|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **文章信息** | **背景、目的及结论** | **结果与讨论** | **文章好在哪里** | **自我想法** |
| **作者**：Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo  **单位**：the Department of Computer Science, University of Rochester, Rochester, NY, 14627.  **来源**：TPAMI  **题目**：Multi-Scale 2D Temporal Adjacent Networks for Moment Localization with Natural Language | **背景**：现存方法没有充分考虑时间点的上下文信息(again)且存在冗余帧问题。另外，若直接将原视频输入模型处理，则显存不够。  **目的**：使用自然语言进行moment精准定位并提升训练效率。  **结论**：提出2D特征映射，将一段视频用2D坐标系表示，横轴表示moment持续时间，纵轴表示moment开始时间。为了扩大感受野和以及获取丰富的上下文信息，对2D特征映射做了尺度变换。 | 在三个benchmark(Charades-STA, ActivityNet Captions, and TACoS)上超过了之前的SOTA模型性能。  暂无讨论：   1. 模型没有使用多模态融合方法(CLIP的混合特征空间) 2. 在提取输入句子特征时使用了LSTM网络，可以试试SA处理。并行化更好。 | 1. 改进了之前的视频表征方法，使用多尺度2D特征映射来表示视频。 2. 抛弃了之前的双阶段处理过程(先使用滑动窗口从全视频中截取moment切片，然后将切片逐一和查询语句对比，找出最佳匹配mo’men’t切片作为结果)。基于2D特征映射来建模而非一维顺序映射。另外，为了进行细粒度查询以及减少计算复杂度，作者还构建了多尺度2D特征映射(改变基础时间步) | **一个思路**：  使用transformer来并行提取语句特征。  **两个图表**： |