计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220560
60(11): 2624−2637, 2023

跨媒体语义关联增强的网络视频热点话题检测

张承德[1] 刘雨宣[1] 肖 霞[2] 梅 凯[1]
[1](中南财经政法大学信息与安全工程学院 武汉 430073)
[2](华中农业大学公共管理学院 武汉 430070)
(chengdezhang@zuel.edu.cn)

Hot Topic Detection of Web Video Based on Cross-Media Semantic Association Enhancement

Zhang Chengde[1], Liu Yuxuan[1], Xiao Xia[2], and Mei Kai[1]
[1](*School of Information and Safety Engineering*, *Zhongnan University of Economics and Law*, *Wuhan* 430073)
[2](*School of Public Administration*, *Huazhong Agricultural University*, *Wuhan* 430070)

Abstract Cross-media web video hot topic detection has become a new research hotspot. However, there is less text information to describe video, which makes the space of text semantic features sparse, resulting in weak correlation between text semantic features, which increases the difficulty of mining hot topics. The existing methods mainly enrich the text semantic feature space through visual information. However, due to the heterogeneity between visual and text information, the semantic features of text and visual are quite different under the same topic. This further reduces the correlation strength between text semantics under the same topic, and also brings great challenges to cross-media hot topic detection based on web videos. Therefore, we propose a new cross-media semantic association enhancement method. Firstly, the core semantic features of the text from the word and sentence levels through double-layer attention are captured; Secondly, by understanding the visual content, a large number of text descriptions highly related to the video content are generated to enrich the text semantic space; Then, through text semantic similarity and visual semantic similarity, the text semantic map and visual semantic map are constructed, and the time decay function is constructed to establish the correlation between cross-media data from the time dimension, so as to enhance the correlation strength between text and visual semantics, and smoothly fuse the two semantic maps into a hybrid semantic map to realize cross-media semantic complementarity; Finally, hot topics are detected by graph clustering method. A large number of experimental results show that the proposed model is superior to the existing methods.

Key words cross-media; web video; topic detection; video caption; semantic association enhancement

收稿日期: 2022-06-16; **修回日期**: 2022-11-15

基金项目: 国家社会科学基金一般项目(22BXW081); 国家自然科学基金面上项目 (G041401); 教育部人文社会科学研究青年基金项目(20YJC860040); 湖北省高等学校哲学社会科学研究重大项目(21ZD018); 2020 年度武汉市科技局应用基础前沿项目(2020010601012183); 湖北省高等学校实验室研究项目(HBSY2021-58); 中南财经政法大学中央高校基本科研业务费专项资金 (202211411, 202151423, 202211415, 202211405)

This work was supported by the General Program of the National Social Science Foundation of China (22BXW081), the General Program of the National Natural Science Foundation of China (G041401), the Project of Humanity and Social Science Youth Foundation of Ministry of Education of China (20YJC860040), the Major Project of Philosophy and Social Science Research in Colleges and Universities of Hubei Province (21ZD018), the Application Foundation Frontier Project of Wuhan Science and Technology Bureau in 2020 (2020010601012183), the Laboratory Research Project of Colleges and Universities in Hubei Province (HBSY2021-58), and the Fundamental Research Funds for the Central Universities of Zhongnan University of Economics and Law (202211411, 202151423, 202211415, 202211405).

摘　要　跨媒体网络视频热点话题检测成为新的研究热点.然而,描述视频的文本信息较少,使得文本语义特征空间稀疏,导致文本语义特征间关联强度较弱,增加了挖掘热点话题的难度.现有方法主要通过视觉信息丰富文本语义特征空间.然而,由于视觉与文本信息间的异构性,导致同一话题下文本与视觉语义特征差异较大,这进一步降低了同一话题下文本语义间的关联强度,也给跨媒体网络视频热点话题检测带来巨大挑战.因此,提出一种新的跨媒体语义关联增强方法.首先,通过双层注意力,从单词和句子2个级别捕捉文本核心语义特征;其次,通过理解视觉内容,生成大量与视频内容高度相关的文本描述,丰富文本语义空间;然后,分别通过文本语义相似性和视觉语义相似性,构建文本语义图和视觉语义图,并构造时间衰减函数,从时间维度建立跨媒体数据间的相关性,以此增强文本与视觉语义间的关联强度,平滑地将2种语义图融合为混合语义图,实现跨媒体语义互补;最后,通过图聚类方法检测出热点话题.大量实验结果表明,提出的模型优于现有方法.

关键词　跨媒体;网络视频;话题检测;视频理解;语义关联增强

中图法分类号　TP37

社交网络和智能手机的普及使得网络视频数量爆炸式增长,并逐渐取代文本成为普通用户信息交流的主要载体[1].据中国互联网络信息中心(CNNIC)发布的最新报告[2],截至2021年6月,国内网络视频用户规模达9.44亿,较2020年增长1 707万.同时,国外最大的视频分享平台YouTube,月活跃用户人数已超过20亿,每分钟上传的视频总时长超过300 h,人们每天在YouTube上花费超过10^{10} h寻找和观看视频[3].当热点话题发生时,用户需要观看数量庞大的网络视频,并花费大量时间和精力梳理和总结其前因后果,才能初步了解热点话题的基本情况.而当遇到完全陌生的话题时,则进一步增加了这一难度.因此,网络视频热点话题检测变得十分必要.

传统网络视频热点话题检测方法,主要通过计算标题、标签间的语义相似度,将视频划分到不同的话题[4].通常,网络视频只有十多个词描述,且文本信息少、噪声多,这很容易引起文本语义特征空间稀疏[5],导致文本间语义关联少且关联强度弱.另外,由于不同的人表达习惯不同,以及多义词、多语言等问题,将进一步降低文本间语义特征的关联强度,难以建立视频间联系.现有方法主要通过引入视频弹幕、评论等外部信息丰富文本语义空间[6−8].但是,这类方法存在2方面问题:一方面,部分媒体平台不支持发送弹幕和评论,导致部分视频缺乏外部信息;另一方面,弹幕、评论有效信息少,内容冗杂,导致文本间语义关联减少.因此,上述方法严重依赖于引入信息与话题的相关度,导致话题检测性能不稳定.由于视频内容丰富、客观,且视频内容与话题往往高度相关.因此,尝试智能理解视频内容,生成大量准确、客观的语义信息,丰富文本语义特征空间,提升热点话题检测效果.

然而,由于视频内容侧重于对视觉具体信息的客观描述,文本信息侧重于对话题内容的抽象表达,带有一定的主观情感,导致同一话题下视觉语义与文本语义特征差异性较大.如图1所示,视频*A*的标题表达了一种主观情感,并未提及任何与话题相关的信息,字幕则直接表达科比坠机这一话题;在视频*B*中,字幕仅表达视频来源,与视频所属话题无关,视频理解则清晰地展示了“士兵、坦克”等与战争主题高度相关的词汇;在视频*C*中,字幕以及视频理解表达的语义都不准确,只有标题清晰地表达2020美国大选这一主题.传统的融合方法无法调和文本与视觉间的语义鸿沟,反而降低了同一话题下文本间的关联强度.因此,如何克服文本语义与视觉语义间的差异,增强跨媒体语义的关联,成为一个巨大的挑战.

为应对上述挑战,提出一种新的跨媒体语义关联增强方法.通过融合视频理解、字幕、标题3种语义特征,增强文本与视觉语义间的关联强度,实现跨媒体信息的互补,解决话题检测中的文本稀疏问题.该方法共分为4步:第1步,通过双层注意力机制,挖掘文本语义特征.第2步,通过视频智能理解以及场景文字识别,生成一系列与视频内容高度相关的文本信息,丰富文本语义空间.第3步,跨媒体语义融合.首先,分别通过文本语义相关度和视觉语义相关度,构建文本语义图和视觉语义图;然后,构造时间衰减函数,通过视频上传的时间差,度量跨媒体数据间的话题相似性;最后,通过时间衰减系数量化文本语义图和视觉语义图间的连接强度,并将其加权叠加形成混合语义图,增强跨媒体语义关联.第4步,话题检测.通过图移位(graph shifts, GS)算法[9]在混合语义图中挖掘出密集子图,并通过TextRank算法[10]检测出热点话题.

Fig. 1 Three semantic differences

图 1 3 种语义差异

本文的主要创新点和贡献总结为 3 个方面:

1)提出一种新的文本信息丰富方法. 通过理解视觉内容,生成大量与话题高度相关的文本描述信息,避免外部数据引入而带来的过多噪声问题,实现网络视频的文本语义特征空间丰富.

2)构建了一种新途径建立视觉与文本信息间的语义关联. 引入时间特征建立跨媒体语义关联,主要通过时间衰减函数,量化视频上传时间差对跨媒体数据话题相似性的影响,以此增强跨媒体语义关联强度,同时避免语义融合所带来的噪声干扰.

3) 提出一种新的跨媒体语义关联增强方法. 通过构建文本语义图,找到网络视频中缺失的话题关联,并通过融合文本语义图与视觉语义图,重建文本与视觉语义间的连接,实现跨媒体语义互补增强. 同时赋予模型良好的可扩展性,能轻松地将其他模态信息融合到语义图中.

