互, 为时序定位任务收集细致而有效的定位线索. 因此, 句子和视频语义关联可以随着时间的推移更好地对齐, 以支持更精确的边界预测.

4.1.3 位置预测网络

与视频动作检测网络 [2, 33] 类似, 在我们的多层时序卷积网络中, 低层和高层的卷积特征图分别用于定位时间跨度较短和较长的事件. 如图 7 所示, 假定我们将一个视频的总长归一化为 1, 那么对

于一个具有 T_k 个特征单元的特征图, 其每个特征单元在视频中对应区域的时间跨度则为 $1/T_k$. 我们对每一个特征单元所对应视频区域的时间跨度都进行一定程度的缩放, 缩放的比例因子为 $r \in R = \{0.25, 0.5, 0.75, 1.0\}$, 那么, 对于此特征图中的第 i 个特征单元, 它将在视频中对应若干具有不同时间跨度的区域, 这些区域的长度为 r/T_k, 中心点为 $(i+0.5)/T_k$. 对于整个特征图, 则总共有 $T_k \cdot |R|$ 个时间区域, 它们中的每一个都对应于一个候选视频片段.

我们在时序卷积特征图上附加一组卷积运算, 以预测目标视频片段的位置. 具体来说, 针对每个候选片段, 我们将利用卷积运算预测一个向量 $p = (p^{\text{over}}, \Delta c, \Delta w)$, 其中 p^{over} 是候选片段与真实目标片段之间的匹配度, Δc 和 Δw 是候选片段相对于真实目标片段的中心和长度偏移. 假设候选片段的中心和长度分别为 μ^c 和 μ^w, 那么根据预测向量, 我们可以计算出由此候选片段调整得到的预测视频片段的中心 ϕ^c 和长度 ϕ^w:

$$\phi^c = \mu^c + \alpha^c \cdot \mu^w \cdot \triangle c, \quad \phi^w = \mu^w \cdot \exp(\alpha^w \cdot \triangle w), \tag{15}$$

其中 α^c 和 α^w 均用于控制位置偏移量 Δc 和 Δw 的影响程度, 使得位置预测过程更为稳定, 我们根据经验将其设置为 0.1. 这样, 对于一个具有 T_k 个特征单元的特征图, 我们可以获得一个预测片段集合 $\Phi_k = \{(p_j^{\text{over}}, \phi_j^c, \phi_j^w)\}_{j=1}^{T_k \cdot |R|}$. 那么, 全局视频的总体预测片段集合表示为 $\mathbf{\Phi} = \{\Phi_k\}_{k=1}^{K}$, 其中 K 是特征图的数量. 最终我们所预测的与句子语义相关的视频片段将在 $\mathbf{\Phi}$ 中选出.

4.2 模型训练与目标片段预测

模型训练. SCDN 模型的训练样本包含 3 个元素: 输入视频、自然语言句子和与句子相关联的目标片段位置. 如果某一候选视频片段与目标视频片段的 "交集除并集系数" (intersection-over-union, IoU) 大于 0.5, 则将其视为正样本片段. 模型的训练损失函数包括匹配度预测损失函数 L_{over} 和位置预测损失函数 L_{loc}. L_{over} 以交叉熵的形式实现, 其定义为

$$L_{\text{over}} = \sum_{z \in \{\text{pos,neg}\}} -\frac{1}{N_z} \sum_{i}^{N_z} g_i^{\text{over}} \log(p_i^{\text{over}}) + (1 - g_i^{\text{over}}) \log(1 - p_i^{\text{over}}), \tag{16}$$

其中, g^{over} 是候选片段和目标视频片段之间的 IoU 值, 而 p^{over} 是模型所预测的二者之间的匹配度. 位置损失函数 L_{loc} 用于衡量正样本片段与目标视频片段之间的位置预测偏差, 以平滑 L_1 损失函数 [32] 的形式实现:

$$L_{\text{loc}} = \frac{1}{N_{\text{pos}}} \sum_{i}^{N_{\text{pos}}} SL_1(g_i^c - \phi_i^c) + SL_1(g_i^w - \phi_i^w), \tag{17}$$

其中 g^c 和 g^w 分别是目标视频片段的的中心位置和长度.

整体模型是在综合考虑这两种损失函数的情况下进行训练的, 其中 λ 和 η 是平衡这两项损失函数贡献的超参数:

$$L_{\text{all}} = \lambda L_{\text{over}} + \eta L_{\text{loc}}. \tag{18}$$

目标片段预测. 模型训练完成后, 我们将测试视频和自然语言句子输入到模型中, 获得预测片段集合 $\mathbf{\Phi}$. 将 $\mathbf{\Phi}$ 内所有的候选片段按其对应的 p^{over} 分数进行排序并使用非最大抑制 (non-maximum suppression, NMS) 进行处理, 排序第一的候选片段即为模型所预测的与句子语义相关的目标视频片段.

5 结合粗粒度语义动态归一化和细粒度边界调整的时序定位

前述语义条件动态归一化机制指导下的时序卷积网络虽然可以在多层时序卷积架构中自然地引入多尺度的视频候选片段, 并在一次前馈中计算出各候选片段相对于目标片段的偏差, 但是这些多尺度的视频候选片段仍然拘泥于一些预定义的尺寸 (如 4.1.3 小节所示), 而自然语言句子所描述的视频活动其时间、位置和跨度通常是不受约束的. 即使先前的工作能够通过位置预测网络对候选片段的时间边界进行调整, 这些调整仍然局限于预定义候选片段的邻近区域, 导致时序定位结果的时间边界仍不够灵活, 是一种粗粒度的 "片段" 级别的预测. 因此, 如何有效地使用句子信息来调整和细化所预测的视频片段的时间边界是十分重要的.

基于上述考虑, 我们在语义条件动态归一化机制指导下的时序卷积网络的基础上, 提出了一个细粒度时间边界调整模块, 它作用在时序卷积网络的底层, 以句子语义信息为指导, 输出每个单位视频片段 (video clip) 位于目标视频片段的起始、终止和中间区域的概率. 在时序卷积网络预测了目标片段位置以后, 我们将利用前述所预测的起始、终止和中间区域的概率来微调目标片段的时间边界位置, 以达到更加精准的时序定位结果. 在 3 个公开数据集上的实验结果表明, 所提出的细粒度时间边界调整模块的确能够进一步提升语义条件动态归一化机制指导下的时序卷积网络的定位准确率, 从而验证了其有效性.

5.1 细粒度时间边界调整模块

5.1.1 网络模块结构

如图 8 所示, 所提出的细粒度时间边界调整模块分别利用 3 个子网来预测每一个单位视频片段位于目标片段的起始、终止和中间区域的概率. 这 3 个子网均作用于多层时序卷积网络的第一个特征图 $\boldsymbol{A}_1$, 并且子网结构相同, 具有不同的网络参数. 每个子网由两层具有语义条件动态归一化机制的时序卷积层组成, 以预测中间区域概率的子网为例, 其第 1 层时序卷积层的卷积核大小为 3, 步幅大小为 1, 过滤器数量为 d_h, 简写为 $\mathrm{Conv}(3,1,d_h)$, 类似地, 第 2 层时序卷积层配置为 $\mathrm{Conv}(3,1,1)$. 通过这样的 3 个子网, 我们将获得 3 个概率序列 $P^{\mathrm{s}}=\{p_k^{\mathrm{s}}\}_{k=1}^{T}$, $P^{\mathrm{e}}=\{p_k^{\mathrm{e}}\}_{k=1}^{T}$, $P^{\mathrm{m}}=\{p_k^{\mathrm{m}}\}_{k=1}^{T}$, 其中 p_k^{s}, p_k^{e} 和 p_k^{m} 分别表示第 k 个单位视频片段位于目标片段的起始点、终止和中间点区域的概率.

上述所预测的 3 个概率序列将进一步用于调整和完善语义条件动态归一化机制指导下的时序卷积网络所预测的视频片段的时间边界. 另外, 由于细粒度时间边界调整模块与时序卷积网络共享底部的时序卷积层, 细粒度时间边界调整模块的学习过程也将促进整体多层时序卷积网络的学习过程.

5.1.2 训练损失函数

由于细粒度时间边界调整模块会为每个单位视频片段预测 3 个概率序列 P^{s}, P^{e} 和 P^{m}, 因此为了方便训练, 我们首先要为每个单位视频片段生成相应的起始点/终止点/中间区域标签. 给定训练集中一个自然语言句子所对应视频片段的位置 $(t^{\mathrm{s}},t^{\mathrm{e}})$, 如果时间戳 $t^{\mathrm{s}}(t^{\mathrm{e}})$ 恰好位于一个单位视频片段中, 则该单位视频片段相对于这一自然语言句子的起始点 (终止点) 概率标签被设置为 1, 否则为 0. 同时, 落在 $(t^{\mathrm{s}},t^{\mathrm{e}})$ 内部的单位视频片段, 其中间区域概率的标签被设置为 1, 否则为 0. 这样, 我们便获得起始点/终止点/中间区域概率标签的序列, 表示为 $G^{\mathrm{s}}=\{g_k^{\mathrm{s}}\}_{k=1}^{T}$, $G^{\mathrm{e}}=\{g_k^{\mathrm{e}}\}_{k=1}^{T}$ 和 $G^{\mathrm{m}}=\{g_k^{\mathrm{m}}\}_{k=1}^{T}$.

