基于视觉 – 文本关系对齐的跨模态视频片段检索

陈卓, 杜昊, 吴雨菲, 徐童*, 陈恩红

中国科学技术大学大数据分析与应用安徽省重点实验室, 合肥 230027

* 通信作者. E-mail: tongxu@ustc.edu.cn

收稿日期: 2019-12-31; 修回日期: 2020-03-22; 接受日期: 2020-04-22; 网络出版日期: 2020-06-10

国家重点研发计划 (批准号: 2018YFB1004300) 和国家自然科学基金 (批准号: 61703386, U1605251) 资助项目

摘要 近年来, 视频数据资源的日益丰富催生了一系列对于视频片段精细检索的需求. 在这样的背景下, 对于跨模态视频片段检索的研究逐渐兴起, 其旨在根据输入的查询文本, 输出一段视频中符合文本描述的片段. 现有的研究工作主要关注于查询文本与视频片段的全局或局部的特征表达, 而忽略了查询文本与视频片段中所蕴含的语义关系在跨模态检索中的匹配. 例如, 给定查询文本“一个人在打篮球”时, 现有检索系统将根据整个查询文本和的视频的特征, 或者关注于文本与视频中所表现的实体 (如“人”, “篮球”) 来计算合适的视频片段, 而缺乏对于“人打篮球”这类语义关系的考虑. 因此, 他们将难以辨别语义关系上的不同, 从而限制了检索质量的提升. 为了解决这个问题, 本文提出跨模态关系对齐的图卷积框架 CrossGraphAlign, 通过分别构建文本关系图 (textual relationship graph) 与视觉关系图 (visual relationship graph) 来建模查询文本与视频片段中的语义关系, 再通过跨模态对齐的图卷积网络来评估文本关系与视觉关系的相似度, 从而帮助构建更加精准的视频片段检索系统. 在公开的跨模态视频片段检索数据集 TACoS 和 ActivityNet Captions 上的实验结果表明, 本文提出的方法可以有效地利用语义关系来提升跨模态视频片段检索的召回率.

关键词 关系对齐, 语言关系, 视觉关系, 图卷积网络, 跨模态视频片段检索

1 引言

在海量视频数据中检索感兴趣的视频是多媒体分析领域的热门议题. 在过去的十年里, 视频检索的相关工作^[1~5]被广泛地研究, 极大地促进了多媒体领域相关技术的发展. 其中, 跨模态视频片段检索^[5,6] (cross-modal video moment retrieval) 因其在细粒度视频检索方面的应用潜力而引起广泛关注. 具体而言, 给定输入的查询文本 (language query), 跨模态视频片段检索期望输出一段视频中符合文本描述的视频片段 (video moment). 如图 1 所示, 给定一个查询文本“一个人在打篮球” (A person is

引用格式: 陈卓, 杜昊, 吴雨菲, 等. 基于视觉 – 文本关系对齐的跨模态视频片段检索. 中国科学: 信息科学, 2020, 50: 862–876, doi: 10.1360/SSI-2019-0292
Chen J, Du H, Wu Y F, et al. Cross-modal video moment retrieval based on visual-textual relationship alignment (in Chinese). Sci Sin Inform, 2020, 50: 862–876, doi: 10.1360/SSI-2019-0292

图 1 (网络版彩图) 跨模态视频片段检索的示例

Figure 1 (Color online) An example for cross-modal video moment retrieval

playing basketball), 跨模态视频片段检索将输出 41.8~53.6 s 这个时间片段, 在这个时间段内视频中的人正在进行篮球运动。

尽管跨模态视频片段检索任务的定义简洁, 应用前景广阔, 它却是一个极具挑战的任务. 已有研究^[5~7]表明, 为了实现视觉信息与文本信息之间的匹配, 视频与查询文本的特征表达方式需要精心设计. 随着以深度神经网络^[8~10]为代表的机器学习方法在物体识别^[11]、动作分类^[12]、词向量^[13,14]、逻辑推理^[15]、视频理解^[16,17]等领域的发展, 研究者们开始关注如何利用其构建的视频与查询文本的特征表达方式. 早期的工作^[5,6]确立了跨模态视频片段检索任务中视频与查询文本特征表达的基本范式: (1) 将视频分割为多个候选视频片段, 采用动作识别的 C3D^[12] 或者双流网络^[18,19] 来提取候选视频片的特征; (2) 采用词向量^[13] 与 LSTM^[20] 来提取查询文本的特征; (3) 将视频与文本的特征联合映射到同一空间来形成检索特征^[5], 或者是计算两者特征之间的距离^[6]. 在上述范式的基础上, 部分研究者开始探索更好的特征表达方式^[7,21~23], 如 Liu 等^[22] 与 Yuan 等^[23] 提出利用注意力机制^[24] 来增强视频片段与查询文本的特征表达; 另一部分研究者关注如何对齐细粒度的视觉信息与文本信息来提升检索质量^[25~27], 如 Jiang 等^[26] 与 Wang 等^[27] 引入目标检测网络 Faster R-CNN^[11] 来检测视觉物体, 从而匹配文本中对应描述的实体. 在流行的跨模态视频片段检索数据集^[5,6,28,29] 上, 基于细粒度表达的检索方法取得了领先的性能水准, 充分说明了该表达方式的必要性.

