JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2020
07
VOL.25
ISSN1006-8961
CN11-3758/TB

俯视深度头肩序列识别 P1393

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第7期（总第291期）
2020年7月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 顾行发
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

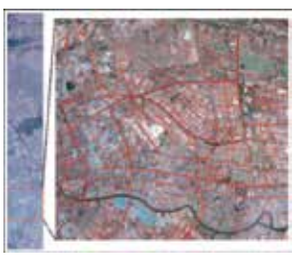
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年7月 第25卷 第7期（总第291期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

非真实感绘制技术研究现状与展望(第1283页)

图卷积网络下牙齿种子点自动选取(第1481页)



迁移学习下高分快视数据道路快速提取(第1501页)

Volume 25, Number 7
Published July 16, 2020

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Research status and progress of non-photorealistic rendering technology(P1283)

Automatic selection of tooth seed point by graph convolutional network(P1481)

Rapid road extraction from quick view imagery of high-resolution satellites combined with transfer learning(P1501)

中图法分类号:TP301.6　　**文献标识码**:A　　**文章编号**:1006-8961(2020)07-1447-12

论文引用格式:Xiong C X, Guo D and Liu X L. 2020. Temporal proposal optimization for temporal action detection. Journal of Image and Graphics, 25(07):1447-1458(熊成鑫,郭丹,刘学亮. 2020. 时域候选优化的时序动作检测. 中国图象图形学报,25(07):1447-1458)[DOI:10.11834/jig.190440]

时域候选优化的时序动作检测

熊成鑫,郭丹,刘学亮

合肥工业大学计算机与信息学院,合肥　230601

摘　要:**目的**　时序动作检测(temporal action detection)作为计算机视觉领域的一个热点课题,其目的是检测视频中动作发生的具体区间,并确定动作的类别。这一课题在现实生活中具有深远的实际意义。如何在长视频中快速定位且实现时序动作检测仍然面临挑战。为此,本文致力于定位并优化动作发生时域的候选集,提出了时域候选区域优化的时序动作检测方法 TPO(temporal proposal optimization)。**方法**　采用卷积神经网络(convolutional neural network,CNN)和双向长短期记忆网络(bidirectional long short term memory,BLSTM)来捕捉视频的局部时序关联性和全局时序信息;并引入联级时序分类优化(connectionist temporal classification,CTC)方法,评估每个时序位置的边界概率和动作概率得分;最后,融合两者的概率得分曲线,优化时域候选区域候选并排序,最终实现时序上的动作检测。**结果**　在 ActivityNet v1.3 数据集上进行实验验证,TPO 在各评价指标,如一定时域候选数量下的平均召回率 AR@100(average recall @ 100),曲线下的面积 AUC(area under a curve)和平均均值平均精度 mAP(mean average precision)上分别达到 74.66、66.32、30.5,而各阈值下的均值平均精度 mAP@IoU(mAP@ intersection over union)在阈值为 0.75 和 0.95 时也分别达到了 30.73 和 8.22,与 SSN(structured segment network)、TCN(temporal context network)、Prop-SSAD(single shot action detector for proposal)、CTAP(complementary temporal action proposal)和 BSN(boundary sensitive network)等方法相比,TPO 的所有性能指标均有提高。**结论**　本文提出的模型兼顾了视频的全局时序信息和局部时序信息,使得预测的动作候选区域边界更为准确和灵活,同时也验证了候选区域的准确性能够有效提高时序动作检测的精确度。

关键词:时序动作检测;时域候选区域;动作概率得分;级联时序分类;卷积神经网络;双向长短期记忆网络

Temporal proposal optimization for temporal action detection

Xiong Chengxin, Guo Dan, Liu Xueliang

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China

Abstract: **Objective** With the ubiquity of electronic equipment, such as cellphones and cameras, massive video data of people's activities and behaviors in daily life are stored, recorded, and transmitted. Increasing video-based applications, such as video surveillance, have attracted the attention of researchers. However, real-world videos are consistently long and untrimmed. Long untrimmed videos in publicly available datasets for temporal action detection consistently contain several ambiguous frames and a large number of background frames. Accurately locating action proposals and recognizing action labels are difficult. Similar to object proposal generation in object detection task, the task of temporal action detection can be resolved into two phases, where the first phase is to determine the specific durations (starting and ending timestamps) of

收稿日期:2019-08-29;**修回日期**:2019-12-14;**预印本日期**:2019-12-21

actions, and the second phase is to identify the category of each action instance. The development of single-action classification in trimmed videos has been extremely successful, whereas the performance of temporal action proposal generation remains unsatisfactory. The phase of candidate action proposal generation experiences time-consuming model training. High-quality proposals contribute to the performance of action detection. The study on temporal proposal generation can effectively and efficiently locate the video content and facilitate video understanding in untrimmed videos. In this work, we focus on the optimization of temporal action proposals for action detection. **Method** We aim to improve the performance of action detection by optimizing temporal action proposals, that is, accurately localizing the boundaries of actions in long untrimmed videos. We propose a temporal proposal optimization (TPO) model for the detection of candidate action proposals. TPO utilizes the advantages of convolutional neural networks (CNNs) and bidirectional long short-term memory (BLSTM) to simultaneously capture the local and global temporal cues. In the proposed TPO model, we introduce connectionist temporal classification (CTC) optimization, which excels at parsing global feature-level classification labels. The global actionness probability calculated by BLSTM and CTC modifies several inexact temporal cues in the local CNN actionness probability. Thus, a probability fusion strategy based on local and global actionness probabilities promotes the accuracy of temporal boundaries of actions in videos and results in the promising performance of temporal action detection. In particular, TPO is composed of three modules, namely, local actionness evaluation module (LAEM), global actionness evaluation module (GAEM), and post processing module (PPM). The extracted features are fed into LAEM and GAEM. Then, LAEM and GAEM generate the global and local actionness probabilities along the temporal dimension, respectively. LAEM is a temporal CNN-based module, and GAEM predicts the global actionness probabilities with the help of BLSTM and CTC losses. LAEM outputs three sequences. Starting and ending probabilities are found in addition to local actionness probabilities. The crossing of starting and ending probability curves builds the candidate temporal proposals. Thus, GAEM captures global actionness probabilities, which is auxiliary to LAEM. Then, the local and global actionness probabilities are fed into PPM to obtain a fused actionness probability curve. Subsequently, we sample the actionness probability curves through linear interpolation to extract proposal-level features. The proposal-level features are fed int a multilayer perceptron) to obtain the confidence score. We use the confidence score to rank the candidate proposals and adopt soft-NMS (non-maximum supression) to remove redundant proposals. Finally, we apply an existing classification model with our generated proposals to evaluate the detection performance of TPO. **Result** We validate the proposed model on two evaluations of action proposal generation and action detection. Experimental results indicate that TPO outperforms other state-of-the-art methods on ActivityNet v1.3 dataset. For the proposal generation, we compare our model with the methods, including SSN (structured segment network), TCN (temporal context network), Prop-SSAD (single shot action detector for proposal), CTAP (complementary temporal action proposal), and BSN (boundary sensitive network). The proposed TPO model performs best and achieves average recall @ average number of proposals of 74.66 and area under a curve of 66.32. For the temporal action detection task, we test the quantitative evaluation metric mean average precision@ intersection over union (mAP@ IoU). Compared with the existing methods, including SCC (semantic cascade context), CDC (convolutional-de-convolutional), SSN and BSN, TPO achieves the best mAPs of 30.73 and 8.22 under the tIoUs of 0.75 and 0.95, respectively, and obtains the best average mAP of 30.5. Notably, the mAP value decreases with the increase in tIoU value. The tIoU metric reflects the overlap between the generated proposals and the ground truth, where a high tIoU value indicates strict constraints on candidate proposals. Thus, TPO achieves the best mAP performance under high tIoU values (0.75 and 0.95). This result validates the detection performance. TPO generates accurate proposals of action instances with high overlap on the ground truth and improves the detection performance. **Conclusion** In this paper, we propose a novel model called TPO for temporal proposal generation that achieves promising performance on ActivityNet v1.3 to resolve the action detection problem. Experimental results demonstrate the effectiveness of TPO. TPO generates temporal proposals with precise boundaries and maintains flexible temporal durations, thereby covering sequential actions in videos with variable-length intervals.