给定预测的概率序列和标签序列, 我们以交叉熵的形式计算各 (G^x,P^x) 对之间的损失函数, 其中 $x\in\{\mathrm{s},\mathrm{e},\mathrm{m}\}$, 并相应地获得起始点损失函数 $L_{\mathrm{clip}}^{\mathrm{s}}$, 终止点损失函数 $L_{\mathrm{clip}}^{\mathrm{e}}$ 和中间区域损失函数 $L_{\mathrm{clip}}^{\mathrm{m}}$.

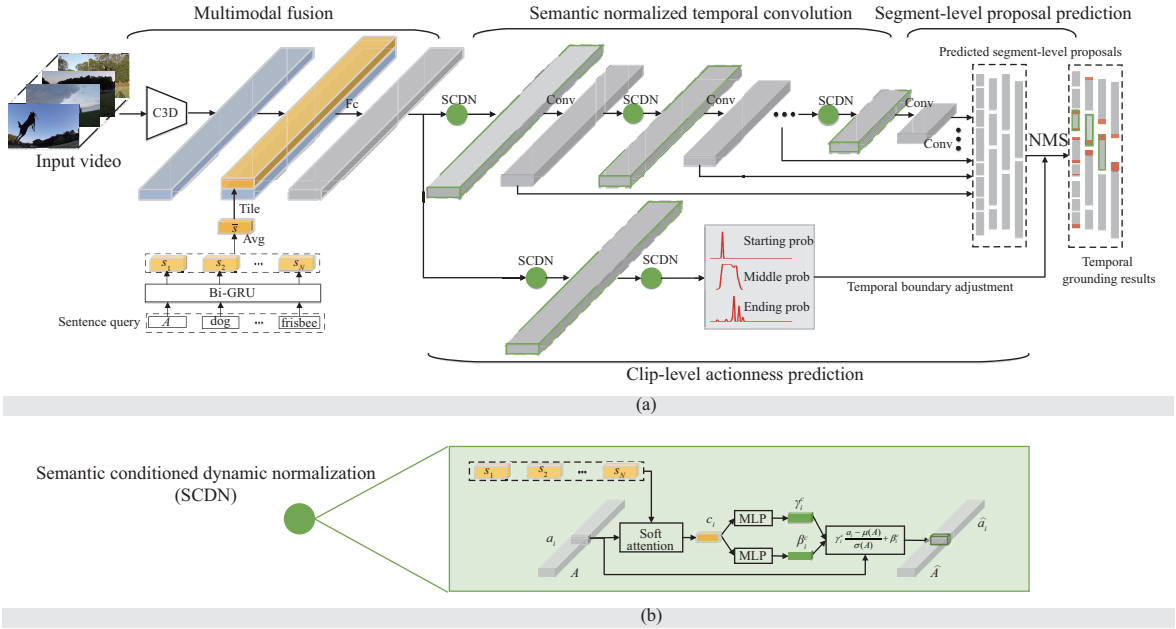

图 8 (网络版彩图) (a) 带有细粒度时间边界调整模块的语义条件动态归一化机制指导下的时序卷积网络模型框架图. 除了语义条件动态归一化机制指导下的时序卷积网络模型中的 3 个子模块 (多模态融合网络, 基于语义条件动态归一化的时序卷积网络和位置预测网络) 外, 本节还额外提出了一个细粒度时间边界调整模块, 用于将位置预测网络所输出的候选视频片段的位置进行微调. (b) 语义条件动态归一化机制示意图.

Figure 8 (Color online) (a) An overview of the temporal convolutional architecture with semantic conditioned dynamic normalization and fine-grained temporal boundary adjustment. Besides the three modules (multimodal fusion, semantic normalized temporal convolution, and segment-level proposal prediction) proposed in the previous sections, we further propose a clip-level actionness prediction module in this section, which will adjust the temporal boundaries of the predicted candidate segments in a fine-grained clip-level. (b) The illustration of the semantic conditioned dynamic normalization.

最终, 我们所提出的细粒度时间边界调整模块其总体损失函数为上述 3 项的加权总和:

$$L_{\text{clip}} = \lambda_{\text{s}} L_{\text{clip}}^{\text{s}} + \lambda_{\text{e}} L_{\text{clip}}^{\text{e}} + \lambda_{\text{m}} L_{\text{clip}}^{\text{m}}, \tag{19}$$

其中 λ_{s}, λ_{e} 和 λ_{m} 是权衡不同损失函数占比的系数.

细粒度时间边界调整模块将与语义条件动态归一化机制指导下的时序卷积网络联合优化, 我们将 4.1 小节时序卷积网络的总体损失函数标记为 L_{seg}, 那么既具有时序卷积网络, 又带有时间边界调整模块的总体模型损失函数为

$$L_{\text{all}} = L_{\text{seg}} + L_{\text{clip}}. \tag{20}$$

5.2 细粒度时间边界调整模块指导下的目标片段位置预测

如前所述, 语义条件动态归一化机制指导下的时序卷积网络所预测的候选视频片段集合 $\boldsymbol{\Phi}$ 可以提供具有不同尺度和位置的时序定位结果, 但这些候选片段仍然拘泥于一些预定义的窗口, 从而使得时序定位的结果不够准确灵活. 而细粒度时间边界调整模块预测了每个单位时间视频片段是否位于目标片段的起始点/终止点/中间区域位置的概率, 这一模块对目标片段边界位置的预测更为敏感. 因此, 我们提出在目标片段位置的预测阶段利用细粒度时间边界调整模块来微调时序卷积网络所预测的候选视频片段的时间边界. 为了帮助理解目标片段预测过程, 本小节中的一些符号汇总在表 1 中.

表 1　本小节中使用到的符号定义

Table 1　The definition of symbols in this subsection

Symbol	Definition
$\boldsymbol{\Phi}$	Segment-level proposal set
Φ_i	One segment-level proposal from $\boldsymbol{\Phi}$
$\phi_i^{\mathrm{s}}/\phi_i^{\mathrm{e}}$	Starting/ending point of Φ_i
p_i^{over}	Overlap score between Φ_i and groundtruth
P^{s}	Starting probability sequence
P^{e}	Ending probability sequence
P^{m}	Middle probability sequence
ξ_i^{s}	Relaxation interval around the starting point of Φ_i
ξ_i^{e}	Relaxation interval around the ending point of Φ_i
$\phi_i^{\mathrm{s,max}}$	Point of maximal starting probability in ξ_i^{s}
$\phi_i^{\mathrm{e,max}}$	Point of maximal ending probability in ξ_i^{e}
Φ_i^*	Proposal after adjusting the boundary of Φ_i
$\phi_i^{\mathrm{s},*}/\phi_i^{\mathrm{e},*}$	Starting/ending point of Φ_i^*
$p_i^{\mathrm{over},*}$	Overlap score between Φ_i^* and groundtruth
$\boldsymbol{\Phi}^{\mathrm{b}}$	Proposal set after boundary adjustment and NMS
$\boldsymbol{\Phi}^{\mathrm{tag}}$	TAG-based proposal set with respect to P^{m}
$\boldsymbol{\Phi}^{\mathrm{b,tag}}$	Proposal set after intergrating $\boldsymbol{\Phi}^{\mathrm{b}}$ and $\boldsymbol{\Phi}^{\mathrm{tag}}$

基于起始点和终止点概率的时间边界调整. 具体来说, 假设 $\Phi_i = (\phi_i^{\mathrm{s}}, \phi_i^{\mathrm{e}}, p_i^{\mathrm{over}})$ 是语义条件动态归一化机制指导下的时序卷积网络所预测的候选视频片段集合 $\boldsymbol{\Phi} = \{\Phi_i\}_{i=1}^{M}$ 中的一个候选片段. 这里 ϕ_i^{s} 和 ϕ_i^{e} 表示候选片段 Φ_i 的起点和终点, 而 p_i^{over} 是候选片段与真实目标片段之间的匹配度, M 是 $\boldsymbol{\Phi}$ 中的候选片段数目. 我们将分别根据预测的起始和终止点概率序列 P^{s} 和 P^{e} 调整 ϕ_i^{s} 和 ϕ_i^{e}. 首先, 我们为候选片段定义两个松弛区间:

$$\xi_i^{\mathrm{s}} = [\phi_i^{\mathrm{s}} - d_i/\varepsilon, \phi_i^{\mathrm{s}} + d_i/\varepsilon], \quad \xi_i^{\mathrm{e}} = [\phi_i^{\mathrm{e}} - d_i/\varepsilon, \phi_i^{\mathrm{e}} + d_i/\varepsilon], \tag{21}$$

其中 $d_i = \phi_i^{\mathrm{e}} - \phi_i^{\mathrm{s}}$ 是候选片段 Φ_i 的长度, 而 ε 是控制松弛区间大小的比例因子. 我们进一步将在 ξ_i^{s} 中具有最大起始概率的时间点定义为 $\phi_i^{\mathrm{s,max}}$, 并在 P^{s} 中找到对应的最大起始概率, 记为 $p_i^{\mathrm{s,max}}$. 类似地, 在区间 ξ_i^{e} 中的最大终止概率和相应的时间点分别被定义为 $p_i^{\mathrm{e,max}}$ 和 $\phi_i^{\mathrm{e,max}}$. 如果 $p_i^{\mathrm{s,max}}$ 或者 $p_i^{\mathrm{e,max}}$ 大于预定义的阈值 ν, 候选片段 Φ_i 的起始点或终止点将调整如下:

$$\phi_i^{\mathrm{s},*} = \delta\phi_i^{\mathrm{s}} + (1-\delta)\phi_i^{\mathrm{s,max}}, \quad \phi_i^{\mathrm{e},*} = \delta\phi_i^{\mathrm{e}} + (1-\delta)\phi_i^{\mathrm{e,max}}, \tag{22}$$

其中 ν 和 δ 都位于 $[0,1]$ 区间内, 并将针对不同的数据集进行设置. 另外, 如果 ϕ_i^{s} 或 ϕ_i^{e} 被更新, 则候选片段与真实目标片段之间的匹配度将被放大: $p_i^{\mathrm{over},*} = \tau p_i^{\mathrm{over}}$ $(\tau > 1)$. 这样, 根据上述操作, 我们将获得调整了起始点和终止点的新的候选片段集合 $\Phi_i^* = (\phi_i^{\mathrm{s},*}, \phi_i^{\mathrm{e},*}, p_i^{\mathrm{over},*})$.