1 相关工作

1.1 热点话题检测与跟踪

热点话题检测与跟踪(topic detection and tracking, TDT)任务源于美国国防高级研究计划局赞助的研究计划[11], 核心任务是从数据流中发现新话题,并收集后续的相关报导[12]. 早期的话题检测主要是面向新闻、博客等长文本,随着社交媒体的快速发展,使得信息的传播不再局限于文本这一种形式,网络图片和视频在信息交流中占据主导地位,话题检测也由面向纯文本扩展为基于文本、图像、视频等多模态融合的话题检测[13].

对于文本信息,早期的话题检测方法大多基于 LDA(latent Dirichlet allocation)主题模型[14]或其改进模型[15]. 受 LDA 固有的局限性,这些方法对长文本效果较好、对短文本效果较差. 文献 [16] 通过搜索引入相关文本以扩展短文本,但此方式过于依赖于辅助文本的质量,当引入文本质量不高或者相关文本不充足时,话题检测效果较差. 文献 [17] 提出了一种 BTM (biterm topic model)主题模型,这是对 LDA 模型的一种改进,利用词共现关系丰富词组,从而缓解短文本的稀疏性问题,此模型对短文本主题检测效果有一定的提升. 但是,当文本特征过于稀疏或者存在噪声、一词多义时模型的检测效果较差. 所以这种基于主题模型的话题检测方法对视频标题并不适用.

对于图像和视频,现有方法主要利用丰富的视觉信息进行话题挖掘. 文献 [18] 通过注意力提取视频局部显著语义特征和全局语义特征,并进行分类表示以区分不同话题. 文献 [19] 提出了一种图像主导的主题模型,将视觉特征作为信息线索进行话题检测. 这种检测方法虽然较好地利用了视觉信息,但由于网络视频大多由用户随意拍摄和剪辑,再加上拍光照、运动、拍摄角度等多种因素的变化,导致视觉语义特征获取并不准确、检测效果不佳.

对多模态数据,现有方法主要通过多种模态数据中的互补信息来检测热点话题. 文献 [20] 对视觉信息和文本信息之间的相关性进行研究,探索不同模态数据间的语义关联. 文献 [21] 利用图片理解技术,挖掘图片表达的语义信息;通过融合文本与图像

信息，实现了短文本语义空间丰富. 受此工作的启发，我们通过视频理解技术，深度挖掘网络视频中的语义信息，并与稀疏的文本语义融合，丰富文本语义.

1.2 视频理解

视频理解旨在理解视觉内容，实现视觉特征向自然语言的转化[22]. 文献 [23] 提出了一种按固定的语法和模板生成句子的方法，但是这种基于模板的方法生成的句子单一，束缚了模型的表达能力. 随着深度学习取得突破性进展，基于深度学习的视频理解方法被广泛应用. 文献 [24] 提出了一种 S2VT（sequence to sequence video to text）模型，引入编码解码器架构，实现特征的编码和解码. 同一时期，文献 [25] 在模型中引入注意力机制，对视频帧的卷积特征进行加权求和，实现更精准的特征选择，但是生成的句子表达依旧不够准确，难以满足实际生活的需要.近年来，许多研究都聚焦于改进视频特征提取的方法或改进编码的循环神经网络的结构. 文献 [26] 提出了一种基于双向时序图的对象感知聚合模型，通过构建双向时序图实现了对视频特征更为精细的捕捉. 文献 [27] 提出了一种视觉特征编码技术，通过傅里叶变换嵌入时间动态，并改用循环门控单元生成丰富的语义.

此外，网络视频中包含丰富的场景文字信息，往往与视频主题高度相关. 已有部分工作结合场景文字与视觉特征，增强图像和视频内容理解. 如文献 [28] 将图像中的场景文字信息与图像卷积特征相结合，增强视觉特征表达. 文献 [29] 利用图片中的场景文字，增强短文本与视觉间的交互关系，并在上述工作的启发下，试图将视频中的字幕信息与视频理解相结合，挖掘视觉语义，丰富文本语义空间.

1.3 跨媒体融合

跨媒体数据具有分散异构、语义关联和多模态的特点，相同语义可以借由不同跨媒体数据表达. 而不同跨媒体数据低层特征表示不一致，导致无法直接通过特征计算数据间的相关性[30]. 早期主要是利用人工标注跨媒体数据，然后通过标签信息实现跨媒体数据间的关联和检索，但是人工标注成本高、速度慢且具有一定的主观性，对海量的跨媒体数据并不适用. 随着深度学习的成功，针对图像和视频的分析技术迅速发展. 通过分析和理解跨媒体数据的内容，结合标题、场景文字等信息，形成语义标签以辅助跨媒体融合成为了主流方法. 文献 [31] 提出一种混合注意力模块，同时利用多模态数据内部的联系以及文本词和图像位置之间的关系进行融合，实现了多模态数据的关联和互补. 文献 [32] 捕捉多模态数据的整体特征，并映射到同一空间实现多模态数据的融合. 这种融合方式在一定程度上实现了跨媒体数据的对齐，但是它们都过于复杂，而且要求文本和视觉信息具有严格的对应关系，这对网络视频并不适用.

然而，报道同一热点话题的跨媒体数据具有相似的语义内容，而且通常在相似的时间上传，因此，同一话题下的跨媒体数据具有很强的语义相似性和时间相似性. 文献 [33] 通过时间窗口改进聚类模型实现新闻文本主题的精准捕捉. 但其主要是基于文本的分析，尤其是针对新闻文章，这对稀疏的视频文本并不适用. 所以为了解决上述问题，构建了一种简单且有效的融合方法. 通过文本语义相似性和视觉语义相似性，分别构建文本语义图和视觉语义图，通过图的融合实现跨媒体数据的融合. 此外通过时间衰减函数，将时间特征嵌入混合语义图，增强文本和视觉语义关联强度，形成更为平滑的密集子图.

2 网络视频热点话题检测方法

本文提出的话题检测框架如图 2 所示，框架包括 4 个步骤，分别为文本语义特征提取、视觉语义特征提取、跨媒体语义融合以及话题检测.

2.1 文本语义特征提取

受文献 [34] 的启发，我们构造了一个双层注意力模型，分别通过单词级注意力和句子级注意力挖掘文本核心语义特征. 假定数据集中共有 J 篇文本，每篇文本有 L 个句子，每个句子包含 T 个单词.其中 $j \in [1, J]$ 表示第 j 篇文本，$i \in [1, L]$ 表示文本中第 i 个句子，$t \in [1, T]$ 表示句子中第 t 个单词.

2.1.1 单词级注意力

对给定的第 i 个句子中的第 t 个单词 $\boldsymbol{w}_{i,t}$，由词向量模型[35]将单词编码为词向量 $\boldsymbol{x}_{i,t}$，接着将词向量输入到双向长短期记忆网络（bi-directional long short-term memory, BiLSTM）[36] 单元，分别获取前向隐藏层状态和后向隐藏层状态. 式（1）和式（2）展示了 LSTM（long short-term memory）细胞单元，将隐藏层状态传递到下一个细胞单元，并获取双向时序信息的过程.

$$\overrightarrow{h_{i,t}} = f_{\mathrm{LSTM}}(\boldsymbol{x}_{i,t}, \overrightarrow{h_{i,t-1}}), \tag{1}$$

$$\overleftarrow{h_{i,t}} = f_{\mathrm{LSTM}}(\boldsymbol{x}_{i,t}, \overleftarrow{h_{i,t-1}}), \tag{2}$$

其中 $\boldsymbol{x}_{i,t}$ 表示 LSTM 单元的输入，$\overleftarrow{h_{i,t-1}}$ 表示前一时刻的隐藏层状态.$\overrightarrow{h_{i,t}}$ 和 $\overleftarrow{h_{i,t}}$ 分别表示 LSTM 单元的前向和后向隐藏层状态. 然后连接前向和后向的隐藏层状态，

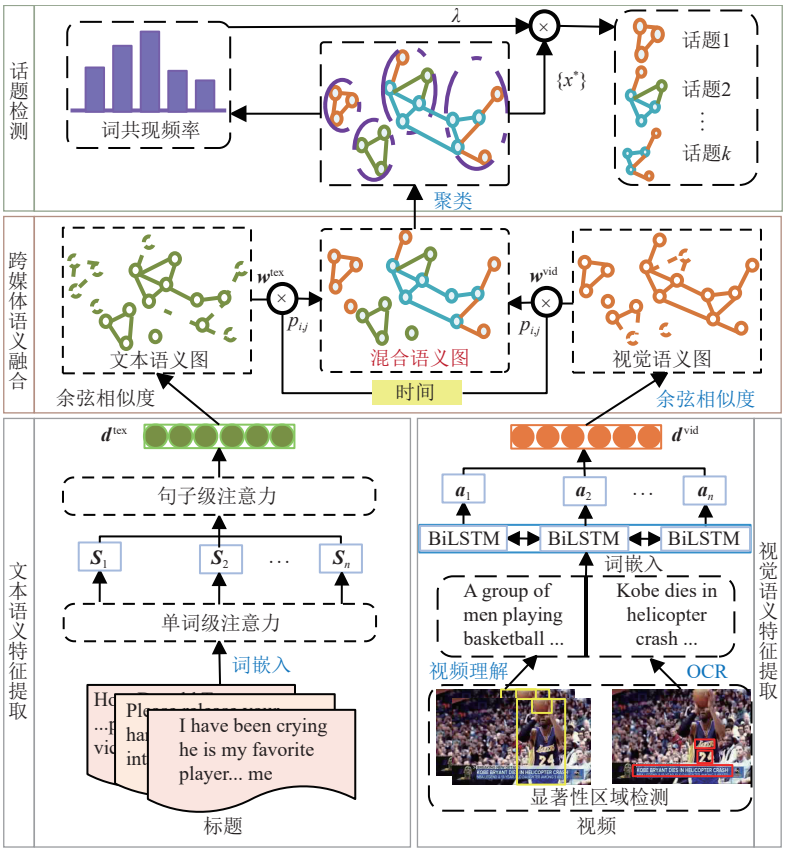


Fig. 2 Topic detection framework

图 2 话题检测框架

得到给定单词$\boldsymbol{w}_{i,t}$的特征向量表示.式(3)展示了整合前向和后向隐藏层状态的过程.