Key words: temporal action detection; temporal action proposals; actionness probability; connectionist temporal classification (CTC); convolutional neural network (CNN); bidirectional long short term memory (BLSTM)

0 引　言

随着电子拍摄设备的普及，用于存储、记录、传输的视频数据海量涌现，其中绝大部分视频都记录着以人为主体的行为或动作，这些长视频已广泛应用于城市监控以及安防等应用领域，衍生出动作检测（Heilbron 等，2017）、行为分析（罗会兰 等，2017，2019；Tu 等，2019）、视频摘要（Yang 等，2018）等诸多视频内容智能理解研究课题。能够有效快速定位视频关键内容并进行智能分析具有深远意义。因此，时序动作检测技术应运而生。作为机器视觉中的一个重要研究分支，时序动作检测旨在精准识别长视频中的动作时域边界和动作类别。然而，面对海量过长时序视频，分析视频内容仍然显得任重道远。事实证明，提高动作时域边界的准确度能够有效提高时序动作检测性能。因此，本文专注于长视频中高质量的时域候选区域提取研究，以提高时序动作检测性能。

早期的时序动作检测方法将动作边界生成和动作类别预测当作一个整体，一并求解并输出两类目标结论，难度较大，效果通常并不理想（Xu 等，2017）。而错误或不够准确的动作边界引入还影响了动作类别预测的精度。鉴于识别单一且短小的动作片段（剪辑过的短视频）类别的优秀模型已大量存在（Xiong 等，2016），动作时域预测的求解已成为长视频中动作精准预测的核心难点之一。

与动作时域边界提取最为接近的研究方向是目标检测（曹诗雨 等，2017；Ren 等，2017）中的目标候选区域提取。不同之处是，后者是在 2 维图像中定位感兴趣的区域坐标，而前者是在 1 维时间维度上定位感兴趣的动作候选区域（帧序列时间子段）。此类研究逐渐引起了研究者的广泛关注（Zhu 等，2018）。但真实自然场景下的视频通常较长，包含大量背景帧或模糊帧等噪音数据。此外，视频中动作的时长差别也较大。这些因素增加了长视频中时序动作的定位及识别难度。

动作时域边界提取的方法大体可分为两类。第 1 类方法基于滑动窗口（Shou 等，2016），这类方法首先按照滑动窗口规则提取时序片段作为输入，再对边界进行微调，最后计算候选区域的置信度。多样性有限的滑动窗口能生成有限种类区间长度的候选区域。然而，视频中的动作区间长度差别较大，这类方法由于预设固定滑动窗口大小，提取的候选区域边界常常不够精确，故而性能仍不理想。另一类方法则基于帧级或单元级的动作得分（Zhao 等，2017；Lin 等，2018）来直接判断有无动作，进而生成时域候选区域。然而，针对长视频，基于帧级预测（Zhao 等，2017）的方法面临着庞大的冗余计算。单元级预测（Lin 等，2018）可以看做是基于滑动窗口的方法与帧级预测方法之间的权衡，既希望得到较为精确的动作边界，又能避免冗余计算。因此，本文采用通过单元级区间判断有无动作概率的研究思路来实现动作边界候选提取生成。

本文提出了一种新的时域候选区域优化的时序动作检测模型 TPO（temporal proposal optimization）。考虑到视频较长，而单个动作实例占比可能过小，在关注局部时序关联性的同时，也兼顾了全局时序信息，预测视频中单元级的边界概率和动作概率得分。具体而言，首先采用卷积神经网络（convolutional neural network，CNN）捕捉视频的局部时序关联性，评估每个时序位置作为动作边界点和包含动作的概率；其次采用双向长短期记忆网络（bidirectional long short term memory，BLSTM）捕捉视频的全局时序信息，引入联级时序分类优化方法（connectionist temporal classification，CTC）（Cui 等，2017），评估每个时序位置的动作概率得分；最后联合利用得到的概率得分曲线，融合预测的动作概率得分，优化时域候选区域并排序，最终实现时序上的动作检测。本文的贡献可以分为 3 个方面：1）兼顾了视频中的局部特征和全局特征，优化了动作候选区域，并得到了可靠的候选区域置信度得分；2）将 CTC 全局优化概念引入到时序动作检测问题中，并证明了有效性；3）本文在候选区域生成和动作检测两方面评估了 TPO，并在 ActivityNet v1.3 数据集上实现了良好的性能。

1 相关工作

1.1 视频特征抽取

卷积神经网络在图像分类 ImageNet 任务中取得巨大成功后，研究者将其引入到视频特征表达的研究上。鉴于视频相对于图像更为复杂，还需考虑时序信息，Simonyan 和 Zisserman（2016）提出了双流的 CNN 网络，分别捕捉来自 RGB 帧的空间信息和

来自光流的时间信息。而 Tran 等人(2015)将卷积核从作用于空间维度的 2 维拓展到了时空维度的 3 维。3D CNN 比 2D CNN 更适用于时空特征学习,能够兼顾视频中的外观(apperarance)信息和运动(motion)信息,使得视频表达更为紧凑和准确,但模型的尺寸也呈指数增长。为此,Tran 等人(2018)采用了伪 3D 结构,将 1 个 3×3×3 的 3 维卷积,分解为空间维上的 2 维卷积(1×3×3)和时间维上的 1 维卷积(3×1×1)。本文采用优秀的双流 CNN 网络(Simonyan 和Zisserman,2016)提取视频的 RGB 视觉特征和光流视觉特征,并将此作为数据预处理步骤。

1.2 时序动作检测

早期的视频数据集包含的动作类别相当有限。而后,研究人员致力于构建包含更多复杂动作的数据集,如 ActivityNet(Heilbron 等,2015)数据集。这些数据集包含大量人类在真实环境中未剪辑的动作视频。针对过长视频,研究者首先将 CNN 引入时序动作检测问题,用以检测短时区间内的关联性。R-C3D(region convolutional 3D network)(Xu 等,2017)将 faster-RCNN(Ren 等,2017)中的目标候选提取网络和分类网络推广到时域,并采用 3D CNN 实现端到端的训练。类似地,Dai 等人(2017)参考 faster-RCNN(Ren 等,2017)的结构,用滑窗机制以等间隔生成不同尺寸的候选区域,再对候选区域进行排序,并将得分最高的候选送入分类网络以判定分类概率。这类方法存在两个问题:1)不准确的候选区域将直接影响后续分类概率的预测性能,最终影响动作检测性能;2)这类方法仅利用了卷积操作来捕捉时序信息,而卷积核的大小有限,因此仅能利用短时区间内的时序关联性。

由于动作检测任务是针对长视频提出的,长范围的时序信息对于动作检测任务也至关重要。Singh 等人(2016)证明了双向长短期记忆网络(BLSTM)用于动作检测任务的有效性,提出 BLSTM 能够有效捕捉视频中的全局信息,并学习到时间维度上相邻动作之间的关系。此外,Lin 等人(2018)和 Huang 等人(2018)发现候选区域的质量对于后续检测的效果有较大影响,因此认为目前改进时序动作检测性能的重点在于提高候选区域的质量,并专注于时序动作候选区域提取任务。

1.3 时序动作候选区域提取

受目标检测(Ren 等,2017)研究思路的启发,时域动作候选区域提取的目标是在 1 维帧序列中确定动作的时序边界。现有方法可分为两类,第 1 类方法基于滑动窗口,第 2 类方法基于帧级或单元级动作概率得分来界定候选区域。

基于滑动窗口的代表性方法有 Shou 等人(2016)提出的 S-CNN。该方法采用 3 个基于片段的 3 维卷积网络来识别滑动窗口生成的片段,然后对候选区域进行排序和微调。这里采用了多种尺度的滑动窗口,意味着要在时间维度上做详尽搜索,不可避免地导致了极高的计算成本。同时,这一方法只是对尺度多样性有限的滑动窗口边界进行微调,时序边界不够灵活,无法应对视频中动作持续时间变化较大的情况,因此性能受到限制。