非极大值抑制处理. 在执行上述时间边界调整之后, 我们进一步根据更新的匹配度分数 $p^{\mathrm{over},*}$ 对候选片段集合 Φ_i^* 进行非极大值抑制 (NMS) 排序处理, 并把边界调整和非极大值抑制处理完毕后的视频片段集合记为 $\boldsymbol{\Phi}^{\mathrm{b}}$.

基于中间区域概率的候选片段调整. 中间区域概率序列 P^{m} 指示每个单位视频片段位于目标片段中的概率, 我们首先遵循时序行动合并网络 (temporal actionness grouping, TAG [36]), 将具有较高中间区域概率的连续单位视频片段合并到一起作为候选片段, 记为基于 TAG 的候选片段集合 $\boldsymbol{\Phi}^{\mathrm{tag}} = \{\Phi_i^{\mathrm{tag}}\}_{i=1}^{M^{\mathrm{tag}}}$, 其中 M^{tag} 是 $\boldsymbol{\Phi}^{\mathrm{tag}}$ 中的片段总数. 由于中间区域概率序列的边界敏感性, 我们将结合 $\boldsymbol{\Phi}^{\mathrm{b}}$ 和 $\boldsymbol{\Phi}^{\mathrm{tag}}$ 来产生更精准全面的时序定位结果. 具体来说, 对于 $\boldsymbol{\Phi}^{\mathrm{b}}$ 中的每个候选片段 Φ_i^{b}, 我们计算其相对于 $\boldsymbol{\Phi}^{\mathrm{tag}}$ 中所有片段的 IoU 值, 如果最大 IoU 值大于 0.95, 我们将用 $\boldsymbol{\Phi}^{\mathrm{tag}}$ 中与 Φ_i^{b} 具有最大 IoU 值的视频片段替换 Φ_i^{b}. 至此, 我们将更新后的片段集合记为 $\boldsymbol{\Phi}^{b,\mathrm{tag}}$, 该集合同时考虑了时序卷积网络和细粒度时间边界调整模块的预测结果, 将作为整体模型的最终输出.

6 实验

本节将对上述提出的基于注意力回归的时序定位模型, 语义条件动态归一化机制指导下的时序卷积网络模型, 和辅以细粒度时间边界调整模块的时序卷积网络模型进行实验分析, 并和目前已有方法进行比较. 同时, 我们还将进行一些消融实验以验证各子模块设计的有效性. 另外, 可视化的实验结果和模型效率也将进行展示, 以获得对所提出模型更进一步的了解.

6.1 数据集介绍

我们利用 3 个公开数据集 Charades-STA, TACoS 和 ActivityNet Captions 对模型进行评估.

Charades-STA [11]: 数据集主要包含描述室内活动的视频. Charades-STA 中的视频平均长 30 s, 每个视频都具有多个自然语言句子描述, 且这些自然语言句子都能够对应于视频中的特定片段. 在官方的数据集分割中, Charades-STA 共有 5338 个视频和 12408 个 (自然语言句子、视频片段) 数据对用于训练, 1334 个视频和 3720 个 (自然语言句子、视频片段) 数据对用于测试.

TACoS [16]: 带有文字注释的烹饪场景数据集 (TACoS) 包含大量的视频描述句子以及它们所描述的视频片段在视频中的位置标注. 该数据集总共有 127 个视频, 这些视频描绘了执行烹饪任务的对象、环境和操作步骤, 总共约有 17000 组 (句子、视频片段) 对. 与已有方法 [11] 一样, 我们将数据集的 50% 用于训练, 将 25% 用于验证, 将 25% 用于测试.

ActivityNet Captions [37]: 此数据集包含 2 万个视频和 1×10^5 个自然语言句子描述, 每个句子描述都对应于某一视频中一个特定视频片段, 这些视频片段的位置也已被明确标注. ActivityNet Captions 数据集的视频数量比 TACoS 数据集高了两个量级, 因此其视频多样性更大, 更具有挑战性. 由于此数据集的测试集被隐去作为 ActivityNet-Challenge 比赛的测试数据, 在我们的实验中, 将其公开的训练集用于模型训练, 验证集用于测试.

6.2 评价指标

遵循现有方法 [11~13, 21, 24], 我们采用 "R@n, IoU@m" 作为模型评价指标. 具体来说, 针对某一个句子, 若被模型排序在前 n 的候选片段中有至少一个与真实目标视频片段的 IoU 大于 m, 我们记该句子为正确预测的句子, "R@n, IoU@m" 则定义为正确预测的句子占所有测试句子的百分比. 值得注意的是, 前述我们所提出的 ABLR 模型直接预测一个视频片段作为时序定位的结果而未对若干候选片段进行排序, 因而对于该模型我们只能报告其 "R@1, IoU@m" 的结果.

6.3 模型实现细节

基于注意力回归的时序定位模型. 根据视频时长分布, 我们将 Charades-STA, ActivityNet Captions 和 TACoS 数据集中的视频分别平均分为 64, 128 和 256 个单位视频片段. 此外, 视频和句子双向 LSTM 的隐藏状态维度 h 被设置为 256, 丢失率 (dropout) 为 0.5. 对于多模态共同注意力交互网络, 我们将隐藏状态大小设置为 $k = 256$, 因此 $\boldsymbol{U}_g, \boldsymbol{U}_z \in \mathbb{R}^{256\times256}$, $\boldsymbol{u}_a, \boldsymbol{b}_a \in \mathbb{R}^{256}$. 我们通过网格搜索, 将超参数 α 和 β 分别设置为 1 和 5. 模型训练的样本批量大小 batch size 被设置为 64. 在使用 ActivityNet Captions 数据集进行训练时, 模型的学习率被设置为 0.001; 在使用 Charades-STA 和 TACoS 数据集进行训练时, 模型的学习率被设置为 0.0001.

语义条件动态归一化机制指导下的时序卷积网络模型. 该模型中所提取的每个特征单元都代表一个时长为 1 s 的视频片段, 根据不同数据集中的视频时长分布, 对于 ActivityNet Captions 和 TACoS 数据集, 输入视频的最大长度设置为 1024 s, 即多层时序卷积网络一次计算最多可以处理 1024 个单位视频片段; 对于 Charades-STA 数据集, 输入视频的最大长度设置为 64 s, 超过最大长度的视频则会被截断, 而较短的视频会被填充零向量. 对于多层时序卷积网络的设计, 用于处理 Charades-STA 数据集中视频的网络共有 6 层, 每一层分别包含 {32, 16, 8, 4, 2, 1} 个特征单元; 用于处理 TACoS 数据集中视频的网络共有 6 层, 每一层分别包含 {512, 256, 128, 64, 32, 16} 个特征单元; 用于处理 ActivityNet Captions 数据集中视频的网络共有 8 层, 每一层分别包含 {512, 256, 128, 64, 32, 16, 8, 4} 个特征单元. 所有的第 1 层卷积特征图都不会用于位置预测, 因为其特征单元的时间感受野太小, 往往无法包含一个视频事件. 为了节省模型内存占用量, 我们仅在直接用于位置预测的特征图上嵌入 SCDN 机制. 对于句子编码, 我们首先使用 Glove [29] 词向量对每个单词进行编码, 然后使用双向门控循环神经网络 (gated recurrent unit, GRU) [38] 对单词特征序列进行进一步的处理. 这样, 句子中的单词最终以其对应的 GRU 隐藏状态表示. 双向 GRU、多模态融合网络的隐层维度 d_f 以及时序卷积运算的过滤器个数 d_h 均设置为 512. 两项损失函数的系数 λ 和 η 分别设置为 100 和 10. 我们使用 Adam [39] 优化器对模型进行训练, 学习率设置为 0.0001, 样本批量大小 batch size 设置为 16.