$$\boldsymbol{h}_{i,t}=\left[\overrightarrow{\boldsymbol{h}_{i,t}}\oplus\overleftarrow{\boldsymbol{h}_{i,t}}\right], \tag{3}$$

其中⊕表示元素求和,$\boldsymbol{h}_{i,t}$表示双向 LSTM 输出的隐藏层状态.

在单词级别,通过单词级注意力区分不同单词的重要性,并将其聚合为句子向量表示.首先,将单词隐藏层向量$\boldsymbol{h}_{i,t}$输入带有激活函数tanh的非线性神经元层,将$\boldsymbol{h}_{i,t}$投影到同一个注意力空间(见式(4));然后,利用$\boldsymbol{u}_{\text{wod}}$作为衡量指标,wod为单词级标记,计算$\boldsymbol{u}_{i,t}$的重要性,并将结果归一化后得到不同单词的权重(见式(5));最后,通过加权求和的方式计算句子向量$\boldsymbol{S}_i$(见式(6)).单词上下文向量$\boldsymbol{u}_{\text{wod}}$在训练过程中被随机初始化并联合学习.

$$\boldsymbol{u}_{i,t}=\tanh(\boldsymbol{W}_{\text{wod}}\boldsymbol{h}_{i,t}+\boldsymbol{b}_{\text{wod}}), \tag{4}$$

$$\boldsymbol{a}_{i,t}=\frac{\exp(\boldsymbol{u}_{i,t}^{\mathrm{T}}\boldsymbol{u}_{\text{wod}})}{\sum_{t}\exp(\boldsymbol{u}_{i,t}^{\mathrm{T}}\boldsymbol{u}_{\text{wod}})}, \tag{5}$$

$$\boldsymbol{S}_i=\sum_{t}\boldsymbol{a}_{i,t}\boldsymbol{h}_{i,t}, \tag{6}$$

其中$\boldsymbol{u}_{i,t}$表示隐藏层状态$\boldsymbol{h}_{i,t}$的投影向量,T 表示转置,$\boldsymbol{W}_{\text{wod}}$和$\boldsymbol{b}_{\text{wod}}$表示网络的权重矩阵和偏置值,$\boldsymbol{a}_{i,t}$表示句子内部每个单词的权重,$\boldsymbol{S}_i$表示文本中第 i 个句子的特征向量.

2.1.2 句子级注意力

在句子级别,通过句子级注意力区分不同句子的重要性,并将其聚合为文本向量表示.首先,通过BiLSTM 对文本中的句子进行编码表示,为了简单起见,我们将 BiLSTM 单元的操作表示为$\boldsymbol{h}_i=f_{\text{BiLSTM}}(\boldsymbol{S}_i)$,即以句子$\boldsymbol{S}_i$为中心,整合句子上下文信息的隐藏层向量表示;其次,将隐藏层状态$\boldsymbol{h}_i$投影到同一个注意力空间(见式(7)),并随机初始化一个句子级别的上下文向量$\boldsymbol{u}_{\text{sen}}$,用于衡量不同句子的重要性;然后,将计算出的句子重要性,归一化后得到每个句子的权重(见式(8));最后,对句子向量进行加权求和,并经过一个全连接层映射得到文本语义特征表示(见式(9)).

$$\boldsymbol{u}_i=\tanh(\boldsymbol{W}_{\text{sen}}\boldsymbol{h}_i+\boldsymbol{b}_{\text{sen}}), \tag{7}$$

$$\boldsymbol{a}_i=\frac{\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_{\text{sen}})}{\sum_{i=1}^{L}\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_{\text{sen}})}, \tag{8}$$

$$\boldsymbol{d}_j^{\text{tex}}=\boldsymbol{W}_{\text{tex}}\sum_{i=1}^{L}\boldsymbol{a}_i\boldsymbol{h}_i+\boldsymbol{b}_{\text{tex}}, \tag{9}$$

式(8)和(9)中，$\boldsymbol{u}_i$表示句子隐藏层状态的投影向量，T表示转置. $\boldsymbol{a}_i$表示文本中第i个句子的权重，$\boldsymbol{u}_{\text{sen}}$表示句子上下文向量，在训练过程中被随机初始化并联合学习，sen为句子级标记. 式(9)中，$\boldsymbol{d}_j^{\text{tex}}$表示蕴含第$j$篇文档结构知识的文本语义特征表示，tex表示文本标记. $\boldsymbol{W}_{\text{sen}}$，$\boldsymbol{W}_{\text{tex}}$和$\boldsymbol{b}_{\text{sen}}$，$\boldsymbol{b}_{\text{tex}}$分别表示训练得到的权重和偏置值.

2.2 视觉语义特征提取

2.2.1 视频理解

由于传统模型在复杂的真实场景中表现不佳，因此通过视频理解结合光学字符识别(optical character recognition, OCR)，实现对视觉语义的精准捕捉. 如图3所示视频理解模型包括4个阶段：

1)特征提取阶段. 利用基于注意力的C3D网络[37]提取视频显著性区域的运动特征，通过VGG16[38]提取视频的全局卷积特征，并通过全连接层连接这2类特征，作为视频的特征输入.

2)特征编码阶段.特征的编码主要由2层LSTM和视觉注意力实现. 在每个时间步长下，2层LSTM将可变长度的输入编码为固定维度的向量，视觉注意力则用于捕捉视觉特征中的显著性区域. 即对于输入特征$\boldsymbol{X}=(x_1, x_2, \cdots, x_n)$经过2层LSTM的编码得到其隐藏层状态序列$\boldsymbol{h}^{\text{emb}}=\left(\boldsymbol{h}_1^{\text{emb}}, \boldsymbol{h}_2^{\text{emb}}, \cdots, \boldsymbol{h}_n^{\text{emb}}\right)$，此序列经过注意力加权得到当前时刻的特征编码(式(10)).

$$\boldsymbol{c}_t=\sum_{i=1}^{n}\boldsymbol{a}_{t,i}\boldsymbol{h}_i^{\text{emb}}, \tag{10}$$

$$\boldsymbol{a}_{t,i}=\frac{\exp(\boldsymbol{e}_{t,i})}{\sum_{k=1}^{n}\exp(\boldsymbol{e}_{t,k})}, \tag{11}$$

$$\boldsymbol{e}_{t,i}=\boldsymbol{w}^{\text{T}}\tanh\left(\boldsymbol{W}_a^{\text{emb}}\boldsymbol{h}_i^{\text{emb}}+\boldsymbol{W}_a^{\text{dec}}\boldsymbol{h}_{t-1}^{\text{dec}}+\boldsymbol{b}_a\right), \tag{12}$$

其中$\boldsymbol{c}_t$表示编码后的特征输出，$\boldsymbol{a}_{t,i}$表示编码器隐藏层向量$\boldsymbol{h}^{\text{emb}}$的权重(式(11))，$\boldsymbol{e}_{t,i}$由LSTM隐藏层状态经过激活函数后得到. $\boldsymbol{W}_a^{\text{emb}}$，$\boldsymbol{W}_a^{\text{dec}}$，$\boldsymbol{b}_a$均表示可学习的参数.

3)特征解码阶段.输入序列经过解码器解码为输出序列$\boldsymbol{Y}=(y_1, y_2, \cdots, y_n)$上的分布(见式(13))，再通过$\boldsymbol{Y}$查询词汇表得到对应的单词. 此外，当前时刻生成的单词会作为特征序列与输入特征连接后传入到解码器，为下一时刻单词的生成提供支持.

$$p(y_1, y_2, \cdots, y_m \mid x_1, x_2, \cdots, x_n)=\prod_{t=1}^{m}p\left(y_t \mid \boldsymbol{h}_{n+t-1}^{\text{dec}}, y_{t-1}, \boldsymbol{c}_t\right), \tag{13}$$

其中$\boldsymbol{c}_t$表示编码器输出，$\boldsymbol{h}^{\text{dec}}$表示解码器隐藏层输出，$y_{t-1}$表示前一时刻解码器输出的序列，$p(y_1, y_2, \cdots, y_m \mid x_1, x_2, \cdots, x_n)$表示对于特定输入$\boldsymbol{X}$获得的输出分布.