另一类方法基于动作概率得分来界定动作实例的边界。CDC(convolutional-de-convolutional)(Shou 等,2017)在传统的 3D CNN 结构后添加反卷积层,将被卷积层压缩的时序维度通过反卷积层恢复并进行帧级预测,再采取时序平滑等策略得到动作实例的边界和类别。这一方法网络结构简单,但反卷积本质上只是一种通过间隔插入 padding 来扩充数据维度的特殊卷积操作,无法完全恢复原始数据,因此依然会丢失时序信息。同时,时序动作检测研究的是在长视频中定位动作边界,而长视频中通常存在大量冗余的背景帧或模糊帧,例如大型数据集 ActivityNet(Heilbron 等,2015)中,平均背景占比为 36%。此外,在视频预处理操作中,每秒提取的帧数(帧/s)通常设置为 25 或 30,因此最后的视频数据中将存在大量内容极为相近的视频帧,而帧级预测将导致大量的冗余计算。为了避免第 1 类方法中候选区域边界不精确的问题和帧级预测中庞大的冗余计算问题,研究者提出进行片段级概率得分预测(Zhao 等,2017;Lin 等,2018)。结构化的段网络(structured segment network,SSN)(Zhao 等,2017)中提出的 TAG(temporal actionness grouping)算法,等距离抽取一些帧作为片段特征,并评估每个片段的动作概率得分,最后合并得分较高的连续时序区域得到最终的候选区域。这类方法是基于滑动窗口和基于帧级别动作得分预测的方法之间的权衡,然而受到有限的感受野大小的限制,这些方法在局部特征响应上表现出色,却可能忽略了全局时序信息。

2　本文方法

本文的主要目标是优化动作候选区域以提高动作检测的性能。本文采用 CNN 捕捉视频的局部时序关联性，并在完整的视频上训练 BLSTM 以捕捉全局时序信息，实现了对视频单元级的边界概率和动作概率的准确预测，进而得到精确的动作候选区域和可靠的候选区域置信度得分。如图 1 所示，本文提出的 TPO 网络模型由 3 个主要部分组成：局部动作概率评估模块（local actionness evaluation module，LAEM）、全局动作概率评估模块（global actionness evaluation module，GAEM）和后处理模块（post processing module，PPM）。LAEM 模块捕捉视频的局部时序特征，分别评估每个时序位置的边界（作为动作起始点和结束点）概率和包含动作的概率，并根据边界概率来构造候选区域。GAEM 模块捕捉视频长时序列中的时序关联性，评估视频中每个时序位置的动作概率。PPM 模块融合了 LAEM 模块和 GAEM 模块预测的动作概率得分曲线，构建了候选区域级别的特征并对候选进行排序。

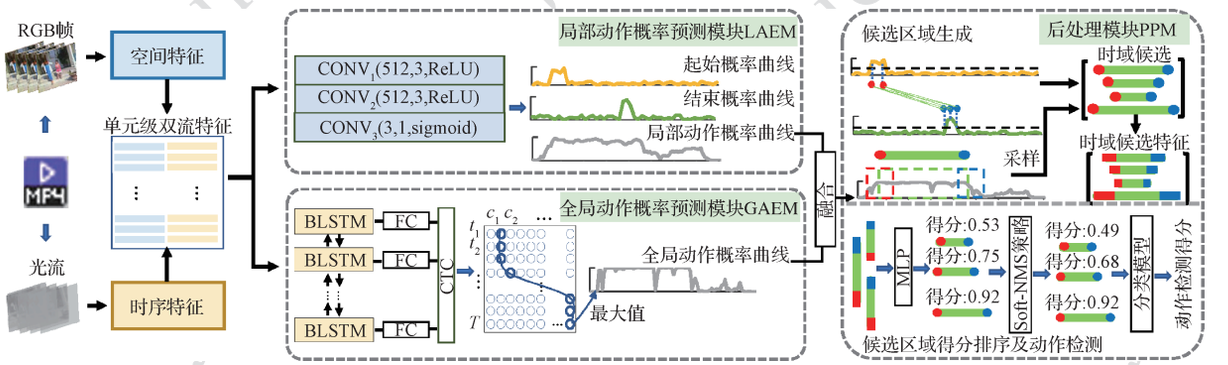


图 1　TPO 网络模型结构

Fig. 1　The basic architecture of the proposed model TPO

2.1　局部动作概率预测模块 LAEM

首先，本文使用双流网络（Simonyan 和 Zisserman，2016）作为特征抽取器，将原视频转换为片段级的特征序列 $\boldsymbol{F}=\{f_t\}_{t=1}^{T}$，其中，$T$ 为特征长度。视频的标签集为 $\boldsymbol{G}=(\boldsymbol{R}_{\mathrm{g}},\boldsymbol{C}_{\mathrm{g}})$，其中，$\boldsymbol{R}_{\mathrm{g}}=\{r_{\mathrm{g},k}\}_{k=1}^{N}$ 为动作实例的时序边界标签，$\boldsymbol{C}_{\mathrm{g}}=\{c_k\}_{k=1}^{N}$ 为动作类别标签，N 是视频中动作实例的数量，$r_{\mathrm{g},k}$ 的具体边界表示为时间刻度值 $[t_{\mathrm{g,s}},t_{\mathrm{g,e}}]$。其次，如图 1 所示，LAEM 模块由 3 个 1 维卷积层组成，所有卷积层的步长均为 1，卷积核大小分别为 3，3，1，前两层的通道数为 512，最后一层通道数为 3，分别对应输出的 3 条概率曲线。具体而言，基于前馈神经网络 CNN，LAEM 模块能够准确捕捉局部特征响应，进而计算输出 3 种类型的概率曲线：1）该视频的起始概率曲线 $\boldsymbol{P}_{\mathrm{S}}=\{p_t^{\mathrm{s}}\}_{t=1}^{T}$；2）该视频的结束概率曲线 $\boldsymbol{P}_{\mathrm{E}}=\{p_t^{\mathrm{e}}\}_{t=1}^{T}$；3）该视频的动作概率曲线 $\boldsymbol{P}_{\mathrm{A}}^{1}=\{p_t^{1}\}_{t=1}^{T}$，即判断有无动作的概率。

获取起始概率 $\boldsymbol{P}_{\mathrm{S}}$ 和结束概率 $\boldsymbol{P}_{\mathrm{E}}$ 后，模型分别遍历两条曲线以选择各自合适候选区域边界点。1）将 $\boldsymbol{P}_{\mathrm{S}}$ 中的最大值记录为 $\boldsymbol{P}_{\mathrm{S}}^{\max}$，则 $\boldsymbol{P}_{\mathrm{S}}$ 中的所有峰值节点或概率得分大于 $0.5\times\boldsymbol{P}_{\mathrm{S}}^{\max}$ 的节点都可作为符合要求的起始节点。符合要求的起始节点都被存入候选区域起始位置集 $\boldsymbol{B}_{\mathrm{S}}=\{t_{\mathrm{s},i}\}_{i=1}^{I}$，$I$ 表示集合中起始节点的数量。2）按照同样的筛选条件，将候选区域结束位置集定义为 $\boldsymbol{B}_{\mathrm{E}}=\{t_{\mathrm{e},j}\}_{j=1}^{J}$，$J$ 是集合中结束节点的数量。3）将起始位置集 $\boldsymbol{B}_{\mathrm{S}}$ 和结束位置集 $\boldsymbol{B}_{\mathrm{E}}$ 中的点两两组合。只要满足 $t_{\mathrm{e},j}-t_{\mathrm{s},i}>0$，则这对起始点和结束点可以构成一个候选区域。记所提取的时域候选区域为 $\boldsymbol{\psi}_{\mathrm{p}}=(t_{\mathrm{s}},t_{\mathrm{e}},p_{\mathrm{s}},p_{\mathrm{e}})$，式中 t_{s} 和 t_{e} 分别是起始时间点和结束时间点，p_{s} 和 p_{e} 是起始点和结束点各自在起始概率曲线 $\boldsymbol{P}_{\mathrm{S}}$ 和结束概率曲线 $\boldsymbol{P}_{\mathrm{E}}$ 上的动作概率得分。

2.2　全局动作概率预测模块 GAEM

GAEM 模块采用双向循环神经网络 BLSTM 来对时序信息进行编码。LSTM 在动作识别（Song 等，2018；Xu 等，2019）和动作检测（Singh 等，2016；Soomro 等，2019）等任务中都有很好的表现，而 BLSTM 可以同时利用过去时刻和未来时刻两个方