然而, 上述细粒度的表达方式仅仅从独立的角度来考虑视频与文本中出现的物体, 缺乏对物体之间语义关系的表达, 这可能会限制跨模态视频片段检索质量的提升. 在视觉层面, 这种被忽略的语义关系刻画了视觉物体间的交互; 在文本层面, 这种被忽略的语义关系描述了语言实体间的联系. 它所具有的典型特点为精确性: 如, 查询文本为“一个人在打篮球”, 期望检索出的视频片段包含“人打篮球”而不是“人抱篮球”, “人打乒乓球”的关系; 最近涌现的一部分工作^[26,27,30,31] 考虑了视觉中的目标检测, 遗憾的是, 它们未能分析文本中的实体, 并将视觉中的语义关系与文本中的语义关系对齐.

借助于图 (graph) 结构在表达关系时的强大能力^[32,33] 与图卷积网络^[34,35] 的近年发展, 本文提出基于视觉 - 文本关系对齐的图卷积框架 CrossGraphAlign, 用以赋予现有跨模态视频片段检索框架中匹配语义关系的能力. 具体而言, CrossGraphAlign 首先为查询文本与待检索视频分别生成文本关系图与视觉关系图; 接着, CrossGraphAlign 中的视觉 - 文本关系对齐的图卷积网络将试图匹配一段时间内的文本关系图与视觉关系图; 最后, 基于匹配结果, CrossGraphAlign 将预测查询文本在视频中的起始时间和结束时间.

本文提出的方法在流行的跨模态视频片段检索数据集 TACoS^[28] 和 ActivityNet Captions^[36] 上进行了实验. 实验结果表明, 在基线模型中加入 CrossGraphAlign 后, 模型在各个交叠比下的召回率得到了明显提升. 在现有的跨模态视频片段检索方法中, 本文提出的模型达到了最佳的效果.

2 相关工作

与本文相关的研究工作包括基于查询文本的图像或视频检索、视频中的动作定位、视觉关系或文本关系建模、基于自然语言的视频时刻定位, 下面分别对其相关工作展开讨论。

2.1 基于查询文本的图像或视频检索

给定一组候选的图像或视频以及自然语言查询语句, 图像或视频检索的目标是选择和查询语句最匹配的图像或视频。在图像检索方面, Wang 等^[37]提出了利用图结构来对图像中的物体之间的关系以及文本进行建模, 着重挖掘图像和文本之间的对齐的关系。Lee 等^[38]将注意力机制应用到图像检索的问题中, 先提取图像和句子的特征, 然后对每个区域和词语对应地使用注意力机制, 然后再计算相似度, 利用注意力机制进行更加准确的对齐。在视频检索方面, Lin 等^[39]提出的模型首先将查询语句解析为语义图, 然后将其与视频中的视觉概念进行匹配。Bojanowski 等^[40]提出了一种策略, 通过为给定视频的每个句子分配时间间隔以及按时间顺序排列一组句子解决视频文本对齐问题。与其不同的是, 我们关注于视频片段上的检索, 这样一个任务更为细粒度与具有挑战性。

2.2 视频中的动作定位

视频中的动作定位旨在解决未裁剪视频中特定动作的定位和识别的问题。Shou 等^[41]用定位损失训练了 3 个基于分段的 C3D^[12], 显式地考虑到视频中的时间重叠。Ma 等^[42]在 RNN 学习目标中引入了排名损失, 以更好地捕捉动作的逐步变化。Singh 等^[43]提出了一种多流双向 RNN, 通过使用动作发出者周围边界框的功能增强全帧图像特征, 实现了细粒度的动作检测。Zhao 等^[44]提出了一种结构的细分网络, 以通过结构化的时间金字塔对动作实例的结构进行建模。Chao 等^[45]通过模仿 Faster R-CNN^[11] 物体检测框架提高动作定位性能。Lin 等^[46]提出了一种基于一维时间卷积层的方法, 该方法跳过了视频片段建议生成步骤, 直接检测未裁剪视频中的动作实例。与其不同的是, 我们关注于面向查询文本的视频片段定位, 需要处理多模态的信息, 因而更具挑战性。

2.3 视觉关系或文本关系建模

在视觉关系建模方面, Dai 等^[47]通过利用两种关系, 即空间关系和统计推断推断一个主谓宾的三元组 (subject, predicate, object)。Yang 等^[48]利用关系推荐网络进行关系过滤, 且作者利用注意力图卷积网络来做信息的融合。Zellers 等^[49]利用语言偏差来模拟符合客观规律的常识, 从而构建符合真实世界分布的场景图。在文本关系建模方面, Schuster 等^[50]对文本进行图结构建模, 挖掘文本的关系, 提高对图像的检索性能。Han 等^[51]不仅利用图结构对文本进行建模, 还考虑了文本中事件发生的先后顺序。本文借鉴了场景图这一结构来表示视觉与文本关系。

2.4 基于自然语言的视频时刻定位

该任务即为跨模态的视频片段检索。给定查询文本与未裁剪的视频, 我们希望定位视频中与其最匹配的片段的开始和结束时间。Gao 等^[5]设计了一种跨模态的时间回归定位器, 估计候选视频片段和文本查询的对齐分数, 然后调整高分视频片段的边界。Hendricks 等^[6]为视频时态的上下文特征和自然语言查询学习共享嵌入, 提出时刻上下文网络通过整合视频的局部和全局特征定位查询语句对应的视频, 使用了排名损失来训练模型。随着视觉和语言交互领域中注意力机制的发展, 解决基于自然语言的视频时刻定位的任务主要分成以下 3 个方面。(1) 充分挖掘视频相邻片段之间的时序关系。Rohrbach 等^[52]对视频采取了相同间隔的划分, 利用一张二维的映射表去建模不同视频片段之间的关

图 2 (网络版彩图) 示意图: 跨模态关系对齐的图卷积框架 CrossGraphAlign. 该方法首先根据查询文本与视频分别生成文本关系图与视觉关系图, 然后通过视觉 - 文本关系图对齐模块来评估与查询文本最符合的视频片段