辅以细粒度时间边界调整模块的语义条件动态归一化机制指导下的时序卷积网络模型. 语义条件动态归一化机制指导下的时序卷积网络模型其实现细节与前述相同. 对于细粒度时间边界调整模块, 其起始点、终止点和中间区域损失函数的系数 $\{\lambda_s, \lambda_e, \lambda_m\}$ 在同一数据集上保持一致, 对于 Charades-STA, TACoS 和 ActivityNet Captions 数据集, 它们都分别设置为 10, 100 和 5. 在目标片段位置预测阶段, 超参数 $\{\varepsilon, \nu, \delta\}$ 的搜索空间根据其在边界调整中的作用凭经验确定, 最终它们在 Charades-STA, TACoS 和 ActivityNet Captions 数据集上分别被设置为 {6.0, 0.9, 0.7}, {5.0, 0.7, 0.8} 和 {5.0, 0.9, 0.1}. τ 在 3 个数据集上均设置为 1.15.

遵循现有方法, 我们统一采用三维卷积网络编码视频, C3D [40] 网络用于提取 TACoS 和 ActivityNet Captions 数据集中的视频特征, I3D [41] 网络用于提取 Charades-STA 数据集中的视频特征.

6.4 基线方法介绍

我们将所提出的模型与以下现有方法进行比较. CTRL [11]: 跨模态时序回归定位器 (cross-modal temporal regression localizer); ACRN [12]: 基于注意力的跨模态检索网络 (attentive cross-modal retrieval network); TGN [13]: 时序定位网络 (temporal ground-net); MCF [21]: 多模态循环融合网络 (multimodal circulant fusion); ACL [20]: 基于事件活动概念检测的定位器 (activity concepts based localizer); SAP [23]: 基于视觉概念挖掘的定位方法; Xu et al. [22]: 基于句子重构的视频片段定位方法; MAN [24]:

表 2 在 TACoS 和 Charades-STA 数据集上的模型性能比较 (%)

Table 2 The performance comparison on the TACoS and the Charades-STA datasets (%)

Method	TACoS				Charades-STA			
	R@1, IoU@0.3	R@1, IoU@0.5	R@5, IoU@0.3	R@5, IoU@0.5	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7
CTRL (C3D) [11]	18.32	13.30	36.69	25.42	23.63	8.89	58.92	29.52
MCF (C3D) [21]	18.64	12.53	37.13	24.73	–	–	–	–
ACRN (C3D) [12]	19.52	14.62	34.97	24.88	–	–	–	–
SAP (VGG) [23]	–	18.24	–	28.11	27.42	13.36	66.37	38.15
ACL (C3D) [20]	24.17	20.01	**42.15**	30.66	30.48	12.20	64.84	35.13
TGN (C3D) [13]	21.77	18.90	39.06	31.02	–	–	–	–
Xu et al. (C3D) [22]	–	–	–	–	35.60	15.80	79.40	45.40
MAN (I3D) [24]	–	–	–	–	46.53	22.72	**86.23**	53.72
TripNet (C3D) [15]	23.95	19.17	–	–	38.29	16.07	–	–
ProFree (I3D) [26]	–	–	–	–	52.02	33.74	–	–
ABLR-aw (*)	18.92	9.33	–	–	42.21	21.42	–	–
ABLR-af (*)	19.54	9.38	–	–	40.78	20.20	–	–
SCDN (*)	26.11	21.17	40.16	32.18	54.44	33.43	74.43	58.08
SCDN_{CAP} (*)	**27.64**	**23.27**	40.06	**33.49**	**54.92**	**34.26**	76.50	**60.02**

片段对齐网络 (moment alignment network); TripNet [15]: 使用门控注意力机制进行时序定位的端到端强化学习框架; ProFree [26]: 基于引导注意力机制的 “去提案化” 时序定位方法.

此外, 我们将前述基于注意力回归的时序定位模型用ABLR 表示, 而ABLR-aw 则表示该模型采用基于注意力权重的回归, ABLR-af 则表示该模型采用基于注意力加权特征的回归. 前述语义条件动态归一化机制指导下的时序卷积网络模型用SCDN 表示, 而在此基础上额外带有细粒度时间边界调整模块的模型用SCDN_{CAP} 表示.

6.5 模型性能比较

表 2 和 3 报告了我们的模型在上述 3 个公开数据集上与现有方法之间的性能比较, 其中各基线方法的结果都直接引用它们原始论文中报告的数据. 总体而言, 我们所提出的 SCDN_{CAP} 在 Charades-STA 和 TACoS 数据集上都具有最高的时序定位准确性, 而 ABLR 模型则在 ActivityNet Captions 数据集上取得了最佳性能, 这些实验结果都证明了我们所提出的模型的优越性.

具体而言, 在没有细粒度时间边界调整模块的情况下, SCDN 模型也已经优于其他基线方法, 而引入细粒度时间边界调整模块还能进一步增加 SCDN 的时序定位准确度. 这是由于更细粒度的起始点、终止点和中间区域概率序列可以进一步用于调整和完善时序卷积网络预测的候选视频片段的时间边界, 从而获得更精确的时序定位结果. 尽管 SCDN 在 Charades-STA 数据集上 R@5, IoU@0.5 的性能较低, 但这主要是由该数据集的标注数据特性导致的. 例如, 在 Charades-STA 数据集中所标注的目标视频片段长度平均为 10 s, 而视频平均时长仅为 30 s, 在此情况下随机选择一个候选片段也可以达到不错的时序定位准确度. 因此为保证模型评估的可靠性, 我们应该在较高 IoU 的阈值条件下计算定位准确度.

对于 ABLR 方法, 我们发现其在 ActivityNet Captions 数据集上取得了最佳性能, 但是在 Charades-

表 3　在 ActivityNet Captions 数据集上的模型性能比较 (%)

Table 3　The performance comparison on the ActivityNet Captions dataset (%)

Method	R@1, IoU@0.3	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.3	R@5, IoU@0.5	R@5, IoU@0.7
TGN (INP) [13]	45.51	28.47	–	57.32	43.33	–
Xu et al. (C3D) [22]	45.30	27.70	13.60	75.70	59.20	38.30
TripNet (C3D) [15]	48.42	32.19	13.93	–	–	–
ProFree (I3D) [26]	51.28	33.04	19.26	–	–	–
ABLR-aw (C3D)	**57.00**	**39.12**	**20.54**	–	–	–
ABLR-af (C3D)	53.56	35.52	18.11	–	–	–
SCDN (C3D)	54.80	36.75	19.86	77.29	64.99	41.53
SCDN_{CAP} (C3D)	55.25	36.90	20.28	**78.79**	**66.84**	**42.92**

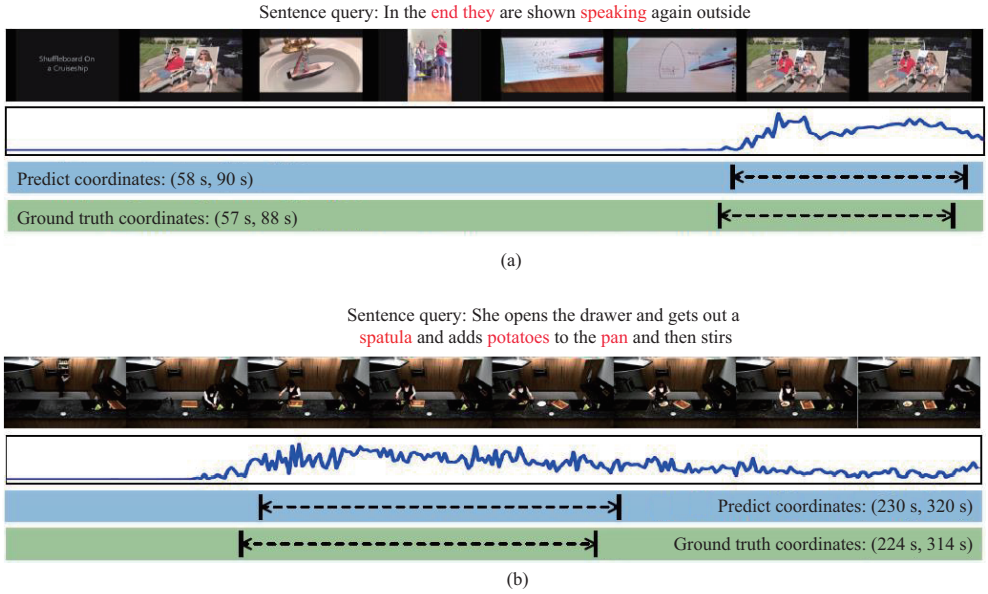

图 9　(网络版彩图) ABLR 模型时序定位的可视化结果. 蓝色背景的长条表示模型所预测的视频片段, 绿色背景的长条表示真实的与句子语义相对应的目标视频片段. ABLR 模型所学习到的视频注意力权重用蓝色波浪线表示, 自然语言句子中注意力权重较高的单词用红色字体突出显示.

Figure 9　(Color online) The qualitative results of the ABLR model. The bars with blue background indicate the predicted segments of ABLR, while the bars with green background indicate the groundtruth segments. The video attention weights learned by the ABLR model are illustrated with blue wavy lines, and the words with higher attention weights are highlighted with red. (a) A temporal sentence grounding example from ActivityNet captions; (b) a temporal sentence grounding example from TACoS.