向上的信息,因此能够准确捕捉长时序列中的时序关联性。如图 1 所示,在 BLSTM 后连接全连接层和 sigmoid 层,并引入 CTC 优化方法来捕捉视频中的高响应点,以获得每个时序位置在所有动作类别上的中间概率。联级分类优化方法 CTC(Cui 等,2017)无需预先对数据对齐,进行动态规划的概率路径解析,从而直接输出最大概率的序列类别预测。将 CTC 引入时序动作检测问题,从全局时序解析的角度,求解单元级特征序列 $\boldsymbol{F}=\{f_t\}_{t=1}^{T}$ 与输出动作类别序列(标签)$\boldsymbol{C}_{\mathrm{g}}=\{c_k\}_{k=1}^{N}$ 之间的对应问题。GAEM 模块最后的输出 $\boldsymbol{P}_{\mathrm{M}}=\{p_{\mathrm{M}}^{t}\}_{t=1}^{T}$ 是每个时序位置在所有动作类别上的中间概率,再将 $\boldsymbol{P}_{\mathrm{M}}$ 转换为动作概率曲线,策略如下:假设有 L 个类别的动作,因此 $\boldsymbol{P}_{\mathrm{M}}$ 的形状将是 $(L+1)\times T$,1 表示空白类(即背景或模糊的内容),T 表示时序长度。对于每个时序位置 t,选择其在 L 个类别上的最高得分作为动作概率得分 p_t^{g},即 $p_t^{\mathrm{g}}=\max(\boldsymbol{P}_{\mathrm{M}}^{t})$。将视频的动作概率曲线定义为 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{g}}=\{p_t^{\mathrm{g}}\}_{t=1}^{T}$,即 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{g}}=\{\max(\boldsymbol{P}_{\mathrm{M}}^{t})\}_{t=1}^{T}$。记 y_c^t 为在时间 t 观察到类 c 输出的概率。

输入和输出的对齐序列表示为 $\boldsymbol{\pi}$,则有 $\boldsymbol{\pi}\in L^T$,$\boldsymbol{\pi}_t$ 是序列 $\boldsymbol{\pi}$ 在第 t 时刻的值。则生成一条路径 $\boldsymbol{\pi}$ 的概率为每个时刻观测到对应输出的概率的连乘积,即

$$P(\boldsymbol{\pi}\mid\boldsymbol{F})=\prod_{t=1}^{T}P(\boldsymbol{\pi}_t\mid\boldsymbol{F})=\prod_{t=1}^{T}y_c^t,\ \forall\boldsymbol{\pi}\in L^T \tag{1}$$

GAEM 的本质是对 $P(\boldsymbol{C}_{\mathrm{g}}\mid\boldsymbol{F})$ 建模,式(1)中 $\boldsymbol{F}$ 为输入的视频特征,$\boldsymbol{C}_{\mathrm{g}}$ 为目标序列,而对齐序列 $\boldsymbol{\pi}$ 与 $\boldsymbol{C}_{\mathrm{g}}$ 存在多对一的映射关系,则观测到目标序列 $\boldsymbol{C}_{\mathrm{g}}$ 的概率是所有路径 $\boldsymbol{\pi}$ 的概率之和,即

$$P(\boldsymbol{C}_{\mathrm{g}}\mid\boldsymbol{F})=\sum_{\boldsymbol{\pi}\in\boldsymbol{\beta}^{-1}(\boldsymbol{C}_{\mathrm{g}})}P(\boldsymbol{\pi}\mid\boldsymbol{F}) \tag{2}$$

式中,$\boldsymbol{\beta}^{-1}(\boldsymbol{C}_{\mathrm{g}})$ 为从 $\boldsymbol{F}$ 到 $\boldsymbol{C}_{\mathrm{g}}$ 的所有对齐方式的集合,对于给定的输入 $\boldsymbol{F}$,希望输出 $\boldsymbol{C}_{\mathrm{g}}$ 的条件概率最高,因此最大化概率 $P(\boldsymbol{C}_{\mathrm{g}}\mid\boldsymbol{F})$ 的路径即为解析路径,本文取最大概率路径作为全局时序序列。

2.3 后处理模块 PPM

PPM 模块融合了 GAEM 模块和 LAEM 模块预测的动作概率得分曲线,构建了候选区域级别的特征,并对候选区域进行重新优化选择。

2.3.1 PPM 的工作流程

PPM 的工作流程如下:

1)GAEM 模块捕捉全局时序信息,LAEM 模块捕捉局部特征响应,两种信息可以彼此互补、相互增强。PPM 模块则针对两者生成的整个视频的动作概率曲线 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{l}}=\{p_t^{\mathrm{l}}\}_{t=1}^{T}$ 和 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{g}}=\{p_t^{\mathrm{g}}\}_{t=1}^{T}$,得到融合后的动作概率曲线 $\boldsymbol{P}_{\mathrm{A}}$,具体为

$$\boldsymbol{P}_{\mathrm{A}}=\{(p_t^{\mathrm{l}}+p_t^{\mathrm{g}})/2\}_{t=1}^{T} \tag{3}$$

式中,p_t^{l} 表示依据局部时序信息预测的在时序位置 t 包含动作的概率,p_t^{g} 表示依据全局时序信息预测的在时序位置 t 包含动作的概率。

2)根据获取的动作概率曲线 $\boldsymbol{P}_{\mathrm{A}}$ 和 2.1 节中获取的候选区域 $\boldsymbol{\psi}_{\mathrm{p}}$ 来构建区域级别特征 $\boldsymbol{f}_{\mathrm{pl}}$。具体而言,对于候选区域 $\boldsymbol{\psi}_{\mathrm{p}}=(t_{\mathrm{s}},t_{\mathrm{e}},p_{\mathrm{s}},p_{\mathrm{e}})$,其区间长度 $d=t_{\mathrm{e}}-t_{\mathrm{s}}$。定义起始区域 $\boldsymbol{r}_{\mathrm{s}}=[t_{\mathrm{s}}-d/5,\ t_{\mathrm{s}}+d/5]$,结束区域 $\boldsymbol{r}_{\mathrm{e}}=[t_{\mathrm{e}}-d/5,\ t_{\mathrm{e}}+d/5]$ 和中心区域为 $\boldsymbol{r}_{\mathrm{c}}=[t_{\mathrm{s}},t_{\mathrm{e}}]$。然后在 $\boldsymbol{r}_{\mathrm{c}}$ 范围内以线性插值法对 $\boldsymbol{P}_{\mathrm{A}}$ 进行采样,本文共取 49 个插值点。为了充分利用候选的时序上下文信息,在 $\boldsymbol{r}_{\mathrm{s}}$ 和 $\boldsymbol{r}_{\mathrm{e}}$ 范围内也各采样 25 个点作为 $\boldsymbol{f}_{\mathrm{s}}$ 和 $\boldsymbol{f}_{\mathrm{e}}$。然后对 $\boldsymbol{f}_{\mathrm{c}}$,$\boldsymbol{f}_{\mathrm{s}}$ 和 $\boldsymbol{f}_{\mathrm{e}}$ 每连续 4 个值求 1 次均值,最终将这 3 个向量连接起来形成候选区域级别的特征,即 $\boldsymbol{f}_{\mathrm{pl}}=(\boldsymbol{f}_{\mathrm{s}},\boldsymbol{f}_{\mathrm{c}},\boldsymbol{f}_{\mathrm{e}})$。

3)将区域级别的特征 $\boldsymbol{f}_{\mathrm{pl}}$ 送入多层感知器模型(multi layer perception,MLP)模块来评估每个候选区域 $\boldsymbol{\psi}_{\mathrm{p}}$ 的置信度得分 p_{conf}。此时每个候选 $\boldsymbol{\psi}_{\mathrm{p}}$ 可以表示为 $\boldsymbol{\psi}_{\mathrm{p}}=(t_{\mathrm{s}},t_{\mathrm{e}},p_{\mathrm{s}},p_{\mathrm{e}},p_{\mathrm{conf}})$。式中 t_{s} 和 t_{e} 是候选的边界,p_{s} 是此候选的起始点在起始概率曲线 $\boldsymbol{P}_{\mathrm{S}}$ 上的得分,p_{e} 是此候选的结束点在结束概率曲线 $\boldsymbol{P}_{\mathrm{E}}$ 上的得分,p_{conf} 表示候选的置信度得分。此时,该候选区域 $\boldsymbol{\psi}_{\mathrm{p}}$ 的概率得分为

$$p_{\mathrm{f}}=p_{\mathrm{s}}\cdot p_{\mathrm{e}}\cdot p_{\mathrm{conf}} \tag{4}$$

记整个视频得到的候选时域区域集合为 $\boldsymbol{\Psi}_{\mathrm{p}}=\{\boldsymbol{\psi}_i\}_{i=1}^{Q}$,式中,$Q$ 为候选区域的数量。