Figure 2 (Color online) An illustration of CrossGraphAlign for cross-modal relationship alignment. The algorithm firstly constructs textural relationship graph and visual relationship graph from language query and video, respectively. Then, the visual-textural graph alignment module estimates the most similar video moments related to language query

系. Liu 等^[22], Zhang 等^[53] 利用注意力机制或图卷积网络来充分理解视频中的关系. (2) 充分关注文本中的关系. Ge 等^[25] 在查询语句中对动词 - 宾语对进行编码. Zhang 等^[54] 在查询语句的基础上构建了句法关系图, 并应用了多层图卷积网络以提取细粒度的文本特征. (3) 关注查询语句中与视频特征紧密联系的单词. Liu 等^[55] 提出语言时间注意模型, 利用视频中的上下文信息学习重点关注某个单词. Chen 等^[21] 提出的模型捕捉了每一帧的视频和文本查询之间细粒度的关系. 然而, 这些工作忽略了视觉物体之间、文本实体之间语义关系的表达与匹配. 受此启发, 我们提出 CrossGraphAlign 进行跨模态的视觉与文本的关系对齐, 帮助提示检索质量.

3 跨模态关系对齐的图卷积框架 CrossGraphAlign

先前的跨模态视频片段检索方法难以辨别语义关系上的不同, 影响检索质量. 为了解决这个问题, 我们提出跨模态关系对齐的图卷积框架 CrossGraphAlign. 如图 2 所示, CrossGraphAlign 主要包含 3 个模块. (1) 文本关系图构建模块 (textural relationship graph constructor): 输入查询文本, 输出以实体特征为节点, 文本关系特征为边的图. (2) 视觉关系图构建模块 (visual relationship graph constructor): 输入目标检测结果, 构建以视觉物体特征为节点, 视觉关系特征为边的图. (3) 视觉 - 文本关系图对齐模块 (visual-textural graph alignment module): 根据文本关系图与视觉关系图进行双方的匹配, 输出视频中符合查询文本描述的片段. 下面将对上述 3 个模块进行分别介绍.

3.1 文本关系图构建模块

输入查询文本 L , 文本关系图构建模块旨在将其解析为文本关系图 $G^L = (O^L, E^L)$, 其中 $O^L =$

图 3 文本关系图构建模块。首先将文本解析为依存树，而后使用场景图解析提取文本关系，最后使用词向量方法构建文本关系图

Figure 3 Building textual relationship graph. The query text is firstly parsed to a dependency tree, then a scene graph is built on it, and finally the word-to-vector method is used to construct the textual relationship graph

$\{o_1^L, o_2^L, \dots, o_n^L\}$ 表示文本中的实体的特征集合，而 E^L 表示实体间的关系的特征集合，其可以表示为 $E^L \subseteq O^L \times R^L \times O^L$ ，其中 R^L 为实体间的谓语动词。如“一个人在打篮球” (A person is playing basketball)，以 $o_1^L, o_2^L, r_{1,2}^L$ 来分别表示“A person”，“basketball”，“playing”的词向量，其对应的文本关系图为 $G^L = (\{o_1^L, o_2^L\}, \{(o_1^L, r_{1,2}^L, o_2^L)\})$ 。

在先前的工作中，文本场景图解析 (scene graph parsing) [50] 与文本关系图构建模块的目的类似，其旨在根据输入查询文本 L 生成场景图 $G^L = (O^L, E^L)$ ，其中 O^L 和 E^L 表示未经特征化的文本实体集合与文本关系集合。如图 3 所示，基于公开的文本场景图解析器¹⁾，我们先将文本转换为依存树 (dependency tree)，再利用它构建场景图。我们发现场景图中的节点中还包含了属性构成的子节点，为了特征化场景图的简便，我们将属性子节点与实体节点合并为短语实体节点。在这之后，我们基于跨模态视频片段检索中常用的词向量方法 GloVe [24] 来特征化场景图，从而构建出文本关系图。

3.2 视觉关系图构建模块

在进行文本关系图的构建后，我们讨论如何进行视觉关系图的构建。仿照文本关系图中的定义，输入图像 V ，视觉关系图构建模块旨在将其解析为视觉关系图 $G^V = (O^V, E^V)$ ，其中 $O^V = \{o_1^V, o_2^V, \dots, o_n^V\}$ 表示视觉中的实体的特征集合，而 E^V 表示视觉物体间关系的特征集合，其可以表示为 $E^V \subseteq O^V \times R^V \times O^V$ ，其中 R^V 可以理解为视觉物体间的“谓语动词”。

图 4 展示了视觉关系图构建模块的主要框架，该框架主要包含对视觉物体进行的特征提取以及对视觉关系进行的特征提取，下面进行分别介绍。

- 视觉物体特征提取：我们基于 Faster R-CNN [11] 进行视觉物体的特征提取。Faster R-CNN 是一个目标检测框架，其分为两个阶段：(1) 区域推荐阶段用于从海量的预设框 (Anchor [11]) 中提取可能包含目标的区域；(2) 逐区域识别阶段分别对建议区域提取特征，并且基于特征预测目标的精确位置和所属类别。为了能够识别出图像中的物体，我们将 Faster R-CNN [11] 在目标检测中流行的 COCO 数据集 [56] 上进行预训练，这一部分将在实验与分析 (第 4 节) 中进行介绍。在逐区域识别阶段中，RoI Pooling 层 [11] 在特征图 (即，图 4 中的 backbone feature) 上提取每一区域的特征，我们将其做为视觉物体特征的一部分，并且命名为实例特征 (instance feature)。同时，我们还将目标识别出的类别经过 GloVe 方法表征为向量，作为标签特征 (label feature)，与实例特征拼接 (concat) 后一起作为视觉物体特征。为了进一步表示物体的位置信息，我们将物体的归一化的坐标作为位置特征 (location feature)，与上述特征一起拼接，形成视觉物体特征。提取后的视觉物体特征将作为视觉关系图的节点特征 (node feature)，如图 4 所示。

- 视觉关系特征提取：我们基于在场景图生成任务中常用的 Union Box [47~49] 概念进行视觉物体

1) <https://nlp.stanford.edu/software/scenegraph-parser.shtml>.

