软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(3):1140−1153 [doi: 10.13328/j.cnki.jos.007078]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向视频的细粒度多模态实体链接*

赵海全[1,2], 王续武[1,2], 李金亮[3], 李直旭[1,2], 肖仰华[1,2]

[1](复旦大学 计算机科学技术学院, 上海 201203)

[2](上海市数据科学重点实验室(复旦大学), 上海 201203)

[3](苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通信作者: 李直旭, E-mail: zhixuli@fudan.edu.cn

摘 要: 随着互联网和大数据的飞速发展, 数据规模越来越大, 种类也越来越多. 视频作为其中重要的一种信息方式, 随着近期短视频的发展, 占比越来越大. 如何对这些大规模视频进行理解分析, 成为学界关注的热点. 实体链接作为一种背景知识补全方式, 可以提供丰富的外部知识. 视频上的实体链接可以有效地帮助理解视频内容, 从而实现对视频内容的分类、检索、推荐等. 但是现有的视频链接数据集和方法的粒度过粗, 因此提出面向视频的细粒度实体链接, 并立足于直播场景, 构建了细粒度视频实体链接数据集. 此外, 依据细粒度视频链接任务的难点, 提出利用大模型抽取视频中的实体及其属性, 并利用对比学习得到视频和对应实体的更好表示. 实验结果表明, 该方法能够有效地处理视频上的细粒度实体链接任务.

关键词: 细粒度; 视频实体链接; 数据集; 大语言模型; 对比学习

中图法分类号: TP391

中文引用格式: 赵海全, 王续武, 李金亮, 李直旭, 肖仰华. 面向视频的细粒度多模态实体链接. 软件学报, 2024, 35(3): 1140−1153. http://www.jos.org.cn/1000-9825/7078.htm

英文引用格式: Zhao HQ, Wang XW, Li JL, Li ZX, Xiao YH. Fine-grained Multimodal Entity Linking for Videos. Ruan Jian Xue Bao/Journal of Software, 2024, 35(3): 1140−1153 (in Chinese). http://www.jos.org.cn/1000-9825/7078.htm

Fine-grained Multimodal Entity Linking for Videos

ZHAO Hai-Quan[1,2], WANG Xu-Wu[1,2], LI Jin-Liang[3], LI Zhi-Xu[1,2], XIAO Yang-Hua[1,2]

[1](School of Computer Science, Fudan University, Shanghai 201203, China)

[2](Shanghai Key Laboratory of Data Science, (Fudan University), Shanghai 201203, China)

[3](School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: With the rapid development of the Internet and big data, the scale and variety of data are increasing. Video, as an important form of information, is becoming increasingly prevalent, particularly with the recent growth of short videos. Understanding and analyzing large-scale videos has become a hot topic of research. Entity linking, as a way of enriching background knowledge, can provide a wealth of external information. Entity linking in videos can effectively assist in understanding the content of video, enabling classification, retrieval, and recommendation of video content. However, the granularity of existing video linking datasets and methods is too coarse. Therefore, this study proposes a video-based fine-grained entity linking approach, focusing on live streaming scenarios, and constructs a fine-grained video entity linking dataset. Additionally, based on the challenges of fine-grained video linking tasks, this study proposes the use of large models to extract entities and their attributes from videos, as well as utilizing contrastive learning to obtain better representations of videos and their corresponding entities. The results demonstrate that the proposed method can effectively handle

* 基金项目: 国家重点研发计划(2020AAA0109302); 国家自然科学基金(62072323, 62102095); 上海市科技创新行动计划(22511105902, 22511104700); 上海市科技重大专项(2021SHZDZX0103); 上海市科学技术委员会资助项目(22511105902)
本文由“面向多模态数据的新型数据库技术”专题特约编辑彭智勇教授、高云君教授、李国良教授、许建秋教授推荐.
收稿时间: 2023-07-18; 修改时间: 2023-09-05; 采用时间: 2023-10-24; jos 在线出版时间: 2023-11-08
CNKI 网络首发时间: 2023-12-22

fine-grained entity linking tasks in videos.

Key words: fine-grained; video entity linking; dataset; large language model; contrastive learning

随着互联网和大数据的飞速发展, 互联网用户群体不断扩大, 数据的规模和种类也越来越庞大. 这些数据包含各种各样的信息, 也包括文字、图片、视频等不同模态. 而近些年来, 移动设备和短视频的飞速发展, 导致视频在多种数据中占据了越来越大的比重, 种类也越来越多. 如何有效地对这些视频进行分析, 成为学界关注的热点. 近些年, 随着深度学习技术的不断发展和应用, 产业对人工智能算法要求越来越高. 深度学习技术通过对大量数据的学习和处理, 自动学习特征并进行分类、聚类等操作, 从而实现对数据的分析和挖掘. 如何有效、合理地利用人工智能算法对大规模、多种类视频进行理解分析, 愈发成为学界研究的重点.