4)采用 soft-NMS(non-maximum suppression)(Bodla 等,2017)算法对候选区域的得分进行处理,实现候选区域的最终排序。具体步骤如下:将候选集 $\boldsymbol{\Psi}_{\mathrm{p}}$ 中得分最高的候选区域记为 $\boldsymbol{\psi}_t$,对于候选集中余下的所有候选区域 $\boldsymbol{\psi}_i$ 都依次与 $\boldsymbol{\psi}_t$ 计算重叠率,对于重叠率高的候选区域,soft-NMS 按高斯衰减函数递归地调整其置信度得分,其中,高斯衰减函数公式表示为

$$p'_{\mathrm{f},i}=\begin{cases}p_{\mathrm{f},i} & iou(\boldsymbol{\psi}_t,\boldsymbol{\psi}_i)^2<\theta\\ p_{\mathrm{f},i}\cdot\mathrm{e}^{-\frac{iou(\psi_t,\psi_i)^2}{\varepsilon}} & iou(\boldsymbol{\psi}_t,\boldsymbol{\psi}_i)^2\geqslant\theta\end{cases} \tag{5}$$

式中，ε 是高斯函数参数，θ 是交并比（intersection-over-union，IoU）阈值，$iou(\boldsymbol{\psi}_t, \boldsymbol{\psi}_i)$ 为两个区间的交集长度与并集长度的比值。

5）在上述筛选后，得到最后确定的候选区域。此时每个候选 $\boldsymbol{\psi}_{\mathrm{p}}$ 可以被表示为 $\boldsymbol{\psi}_{\mathrm{p}} = (t_{\mathrm{s}}, t_{\mathrm{e}}, p'_{\mathrm{f}})$，$p'_{\mathrm{f}}$ 表示候选区域的最终得分，将用于评估动作时域区域生成任务时的候选区域排序。为了评估时序动作检测效果，按照惯例（Lin 等，2018），本文将生成的候选区域与 Xiong 等人（2016）的分类模型相结合。具体而言，统计视频得分最高的类别得分为 p_{c1}，用于评估时序动作检测性能的动作类别检测得分的计算式为

$$p_{\mathrm{F}} = p'_{\mathrm{f}} \cdot p_{c1} \tag{6}$$

式中，p'_{f} 为时域候选经过式（5）处理后的置信度得分。

2.3.2　模型优化函数

本文中，时域候选优化采用 LAEM 模块和 GAEM 模块进行联合训练，即

$$L = L_{\mathrm{LAEM}} + L_{\mathrm{GAEM}} \tag{7}$$

式中，L_{LAEM} 和 L_{GAEM} 分别为 LAEM 模块和 GAEM 模块的目标函数，L 为这两部分的总代价函数。时序动作检测采用基于区域级别的特征 $\boldsymbol{f}_{\mathrm{pl}}$ 和 Xiong 等人（2016）的分类模型来进行分类，模型中采用 MLP 优化目标函数 L_{MLP} 来训练并预测候选区域的置信度得分。

1）LAEM 模块优化函数。为了实现 LAEM 模块中边界标签的优化函数，本文对动作的边界标签 $R_{\mathrm{g}} = \{\boldsymbol{r}_{\mathrm{g},k}\}_{k}^{N} = 1$ 进行扩展。对于动作区间 $\boldsymbol{r}_{\mathrm{g},k}$，持续时间为 $d = t_{\mathrm{g,e}} - t_{\mathrm{g,s}}$。数据预处理如下：根据动作中心区域 $\boldsymbol{r}_{\mathrm{g},k}$，记动作起始区域为 $\boldsymbol{r}_{\mathrm{g}}^{\mathrm{s}} = [t_{\mathrm{g,s}} - d/20, t_{\mathrm{g,s}} + d/20]$，动作结束区域为 $\boldsymbol{r}_{\mathrm{g}}^{\mathrm{e}} = [t_{\mathrm{g,e}} - d/20, t_{\mathrm{g,e}} + d/20]$。此外，对于任意时序位置 $t \in \boldsymbol{r}_{\mathrm{g},k}$，计算中心区域 $\boldsymbol{r}_{\mathrm{g},k}$ 的最大交叠率（intersection-over-anchor，IoA），并记做动作得分标签 $p_{\mathrm{g},t}^{\mathrm{a}}$；其中，IoA 值定义为两个区间交集长度与时序位置 t 的区间长度的比值。因此视频动作得分标签为 $\boldsymbol{P}_{\mathrm{G,A}} = \{p_{\mathrm{g},t}^{\mathrm{a}}\}_{t=1}^{T}$。类似地，对于起始区域 $\boldsymbol{r}_{\mathrm{g}}^{\mathrm{s}}$ 和结束区域 $\boldsymbol{r}_{\mathrm{g}}^{\mathrm{e}}$ 也分别与每个时序位置 t 计算 IoA 值，得到动作起始得分标签 $\boldsymbol{P}_{\mathrm{G,S}} = \{p_{\mathrm{g},t}^{\mathrm{s}}\}_{t=1}^{T}$ 和结束得分标签 $\boldsymbol{P}_{\mathrm{G,E}} = \{p_{\mathrm{g},t}^{\mathrm{e}}\}_{t=1}^{T}$。

LAEM 模块输出的 3 条曲线为 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{l}}$、$\boldsymbol{P}_{\mathrm{S}}$ 和 $\boldsymbol{P}_{\mathrm{E}}$，再以 $\boldsymbol{P}_{\mathrm{G,A}}$、$\boldsymbol{P}_{\mathrm{G,S}}$ 和 $\boldsymbol{P}_{\mathrm{G,E}}$ 作为标签，则代价函数定义为

$$L_{\mathrm{LAEM}} = 2 \times L_{\mathrm{bl}}^{\mathrm{action}} + L_{\mathrm{bl}}^{\mathrm{start}} + L_{\mathrm{bl}}^{\mathrm{end}} \tag{8}$$

式中，L_{bl} 是一个交叉熵函数，具体定义为

$$L_{\mathrm{bl}} = \frac{1}{l_{\omega}} \sum_{i=1}^{l_{\omega}} \left(\frac{1}{\alpha^{+}} \cdot b_i \cdot \ln(p_i) + \frac{1}{\alpha^{-}} \cdot (1 - b_i) \cdot \ln(1 - p_i) \right) \tag{9}$$

式中，p_i 是 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{l}}$、$\boldsymbol{P}_{\mathrm{S}}$ 和 $\boldsymbol{P}_{\mathrm{E}}$ 曲线上的值，b_i 是一个二值函数，具体为

$$b_i = \mathrm{sign}(p_{\mathrm{g},i} - \tau) \tag{10}$$

式中，$p_{\mathrm{g},i}$ 是 $\boldsymbol{P}_{\mathrm{G,A}}$、$\boldsymbol{P}_{\mathrm{G,S}}$ 和 $\boldsymbol{P}_{\mathrm{G,E}}$ 曲线上的值，并以 τ 为阈值将标签中的概率得分转换为 $\{0,1\}$，l_{ω} 表示样本点总数，而 α^{+} 和 α^{-} 分别代表标签值大于 τ 的样本点和标签值小于 τ 的样本点占总样本的比例，这些参数是为了平衡正负样本的影响而设置的。本文中 τ 设为 0.5。

2）GAEM 模块优化函数。GAEM 模块以动作的类别信息 $\boldsymbol{C}_{\mathrm{g}} = \{c_k\}_{k=1}^{N}$ 作为标签，并使用 CTC 作为代价函数来预测每个时序位置上的动作概率得分，具体为

$$L_{\mathrm{GAEM}} = -\ln P(\boldsymbol{C}_{\mathrm{g}} \mid \boldsymbol{F}) \tag{11}$$

式中，$\boldsymbol{F}$ 为输入特征序列，$\boldsymbol{C}_{\mathrm{g}}$ 为目标标签序列，$P(\boldsymbol{C}_{\mathrm{g}} \mid \boldsymbol{F})$ 表示基于输入序列 $\boldsymbol{F}$ 观测到 $\boldsymbol{C}_{\mathrm{g}}$ 的条件概率。CTC 将针对输入序列到输出序列的所有对齐方式进行动态规划的概率路径解析，并得到全局最优路径，即全局时序序列。

3）MLP 模块优化函数。MLP 预测候选区域的置信度得分 p_{conf}，采用均方误差作为代价函数，将候选区域 $\boldsymbol{\psi}_{\mathrm{p}}$ 与所有动作区间的最大 IoU 值作为这个候选的标签 p_{g}，即

$$L_{\mathrm{MLP}} = \frac{1}{N_{\mathrm{train}}} \sum_{i=1}^{l_{\omega}} (p_{\mathrm{conf}} - p_{\mathrm{g}})^2 \tag{12}$$