4)训练阶段. 模型以MSVD数据集[39]作为知识库，采用端到端方式训练. 训练目标是：使得预测句子的对数释然估计值最大. 即通过最大化对数释然估计值，不断更新参数θ让模型找到最优解码序列$\boldsymbol{Y}=(y_1, y_2, \cdots, y_m)$，见式(14).

$$\theta^{*}=\arg\max_{\theta}\sum_{t=1}^{m}\ln p(y_t|\boldsymbol{h}_{n+t-1}^{\text{dec}}, y_{t-1}, \boldsymbol{c}_t; \theta), \tag{14}$$

其中$\boldsymbol{h}_{n+t-1}^{\text{dec}}$表示解码器隐藏层状态，$y_{t-1}$表示前一时刻输出词序列，$\theta^{*}$表示参数为$\theta$的对数释然估计值.

2.2.2 视觉语义嵌入

视觉语义嵌入分为2步：第1步，将生成的字幕和视频描述编码为词向量，并将该词向量输入到BiLSTM模型中，得到蕴含上下文语义信息的隐藏层特征向量；第2步，通过自注意力对隐藏层状态进行加权(见式(15)和式(16))，并将加权融合后的特征经向量全连接层得到视觉语义特征表示(见式(17)).

$$\boldsymbol{e}_i=\tanh(\boldsymbol{W}\boldsymbol{h}_i^{\text{vid}}+\boldsymbol{b}), \tag{15}$$

$$\boldsymbol{a}_i=\frac{\exp(\boldsymbol{u}^{\text{T}}\boldsymbol{e}_i)}{\sum_{i=1}^{L}\exp(\boldsymbol{u}^{\text{T}}\boldsymbol{e}_i)}, \tag{16}$$

$$\boldsymbol{d}_j^{\text{vid}}=\boldsymbol{W}_{\text{vid}}\sum_{i=1}^{L}\boldsymbol{a}_i\boldsymbol{h}_i^{\text{vid}}+\boldsymbol{b}_{\text{vid}}, \tag{17}$$

其中$\boldsymbol{e}_i$表示隐藏层状态$\boldsymbol{h}_i^{\text{vid}}$经过全连接层后的输出，$\boldsymbol{u}$

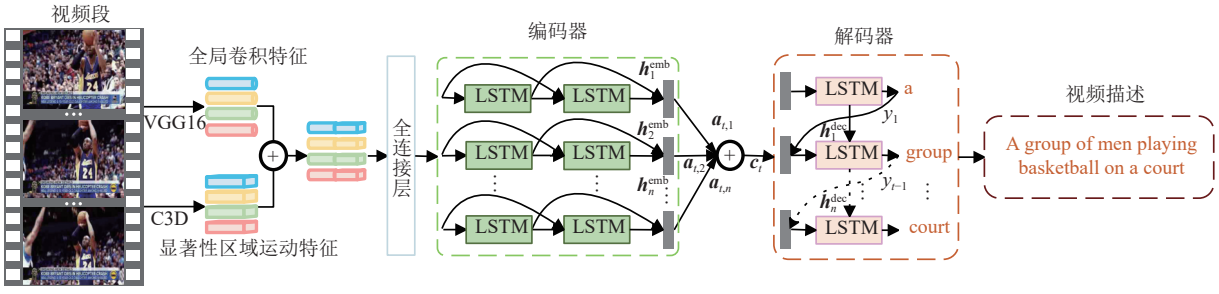


Fig. 3 Video caption model

图3 视频理解模型

表示经过训练后得到的权重参数矩阵，$\boldsymbol{a}_i$表示注意力权重，$\boldsymbol{d}_j^{\text{vid}}$表示第$j$个视频的视觉语义特征表示，$L$表示文本长度. $\boldsymbol{W}$, $\boldsymbol{W}_{\text{vid}}$, $\boldsymbol{b}$, $\boldsymbol{b}_{\text{vid}}$分别表示训练权重和偏置值，vid表示视频标记.

2.3 跨媒体语义融合

2.3.1 文本语义图

文本语义图$\boldsymbol{G}^{\text{tex}}=\left(\left\{n_i^{\text{tex}}\right\},\left\{w_{i,j}^{\text{tex}}\right\}\right)$由文本数据集构建. 其中$\left\{n_i^{\text{tex}}\right\}$代表图的结点集(见式(18))，$\left\{w_{i,j}^{\text{tex}}\right\}$表示结点间连接权重由文本语义相似度确定.

$$N^{\text{tex}}=\{n_i^{\text{tex}}\},\tag{18}$$

$$w_{i,j}^{\text{tex}}=\text{cosine}\left(\boldsymbol{d}_i^{\text{tex}},\boldsymbol{d}_j^{\text{tex}}\right)=\frac{\boldsymbol{d}_i^{\text{tex}}\cdot\boldsymbol{d}_j^{\text{tex}}}{\left\|\boldsymbol{d}_i^{\text{tex}}\right\|\cdot\left\|\boldsymbol{d}_j^{\text{tex}}\right\|},\tag{19}$$

式(18)中n_i^{tex}表示文本语义图的第i个结点，i指明此结点由第i条文本数据表达. 式(19)中$w_{i,j}^{\text{tex}}$表示结点i与结点j之间的连接权重，cosine 表示计算余弦相似度.$\boldsymbol{d}_i^{\text{tex}}$为 2.1.2 节得到的文本语义特征.

2.3.2 视觉语义图

视觉语义图$\boldsymbol{G}^{\text{vid}}=(\{n_i^{\text{vid}}\},\{w_{i,j}^{\text{vid}}\})$的结点由视频数据集构成，结点间的连接权重由视觉语义特征间的余弦相似度衡量，见式(20)和式(21).

$$N^{\text{vid}}=\{n_i^{\text{vid}}\},\tag{20}$$

$$w_{i,j}^{\text{vid}}=\text{cosine}\left(\boldsymbol{d}_i^{\text{vid}},\boldsymbol{d}_j^{\text{vid}}\right)=\frac{\boldsymbol{d}_i^{\text{vid}}\cdot\boldsymbol{d}_j^{\text{vid}}}{\left\|\boldsymbol{d}_i^{\text{vid}}\right\|\cdot\left\|\boldsymbol{d}_j^{\text{vid}}\right\|},\tag{21}$$

式(20)中n_i^{vid}表示视觉语义图的第i个结点.式(21)中$w_{i,j}^{\text{vid}}$表示结点i与结点j之间的连接权重，由结点语义特征的余弦相似度计算得出.$\boldsymbol{d}_i^{\text{vid}}$为视觉语义特征.

2.3.3 混合语义图

通过视频上传时间差，构造时间衰减函数，以衡量时间对跨媒体数据间话题相似性的影响，从时间维度增强话题结点间的关联.时间衰减函数为：

$$p_{i,j}=\exp\left(-\beta\left(\frac{\lfloor t_i-t_j\rfloor}{L}\right)^2\right),\tag{22}$$

其中β是控制衰减速率的正标度参数，L用于控制衰减周期，t_i和t_j表示视频i和视频j的时间戳，⌊⌋表示向下取整. 从式(22)中可以看出时间衰减函数$p_{i,j}$随时间间隔$\lfloor t_i-t_j\rfloor$的增加而降低，这表明视频i和j属于同一话题的可能性较小. 不同数据的主题相似度随它们的时间间隔呈指数下降.

如式(23)所示，通过时间衰减函数$p_{i,j}$赋予文本语义图和视觉语义图不同的连接权重，并将其加权叠加形成混合语义图$\boldsymbol{G}=(\{n_i\},\{w_{i,j}\})$. 融合过程中，将$\boldsymbol{G}^{\text{vid}}$和$\boldsymbol{G}^{\text{tex}}$的结点合并，得到混合语义图的结点集. 这样，缺失的结点信息在融合过程中得到丰富和补充，从而建立起更丰富的语义关联. 结点间的连接权重通过时间衰减函数加权求和并归一化后得到融合后的连接权重$w_{i,j}$.

$$\begin{aligned}\{n_i\}&=\{n_i^{\text{tex}}\}\cup\{n_i^{\text{vid}}\},\\\{w_{i,j}\}&=\{p_{i,j}w_{i,j}^{\text{tex}}+p_{i,j}w_{i,j}^{\text{vid}}\},\end{aligned}\tag{23}$$

其中$\boldsymbol{G}$表示构造的混合语义图，$\{n_i\}$表示结点集，$\{w_{i,j}\}$表示边的权重集. 时间信息的嵌入增强文本语义与视觉语义间的连接强度，若混合语义图中的文本语义、视觉语义以及上传时间都相似，那么所对应结点的连接较强(即边权重大)，会在图中形成一个稠密的子图. 这种融合方式使模型具有较强的可扩展性，能轻松地将其他模态的信息(如地理位置、情感倾向等)也融合到语义图中. 而且基于图的方式可以在话题个数未知的情况下进行话题聚类，使得模型更加鲁棒.