STA 和 TACoS 数据集上的性能相比 SCDN 和 SCDN_{CAP} 较低. 我们认为这种现象也是由数据集的特性引起的. 如图 9 所示, ActivityNet Captions 数据集中的视频包含各种场景和活动, 即使在单个视频中, 不同片段之间的差异也很明显. 与之相反的是, TACoS 数据集中的所有视频共享一个共同的场景, 只有一些视觉上占比较小的烹饪对象 (例如砧板、刀、操作者等) 有细微改变, 这些视觉上难以区分的视频片段会造成相对较为扁平的注意力曲线. 在这种情况下, ABLR 模型可以有效地定位句子所对应的视频片段的近似位置, 在较低的 IoU 值下获得较好的 R@1 值, 但在较高的 IoU 阈值条件下对确定

精确的片段边界则较为含混. 此外, 与具有 2 万个视频的 ActivityNet Captions 数据集相比, TACoS 数据集只有 127 个视频, TACoS 有限的数据集规模限制了定位方法的学习与训练, 这也将影响模型性能. 此外, 在 ActivityNet Captions 数据集上, 基于注意力权重的回归 ABLR-aw 的表现优于基于注意力加权特征的回归 ABLR-af, 而在 TACoS 上, ABLR-af 则稍显优势. 由于 ABLR-aw 直接从视频注意力权重中回归时间坐标, 因此 TACoS 中平坦的注意力曲线使 ABLR-aw 难以准确预测目标片段位置. 与 ABLR-aw 相比, ABLR-af 还融合了视频内容信息, 这进一步增强了位置回归网络输入的可分辨性, 从而带来了更好的结果.

6.6 可视化实验结果展示

我们在图 9 中提供了一些 ABLR 模型的时序定位可视化结果. 具体来说, 如图 9(a) 所示, 描述人们说话行为的场景在视频中出现了两次, 而自然语言句子指出了 “最后 (End)” 的这一重要提示, 因此后者是一个正确的位置. 但是, 基线方法很难作出正确的决策, 因为它们独立地处理每个候选片段, 并且不探究局部视频内容与全局视频环境之间的相对关系. 相比之下, 我们的 ABLR 模型不仅通过双向 LSTM 灵活地维护了上下文信息, 而且还通过视频注意力权重刻画视频的全局时间结构. 因此, ABLR 的时序定位决策更加准确和全面. 此外, 通过多模态共同注意力机制, 我们还可以从图 9 中看到, 模型学习到的句子注意力权重突出了自然语言句子中的一些关键词, 例如某些对象、动作甚至具有时间含义的单词. 这些突出显示的词为时序定位提供了清晰的线索, 并增强了时序定位模型的可解释性.

为了进一步展示我们提出的细粒度时间边界调整模块的优势, 我们在图 10 中可视化了边界调整前后的一些时序定位结果, 以及模型所预测的起始点/中间区域/终止点概率值. 在这些示例中, 时间边界调整后的粉红色所示的预测视频片段比棕色的视频片段更接近真实的目标视频片段. 同时, 通过观察模型所预测的起始点/中间区域/终止点概率值序列, 我们发现概率值较高的区域确实对应于视频中特定活动的起始/发生/终止区域. 这样的结果表明这些细粒度的概率预测序列可以很好地捕获视频中的活动演变, 从而可以将时序卷积网络预测的候选片段的边界调整到更为精确的位置. 值得注意的是, 边界调整前由 SCDN 模型预测的视频片段位置也已经十分接近真实的目标片段位置, 这也验证了我们所提出的语义条件动态归一化机制指导下的时序卷积网络模型的能力.

6.7 消融实验分析

6.7.1 ABLR 消融实验分析

为了验证我们的 ABLR 设计的有效性, 我们还使用以下不同的配置来构造一些 ABLR 的消融模型.

- ABLP. 通过对视频注意力权重后处理来进行时序定位. 此模型中略去了原 ABLR 模型中的基于注意力的位置预测网络. 最终所预测的视频片段的时间坐标是通过对视频注意力权重进行时间边界细化 [5] 确定的.
- ABLR-reg-aw/ABLR-reg-af. 在此模型中, 我们仅使用注意力回归损失函数进行训练, 省略了注意力校准损失函数. 另外, “aw” 表示采用基于注意力权重的回归策略, 而 “af” 表示采用基于注意力加权特征的回归策略.
- ABLR-c3d-aw/ABLR-c3d-af. 在此模型中, 我们在视频编码中删除了双向 LSTM, 因而视频片段仅由三维卷积特征表示, 而没有融合上下文信息.
- ABLR-stv-aw/ABLR-stv-af. 在此模型中, 用于句子编码的双向 LSTM 被 Skip-thought [42] 句子嵌入表征所替代. 因此, 每个句子描述都由单个特征向量表示, 句子中的关键词信息将不会被隐含学习,

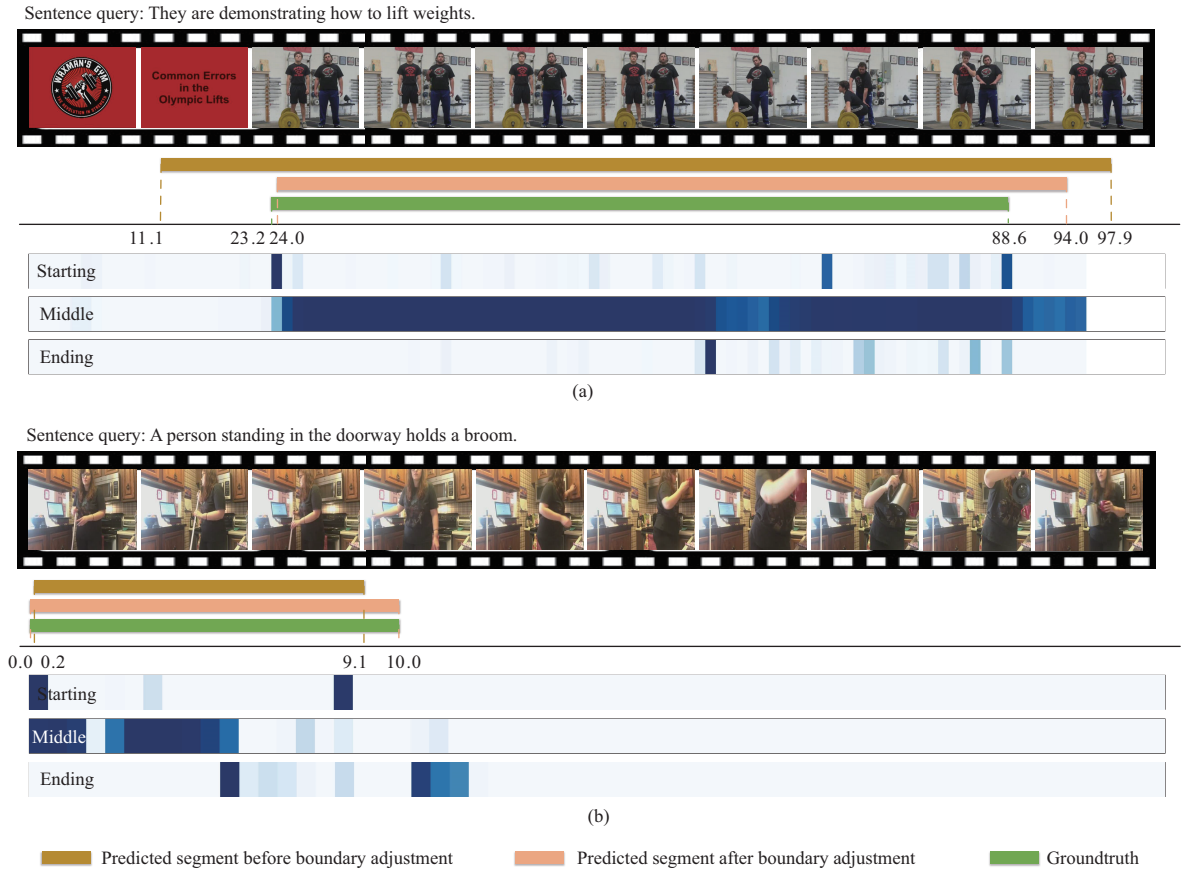

图 10 (网络版彩图) 带有细粒度时间边界调整模块的时序定位模型的可视化实验结果. 绿色长条标注了与给定的自然语言句子语义相关的目标视频片段的位置, 棕色和粉色长条分别标注了时间边界调整前后的模型时序定位结果 (即, 边界调整前的结果由 **SCDN** 模型输出, 边界调整后的结果由 $\mathbf{SCDN_{CAP}}$ 输出). 每幅子图最底端的 **3** 个长条形热度图显示了细粒度时间边界调整模块所预测的起始点、中间区域和终止点概率序列. 颜色越深, 代表概率值越高. **(a)** 可视化结果 **1**; **(b)** 可视化结果 **2**

Figure 10 (Color online) The qualitative results of the temporal convolutional architecture with semantic conditioned dynamic normalization and fine-grained temporal boundary adjustment. The green bars indicate the groundtruth video segments of the sentence queries, and the brown and pink bars indicate the model predicted segments before and after the temporal boundary adjustment (The segments indicated by brown bars are predicted by SCDN, and the segments indicated by pink bars are predicted by SCDN_{CAP}), respectively. The three heat maps at the bottom of each subgraph show the starting point, middle point, and ending point probability sequences predicted by the clip-level actionness prediction module. The darker the color, the higher the probability. (a) Qualitative example 1; (b) qualitative example 2

多模态共同注意力交互网络被降级, 只保留了视频注意力权重的计算.