2.4　本文算法

本文算法的具体步骤如下：

输入：视频特征序列 $\boldsymbol{F} = \{f_t\}_{t=1}^{T}$，视频标签集 $\boldsymbol{G} = (\boldsymbol{R}_{\mathrm{g}}, \boldsymbol{C}_{\mathrm{g}})$

输出：动作候选区域 $\boldsymbol{\psi}_{\mathrm{p}}$，及相应候选得分 $\boldsymbol{p}_{\mathrm{F}}$

1）for *epoch* = 1, 2, …, *E* do

2）　for *video* = 1, 2, …, *N* do

3）　局部动作概率预测模块 LAEM 预测该视频的起始概率曲线 $\boldsymbol{P}_{\mathrm{S}}$，结束概率曲线 $\boldsymbol{P}_{\mathrm{E}}$，局部动作概率曲线 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{l}}$；

4) 全局动作概率预测模块 GAEM 依据 CTC 优化解码的结果预测该视频的全局动作概率曲线 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{g}}$;

5) 依据式(7)—(10),计算代价函数 loss,反向传播并更新模型权重;

6) end for

7) end for

8) 结合起始和结束概率曲线 $\boldsymbol{P}_{\mathrm{S}}$, $\boldsymbol{P}_{\mathrm{E}}$构建候选集 $\boldsymbol{\Psi}_{\mathrm{p}}$;

9) $\boldsymbol{P}_{\mathrm{A}} = (\boldsymbol{P}_{\mathrm{A}}^{\mathrm{l}} + \boldsymbol{P}_{\mathrm{A}}^{\mathrm{g}})/2$; //$\boldsymbol{P}_{\mathrm{A}}$为融合后的动作概率曲线;

10) $\boldsymbol{f}_{\mathrm{pl}} = PPM(\boldsymbol{P}_{\mathrm{A}}, \boldsymbol{\Psi}_{\mathrm{p}})$; //*PPM* 为后处理模块,$\boldsymbol{f}_{\mathrm{pl}}$为候选区域级特征;

11) $\boldsymbol{p}_{\mathrm{f}} = MLP(\boldsymbol{f}_{\mathrm{pl}})$; //*MLP* 为 *PPM* 中的多层感知器,$\boldsymbol{p}_{\mathrm{f}}$为候选区域得分;

12) soft-NMS 筛选策略,调整候选得分并排序;

13) 结合分类模型得到最终的类别检测得分 $\boldsymbol{p}_{\mathrm{F}}$。

3 实验及结果分析

3.1 数据集

在 ActivityNet v1.3 数据集上验证本文方法的有效性,包括评估动作候选区域生成和动作检测的性能。ActivityNet v1.3 数据集由从 YouTube 收集的 19 994 个视频组成,共 200 个动作类别。该数据集分为训练、验证和测试 3 个子集,所占比例分别为 50%、25% 和 25%。每个视频至少包含 1 个动作实例,最多为 24 个动作实例数,动作分布差异较大。视频的平均长度为 116.7 s,最短长度为 1.579 s,最长视频为 975 s,平均每个视频至少有超过 36% 的背景帧。

3.2 评价指标

采用 AR@AN(average recall@average number of proposals)、AUC(area under curve)、平均 mAP(mean average precision)和不同阈值下的均值平均精度(mAP@IoU)等 4 个指标对本文方法进行评估,前两个指标评估生成动作候选区域任务,后两个指标评估动作类别检测任务。AR@AN 评估召回率和候选区域数量的关系。平均召回率(average recall, AR)可以看做是候选区域的平均数量(average number, AN)的函数。按照评估惯例(Zhao 等,2017),本文设置 AN 为 100, IoU 阈值集为[0.5∶0.05∶0.95];计算不同阈值下的 AR,取最后的平均值作为AR@AN。AUC 即为AR@AN曲线下的面积。mAP@IoU 中使用的 IoU 阈值集为{0.5, 0.75, 0.95},平均 mAP 中使用的阈值集为[0.5∶0.05∶0.95]。

3.3 模型设置

本文将视频以帧间隔 δ 切割成长度为若干片段序列,δ 设置为 16;再使用双流网络(Simonyan 和 Zisserman,2016)提取视频特征;最后,根据 Lin 等人(2017)提出的线性插值法设置将视频特征长度统一调节到 100,即 $T = 100$。2.3.1 节中设置的 MLP 为只含一层神经元隐藏层的 MLP。模型训练过程中,采用 Adam(adaptive moment estimation)模型优化器;每个训练批次包含 16 组数据,训练 20 轮。LAEM 模块和 GAEM 模块进行联合训练,初始学习率设为 0.001,LAEM 模块每 7 轮学习率调整为初始值的 1/10,GAEM 模块每 10 轮将学习率调整为初始值的 1/10。PPM 模块中 MLP 的初始学习率设为 0.01,在第 10 轮训练后调整为 0.001。此外,soft-NMS 算法中参数 ε 设置为 0.75,阈值 θ 设置为 0.8。

3.4 实验结果

3.4.1 动作时域区域生成评估

将本文方法与 SSN(structured segment network)(Zhao 等,2017)、TCN(temporal context network)(Dai 等,2017)、Prop-SSAD(single shot action detector for proposal)(Lin 等,2017)、CTAP(complementary temporal action proposal)(Gao 等,2018)和 BSN(boundary sensitive network)(Lin 等,2018)等方法在 ActivityNet v1.3 数据集上进行时域区域生成性能对比实验,结果如表 1 所示。可以看出,本文方法在评价指标 AR@100 和 AUC 中均优于其他方法。与其他方法中效果最好的 BSN 相比,本文方法将 AR@100 提高了 0.5,将 AUC 指标从 66.17 提高到 66.32。虽然其他方法同样采用双流网络提取视频特征,但本文方法表现更好,因为其他方法只利用了卷积层,有限的感受野使得模型一次只能注意到视频中的一小段区域,丢失了视频的全局时序信息及候选的上下文信息。而在本文方法中,LAEM 模块基于 CNN,强调时序信息的局部特征响应,并利用局部信息预测候选区域的边界概率和动作概率;GAEM 模块基于 BLSTM,利用前后回顾的记忆单元处理长视频中的时序关联性,捕捉视频的全局信息

表 1　在 ActivityNet v1.3 数据集上的时域区域生成性能对比

Table 1　Comparison on ActivityNet v1.3 in temporal proposal generation

/%

方法	AR@100	AUC
SSN	63.52	53.02
TCN	—	59.58
Prop-SSAD	73.01	64.40
CTAP	73.17	65.72
BSN	74.16	66.17
本文	**74.66**	**66.32**

注：加粗字体表示最优结果，“—”表示无法获取数据。

并准确预测候选区域的动作概率。基于这两个模块对局部信息和全局信息的整合，本文方法得以结合概率曲线得到具有更准确、更灵活的动作时序边界的候选区域，可以覆盖视频中多种长度间隔的动作检测。

3.4.2　时序动作检测评估

将本文方法与 SCC（semantic cascade context）（Heilbron 等，2017）、CDC（convolutional de-convolutional network）（Shou 等，2017）、SSN（Zhao 等，2017）和 BSN（Lin 等，2018）等方法在 ActivityNet v1.3 数据集上进行时序动作检测性能对比实验，结果如表 2 所示。表中阈值表示提取的候选与真实动作边界的重叠度，即阈值为 0.5 时，只有与真实动作区间的交集和并集之比大于 1/2 的候选区域才能被判定为正样本；而在阈值为 0.95 时，只有重叠率大于 0.95 的候选区域才能被判定为正样本。阈值越高，筛选条件越严格。

表 2　在 ActivityNet v1.3 数据集上的时序动作检测性能对比

Table 2　Comparison on ActivityNet v1.3 in temporal action detection

/%

方法	阈值			平均 mAP
	0.5	0.75	0.95	
SCC	40	17.9	4.7	21.7
CDC	43.83	25.88	0.21	22.77
SSN	39.12	23.48	5.49	23.98
BSN	**46.45**	29.96	8.02	30.03
TPO	46.23	**30.73**	**8.22**	**30.5**