图 4 (网络版彩图) 视觉关系图构建模块. 改造 Faster R-CNN 框架用于提取物体特征与关系特征

Figure 4 (Color online) Building visual relationship graph. Modifying Faster R-CNN to extract object features and relationship features

间关系的提取. 给定两个目标的边界框 (x_1, y_1, w_1, h_1) 与 (x_2, y_2, w_2, h_2) , Union Box 用于表示两个物体边界框的最小外接框 $(\min(x_1, x_2), \min(y_1, y_2), \max(w_1, w_2), \max(h_1, h_2))$. 在获得 Union Box 后, 我们仿照视觉物体特征提取的方式, 利用 RoI Pooling 提取关系特征为视觉关系图的边特征 E^V . 请注意, 在这里视觉关系特征并没有被显式地表达, 但是它可以从中学习到合适的表达.

3.3 视觉 – 文本关系图对齐模块

在构建了视觉关系图与文本关系图后, 我们将进行它们之间的对齐, 以利用视觉 – 文本关系来提升检索质量. 我们首先改造图卷积网络^[34], 以期对关系图进行自身信息的更新; 而后对视觉关系图与文本关系图进行特征表示, 用于计算匹配度分数. 具体而言, 我们提出以关系为中心的更新 (relationship-centric update) 机制和关系图特征嵌入 (relationship graph embedding), 下面分别进行介绍.

- 以关系为中心的更新: 我们首先对关系图进行信息更新, 使得关系的特征表达更加丰富. 对于节点 o_i , 我们可以利用其邻接节点 $\{o_j \mid j \in \text{adj}(i)\}$ 与邻接边 $\{e_{i,j} \mid j \in \text{adj}(i)\}$ 的特征进行更新; 对于边特征 $e_{i,j}$, 我们可以利用其连接的首尾节点的特征进行更新. 然而, 在节点较多的情况下, 这会引入较大的计算开销. 因此, 我们假设在不同关系下, 实体表示的意义相同, 即“一个人在打篮球”和“一个人在抱篮球”两个句子, “一个人”这一节点表达的意思相同, 而其不同之处主要在于“打”和“抱”的语义关系的区别. 因此, 我们提出以关系为中心的更新方法, 忽略节点特征的更新, 而边特征引入节点

图 5 (网络版彩图) 视觉 - 文本关系图对齐模块. 使用图卷积网络和注意力机制去寻找视觉和文本的对齐关系

Figure 5 (Color online) Align visual-textual relationship graph. Utilizing GCN and attention to find visual-textual relationship feature

特征进行更新:

$$e_{i,j}^U = e_{i,j} + \text{RELU}(W_i o_i + W_j o_j), \quad (1)$$

其中 $e_{i,j}^U$ 表示更新后的边特征, W_i 与 W_j 表示可学习的参数, $\text{RELU}()$ 表示非线性激活. 在经过上述计算后, 关系的特征利用连接节点特征进行了更新, 从而可以更好地表达具体语境中的关系特征. 如图 5 所示, 我们对视觉关系图与文本关系图进行同样方式的更新, 在这之后进行关系图特征嵌入. 由于我们忽略了节点的更新, 因此该处的图卷积网络仅有一层, 从而使得计算过程较为轻量.

• 关系图特征嵌入: 在进行信息更新后, 我们分别讨论视觉关系图以及文本关系图的特征表达. 由于视觉关系图为无向图, 因此我们将视频候选片段中的第 n 帧中每条边 $e_{i,j}^{V_n}$ 以及其连接的节点 $o_i^{V_n}$, $o_j^{V_n}$ 表达为 $f_{i,j}^{V_n} = \langle o_i^{V_n}, e_{i,j}^{V_n}, o_j^{V_n} \rangle$, $f_{j,i}^{V_n} = \langle o_j^{V_n}, e_{j,i}^{V_n}, o_i^{V_n} \rangle$ 两个特征, 其中 $\langle \rangle$ 代表拼接 (concat). 与之不同的是, 文本关系图为有向图, 我们只需要将每条弧 $e_{i,j}^L$ 以及其连接的节点 o_i^L , o_j^L 表达为特征 $f_{i,j}^L = \langle o_i^L, e_{i,j}^L, o_j^L \rangle$. 为了对齐文本关系与视觉关系, 以某一文本特征 $f_{i,j}^L$ 为例, 我们令 $f_{i,j}^{L-V_n} = \langle f_{i,j}^{V_n}, f_{i,j}^L \rangle$ 表示为文本关系特征与视觉关系特征的联合映射, 并且将一帧中的所有视觉关系都与文本关系联合映射定义为 f^{L-V_n} . 我们基于它来进行注意力机制的计算:

$$\alpha^{L-V_n} = \text{softmax}(v_\alpha^T \tanh(W_\alpha f^{L-V_n}) + b_\alpha), \quad (2)$$