2.4 话题检测

通过 GS 算法在混合语义图中由聚类分析找出密集子图(即话题簇). GS 算法的输入是混合语义图$\boldsymbol{G}$的邻接矩阵$\boldsymbol{M}$，矩阵$\boldsymbol{M}$中的每一个元素都由混合语义图的连接权值$w_{i,j}$决定. 混合语义图的子图由概率簇$\boldsymbol{x}\in\Delta^m$表示. 其中$\Delta^m=\{\boldsymbol{x}|\boldsymbol{x}\in\mathbb{R}^m,\ x_i\geqslant 0,|\boldsymbol{x}|_1=1\}$，$m$表示图中的结点总数，$\boldsymbol{x}$表示一个映射向量，$\boldsymbol{x}=(x_i)$实现了图中的结点集到单一标准形$\mathbb{R}^m$的映射，每一个$\boldsymbol{x}\in\Delta^m$表示了各结点的组合概率，称为概率簇. 其中$x_i$表示$\boldsymbol{x}$的第$i$个分量包含结点$n_i$的概率，$x_i=0$表示该概率簇不包含顶点$n_i$. 如式(24)所示，GS 算法通过衡量子图$\boldsymbol{x}$中的平均连接强度，找到$g(\boldsymbol{x})$所有的局部极大值$\{x^*\}$. 每一个局部极大值$x^*$代表了混合语义图的一个密集子图，也就是我们所探寻的热点话题.

$$\begin{aligned}g(\boldsymbol{x})&=\boldsymbol{x}^{\text{T}}\boldsymbol{M}\boldsymbol{x},\\x^*&=\max_{\boldsymbol{x}}g(\boldsymbol{x}),\ \boldsymbol{x}\in\Delta^m,\end{aligned}\tag{24}$$

其中$\boldsymbol{M}$表示存储混合语义图的邻接矩阵，$\boldsymbol{x}$表示混合语义图中结点到话题簇的映射，$g(\boldsymbol{x})$表示子图的平均连接强度，x^*表示$g(\boldsymbol{x})$的局部极大值.挖掘出密集子图后，利用 TextRank 算法[10]在密集子图中抽取出 K 个关键词表示热点话题.

3 实验结果与分析

3.1 数据集

以国内新华网、新浪网，国外视频分享网站 YouTube、美国有线电视新闻网 CNN 等主流媒体上最热门的话题为依据，爬取了 10 个热点话题. 实验数据集详细信息如表 1 所示，包括 10 021 条文本、10 971

个视频，利用基于颜色直方图的方法[40]切分视频得到 89 348 个视频段，提取 227 930 个关键帧. 实验所用数据集涵盖多个角度、覆盖不同的时间跨度，范围从 1 个月到 3 年不等，涉及经济、政治、体育、生活等多个领域. 因此，实验数据具有足够的代表性，充分验证了实验的有效性.

Table 1 Experimental Dataset

表 1 实验数据集

编号	话题	文本的单词数	视频数	视频段	关键帧
1	英国脱欧	1 053	1 139	9 455	21 049
2	新冠疫情	992	1 068	7 546	19 229
3	香港暴乱	1 051	1 175	9 568	25 356
4	伊朗危机	1 042	1 101	8 659	22 513
5	马拉多纳逝世	910	917	7 398	15 846
6	2016 里约奥运会	981	1 071	8 956	22 927
7	叙利亚战争	1 041	1 161	8 735	23 584
8	科比坠机	1 012	1 161	9 648	25 339
9	全球金融危机	919	991	9 821	26 946
10	2020 美国大选	1 020	1 187	9 562	25 141
总计		10 021	10 971	89 348	227 930

3.2 评价指标

实验采用标准的精确率(*Precision*)、召回率(*Recall*)和 *F*1 值评价所提出方法的有效性.

$$Precision=\frac{|B_+|}{|A|}, \tag{25}$$

$$Recall=\frac{|B_+|}{|B|}, \tag{26}$$

$$F1=\frac{2\cdot Precision\cdot Recall}{Precision+Recall}=\frac{2|B_+|}{|A|+|B|}, \tag{27}$$

其中 A 表示检测到的话题数据集，B 表示与 A 最匹配的真实话题数据集，B_+ 表示正确检测的话题数据集. *Precision* 表示所有检测出的信息中，正确样本数目的占比；*Recall* 表示正确检测出的样本在所有样本中的占比. 由于 *F*1 值兼顾了精确率和召回率的优点，因此 *F*1 值被作为评价实验结果的主要方式.

3.3 参数设置及其分析

实验中，我们在数据集上无监督地训练一个 word-2vec 模型[35]，得到 128 维的词嵌入向量.在视觉特征提取方面，我们在 ImageNet 数据集上预训练一个 VGG16 模型[38]，并将其第 1 个全连接层的输出作为视觉特征，维度大小为 4 096. 对于文本特征，2 层 LSTM 维度相同，设置为 64 维. 上下文向量 $\boldsymbol{u}_{wod}$ 和 $\boldsymbol{u}_{sen}$ 设置为 128 维，并且随机初始化.

此外，时间衰减函数(见式(22))中设置了 2 个参数，固定的时间单位因子 L 和正标度参数 β. L 控制时间函数的衰减周期，对实验结果影响较小，在我们的模型中设置 L=10；标度参数 β 控制函数的衰减速率，即控制检测话题的感兴趣粒度(热点话题的持续时间)，较小的 β 使得结点之间的连接权重更大，加强数据之间的联系，使得大粒度的话题更容易形成密集子图. 例如话题 6(2016 里约奥运会)的粒度大于话题 5(马拉多纳逝世)，所以在话题检测时相似性较弱的数据结点更容易被划分到话题 6，从而带来一些噪声. 相反，较大的 β 值，削弱了结点之间的联系，有利于小粒度的话题，但可能会丢失一些大粒度话题的正确结点. 我们设置不同的 β 值，以探究 β 对实验结果的影响. 如图 4 所示，话题检测的 *F*1 值随着 β 的增长，先增大后减小. 这是因为我们选取的话题粒度不一，β 较小时，大粒度的话题 *F*1 值较高，小粒度的话题 *F*1 值较低，而当 β 较大时，则小粒度话题的 *F*1 值高，大粒度话题的 *F*1 值低. 所以在我们的数据集中选取 β=0.25 时能获得最佳的实验效果.

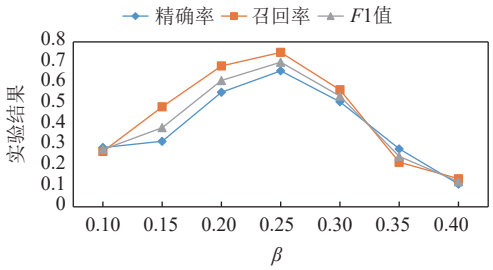


Fig. 4 Effect of β on topic detection

图 4 β 对话题检测的影响

3.4 实验结果对比分析

3.4.1 基线方法

为了全面评估本文所提出的模型，我们将本文模型与 8 种基线方法进行对比，以验证模型的有效性.

1)LDA(latent Dirichlet allocation)[14]. 该方法利用视频文本进行主题挖掘，通过推测文档的主题分布，得到视频的主题分布. 实验中，我们将文本输入到 LDA，并设置挖掘的主题数量，得到文本-主题分布向量，最后通过聚类方法得到主题.

2)BTM(biterm topic model)[17]. 该方法与 LDA 主题模型类似，BTM 模型对多个词组建模更适用于短文本. 实验中，我们将文本输入到 BTM，并设置挖掘的主题数量，得到文本-主题分布向量，最后通过聚类方法得到视频主题.

3)Doc2vec(document to vector)[41]. 该方法属于一种无监督算法，通过预训练的 Doc2vec 模型，从变长的文本中学习得到固定长度的特征表示. 在实验中，

我们将变长的文本信息编码为128维的特征向量，再通过对文档向量的聚类得到视频主题.

4)BiLSTM_Attn(attention-based bidirectional LSTM)[42]. 该方法通过双向长短期记忆网络挖掘文本上下文语义，并利用注意力对句子加权求和得到文本特征表示，再经过分类得到不同主题.

5)VSD_Attn(attention for video saliency detection)[43].该方法通过视觉显著性的注意力模型，提取视频局部显著语义特征和全局语义特征，并进行分类表示以区分不同话题.

6)CSG(capsule semantic graph)[44]. 该方法利用文本中词共现关系构建关键字子图，并通过相似度连接形成胶囊语义图，最后在语义图上进行聚类得到不同的话题.

7)TopicBERT(topic detection using BERT)[45]. 该方法通过多模态命名实体识别和Bert挖掘单词间的语义关联，并通过结构规则和实体类别增强话题信息.

8)SMMTM(unsupervised multimodal topic model)[46]. 该方法通过图像信息丰富文本语义，解决文本稀疏问题，并利用无监督的多模态主题模型对文本和图像信息进行建模和分类.

3.4.2 实验结果分析

表2展示了仅利用文本信息(标题+标签)、仅利用视觉信息(视频理解+字幕)以及文本与视觉信息联合的实验结果. 通过前10个主题的平均精确率、平均召回率和平均$F1$值评估话题检测方法的总体性能. 实验结果表明，提出的模型比其他话题检测方法表现更优，与基线模型相比提升20%~30%(平均$F1$值).