- ABLR-aw/ABLR-af. 完整的 ABLR 模型.

我们在 ActivityNet Captions 数据集上对这些消融模型进行了实验, 结果如表 4 所示. 在基于注意力权重的回归和基于注意力加权特征的回归策略中, 完整的 ABLR 明显优于其他消融模型. 具体而言, ABLR-af 的 R@1, IoU@0.5 比 ABLR-c3d-af 有了 3.91% 的改进, 这证明了视频上下文信息在时间定位中的重要性. 另外, 我们将 ABLR-stv-aw 与 ABLR-aw 进行比较, 可以看到句子注意力的引入使得R@1, IoU@0.5 指标提高了 8.18%. 它表明, 考虑句子细节并在自然语言句子中强调关键词可以提高定位精度. 通过引入注意力校准损失函数, ABLR 的性能要优于 ABLR-reg, 这展示了注意力校准的优

表 4 ABLR 消融模型结果比较 (%)

Table 4 The performances of the ABLR ablation models (%)

Method	R@1, IoU@0.1	R@1, IoU@0.3	R@1, IoU@0.5
ABLP	68.04	45.03	23.04
ABLR-reg-af	69.22	51.84	33.53
ABLR-reg-aw	69.82	51.87	33.39
ABLR-c3d-af	68.10	51.13	31.61
ABLR-c3d-aw	55.78	39.43	18.35
ABLR-stv-af	68.41	50.83	31.53
ABLR-stv-aw	71.77	54.42	30.94
ABLR-af	70.91	53.56	35.52
ABLR-aw	**75.04**	**57.00**	**39.12**

表 5 SCDN 消融模型结果比较 (%)

Table 5 The performances of the SCDN ablation models (%)

Method	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7
w/o-SCDN	47.52	26.91	69.85	49.35
FC	46.33	25.94	68.96	49.81
MUL	49.08	28.77	72.68	51.02
SCM	53.07	31.41	71.71	54.57
SCDN	**54.44**	**33.43**	**74.43**	**58.08**

势, 即它鼓励多模态共同注意力机制学习与时间坐标一致的视频注意力权重, 并进一步为位置回归网络提供更准确的输入. 此外, ABLP 相比于 ABLR 有性能下降, 这证实了基于注意力的位置回归优于后处理策略.

6.7.2 SCDN 消融实验分析

为了验证所提出的 SCDN 机制的贡献, 我们也进行消融实验, 并使用以下 4 种设定重新训练模型.

- **w/o-SCDN.** 将模型中的 SCDN 机制用普通批处理规范化 [43] 替代;
- **FC.** 不在多层时序卷积网络中嵌入 SCDN, 而是用一个全连接层级联在每个时序卷积层之后, 将每个特征单元与全局句子表征 $\bar{\boldsymbol{s}}$ 融合.
- **MUL.** 不在多层时序卷积网络中嵌入 SCDN, 在每个时序卷积层之后将每个特征单元与全局句子表征 $\bar{\boldsymbol{s}}$ 相乘以完成不同尺度下的多模态特征融合.
- **SCM.** 使用全局句子表示 $\bar{\boldsymbol{s}}$ 来生成 γ^c 和 β^c 参数, 即语义条件归一化过程不会随不同视频内容而动态演变.

表 5 报告了上述不同设定下的 SCDN 消融模型在 Charades-STA 数据集上的性能. 如果不考虑 SCDN, w/o-SCDN模型的性能将大大降低, 这表明仅依靠多模态融合来建模视频和句子之间的语义关系不足以支撑句子在视频中的时序定位. 我们应该利用句子的语义加强指导时序卷积操作, 以便随着时间的推移更好地链接与句子相关的视频内容. 但是, 像 MUL 和 FC 这样在时序卷积体系中粗暴地引入句子信息并不能获得令人满意的结果. 由于句子信息已在多模态融合网络中与视频信息进行了交互, 因此模型中的卷积特征图已经是多模态表征. 将全局句子表征 $\bar{\boldsymbol{s}}$ 与时序卷积特征单元直接耦合

表 6 $\mathbf{SCDN_{CAP}}$ 消融模型结果比较 (%)

Table 6 The performances of the $\mathrm{SCDN_{CAP}}$ ablation models (%)

Method	TACoS				Charades-STA			
	R@1, IoU@0.3	R@1, IoU@0.5	R@5, IoU@0.3	R@5, IoU@0.5	R@1, IoU@0.5	R@1, IoU@0.7	R@5, IoU@0.5	R@5, IoU@0.7
SCDN	26.11	21.17	**40.16**	32.18	54.44	33.43	74.43	58.08
CAP	26.28	22.72	37.89	31.79	42.72	23.49	69.36	43.31
$\mathrm{SCDN_{CAP}}$(trainOnly)	27.49	22.33	39.92	33.30	54.36	33.81	76.21	59.42
$\mathrm{SCDN_{CAP}}$(TBA)	27.61	23.17	40.01	33.42	54.82	34.15	76.53	59.99
$\mathrm{SCDN_{CAP}}$(TAG)	27.53	22.94	39.99	33.38	54.52	33.80	76.23	59.42
$\mathrm{SCDN_{CAP}}$	**27.64**	**23.27**	40.06	**33.49**	**54.92**	**34.26**	**76.50**	**60.02**

可能会破坏视频的视觉相关性和时间依存关系, 这将对时序定位造成负面影响. 相比之下, 所提出的 SCDN 机制通过在句子指导下控制卷积特征图的归一化操作, 微调其数据分布情况, 这种操作比较谨慎且轻量级, 可以达到不错的效果.

6.7.3 $\mathbf{SCDN_{CAP}}$ 消融实验分析

本小节设计了如下几个消融模型来进一步验证和分析所提出的细粒度时间边界调整模块.

• **CAP.** 在此模型中, 我们仅使用细粒度时间边界调整模块来处理自然语言句子在视频中的时序定位任务. 具体来说, 我们首先根据所预测的中间区域概率序列生成基于 TAG 的候选片段集合 $\mathbf{\Phi}^{\mathrm{tag}}$. 然后, 我们使用起始和终止概率序列调整 $\mathbf{\Phi}^{\mathrm{tag}}$ 中候选片段的边界. 每个调整后的候选片段将获得 $p_i^{\mathrm{CAP}} = p_i^{s,\max} \times p_i^{e,\max} \times p_i^{\mathrm{tag}}$ 的匹配度得分, 其中 $p_i^{s,\max}$ 和 $p_i^{e,\max}$ 的定义详见 5.2 小节. p_i^{tag} 是 TAG 模型输出的候选片段的置信度得分. 根据 p^{CAP} 分数, 我们使用 NMS 机制对边界调整后的候选片段进行处理和排序. 该消融模型仅使用 L_{clip} 损失函数进行训练.

• **$\mathbf{SCDN_{CAP}}$(trainOnly).** 使用 L_{seg} 和 L_{clip} 损失函数训练整体模型. 但是, 基于起始点和终止点概率的时间边界调整和基于中间区域概率的候选片段调整将不会被执行. 我们直接使用语义条件动态归一化机制指导下的时序卷积网络模型的输出作为时序定位的结果, 细粒度时间边界调整模块仅用于辅助模型训练, 其输出将不会用于调整候选片段的时间边界.

• **$\mathbf{SCDN_{CAP}}$(TBA).** 在$\mathrm{SCDN_{CAP}}$(trainOnly) 模型的基础上, 进一步执行基于起始点和终止点概率的时间边界调整.

• **$\mathbf{SCDN_{CAP}}$(TAG).** 在$\mathrm{SCDN_{CAP}}$(trainOnly) 模型的基础上, 进一步执行基于中间区域概率的候选片段调整.

• **$\mathbf{SCDN_{CAP}}$.** 我们所提出的带有细粒度时间边界调整模块的语义条件动态归一化机制指导下的时序卷积网络模型. 该模型在 L_{seg} 和 L_{clip} 损失函数下进行了训练. 在目标片段预测阶段, 如第 5.2 小节所述, 我们同时引入了基于起始点和终止点概率的时间边界调整和基于中间区域概率的候选片段调整. 请注意, 非极大值抑制机制 NMS 始终在所有消融模型中执行.