其中 $\text{softmax}()$ 为归一化指数函数, v_α^T , W_α 和 b_α 为注意力机制的学习参数, α^{L-V_n} 为注意力权重. 之后, 我们利用该权重生成一帧中文本关系和视觉关系对齐的特征:

$$f^{L-V_n} = \alpha^{L-V_n} \odot f^{L-V_n}, \quad (3)$$

其中 $\odot$ 表示 Hadamard 积. 随后, 我们对于视频候选片段均匀地采样 N 帧, 并将每一帧的对齐特征求和平均得到该视频片段的关系特征:

$$f^{L-V} = \frac{1}{N} \sum_{n=1}^N f^{L-V_n}. \quad (4)$$

之后, 我们利用对齐的特征计算文本关系在该片段中出现的置信度得分:

$$s = \sigma(W_{fc} f^{L-V} + b_{fc}), \quad (5)$$

其中 $\sigma()$ 表示 sigmoid 激活函数, W_{fc} 和 b_{fc} 为全连接层的学习参数. s 评估了文本关系与视觉关系对齐的置信度得分. 该得分可以用于表示视频片段与查询文本的相似度, 并且可以通过交叉熵损失进行训练.

在介绍了文本关系图构建、视觉关系图构建、视觉-文本关系图对齐模块后, 我们阐述如何将其应用于整个跨模态视频片段检索框架中. 首先, 上述 3 个模块结合后可用于计算视频片段与查询文本的匹配程度, 且是加入了关系对齐思想后的匹配程度. 因此, 该模块可以无缝衔接任何跨模态视频片段检索框架. 我们借助于最近提出的 2D-Tan^[52] 时序视频处理方法, 在每一个视频片段与查询文本的计算上加入我们的方法, 即可形成 CrossGraphAlign 跨模态视频片段检索框架. 在训练阶段, 我们沿用 2D-Tan 的定义, 采用视频候选片段与真实视频片段的交叠比 (intersection-over-union) 为标签, 定义交叉熵损失进行回归学习.

4 实验与分析

目前, 在跨模态视频片段检索的领域中, 常用的数据集主要有 TACoS^[28], ActivityNet Captions^[36], Charades-STA^[5], DiDeMo^[6] 等, 为了快速评估, 我们采用 TACoS 数据集来对我们提出的 CrossGraphAlign 进行消融实验. 为了验证方法的有效性, 我们还采用了目前在该领域最大的数据集 ActivityNet Captions 来进行实验. 本节首先介绍这些数据集和我们的实验细节, 其次我们通过消融实验看不同成分的影响, 最后, 我们比较 CrossGraphAlign 和当前的最好方法的结果.

4.1 训练数据集

TACoS^[28]: TACoS 是在 MPII-Composative^[57] 数据集的基础上建立起来的. 它包含 127 个视频, 每个视频都是关于烹饪任务的. 每一个视频都包含 2 种注释, (1) 包含起始和结束时间的动作标签, (2) 在起始和结束时间的自然语言描述. 它包含 17344 个视频片段和相对应的文本描述. 我们采用和 Gao 等^[5] 相同的划分数据集的策略, 其中, 训练集、验证集、测试集的查询文本数量分别为 10146, 4589, 4083.

ActivityNet Captions^[36]: ActivityNet Captions 由 19209 个视频组成, 内容多样且开放. 它最初是为视频字幕任务而设计的, 最近被引入到自然语言的视频片段定位任务中. 仿照 2D-Tan^[52] 中的实验策略, 我们使用 val1 作为验证集, val2 作为测试集, 其中分别有 37417, 17505, 和 17031 个语句用于训练、验证和测试. 目前, 这是此任务中最大的数据集.

4.2 实验设置

- 评估方法: 我们采用和文献 [55] 中相似的指标来计算 $R@n, IOU = m$, 其中 R 表示召回率, n 表示预测视频片段的数量, IOU 为预测视频片段和真实视频片段的交叠比. 假设总共有 K 个查询文本, 那么我们的方法的整体表现就是这 K 个句子的平均值, 即 $\frac{1}{K} \sum_{i=1}^K r(n, m, s_i)$, 其中 $r(n, m, s_i)$ 是对于一个查询语句 s_i 的召回率.

- 实验细节: 目标检测. 本文的目标检测模型 Faster R-CNN^[11] 采用在 COCO^[56] 上训练得到. 因此, 我们首先将 ResNet-101^[9] 的 Faster R-CNN^[11] 在 ImageNet 数据集^[55] 上进行分类任务的预训练, 然后在数据集 COCO 上进行微调. 对本文进行建模. 我们采用和文献 [50] 一样的操作, 利用公开的文本场景图解析器对文本进行图结构的建模. 视频特征. 我们采用和文献 [12] 中相同的处理方法, 利用 C3D^[12] 在 Sports1M^[57] 上预训练得到, 每 16 帧作为一个视频片段, 对视频提取特征.