如表2所示，LDA模型在3组数据中的表现都不好. 因为LDA是基于概率的主题模型，具有主题模型固有的局限性，导致LDA对长文本较为敏感，对短文本的检测效果较差. 而数据集中的文本由视频标题组成，通常只含有10多个单词，长度短、词共现频率低，导致LDA检测的效果较差. 此外，视觉信息由视频理解和字幕信息组成，是对视频底层细节的客观描述，包含较多的噪声以及与主题无关的信息，导致LDA检测效果较差. 另一方面，直接合并文本和视觉信息虽然丰富了文本，但引入了大量的噪声，导致LDA模型的检测效果降低.

BTM模型也是一种基于概率的主题模型，通过共现词对预测文档主题，比LDA模型更适用于短文本. 实验结果表明，BTM模型对文本的检测效果相比LDA有一定的提升，但是在视觉信息上实验效果仍然较差，因为BTM模型以词袋模型编码单词作为输入，忽视了词汇间的语义关联，再加上视觉信息中所含的噪声较多，因此实验效果不佳.

Table 2 Average Experimental Results Under 10 Topics In Multiple models

表2 各模型中在10个话题下的平均实验结果

数据类型	模型	平均精确率	平均召回率	平均$F1$值
文本	LDA[14]	0.373 0	0.422 1	0.396 0
	BTM[17]	0.578 5	0.528 1	0.552 1
	Doc2vec[41]	0.555 8	0.561 4	0.558 5
	BiLSTM-Attn[42]	0.723 0	0.511 1	0.598 9
	CSG[44]	0.607 1	0.613 2	0.610 1
视觉	LDA[14]	0.288 2	0.380 3	0.327 9
	BTM[17]	0.333 3	0.454 6	0.384 6
	Doc2vec[41]	0.253 7	0.455 6	0.325 9
	BiLSTM_Attn[42]	0.380 6	0.637 5	0.476 6
	VSD_Attn[43]	0.672 3	0.367 5	0.475 2
视觉+文本	LDA[14]	0.306 8	0.402 6	0.348 2
	BTM[17]	0.289 1	0.326 4	0.306 6
	Doc2vec[41]	0.451 8	0.321 5	0.375 6
	BiLSTM_Attn[42]	0.574 1	0.478 6	0.522 0
	TopicBert[45]	0.742 8	0.664 1	0.701 2
	SMMTM[46]	0.627 8	0.671 4	0.648 8
	本文	0.770 3	0.739 6	0.754 6

Doc2ve是一种无监督文本向量模型，相比LDA在3个评价指标上都有较大的提升，相比BTM模型在平均召回率上有一定的提升. 这是因为Doc2vec方法克服了词袋模型的缺点，考虑到词汇间的语义关联. 但Doc2vec在视觉信息上的实验效果也较差，因为其仅仅是文本中多个词向量的平均，没有区分不同单词的重要性，并且模型忽视了单词间的时序信息，导致模型的平均精确率不高.

BiLSTM_Attn模型基于循环神经网络，对于文本信息，获得了较高的平均精确率. 因为模型利用长短期记忆网络捕捉时序信息，此外注意力机制区分单词的重要性，获得精确的语义表达，故得到较高的平均精确率. 但受限于文本信息较少、文本间语义关联较少，导致模型的平均召回率较低、平均$F1$值也较低. 对于视觉信息，注意力机制的加入使得模型的平均召回率有一定程度的提高，但受噪声影响，平均精确率提升并不明显，这导致模型平均$F1$值不高.

VSD_Attn方法获得较高的平均精确率，这表明若视频包含了相同的显著性区域，则这些视频通常属于同一话题. 但是，此方法的平均召回率和平均

*F*1 值较低，因为与文本信息相比视觉内容的噪声更少，平均精准率更高，但是更容易受到视频编辑、光照、拍摄角度变化的影响. 导致显著性区域相似性检测不准确，因此当只有视觉信息时，平均召回率非常低. 事实上，同一个话题可以被分割成多个视觉场景，而基于显著性区域检测的方法只能将视觉特征相似的场景组合到一起，导致同一话题下的视频缺失.

CSG 方法通过词共现关系构建胶囊语义图，从而挖掘话题信息，相比传统的主题模型获得更高的平均召回率和平均 *F*1 值. 但由于稀疏文本中词共现频率低，再加上个人语言表达习惯问题，相同的视频可能由不同的词描述，导致相同的语义无法关联，在社区检测中相关结点无法被准确地划分到同一个子图中，导致准确率较低.

TopicBert 方法通过 Bert 和多模态命名实体识别增强文本间的语义关联，获得较高的平均精确率，但平均召回率提高并不明显. 事实上，一个话题包含多种类别的视频，可能由完全不同或部分不同的场景组成，通过嵌入实体类别划分视频只能将内容相似的场景组合在一起，更多不同场景的视频将会丢失. 另外基于规则的主题优化方法，也很难将所有的视频都挖掘形成话题.

SMMTM 方法结合短文本与图像信息，通过视觉相似性，增强文本间的语义关联，但是对视觉特征的简单处理，并没有解决视觉相似性检测不精确问题，因此模型的平均精确率不高. 这说明 SMMTM 虽然可以增强文本间的语义关联，但是不可避免地引入噪声.

此外，通过分析数据我们得出 2 个结论：

1）文本比视觉信息在多个模型中表现得更好，这说明文本特征比视觉特征更重要. 通过文本与视觉信息在 LDA, BTM, Doc2vec, BiLSTM_Attn 中的结果可以发现，将视觉特征转化为文本信息后，在一定程度上提升了平均召回率，但平均精确率受到了较大影响，导致平均 *F*1 值降低. 因为视频理解将视觉信息翻译成自然语言的过程中丢失了视觉特征的特异性. 例如：在科比坠机的话题下，会出现科比打篮球的场景，在里约奥运会这个话题下同样会出现打篮球的场景. 但通过视频理解它们会被翻译为相同的描述：a group of men playing basketball，相似但不相关视觉信息被翻译成相同的语言描述，且在后续的处理中它们被认为是完全相同的特征，导致纯视觉信息检测的平均精确率较低. 相反，视觉信息检测的平均召回率高于文本检测，一方面是因为文本较为稀疏，所含的有效信息少、噪声多，使得文本检测的平均召回率低，另一方面 OCR 从视频中获取的字幕信息与视频内容高度相关，很容易与视频理解得到的语义信息产生联系，从而将更多相关的数据聚集到一起，使得视觉信息检测的平均召回率高于文本检测.

2）视觉特征与文本特征的直接合并效果比单一文本更差，这表明实现视觉信息与文本信息的互补，需要找到适合的融合方法. 以 BiLSTM_Attn 模型为例，文本与视觉信息的直接合并相比于纯文本，平均精确率下降 15%，平均召回率下降 4%，平均 *F*1 值下降 8%. 这是因为视频特征经由视频理解会引入部分的无意义词汇，简单地合并造成了文本信息的泛化、文本间的语义关联强度降低，导致模型检测效果较差. 提出的方法在 3 种评价指标上都获得了最优的实验性能，相比于单一文本或视觉的最优值，本文方法的平均精确率提升约 5%，平均召回率提升约 10%，平均 *F*1 值提升约 15%. 因为通过双层注意力，区分了不同单词和句子的重要性，极大地降低噪声的干扰. 此外，将文本和视觉语义融合到同一个混合语义图中，保证文本和视觉语义优势互补，避免相互干扰，使得模型的平均精确率和平均召回率都有较大的提高.

此外，通过消融实验验证本文所提出方法的有效性. 如表 3 所示，文本、视觉分别表示仅利用文本、视觉数据构建语义图，文本+视觉表示文本与视觉信息融合构建语义图，文本+视觉+时间表示嵌入时间的文本和视觉语义图融合.

Table 3 Ablation Experiment

表 3 消融实验

数据类型	平均精确率	平均召回率	平均 *F*1 值
文本	0.437 2	0.658 2	0.525 4
视觉	0.752 4	0.501 5	0.601 8
文本+时间	0.526 4	0.657 7	0.584 7
视觉+时间	0.751 5	0.577 5	0.653 1
文本+视觉	0.741 4	0.621 2	0.675 9
文本+视觉+时间	0.770 3	0.739 6	0.754 6

通过对比实验数据可以得出 2 个结论：

相比单一文本或视觉，跨媒体融合的方式在话题检测效果上有较大的提升. 虽然文本+视觉相比于单一视觉信息，平均精确率略有降低（降低 1%），但平均召回率有较大的提高（提高 12%），平均 *F*1 值提高 7%. 这是由于视觉特征会受到光照、编辑、拍摄角度等的影响，导致同一场景下的视觉特征可能完全不同，因此单一视觉信息具有精确率高、召回率低的

特点. 融合文本信息后, 文本语义相似的视频也会被纳入同一话题, 这个过程中带入部分噪声数据, 导致平均精确率略有降低, 但同时平均召回率有较大的提升. 所以我们在混合语义图中嵌入时间特征, 通过时间相似性, 从时间维度给文本和视觉语义关联增加权重, 使得语义上难以区分的数据可以通过时间特征被轻松区分, 很好地解决了这一问题.