上述消融模型在 TACoS 和 Charades-STA 数据集上的实验结果如表 6 所示. 对于 CAP 的性能, 它在 TACoS 数据集上的性能仅略低于 SCDN, 这表明起始点/终止点/中间区域概率预测可以很好地捕获单位视频片段内自然语言句子与视频内容的相关性, 由此为后续精细地调整候选片段时间边界提供了良好的基础. 通过将更为细粒度的单位视频片段级别的损失函数 L_{clip} 纳入 $\mathrm{SCDN_{CAP}}$(trainOnly) 的

表 7 模型的运行时效率与模型大小比较

Table 7 The comparison of run-time efficiency and model size

Method	Run-Time (s)	Model size (M)
CTRL [11]	2.23	22
ACRN [12]	4.31	128
ABLR	**0.15**	**7.6**
SCDN	0.78	15
SCDN_{CAP}(trainOnly)	0.80	18
SCDN_{CAP}	1.48	18

训练过程中, 其模型性能在 Charades-STA 和 TACoS 数据集上均优于 SCDN 模型. 这是因为 L_{clip} 损失函数将指导在更精细的单位视频片段级别上建立自然语言句子与视频的语义相关性, 且细粒度时间边界调整模块与语义条件动态调制机制指导下的时序卷积网络共享底层特征图, 那么底层细粒度的单位视频片段级别的预测也将促进粗粒度的具有更长时间跨度的视频片段预测. 在 SCDN_{CAP}(trainOnly) 模型基础上, 我们观察到执行基于起始点和终止点概率的时间边界调整 (即 SCDN_{CAP}(TBA) 模型) 或执行基于中间区域概率的候选片段调整 (即 SCDN_{CAP}(TAG) 模型) 均可以进一步带来时序定位的性能提升. 同时考虑上述两项, 我们提出的完整模型 SCDN_{CAP} 达到了最佳性能, 这证明了在目标片段的预测过程中根据起始点/终止点/中间区域概率分数来微调候选视频片段边界的有效性.

6.8 模型效率比较

表 7 比较了不同方法在 TACoS 数据集上进行实验时的运行时效率和模型大小. 具体来说, "运行时间 (Run-Time)" 表示在视频中定位一个句子的平均时间. 我们将已经公开发布代码的 CTRL 和 ACRN 方法以及我们所提出的 ABLR, SCDN, 和 SCDN_{CAP} 方法在一个具有 Nvidia TITAN XP GPU 的服务器上进行测试. 可以看出, ABLR 方法具有最小的模型大小, 最快的运行时间, 这是因为 ABLR 模型在视频编码过程中仅需要通过双向 LSTM 两次遍历视频, 避免了冗余的计算, 从而获得了较好的时间效率性能. CTRL 和 ACRN 方法都需要在视频中使用滑动窗口对候选片段进行采样, 然后将句子分别与每个片段进行匹配. 由于基于滑动窗口的匹配过程非常耗时, 这种做法将不可避免地影响时序定位的效率. SCDN 采用了多层时序卷积网络, 其层序体系架构自然地覆盖了多尺度的视频片段. 因此, 我们只需要在多层时序卷积网络中前馈计算一次, 就可以获得时序定位结果, 从而取得更高的效率. 另外, SCDN 只需要控制特征归一化参数, 而不需要直接更改卷积核, 这对整个卷积体系架构来说带来的额外开销较小. 因此, SCDN 的模型大小也相对更具优势, 是一个较为轻量级的方法.

通过增加细粒度时间边界调整模块, 模型 SCDN_{CAP}(trainOnly) 和 SCDN_{CAP} 在训练阶段的模型大小有所增加. 但在预测阶段 SCDN_{CAP}(trainOnly) 模型定位一个句子的平均运行时间仍然与 SCDN 近似. 这是由于我们在 SCDN_{CAP}(trainOnly) 中仅增加了损失函数 L_{clip} 用于模型训练, 而在预测阶段则不会执行时间边界调整, 因此 SCDN_{CAP}(trainOnly) 模型的预测过程与 SCDN 模型的预测过程完全相同. 但是, 如果执行了时间边界调整, 则完整模型 SCDN_{CAP} 的预测时间会更长. 这是由于边界调整将搜索起始点/终止点/中间区域概率序列, 并且 TAG 操作还需要进行额外的 NMS 操作, 这都是比较费时的. 如表 6 所示, 由于 SCDN 和 SCDN_{CAP}(trainOnly) 模型也都具有良好的性能, 因此如果在一些情况下需要考虑模型的时间效率, 则可以直接将这两个模型应用于时序定位任务而不必过分追求定位准确度更高的 SCDN_{CAP} 模型.

7 未来研究方向

本节讨论自然语言句子在视频中的时序定位问题的未来研究方向, 包括 (1) 引入视频中的音频信息作为视觉信息的补充, 帮助更好地关联句子和视频的语义内容; (2) 考虑在弱监督的环境下解决时序定位问题, 即仅利用成对的视频和自然语言句子作为模型学习的监督数据, 而不借助自然语言句子在视频中的时间定位信息; (3) 当前时序定位数据集中存在严重的时间位置标注偏见 (bias), 使得许多方法陷入了学习数据偏见的陷阱, 而忽略了视频和自然语言句子多模态匹配关系. 构建无偏见的数据集和不受数据偏见影响的时序定位方法也是未来一个重要的研究方向.

7.1 利用音频信息辅助自然语言句子在视频中的时序定位问题

当前自然语言句子在视频中的时序定位方法仅考虑视频中的视觉信息和自然语言句子信息的语义匹配关系, 而忽略了视频中的音频信息. 事实上, 音频信息也能够为时序定位提供关键的线索. 例如, 观众的欢呼声常常伴随着足球运动员射门这一动作出现, 那么当给定自然语言句子描述的是与射门相关的内容时, 有观众欢呼声出现的视频片段则可以被重点关注. 此外, 视频中人们的言语交谈也可以通过语音识别的方法转化为字幕, 而字幕可以被认为是视频在文本模态的表达, 那么在此同一模态进行语义匹配和对齐, 将为视觉和文本信息的跨模态语义匹配提供额外的辅助. 在视频字幕生成 (video captioning) 问题中, 已经有研究将音频信息与视觉信息相融合, 有效提高了模型所生成的视频文本描述的准确性 [44~46]. 在未来工作中, 如何更好地利用音频信息辅助自然语言句子在视频中的时序定位问题, 也是一个很重要的研究方向.

7.2 弱监督环境下的自然语言句子在视频中的时序定位问题

为了训练自然语言句子在视频中的时序定位问题的方法, 需要的监督信息主要有 3 项: 未剪辑的视频、与视频语义内容相关的自然语言句子, 以及自然语言句子所对应的视频片段的时间标注信息. 在构建这一类训练数据时, 标注者需要领会句子的含义, 完整观看视频再决定正确的标注位置, 这一过程是十分耗时的, 因而带来的人力成本十分巨大. 事实上, 获取视频和与其匹配的自然语言信息并不困难 (参考网络视频和其视频标题、描述和评论), 数据标注的成本主要集中在为自然语言句子明确时间信息. 那么, 为了省去这一标注开销, 人们开始思考在弱监督的环境下解决自然语言句子在视频中的时序定位问题, 即仅提供成对的视频和自然语言句子信息, 而不借助自然语言句子在视频中的时间定位信息来完成模型的训练.

在弱监督条件下, 有研究将视频描述生成问题和时序定位问题相关联, 构建了一个从 “在视频中定位自然语言句子”, 再到 “从时序定位的片段生成给定的自然语言句子” 的回路, 利用自然语言句子的生成损失函数来帮助时序定位模型的学习 [47, 48]. 还有研究借助 “准确的时序定位结果会为更精准的视频 – 文本整体匹配关系提供指导” 这一思想, 将全局的视频和文本的匹配关系依托到局部视频内容和文本的匹配, 再以全局的匹配损失函数作为学习信号, 帮助弱监督条件下时序定位模型的学习 [49]. 尽管前述的方法都取得了一定的进展, 但是目前弱监督时序定位方法的性能仍然与全监督的方法存在较大差距, 因而不断探究弱监督环境下的时序定位问题, 提升其模型准确度也是一个十分有意义且值得继续深入探索的研究方向.

7.3 无偏见的时序定位数据集构建和方法研究

目前自然语言句子在视频中的时序定位问题主要依托的是 ActivityNet Captions [37], Charades-

STA [11] 和 TACoS [16] 数据集, 但有研究表明, 这些数据集中所标注的目标视频片段的时间位置存在着明显的标注偏见 [50]. 具体而言, 若我们分别以标注片段的起始点和终止点为横坐标和纵坐标 (将时间坐标除以视频全长, 使其归一化到 0∼1 区间内), 画出训练集和测试集中所有标注片段在此二维空间中的分布, 可以观察到训练集和测试集上标注片段的时间坐标分布十分相似, 都聚集在类似的时间窗口中. 这一现象导致我们可以设计一些特定的方法, 它们不考虑视频和自然语言句子的信息, 仅仅学习视频片段在训练集上的时间标注分布, 再从学习到的分布中采样一些时间坐标作为测试集上的时序定位结果, 依然能够达到不错的定位性能. 另外, 近期还有研究将现有数据集进行了重新组织, 人为拉大训练集和测试集中时间标注分布的差异, 并将现有的时序定位方法在重新组织的数据集上进行模型学习和性能测试 [51]. 结果表明, 这些现有方法的模型性能在重新组织的数据集上均有大幅度下降, 远低于其在原始数据集中的时序定位准确度. 上述研究表明, 现有的时序定位方法都陷入了数据集标注的陷阱, 它们过多地学习了时间位置标注的分布, 而忽略了视频和自然语言文本的语义匹配关系, 这样的模型只能在存在相同标注特征的场景下取得不错的时序定位效果, 而无法推广到真实的自然语言句子在视频中的时序定位任务中.

因此, 在未来的工作中改进现有数据集, 或构建新的数据集以缓解时序定位问题中的时间标注偏见现象, 并提出不过分依赖时间标注偏见的时序定位方法, 对引导自然语言句子在视频中的时序定位问题的良性发展起着至关重要的作用.