表 1 CrossGraphAlign 在 TACoS 数据集上的表现
Table 1 Performances of CrossGraphAlign on TACoS dataset

Method	$R@1,$ IOU = 0.1	$R@1,$ IOU = 0.3	$R@1,$ IOU = 0.5	$R@5,$ IOU = 0.1	$R@5,$ IOU = 0.3	$R@5,$ IOU = 0.5
MCN [6]	14.4	–	5.9	37.4	–	10.3
CTRL [5]	24.3	18.3	13.3	48.7	36.7	25.4
CMIN [54]	32.5	24.6	18.1	62.1	38.5	27.0
SLTA [26]	23.1	17.1	11.9	46.5	32.9	20.9
ABLR [23]	34.7	19.5	9.4	–	–	–
QSPN [7]	25.3	20.2	15.2	53.2	36.7	25.3
TGN	41.9	21.8	18.9	53.4	39.1	31.0
2D-Tan [52]	47.6	37.3	25.3	70.3	57.8	45.0
Ours	51.9	39.8	26.4	74.5	60.0	47.2

4.3 实验结果

• TACoS 数据集: 如表 1 [5~7, 23, 26, 52, 54] 所示, 在 TACoS 数据集上评估我们的 CrossGraphAlign 方法, 并与当前较流行的几个 baseline 进行比较, 分别是 MCN [6], CTRL [5], CMIN [54], SLTA [26], ABLR [23], QSPN [7], TGN [21], 2D-Tan [52]. 其中, MCN [6], CTRL [5], QSPN [7], SLTA [26] 方法主要利用了视频和文本的全局特征进行视频片段检索, TGN [21], CMIN [54], ABLR [23] 以及 2D-Tan [52] 方法主要考虑了文本或视频片段之间的相互关系. 结果表明, 我们的 CrossGraphAlign 在 TACoS 上表现最佳, 特别需要注意的是, 我们的 CrossGraphAlign 在 $R@1, IOU = 0.1$ 和 $R@5, IOU = 0.1$ 的情况下, 超越了当前最好的方法 2D-Tan [52] 4 个百分点. 通过将 CrossGraphAlign 和其他相关方法进行对比, 我们获得了一些观察结果. 首先, 我们将 CrossGraphAlign 与使用视频和文本的全局特征的方法 MCN, CTRL, QSPN, SLTA 进行对比, 可以发现, 我们的 CrossGraphAlign 优于这些方法, 原因是我们的方法更加关注文本的关系以及视频片段之间的关系, 也可以认为, 我们的方法不仅关注模态之间的关系, 还关注了模态内的关系. 其次, 我们将 CrossGraphAlign 与考虑了文本或视频片段之间的相互关系的方法 TGN, ABLR, CMIN 进行比较, 可以发现, 我们的方法还是远远好于这些方法, 原因是上述的方法利用注意力机制来查找视频片段或文本之间的隐式的关系, 而我们的方法更加关注视频片段或文本之间的显式的关系. 最后, 我们将 CrossGraphAlign 与当前最好的方法 2D-Tan 进行比较, 可以发现我们的方法好于 2D-Tan, 显示了利用图结构充分地挖掘视频片段的关系和文本的关系, 以及利用注意力机制来捕捉视频和文本之间的对齐的优势.

• ActivityNet Captions 数据集: 为了进一步验证提出方法的泛化性, 如表 2 [5~7, 52] 所示, 我们在 ActivityNet Captions 数据集上评估我们的 CrossGraphAlign 方法. 相较于先前的工作, 我们的方法在 ActivityNet Captions 数据集上取得了更好的性能.

4.4 消融实验

本小节主要评估不同的因素对我们提出的 CrossGraphAlign 的影响.

图卷积网络中的更新信息 (node update vs. edge update): 图卷积网络中对节点进行信息更新与对边进行信息更新对结果有明显的影响. 在这里我们探究它们引起 CrossGraphAlign 的性能变化. 可以看到, 如果节点和边都更新的话, 效果将达到最佳. 正如我们在第 3.3 小节所述, 这将带来巨大的时

表 2 CrossGraphAlign 在 ActivityNet Captions 数据集上的表现

Table 2 Performances of CrossGraphAlign on 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
MCN [6]	39.4	21.4	6.4	68.1	53.2	29.7
CTRL [5]	47.4	29.0	10.3	75.3	59.2	37.5
QSPN [7]	52.1	33.3	13.4	77.7	62.4	40.8
2D-Tan [52]	59.5	44.5	26.5	85.5	77.1	62.0
Ours	62.7	47.2	27.9	88.1	79.1	64.2

表 3 图卷积网络更新信息和注意力机制的消融实验. 框架运行时间在 Nvidia Titan Xp GPU 上进行衡量

Table 3 Ablation study of GCN update information and attention mechanism. The speed of our framework is evaluated on a Nvidia Titan Xp GPU

Model in GCN			$R@1$, IOU = 0.1	$R@1$, IOU = 0.3	$R@1$, IOU = 0.5	$R@5$, IOU = 0.1	$R@5$, IOU = 0.3	$R@5$, IOU = 0.5	Speed (ms)
Node update	Edge update	Attention							
✓	✓	✓	52.3	40.1	26.5	74.9	60.3	47.4	752
✓	✗	✓	50.1	38.4	25.1	72.2	58.4	45.7	568
✓	✓	✗	48.1	37.5	25.9	71.7	58.7	45.3	503
✗	✓	✓	51.9	39.8	26.4	74.5	60.0	47.2	352

间消耗. 因此, 我们使用以关系为中心的更新方法 (性能见表 3), 只对边的信息进行更新. 在这样的情况下, 召回率只有少许的下降, 而框架的处理速度有了 2 倍以上的提升.

注意力机制 (attention vs. without attention): 我们最后评估注意力机制对整个框架的影响. 如表 3 的第 3 行所示, 若不加入注意力机制, 性能将出现显著下降. 这是由于, 注意力机制期望将文本关系与视觉关系进行对齐, 并且进行相似度的计算, 用于整个视频片段与查询文本相似度的评估. 我们认为, 这一过程对特定的查询文本关系匹配到特定的视觉关系作出贡献, 因此显著提升了召回率.