嵌入时间特征后, 本文模型的平均精确率提升3%, 平均召回率提升 11%, 平均 $F1$ 值提升 8%, 表明在混合语义图中嵌入时间特征能增强混合语义图的区分能力. 这是因为热点话题的爆发都具有很强的时间聚集性, 所以时间特征对于不同的热点话题具有很强的区分能力. 通过时间衰减函数将时间分布特征嵌入到混合语义图中, 增强相同时间跨度内文本与视觉语义间的连接强度, 从而避免具有相似文本和视觉特征的噪声干扰. 因此在本文模型下, 文本和视觉难以区分的数据可通过时间分布特征轻松区分. 此外, 这种时间系数的软量化在一定程度上保证话题在时间上的连续性, 防止由于时间线硬分割而导致的同一话题被错误分割.

为了进一步证明时间特征可以有效增强话题检测效果, 我们统计了短期爆发和周期型爆发话题的数据分布情况. 如图 5 所示, 对于“科比坠机”这一话题, 网络视频在短期内集中爆发, 文本和视觉特征具有很强的时间相似性. 通过时间戳的嵌入, 可以在时间维度建立跨媒体数据间的语义关联, 增强密集子图中结点的连接强度, 所以模型对于短期爆发的热点话题具有很强的检测能力. 对于“新冠疫情”这一话题, 在不同时段内形成多个密集子图, 导致热点话题被分割成多个密集子图. 但是, 同一话题下密集子图的语义相似性远高于不同话题下的密集子图. 因此, 通过计算密集子图间的平均欧氏距离(Euclidean distance)能有效实现密集子图的合并, 得到完整的热点话题.

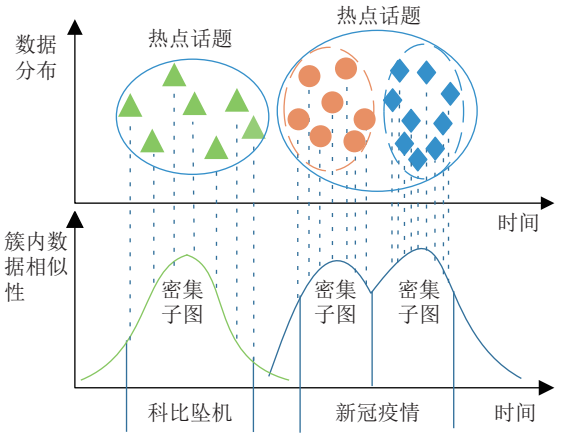


Fig. 5 Time distribution characteristics of topics

图 5 话题的时间分布特征

表 4 展示了从 10 个话题中挖掘出的高频话题词. 以话题 8 为例, 我们不仅挖掘出了 Kobe, death, NBA 等文本中出现的高频词, 还挖掘出 playing, running, basketball 等低频词. 视觉特征向文本信息转换的过程中生成了大量新词, 增加了 playing, running, basketball 这类低频词的出现频率, 丰富了文本语义空间. 此外, 视频理解生成的 basketball, player 与标题和字幕中的 NBA, playing 高度相关, 增强了跨媒体数据间的语义关联强度, 使得热点话题更容易形成密集子图, 从而被准确地检测.

表 5 展示了 3 个具体的例子, 以表明本文提出方法的有效性. 例 1 标题中并没有出现与“科比坠机”相关的单词, 导致语义关联缺失, 基线方法仅通过标题无法准确划分. 而通过视频理解和字幕, 从视觉语义中挖掘出 Kobe, bryant, crash, accident 等词, 建立起

Table 4 Hot Topic Words

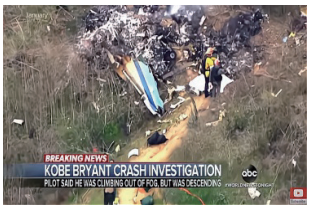
表 4 热点话题词

编号	话题	话题词
1	英国脱欧	UK, EU, brexit, trade, deal, politics, laws, johnson, market, vote
2	新冠疫情	covid, virus, death, masks, doctor, Wuhan, flight, hospital, oxygen, economic
3	香港暴乱	Hong Kong, protests, China, police, law, violent, strikes, clash, support
4	伊朗危机	crisis, Iran, hostages, nuclear, history, economic, financial, America, market, oil
5	马拉多纳逝世	Maradona, death, football, playing, running, heart, attack, tribute, legend, grassland
6	2016 里约奥运	Rio, Olympics, sports, champion, gold, games, vs, player, Brazil, highlights
7	叙利亚战争	Syrian, war, conflict, military, children, airstrikes, battle, civil, homeless, years
8	科比坠机	Kobe, bryant, death, basketball, NBA, helicopter, playing, daughter, crash, running
9	全球金融危机	financial, global, recession, economic, world, collapse, subprime, crash, stock, dumping
10	2020 美国大选	America, election, presidential, Trump, Biden, results, campaign, vote, states, debate

该视频结点与“科比坠机”话题的联系，这有效提高模型的召回率；例 2 标题中出现“2020 presidential victory”和“Brexit”，极易被错分为“2020 美国大选”. 而通过融合视觉语义以及时间特征，增强了视觉语义 London 与 Brexit 的语义关联，这使得该结点在聚类过程中被正确划分. 同样地，例 3 中语义融合时，受文本噪声 financial markets 的影响，该结点极易与话题“全球金融危机”混淆，而融合方法从时间维度对该结点进行相似度加权，因为此视频上传时间处于新冠疫情大爆发期间，并且视觉语义 masks 与新冠疫情密切相关，这增强了该视频结点与新冠疫情的语义关联强度，从而使得该视频被正确划分，因此模型具有较高的精确率.

Table 5 Three Detection Examples of Typical Topics

表 5 3 个典型话题检测示例

示例	视频	标题	字幕	视频理解	真实标签	方法	预测标签
例 1		My world collapsed. I can't believe my eyes. He is my favorite player	Kobe bryant crash investigation plot said he was climbing out of fog but was descending	One person is handling the accident scene	科比坠机	LDA	马拉多纳逝世
						BTM	马拉多纳逝世
						Doc2vec	马拉多纳逝世
						BiLSTM_Attn	2016 里约奥运
						VSD_Attn	叙利亚战争
						CSG	叙利亚战争
						TopicBert	叙利亚战争
						SMMTM	叙利亚战争
						本文	科比坠机
例 2		How will Biden's 2020 presidential victory affect Brexit	I am delighted to be back in London	A man in a suit was walking down the street	英国脱欧	LDA	2020 美国大选
						BTM	2020 美国大选
						Doc2vec	2020 美国大选
						BiLSTM_Attn	2020 美国大选
						VSD_Attn	2020 美国大选
						CSG	2020 美国大选
						TopicBert	2020 美国大选
						SMMTM	2020 美国大选
						本文	英国脱欧
例 3		How coronavirus will impact financial markets	States	A man wearing a mask	新冠疫情	LDA	全球金融危机
						BTM	全球金融危机
						Doc2vec	全球金融危机
						BiLSTM_Attn	全球金融危机
						VSD_Attn	2020 美国大选
						CSG	全球金融危机
						TopicBert	全球金融危机
						SMMTM	全球金融危机
						本文	新冠疫情

4 总 结

本文提出一个新的跨媒体语义关联增强方法，解决文本特征稀疏导致的同一话题下文本语义特征差异大且关联强度弱的问题. 首先，构造一个双层注意力模型，通过单词级注意力和句子级注意力捕捉文本核心语义特征；然后，通过视频理解和场景文字识别，生成与视频内容高度相关的文本描述信息，丰富文本语义空间；最后，将文本语义和视觉语义通过时间衰减函数，加权融合形成混合语义图，增强跨媒体语义关联，使得话题的分割更为准确. 实验结果表

明,本文提出的方法能有效提高网络视频热点话题检测效果.

作者贡献声明:张承德提出论文思路,并指导论文撰写和修改;刘雨宣负责方法设计,完成实验并撰写论文;肖霞对实验设计和论文撰写提出指导意见;梅凯为实验数据提供支持,并对数据进行处理.