8 总结

本文研究了自然语言句子在视频中的时序定位问题, 指出现有方法忽略了自然语言句子中的关键定位线索, 忽视了自然语言句子对于关联视频内部相关内容的重要指导作用, 因而导致时序定位准确率不高的问题. 为解决上述难题, 本文首先提出了基于注意力回归的时序定位方法, 该方法利用多模态共同注意力机制, 能够挖掘自然语言句子中与时序定位相关的重要语义细节, 精细地构建句子中各单词和视频之间的语义关系, 在考虑全局视频结构的情况下高效地回归预测出目标视频片段的位置. 在此基础上, 我们还进一步提出了一种语义条件动态归一化机制, 并将其嵌入到多层时序卷积网络中. 该机制建立了不同位置的视频内容与句子的语义交互关系, 突出强调不同的句子细节并产生对应的注意力加权句子表征, 以此动态地引导视频中与句子语义相关联的局部视频内容紧密耦合, 形成更明确的时序定位边界从而提升时序定位的准确率. 此外, 基于语义条件动态归一化机制指导下的时序卷积网络模型, 我们提出了一个细粒度时间边界调整模块用于微调卷积网络所预测的视频片段的边界区域, 使模型能够摆脱时序卷积网络预设的时间窗口的限制, 获得更为精准灵活的定位结果. 在 3 个公开数据集上的实验中, 我们验证了所提出的方法的有效性和优越性. 最后, 本文对未来方向进行了展望, 指出了利用音频信息作为辅助, 考虑弱监督环境下的时序定位问题, 和构建无偏见时序定位数据集方面的潜在研究价值.

参考文献

1 Yeung S, Russakovsky O, Mori G, et al. End-to-end learning of action detection from frame glimpses in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2678–2687

2 Lin T, Zhao X, Shou Z. Single shot temporal action detection. In: Proceedings of the 25th ACM International Conference on Multimedia, 2017. 988–996

3 Singh B, Marks T K, Jones M, et al. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1961–1970

4 Shou Z, Wang D, Chang S F. Action temporal localization in untrimmed videos via multi-stage CNNs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016
5 Shou Z, Chan J, Zareian A, et al. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017
6 Yuan Z, Stroud J, Lu T, et al. Temporal action localization by structured maximal sums. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017
7 Escorcia V, Heilbron F C, Niebles J C, et al. Daps: deep action proposals for action understanding. In: Proceedings of European Conference on Computer Vision, 2016. 768–784
8 Heilbron F C, Niebles J C, Ghanem B. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1914–1923
9 Gao J, Yang Z, Sun C, et al. Turn tap: temporal unit regression network for temporal action proposals. In: Proceedings of the IEEE International Conference on Computer Vision, 2017
10 Hendricks L A, Wang O, Shechtman E, et al. Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision, 2017
11 Gao J, Sun C, Yang Z, et al. Tall: temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 5267–5275
12 Liu M, Wang X, Nie L, et al. Attentive moment retrieval in videos. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018. 15–24
13 Chen J, Chen X, Ma L, et al. Temporally grounding natural sentence in video. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. 162–171
14 Zhang S, Peng H, Fu J, et al. Learning 2D temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 12870–12877
15 Hahn M, Kadav A, Rehg J M, et al. Tripping through time: efficient localization of activities in videos. In: Proceedings of British Machine Vision Conference (BMVC), 2020
16 Regneri M, Rohrbach M, Wetzel D, et al. Grounding action descriptions in videos. Trans Assoc Comput Linguist, 2013, 1: 25–36
17 Naim I, Song Y C, Liu Q, et al. Unsupervised alignment of natural language instructions with video segments. In: Proceedings of AAAI Conference on Artificial Intelligence, 2014. 1558–1564
18 Tapaswi M, Bauml M, Stiefelhagen R. Book2movie: aligning video scenes with book chapters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1827–1835
19 Bojanowski P, Lajugie R, Grave E, et al. Weakly-supervised alignment of video with text. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 4462–4470
20 Ge R, Gao J, Chen K, et al. MAC: mining activity concepts for language-based temporal localization. In: Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision, 2019. 245–253
21 Wu A, Han Y. Multi-modal circulant fusion for video-to-language and backward. In: Proceedings of International Joint Conference on Artificial Intelligence, 2018. 1029–1035
22 Xu H, He K, Sigal L, et al. Multilevel language and vision integration for text-to-clip retrieval. In: Proceedings of AAAI Conference on Artificial Intelligence, 2019
23 Chen S, Jiang Y G. Semantic proposal for activity localization in videos via sentence query. In: Proceedings of AAAI Conference on Artificial Intelligence, 2019
24 Zhang D, Dai X, Wang X, et al. MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 1247–1257
25 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of International Conference on Learning Representations, 2017
26 Rodriguez C, Marrese-Taylor E, Saleh F S, et al. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2020. 2464–2473
27 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

28 Karpathy A, Li F-F. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3128–3137

29 Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014. 1532–1543

30 Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering. In: Proceedings of Neural Information Processing Systems, 2016. 289–297

31 Rohrbach A, Rohrbach M, Hu R, et al. Grounding of textual phrases in images by reconstruction. In: Proceedings of European Conference on Computer Vision, 2016. 817–834

32 Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 1440–1448

33 Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Proceedings of European Conference on Computer Vision, 2016. 21–37

34 de Vries H, Strub F, Mary J, et al. Modulating early visual processing by language. In: Proceedings of Neural Information Processing Systems, 2017. 6594–6604

35 Dumoulin V, Shlens J, Kudlur M. A learned representation for artistic style. In: Proceedings of International Conference on Learning Representations, 2017

36 Xiong Y, Zhao Y, Wang L, et al. A pursuit of temporal accuracy in general activity detection. 2017. ArXiv:1703.02716

37 Krishna R, Hata K, Ren F, et al. Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 706–715

38 Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014. ArXiv:1412.3555

39 Kingma D P, Ba J. ADAM: a method for stochastic optimization. 2014. ArXiv:1412.6980

40 Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 4489–4497

41 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 6299–6308

42 Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors. In: Proceedings of Neural Information Processing Systems, 2015. 3294–3302

43 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning, 2015. 448–456

44 Xu J, Yao T, Zhang Y, et al. Learning multimodal attention LSTM networks for video captioning. In: Proceedings of the 25th ACM International Conference on Multimedia, 2017. 537–545

45 Iashin V, Rahtu E. Multi-modal dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020. 958–959

46 Xu Y C, Yang J F, Mao K Z. Semantic-filtered soft-split-aware video captioning with audio-augmented feature. Neurocomputing, 2019, 357: 24–35

47 Duan X, Huang W, Gan C, et al. Weakly supervised dense event captioning in videos. In: Proceedings of Advances in Neural Information Processing Systems, 2018

48 Song Y, Wang J, Ma L, et al. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. 2020. ArXiv:2003.07048

49 Mithun N C, Paul S, Roy-Chowdhury A K. Weakly supervised video moment retrieval from text queries. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019

50 Otani M, Nakashima Y, Rahtu E, et al. Uncovering hidden challenges in query-based video moment retrieval. 2020. ArXiv:2009.00325

51 Yuan Y, Lan X, Chen L, et al. A closer look at temporal sentence grounding in videos: datasets and metrics. 2021. ArXiv:2101.09028

Temporal sentence grounding in videos with fine-grained multi-modal correlation

Yitian YUAN, Xin WANG & Wenwu ZHU*

Department of Computer Science and Technology, Tsinghua University, Beijing 100084*, China*
* Corresponding author. E-mail: wwzhu@tsinghua.edu.cn

Abstract Videos have become a new way of communication among Internet users with the proliferation of sensor-rich mobile devices. Due to the redundant background information in video data, people usually spend much time browsing and analyzing video content. This necessity motivates us to investigate the temporal sentence grounding task in videos. Formally, given an untrimmed video and a natural language sentence query, the task is to identify the start and end points of the video segment in response to the given sentence query. With such a technique, people can quickly find specific content of interest in the video by providing a clear and concise text description, thereby improving users' video browsing experience and search efficiency. Previous methods often formulate the temporal grounding task as a multimodal matching problem. Doing so ignores the important sentence details for grounding and neglects the important guiding role of sentences to compose and correlate video contents over time, causing limited temporal grounding accuracy. To solve the above problems, we first propose a multimodal co-attention mechanism to mine important semantic details for temporal grounding in the given query and finely construct the semantic correlation between each word in the sentence and the video content. On this basis, we then propose a semantic condition dynamic normalization mechanism to tightly compose the sentence-related video content over time, including a clip-level actionness prediction module for fine-grained temporal boundary adjustment, thus making the temporal grounding results in the video clearer, more flexible, and more accurate than usual. Experiments on public datasets also verify our effectiveness and superiority over the state-of-the-arts. Last but not least, we present our insights on future research directions that deserve further investigations in the areas of audio-enabled temporal grounding techniques, weakly supervised grounding problem formulation, and debiased temporal grounding dataset construction.

Keywords temporal sentence grounding in videos, semantic correlation, multimodal co-attention mechanism, temporal convolutional network, semantic conditioned dynamic normalization