注：加粗字体表示最优结果。

从表 2 可以看出，当阈值为 0.5 时，本文方法的行为分类性能低于 BSN，意味着在对重叠度要求较低时，BSN 能够较为准确地捕捉局部时序信息，提取更多符合条件的候选区域。当阈值为 0.75 及 0.95 时，本文方法实现的效果优于其他所有方法，意味着在筛选条件严格时，BSN 只利用局部时序信息的弊端逐渐显现，而本文方法结合了局部时序信息和全局时序信息，能够预测更为准确的动作候选区域。

精度平均指标 mAP 是模型方法在[0.5∶0.05∶0.95]共 10 个阈值上的 mAP 均值，如图 2 所示，随着阈值的提高，mAP 的值在降低。

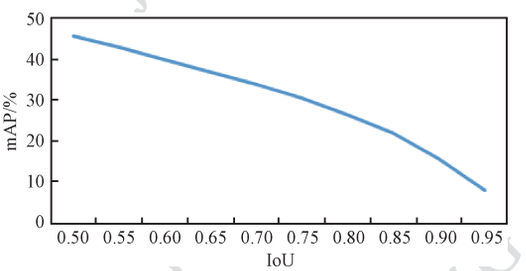


图 2　mAP 与 IoU 关系图

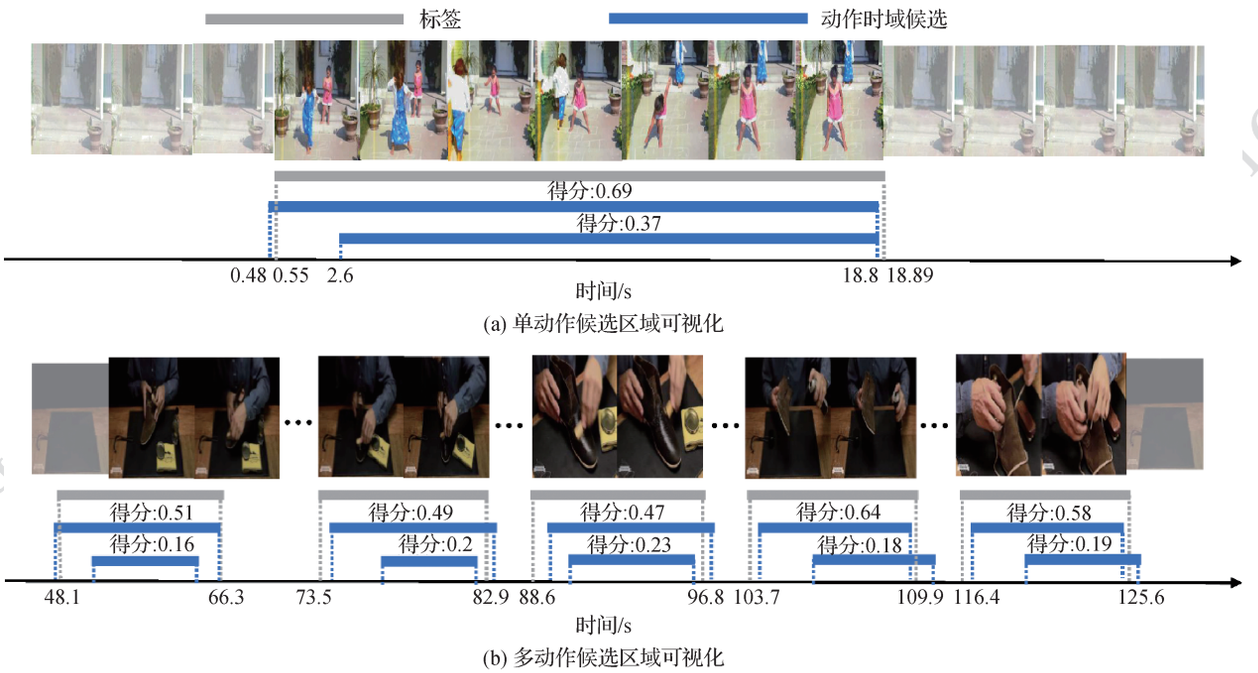
Fig. 2　The relation between mAP and IoU

平均 mAP 综合评估了模型在不同情况下的总体表现。与 SCC、CDC、SSN 及 BSN 等方法相比，本文方法的平均 mAP 分别提高了 8.8%、7.73%、6.52%、0.47%。其中，SCC 通过探索动作与物体、动作与场景之间的关系来提高精度，但忽略了动作与动作之间的关系；CDC 是在传统 3D CNN 上的改进，一次只输入一个视频段，而不是完整的视频，因此丢失了长时序列中的时序关联性；SSN 用一帧的信息代表一个片段的信息，丢失过多时序信息；BSN 受限于有限的 CNN 局部感受野大小，虽然捕捉了视频的局部时序信息，却在一定程度上丢失了视频的全局信息；而本文方法兼顾了视频的局部特征响应和全局的时序关联性，因此在对候选区域与真实动作边界重叠度要求更高，即条件更苛刻时，能够实现比其他方法更好的效果。进一步表明本文方法能够生成高质量的候选区域，并显著提高了动作检测的效果。

3.4.3　动作候选区域结果可视化示例

为了进一步表明本文方法在动作候选区域的有效性，图 3 给出了本文方法生成的候选区域的可视

化示例图。图3(a)整段视频只包含一段动作实例，动作是游戏跳房子；图3(b)整段视频包含了多段动作实例，动作是清洁鞋。具体而言，游戏跳房子的动作幅度较大，相对比较容易识别，而清洁鞋动作幅度较小，且室内光线较暗，使得模型难以捕获有效信息。实验结果表明，本文方法在这两个样例中都生成了与真实标签高度重合的候选区域，并且给重合度低的候选区域打了相对低分。这意味着本文方法具有良好的鲁棒性，能够有效捕获具有精确时间边界的动作候选区域，可以适应动作分布不同的视频。此外，各区域置信度得分也同样验证了本文方法在候选区域划分上的可靠性。



(a) 单动作候选区域可视化

(b) 多动作候选区域可视化

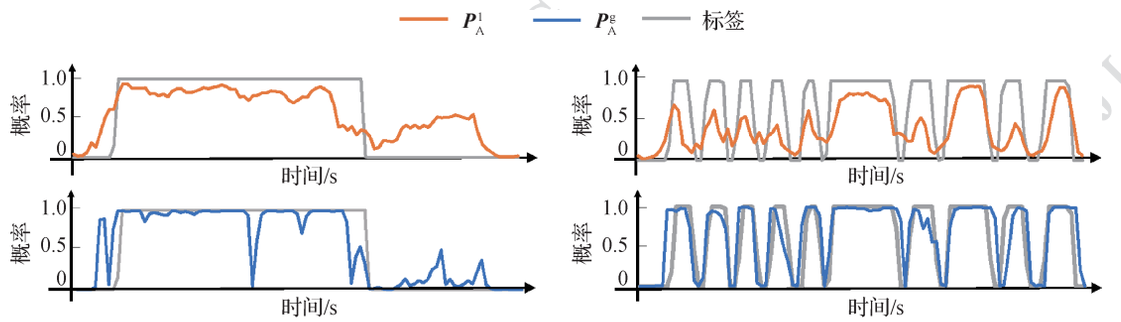
图3　TPO在ActivityNet v1.3数据集上生成的候选区域样例

Fig. 3　Visualization examples of candidate action proposals generated by TPO on ActivityNet v1.3 dataset

((a) visualization of proposals in single action instance; (b) visualization of proposals in multiple action instances)

3.4.4　动作概率曲线结果可视化示例

为进一步验证本文方法对时序动作检测的有效性，图4给出了本文方法单动作和多动作实例下的动作概率曲线，图中灰色、橘色、蓝色的线段分别表示时间维度上动作概率曲线真实标签、LAEM模块生成的动作概率曲线 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{l}}$ 以及GAEM模块生成的动作概率曲线 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{g}}$。图4(a)(b)分别针对视频中只有一段动作实例和有多段动作实例的情况。概率曲线上的概率值均在0～1范围内，概率值越高，意味着此时序位置包含动作的可能性越大。



(a) 单动作实例　　(b) 多动作实例

图4　单动作实例及多动作实例下的动作概率曲线

Fig. 4　Visualization examples of actionness probability curves

((a) single action instance; (b) multiple action instances)