参 考 文 献

[1] Peng Yuxin, Qi Jinwei, Huang Xin. Current research status and prospects on multimedia content understanding[J]. Journal of Computer Research and Development, 2019, 56(1): 183−208 (in Chinese)
(彭宇新, 綦金玮, 黄鑫. 多媒体内容理解的研究现状与展望[J]. 计算机研究与发展, 2019, 56(1): 183−208)

[2] China Internet Network Information Center. The 48th statistical report on the development status of China's Internet[EB/OL]. Beijing: CNNIC, 2021 [2022-09-11]. http://n2.sinaimg.cn/finance/a2d36afe/20210827/FuJian1.pdf (in Chinese)
(中国互联网络信息中心(CNNIC). 第48次中国互联网络发展现状统计报告[EB/OL]. 北京: 中国互联网络信息中心, 2021 [2022-09-11]. http: //n2.sinaimg.cn/finance/a2d36afe/20210827/FuJian1.pdf)

[3] Amudha S, Niveditha V R, Kumar P S R, et al. YouTube trending video metadata analysis using machine learning[J]. International Journal of Advanced Science and Technology, 2020, 29(7): 3028−3037

[4] Pervaiz R, Aloufi K, Zaidi S S R, et al. A methodology to identify topic of video via n-gram approach[J]. International Journal of Computer Science and Network Security, 2020, 20(1): 79−94

[5] Pang Junbiao, Hu Anjing, Huang Qingming, et al. Increasing interpretation of web topic detection via prototype learning from sparse poisson deconvolution[J]. IEEE Transactions on Cybernetics, 2018, 49(3): 1072−1083

[6] Shi Cunhui, Hu Yaokang, Feng Bin, et al. A hierarchical knowledge based topic recommendation method in public opinion scenario[J]. Journal of Computer Research and Development, 2021, 58(8): 1811−1819 (in Chinese)
(史存会, 胡耀康, 冯彬, 等. 舆情场景下基于层次知识的话题推荐方法[J]. 计算机研究与发展, 2021, 58(8): 1811−1819)

[7] Xu Xiaoying, Dutta K, Ge C. Do adjective features from user reviews address sparsity and transparency in recommender systems[J]. Electronic Commerce Research and Applications, 2018, 29: 113−123

[8] Cui Wanqiu, Du Junping, Kou Feifei, et al. The social and conceptual semantic extended search method for microblog short text[J]. Journal of Computer Research and Development, 2018, 55(8): 1641−1652 (in Chinese)
(崔婉秋, 杜军平, 寇菲菲, 等. 面向微博短文本的社交与概念化语义扩展搜索方法[J]. 计算机研究与发展, 2018, 55(8): 1641−1652)

[9] Liu Hairong, Yan Shuicheng. Robust graph mode seeking by graph shift[C]// Proc of the Int Conf on Machine Learning. New York: ACM, 2010: 671−678

[10] Li Wengen, Zhao Jiabao. TextRank algorithm by exploiting Wikipedia for short text keywords extraction[C]// Proc of the 3rd Int Conf on Information Science and Control Engineering (ICISCE). Piscataway, NJ: IEEE, 2016: 683−686

[11] Cao Juan, Ngo C W, Zhang Yongdong, et al. Tracking web video topics: Discovery, visualization, and monitoring[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(12): 1835−1846

[12] Liu Wei, Li Weimin, Lei Jiang, et al. Topic detection and tracking based on event ontology[J]. IEEE Access, 2020, 8: 98044−98056

[13] Liu Tianpeng, Xue Feng, Sun Jian, et al. A survey of event analysis and mining from social multimedia[J]. Multimedia Tools and Applications, 2020, 79(45): 33431−33448

[14] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993−1022

[15] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical Dirichlet processes[J]. Journal of the American Statistical Association, 2006, 101(476): 1566−1581

[16] Jin Ou, Liu N N, Zhao Kai, et al. Transferring topical knowledge from auxiliary long texts for short text clustering[C]//Proc of the 20th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2011: 775−784

[17] Yan Xiaohui, Guo Jiafeng, Lan Yanyan, et al. A biterm topic model for short texts[C]//Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 1445−1456

[18] Zhao Zhicheng, Xiang Rui, Su Fei. Complex event detection via attention-based video representation and classification[J]. Multimedia Tools and Applications, 2018, 77(3): 3209−3227

[19] Zhang Jiyong, Li Wenchao, Li Liang wenchao, et al. Enabling 5G: Sentimental image dominant graph topic model for cross-modality topic detection[J]. Wireless Networks, 2020, 26(3): 1549−1561

[20] Zhao Sicheng, Gao Yue, Ding Guiguang, et al. Real-time multimedia social event detection in microblog[J]. IEEE Transactions on Cybernetics, 2017, 48(11): 3218−3231

[21] Zhang Chengde, Lu Shaozhen, Zhang Chengming, et al. A novel hot topic detection framework with integration of image and short text information from Twitter[J]. IEEE Access, 2018, 7: 9225−9231

[22] Kojima A, Izumi M, Tamura T, et al. Generating natural language description of human behavior from video images[C]//Proc of the 15th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2000: 728−731

[23] Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of Actions[J]. International Journal of Computer Vision, 2002, 50(2): 171−184

[24] Venugopalan S, Xu Huijuan, Donahue J, et al. Translating videos to natural language using deep recurrent neural networks[C] //Proc of the 2015 Conf of the North American Chapter of The Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2015: 1494−1504

[25] Li Yao, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure[C]//Proc of the IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2015: 4507−4515

[26] Zhang Junchao, Peng Yuxin. Object-aware aggregation with bidirectional temporal graph for video captioning[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition.

Piscataway, NJ: IEEE, 2019: 8327−8336

[27] Aafaq N, Akhtar N, Liu Wei, et al. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 12487−12496

[28] Mishra A, Shekhar S, Singh A K, et al. Ocr-vqa: Visual question answering by reading text in images[C] // Proc of the 2019 Int Conf on Document Analysis and Recognition (ICDAR). Piscataway, NJ: IEEE, 2019: 947−952

[29] Wang Yue, Li Jing, Lyu M R, et al. Cross-Media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings[C]//Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 3311−3324

[30] Zhong Qinghong, Qiao Xiaodong, Zhang Yunliang, et al. Cross-media fusion method based on LDA2Vec and residual network[J]. Data Analysis and Knowledge Discovery, 2019, 3(10): 78−88

[31] Ying Long, Yu Hui, Wang Jinguang, et al. Fake news detection via multi-modal topic memory network[J]. IEEE Access, 2021, 9: 132818−132829

[32] Xue Junxiao, Wang Yabo, Tian Yichen, et al. Detecting fake news by exploring the consistency of multimodal data[J/OL]. Information Processing & Management, 2021 [2022-09-11].https://doi.org/10.1016/j.ipm.2021.102610

[33] Li Chuanzhen, Liu Minqiao, Cai Juanjuan, et al. Topic detection and tracking based on windowed DBSCAN and parallel KNN[J]. IEEE Access, 2020, 9: 3858−3870

[34] Yang Zichao, Yang Diyi, Dyer C, et al. Hierarchical attention networks for document classification[C]//Proc of the Int Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2016: 1480−1489

[35] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C/OL]//Proc of the 1st Int Conf on Learning Representations. Ithaca, NY, Cornell University, 2013 [2022-09-11].https://doi.org/10.48550/arXiv.1301.3781

[36] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673−2681

[37] Huang Jie, Zhou Wengang, Li Houqiang, et al. Attention-based 3D-CNNs for large-vocabulary sign language recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(9): 2822−2832

[38] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C/OL] //Proc of the 3rd Int Conf on Learning Representations(ICLR). Ithaca, NY, Cornell University, 2015 [2022-09-11].https://doi.org/10.48550/arXiv.1409.1556

[39] Chen D L, Dolan W B. Collecting highly paralleldata for paraphrase evaluation[C]//Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2011: 190−200

[40] Guru D S, Suhil M. Histogram based split and merge framework for shot boundary detection[M]//Mining Intelligence and Knowledge Exploration. Cham, Switzerland: Springer, 2013: 180−191

[41] Vahidnia S, Abbasi A, Abbass H A. Embedding-based detection and extraction of research topics from academic documents using deep clustering[J]. Journal of Data and Information Science, 2021, 6(3): 99−122

[42] Zhou Peng, Shi Wei, Tian Jun, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 207−212

[43] Jian Muwei, Wang Jiaojin, Yu Hui, et al. Integrating object proposal with attention networks for video saliency detection[J]. Information Sciences, 2021, 576: 819−830

[44] Yang Shuang, Tang Yan. News topic detection based on capsule semantic graph[J]. Big Data Mining and Analytics, 2022, 5(2): 98−109

[45] Asgari-Chenaghlu M, Feizi-Derakhshi M R, Balafar M A, et al. TopicBERT: A cognitive approach for topic detection from multimodal post stream using BERT and memory –graph[J/OL]. Chaos, Solitons & Fractals, 2021 [2022-09-11].https://doi.org/10.1016/j.chaos.2021.111274

[46] Zhang Huakui, Cai Yi, Zhu Bingshan, et al. Multimodal topic modeling by exploring characteristics of short text social media[J/OL]. IEEE Transactions on Multimedia, 2022([2022-09-11].https://ieeexplore.ieee.org/abstract/document/9696359

Zhang Chengde, born in 1982. PhD, associate professor. Member of AAIA, CCF, IEEE. His main research interests include multimedia information retrieval, data analysis, and machine learning.

张承德, 1982 年生. 博士, 副教授. AAIA, CCF, IEEE 会员. 主要研究方向为多媒体信息检索、数据分析和机器学习.

Liu Yuxuan, born in 1996. Master candidate. Member of CCF. His main research interests include multimedia information retrieval, data mining, and image processing.

刘雨宣, 1996 年生. 硕士研究生. CCF 会员. 主要研究方向为多媒体信息检索、数据挖掘和图像处理.

Xiao Xia, born in 1986. PhD. Her main research interests include multimedia information retrieval, data analysis, and intergenerational mobility.

肖　霞, 1986 年生. 博士. 主要研究方向为多媒体信息检索、数据分析和代际流动.

Mei Kai, born in 1996. Master candidate. Member of CCF. His main research interests include multimedia information retrieval, data analysis, and machine learning.

梅　凯, 1996 年生. 硕士研究生. CCF 会员. 主要研究方向为多媒体信息检索、数据分析和机器学习.