4.5 可视化结果

如图 6 所示, 我们可视化了部分 CrossGraphAlign 在 TACoS 上的预测结果. 当关系图建立成功时, 它可以顺利地预测出查询文本所对应的视频片段. 当关系图建立不成功时, CrossGraphAlign 仍然可以较为合理地预测出查询文本所对应的视频片段. 这是由于 CrossGraphAlign 框架中同样具有查询文本与视频全局特征的融入, 能够与视觉-文本关系匹配综合做出最佳的视频片段预测.

5 总结

本文对跨模态的视频片段检索进行了研究, 并发现现有的检索系统中, 忽略了查询文本中的文本关系与视频片段中视觉关系的对齐, 限制了检索质量的提升. 基于此发现, 我们提出跨模态视觉-文本对齐的 CrossGraphAlign 方法, 通过分别构建文本关系图与视觉关系图来建模查询文本与视频片段中的语义关系, 再通过跨模态对齐的图卷积网络来评估文本关系与视觉关系的相似度, 从而帮助构建更加精准的视频片段检索系统. 在流行的 TACoS 数据集上的实验结果表明, 本文提出的方法可以有

图 6 (网络版彩图) CrossGraphAlign 预测的可视化样例。即使关系图建立不成功时, CrossGraphAlign 仍然可以较为合理地预测查询文本与视频片段的相似度

Figure 6 (Color online) The visualization results from our CrossGraphAlign. Even when the relationship graph is not established successfully, CrossGraphAlign can still reasonably predict the similarity between language query and video moment. (a) Example 1: When the relationship graph is successfully established, CrossGraphAlign can successfully perform visual-text matching. (b) Example 2: When the relationship graph is not established successfully, CrossGraphAlign can still reasonably predict the similarity

效地利用语义关系来提升跨模态视频片段检索的召回率.

参考文献

- 1 Zhang H J, Wu J H, Zhong D, et al. An integrated system for content-based video retrieval and browsing. Pattern Recogn, 1997, 30: 643-658
- 2 Jiang Y G, Ngo C-W, Yang J. Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proceedings of the ACM International Conference on Image and Video Retrieval, Amsterdam, 2007. 494-501

- 3 Snoek C G, Worring M. Concept-based video retrieval. *Found Trends Inform Ret*, 2009, 2: 215–322
- 4 Liu Y, Albanie S, Nagrani A, et al. Use what you have: video retrieval using representations from collaborative experts. In: *Proceedings of British Machine Vision Conference*, 2019
- 5 Gao J Y, Sun C, Yang Z H, et al. TALL: temporal activity localization via language query. In: *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 2017. 5277–5285
- 6 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*, Venice, 2017. 5803–5812
- 7 Xu H J, He K, Plummer B A, et al. Multilevel language and vision integration for text-to-clip retrieval. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, 2019. 9062–9069
- 8 Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: *Proceedings of Advances in Neural Information Processing Systems*, Lake Tahoe, 2012. 1106–1114
- 9 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016. 770–778
- 10 LeCun Y, Bengio Y, Hinton G E. Deep learning. *Nature*, 2015, 521: 436–444
- 11 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Proceedings of Advances in Neural Information Processing Systems*, Quebec, 2015. 91–99
- 12 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*, Santiago, 2015. 4489–4497
- 13 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of Advances in Neural Information Processing Systems*, Lake Tahoe, 2013. 3111–3119
- 14 Devlin J, Chang M-W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, 2019. 4171–4186
- 15 Zhou Z-H. Abductive learning: towards bridging machine learning and logical reasoning. *Sci China Inf Sci*, 2019, 62: 076101
- 16 Lv G Y, Xu T, Chen E H, et al. Reading the videos: temporal labeling for crowdsourced time-sync videos based on semantic embedding. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Phoenix, 2016. 3000–3006
- 17 Zhou P L, Xu T, Yin Z Z, et al. Character-oriented video summarization with visual and textual cues. *IEEE Trans Multim*, 2019. doi: 10.1109/tmm.2019.2960594
- 18 Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition. In: *Proceedings of European Conference on Computer Vision*, Amsterdam, 2016. 20–36
- 19 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: *Proceedings of Advances in Neural Information Processing Systems*, Montreal, 2014. 568–576
- 20 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, 9: 1735–1780
- 21 Chen J Y, Chen X P, Ma L, et al. Temporally grounding natural sentence in video. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Brussels, 2018. 162–171
- 22 Liu M, Wang X, Nie L, et al. Attentive moment retrieval in videos. In: *Proceedings of International ACM SIGIR Conference on Research & Development in Information Retrieval*, Ann Arbor, 2018. 15–24
- 23 Yuan Y T, Mei T, Zhu W W. To find where you talk: temporal sentence localization in video with attention based location regression. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, 2019. 9159–9166
- 24 Pennington J, Socher R, Manning C D. GloVe: global vectors for word representation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, 2014. 1532–1543
- 25 Ge R Z, Gao J Y, Chen K, et al. MAC: mining activity concepts for language-based temporal localization. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Waikoloa Village, 2019. 245–253
- 26 Jiang B, Huang X, Yang C, et al. Cross-modal video moment retrieval with spatial and language-temporal attention. In: *Proceedings of the International Conference on Multimedia Retrieval*, Ottawa, 2019. 217–225
- 27 Wang W, Huang Y, Wang L. Language-driven temporal activity localization: a semantic matching reinforcement learning model. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 2019. 334–343
- 28 Regneri M, Rohrbach M, Wetzell D, et al. Grounding action descriptions in videos. *Trans Assoc Comput Linguist*,