图 4(a)为单动作样例的动作概率曲线。在动作持续期间内，$\boldsymbol{P}_{\mathrm{A}}^{\mathrm{l}}$ 的概率值较高，但在动作结束后响应值依然较高，意味着 LAEM 模块能够识别出动作实例，但是基于局部时序信息的学习，缺乏对近似延续动作的判别，无法准确界定动作区间。而 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{g}}$ 在动作持续期间内表现出了与真实标签一致的接近于 1 的概率值，意味着 GAEM 模块通过对全局信息的学习，能够较为准确地切分动作区间。图 4(b)为多动作样例的动作概率曲线，动作实例出现较为密集。$\boldsymbol{P}_{\mathrm{A}}^{\mathrm{l}}$ 的概率值起伏与标签基本一致，但响应值不高，意味着 LAEM 模块能够察觉多动作实例及动作之间的间隔，但依然无法准确界定区间。而 $\boldsymbol{P}_{\mathrm{A}}^{\mathrm{g}}$ 的概率值起伏与标签基本完全一致，意味着 GAEM 模块能够非常敏锐地察觉到动作的出现，即证明了通过对全局信息的学习，GAEM 模块能够学习到动作之间的关系，并证明了 CTC 的有效性。为了兼顾局部信息和全局信息的利用，本文最后采用的设置是融合 GAEM 模块和 LAEM 模块的动作概率曲线，并实现了更好的效果。

4　结　论

本文提出了一种优化时序动作候选区域，进而提高时序动作检测精度的新方法。TPO 模型采用 CNN 和 BLSTM 来兼顾视频的局部时序关联性和全局时序信息，引入全局时域联级分类优化方法 CTC 实现了片段级别的边界概率和动作概率预测，最后得到了准确的动作候选区域和可靠的置信度得分。在 ActivityNet v1.3 数据集上的实验验证本文方法具有良好性能，在多个评价指标上均有提升。此外在基于候选区域的动作检测方面，同样验证了本文方法的有效性。

本文提出了能够提取高质量动作候选区域的时序动作检测模型 TPO，但仍存在不足之处。未来将从以下几个方面进一步改进：改进提取的全局时序序列与局部时序序列的融合方式；增加对提取的动作候选区域的完整性检查；尝试采用其他方式提取视频的全局时序序列，如扩张卷积。

参考文献(References)

Bodla N, Singh B, Chellappa R and Davis L S. 2017. Soft-NMS—improving object detection with one line of code//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5562-5570 [DOI: 10.1109/ICCV.2017.593]

Cao S Y, Liu Y H and Li X Z. 2017. Vehicle detection method based on Fast R-CNN. Journal of Image and Graphics, 22(5): 671-677 (曹诗雨, 刘跃虎, 李辛昭. 2017. 基于 Fast R-CNN 的车辆目标检测. 中国图象图形学报, 22(5): 671-677) [DOI: 10.11834/jig.160600]

Cui R P, Liu H and Zhang C H. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1610-1618 [DOI: 10.1109/CVPR.2017.175]

Dai X Y, Singh B, Zhang G Y, Davis L S and Chen Y Q. 2017. Temporal context network for activity localization in videos//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5727-5736 [DOI: 10.1109/ICCV.2017.610]

Gao J Y, Chen K and Nevatia R. 2018. CTAP: complementary temporal action proposal generation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 70-85 [DOI: 10.1007/978-3-030-01216-8_5]

Heilbron F C, Barrios W, Escorcia V and Ghanem B. 2017. SCC: Semantic context cascade for efficient action detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3175-3184 [DOI: 10.1109/CVPR.2017.338]

Heilbron F C, Escorcia V, Ghanem B and Niebles J C. 2015. Activitynet: a large-scale video benchmark for human activity understanding//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 961-970 [DOI: 10.1109/CVPR.2015.7298698]

Huang S, Wang W Q, He S F and Lau R W H. 2018. Egocentric temporal action proposals. IEEE Transactions on Image Processing, 27(2): 764-777 [DOI: 10.1109/tip.2017.2772904]

Lin T W, Zhao X and Shou Z. 2017. Temporal convolution based action proposal: submission to ActivityNet 2017 [EB/OL]. [2019-08-14]. https://arxiv.org/pdf/1707.06750.pdf

Lin T W, Zhao X, Su H S, Wang C J and Yang M. 2018. BSN: boundary sensitive network for temporal action proposal generation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 3-21 [DOI: 10.1007/978-3-030-01225-0_1]

Luo H L, Lai Z Y and Kong F S. 2017. Action recognition in videos based on action segmentation and manifold metric learning. Journal of Image and Graphics, 22(8): 1106-1119 (罗会兰, 赖泽云, 孔繁胜. 2017. 动作切分和流形度量学习的视频动作识别. 中国图象图形学报, 22(8): 1106-1119) [DOI: 10.11834/jig.170032]

Luo H L, Tong K and Kong F S. 2019. The progress of human action

recognition in videos based on deep learning: a review. Acta Electronica Sinica, 47(5): 1162-1173 (罗会兰, 童康, 孔繁胜. 2019. 基于深度学习的视频中人体动作识别进展综述. 电子学报, 47(5): 1162-1173) [DOI: 10.3969/j.issn.0372-2112.2019.05.025]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/tpami.2016.2577031]

Shou Z, Chan J, Zareian A, Miyazawa K and Chang S F. 2017. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1417-1426 [DOI: 10.1109/cvpr.2017.155]

Shou Z, Wang D G and Chang S F. 2016. Temporal action localization in untrimmed videos via multi-stage CNNs//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1049-1058 [DOI: 10.1109/cvpr.2016.119]

Simonyan K and Zisserman A. 2016. Two-stream convolutional networks for action recognition in videos//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press: 568-576

Singh B, Marks T K, Jones M, Tuzel O and Shao M. 2016. A multi-stream bi-directional recurrent neural network for fine-grained action detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1961-1970 [DOI: 10.1109/cvpr.2016.216]

Song S J, Lan C L, Xing J L, Zeng W J and Liu J Y. 2018. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. IEEE Transactions on Image Processing, 27(7): 3459-3471 [DOI: 10.1109/tip.2018.2818328]

Soomro K, Idrees H and Shah M. 2019. Online localization and prediction of actions and interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 459-472 [DOI: 10.1109/tpami.2018.2797266]

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/iccv.2015.510]

Tran D, Wang H, Torresani L, Ray J, LeCun Y and Paluri M. 2018. A closer look at spatiotemporal convolutions for action recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6450-6459 [DOI: 10.1109/cvpr.2018.00675]

Tu Z G, Li H Y, Zhang D J, Dauwels J, Li B X and Yuan J S. 2019. Action-stage emphasized spatiotemporal VLAD for video action recognition. IEEE Transactions on Image Processing, 28(6): 2799-2812 [DOI: 10.1109/TIP.2018.2890749]

Xiong Y J, Wang L M, Wang Z, Zhang B W, Song H, Li W, Lin D H, Qiao Y, van Gool L and Tang X O. 2016. CUHK and ETHZ and SIAT submission to ActivityNet challenge 2016 [EB/OL]. [2019-08-14]. https://arxiv.org/pdf/1608.00797.pdf

Xu B H, Ye H, Zheng Y B, Wang H, Luwang T Y and Jiang Y G. 2019. Dense dilated network for video action recognition. IEEE Transactions on Image Processing, 28(10): 4941-4953 [DOI: 10.1109/tip.2019.2917283]

Xu H J, Das A and Saenko K. 2017. R-C3D: region convolutional 3D network for temporal activity detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5794-5803 [DOI: 10.1109/ICCV.2017.617]

Yang Y, Zhou J, Ai J B, Hanjalic A, Shen H T and Ji Y L. 2018. Video captioning by adversarial LSTM. IEEE Transactions on Image Processing, 27(11): 5600-5611 [DOI: 10.1109/TIP.2018.2855422]

Zhao Y, Xiong Y J, Wang L M, Wu Z R, Tang X O and Lin D H. 2017. Temporal action detection with structured segment networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2933-2942 [DOI: 10.1109/ICCV.2017.317]

Zhu H Y, Vial R, Lu S J, Peng X, Fu H Z, Tian Y H and Cao X B. 2018. YoTube: searching action proposal via recurrent and static regression networks. IEEE Transactions on Image Processing, 27(6): 2609-2622 [DOI: 10.1109/tip.2018.2806279]

作者简介

熊成鑫,1995 年生,女,硕士研究生,主要研究方向为时序动作检测。
E-mail: XiongCx@mail.hfut.edu.cn

郭丹,通信作者,女,副研究员,主要研究方向为视频分析、模式识别、深度学习。
E-mail: guodan@hfut.edu.cn

刘学亮,男,副教授,主要研究方向为多媒体信息检索。
E-mail: Liuxueliang1982@gmail.com