- 2013, 1: 25–36
- 29 Heilbron F C, Escorcia V, Ghanem B, et al. A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 961–970
- 30 Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 6077–6086
- 31 Yang X, Tang K, Zhang H, et al. Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 10685–10694
- 32 Johnson J, Krishna R, Stark M, et al. Image retrieval using scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 3668–3678
- 33 Xu D F, Zhu Y K, Choy C B, et al. Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 3097–3106
- 34 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of International Conference on Learning Representations, Toulon, 2017
- 35 Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, 2018. 7444–7452
- 36 Krishna R, Hata K, Ren F, et al. Dense-captioning events in videos. In: Proceedings of the IEEE Conference on Computer Vision, Venice, 2017. 706–715
- 37 Wang S J, Wang R P, Yao Z W, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval. In: Proceedings of Winter Conference on Applications of Computer Vision, Colorado, 2020
- 38 Lee K-H, Chen X, Hua G, et al. Stacked cross attention for image-text matching. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 212–228
- 39 Lin D, Fidler S, Kong C, et al. Visual semantic search: retrieving videos via complex textual queries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 2657–2664
- 40 Alayrac J-B, Bojanowski P, Agrawal N, et al. Unsupervised learning from narrated instruction videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 4575–4583
- 41 Shou Z, Wang D, Chang S F. Temporal action localization in untrimmed videos via multi-stage CNNs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 1049–1058
- 42 Ma S, Sigal L, Sclaroff S. Learning activity progression in LSTMs for activity detection and early detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 1942–1950
- 43 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, Las Vegas, 2016. 1961–1970
- 44 Zhao Y, Xiong Y J, Wang L M, et al. Temporal action detection with structured segment networks. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017. 2933–2942
- 45 Chao Y-W, Vijayanarasimhan S, Seybold B, et al. Rethinking the faster R-CNN architecture for temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 1130–1139
- 46 Lin T W, Zhao X, Shou Z. Single shot temporal action detection. In: Proceedings of the ACM on Multimedia Conference, Mountain View, 2017. 988–996
- 47 Dai B, Zhang Y Q, Lin D H. Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 3298–3308
- 48 Yang J W, Lu J S, Lee S, et al. Graph R-CNN for scene graph generation. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 690–706
- 49 Zellers R, Yatskar M, Thomson S, et al. Neural motifs: scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 5831–5840
- 50 Schuster S, Krishna R, Chang A, et al. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Proceedings of the 4th Workshop on Vision and Language, Lisbon, 2015. 70–80
- 51 Han R J, Ning Q, Peng N Y. Joint event and temporal relation extraction with shared representations and structured prediction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the

- International Joint Conference on Natural Language Processing, Hong Kong, 2019. 434–444
- 52 Rohrbach M, Regneri M, Andriluka M, et al. Script data for attribute-based recognition of composite activities. In: Proceedings of European Conference on Computer Vision, Florence, 2012. 144–157
 - 53 Zhang D, Dai X Y, 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, Long Beach, 2019. 1247–1257
 - 54 Zhang S Y, Peng H W, Fu J L, et al. Learning 2D temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence, New York, 2020
 - 55 Liu M, Wang X, Nie L, et al. Cross-modal moment localization in videos. In: Proceedings of ACM Multimedia Conference on Multimedia Conference, Seoul, 2018. 843–851
 - 56 Lin T-Y, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 740–755
 - 57 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*, 2015, 115: 211–252

Cross-modal video moment retrieval based on visual-textual relationship alignment

Joya CHEN, Hao DU, Yufei WU, Tong XU* & Enhong CHEN

Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China, Hefei 230027, China

* Corresponding author. E-mail: tongxu@ustc.edu.cn

Abstract In recent years, increasing amounts of video resources have created a series of demands for fine retrieval of video moments, such as highlight moments in sports events and the re-creation of specific video content. In this context, research on cross-modal video segment retrieval, which attempts to output a video moment that matches the input query text, is gradually emerging. Existing solutions primarily focus on global or local feature representation for query text and video moments. However, such solutions ignore matching semantic relations contained in query text and video moments. For example, given the query text “a person is playing basketball”, existing retrieval systems may incorrectly return a video moment of “a person holding a basketball” without the considering the semantic relationship of “a person playing basketball”. Therefore, this paper proposes a cross-modal relationship alignment framework, which we refer to as CrossGraphAlign, for cross-modal video moment retrieval. The proposed framework constructs a textual relationship graph and a visual relationship graph to model the query semantics in text and video segment relations, and then evaluates the similarity between text relations and visual relations through cross-modally aligned graph convolutional networks to help construct a more accurate video moment retrieval system. Experimental results on the publicly available cross-modal video retrieval datasets TACoS and ActivityNet Captions demonstrate that the proposed method can effectively utilize the semantic relationships to improve the recall rate in cross-modal video moment retrieval.

Keywords relationship alignment, textual relationship, visual relationship, graph convolutional network, cross-modal video moment retrieval

Joya CHEN is now a graduate student in computer science at University of Science and Technology of China, Hefei, China. His research interests include visual recognition, and multimedia analysis. Furthermore, he has contributed to several well-known computer vision implementations, e.g., maskrcnn-benchmark and insightface.

Hao DU received his B.S. degree in communication engineering from Southwest Jiaotong University, Chengdu, China, in 2019. He is currently a master student in the School of Software Engineering at University of Science and Technology of China, Hefei, China. His main research interests include computer vision and multimedia analysis.

Tong XU received a Ph.D. degree from University of Science and Technology of China, Hefei, China, in 2016. Currently, he is working as an associate researcher at the Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China.

Enhong CHEN is a professor and vice dean of the School of Computer Science, University of Science and Technology of China. His general areas of research include data mining and machine learning, social network analysis, and recommender systems.