实体链接[1]技术是将实体提及链接到知识库中, 通过提供更加丰富的背景信息, 加强计算机对内容的理解. 而视频实体链接[2]技术是将视频中的实体与知识库中的实体进行关联, 从而为视频内容提供更加丰富的背景信息. 例如: 在一段视频中出现了“武球王”, 将“武球王”与知识库中的“武磊”链接到一起, 这段视频很大可能与足球有关, 从而为视频提供更多关于足球方面的背景知识. 因此, 视频实体链接技术可以作为视频内容的分类、检索、推荐的基础, 从而为用户提供更加智能化、个性化的服务.

然而, 现有视频实体链接的任务粒度过粗[2−4], 并不适用于当前大规模视频理解分析的要求. Li 等人[2]从 YouTube 等平台通过关键字检索构建出视频实体链接的数据集, 例如, 从“科比·布莱恩特生涯高光集锦”的视频中识别出“科比·布莱恩特”并链接到知识库中. 还有研究[3]从纪录片中构建视频链接数据集, 识别出视频中的动物链接到狮子或者老虎上. 一方面, 现有的视频实体链接任务中, 数据集的构建来源于纪录片或名人职业生涯集锦等, 视频的噪声较小, 语音模态表述较为规范, 视觉模态中的主体也较为突出, 此类任务的难度并不大; 另一方面, 现有方法链接的粒度较粗, 例如识别出视频中的实体是一只鸟, 但并没有分辨出是杜鹃还是黄鹂, 面对大规模、多种类的视频分析时, 仍然有一定的局限性. 近几年, 多模态实体链接[5−16]受到了学界的广泛关注, 现有研究基于多模态实体链接任务构建了多个数据集, 并提出了解决这些任务的方案. 但是, 现有的方法主要针对静态图像和文本, 在其中识别出实体提及, 然后利用多模态对齐和融合等方法进行实体消歧, 最后将实体链接到多模态知识库中. 其中不乏细粒度的多模态实体链接, 但还没有推广到视频领域.

因此, 本文立足于对视频内容进行细粒度理解, 提出了视频实体链接任务, 并立足于直播场景, 构建了细粒度实体链接数据集, 数据集中包含有 1 838 个直播视频、11 248 个细粒度商品. 如图 1 所示, 给定一个讲解商品的视频, 视频实体链接任务需要将视频链接到与知识库中对应的商品上. 商品本身包含种类、品牌、型号等各种信息, 同一种商品会有不同的品牌, 例如, “华为手机”和“小米手机”是不同品牌的同一实体; 同一品牌同一商品还会有不同的型号, 在“华为手机”中, 还会出现“华为 mate50”和“华为 P50”两款相似手机. 同一类商品的不同品牌和型号之间的复杂差异, 给实体消歧带来了很大的挑战.

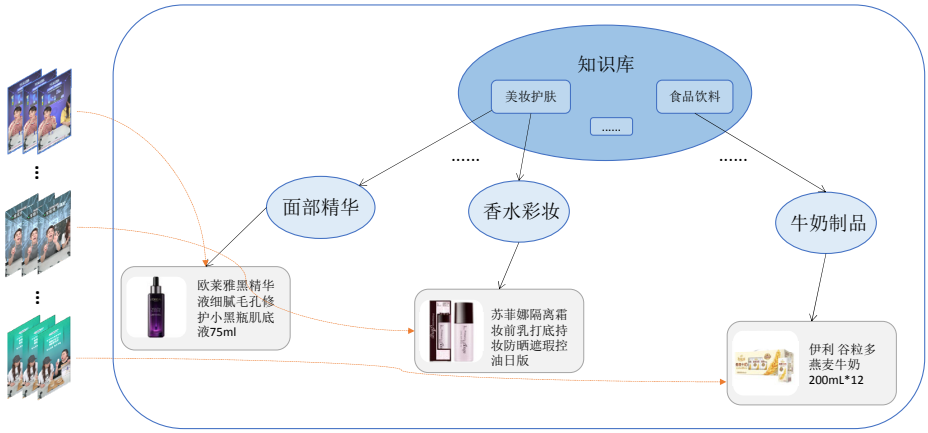

图 1 视频实体链接任务

视频作为一种多模态数据, 包含图像、语音等多种信息. 如何利用多种模态之间的信息交互, 提升不同模态之间信息表达的能力, 是当前学术界研究的一个重要方向. 在直播中, 主播讲解场景较为固定, 现有基于图像预训练的模型应用到直播视频时, 会得到较为相似的表示. 本文为了解决这一问题, 利用对比学习的方法, 拉远特定领域中的相似视频或知识库中的相似实体的特称表示, 以更好地区分易混淆的实体.

此外, 真实场景中的视频与上映的电影、纪录片等有较大差异. 后者为了方便受众更好地理解电影情节和纪录片主题, 通常采用规范的书面表达, 语音噪声比较少、主题表达明确、观点突出. 真实场景下的视频通常采用口语化表达, 含有大量语气词、不规则的停顿以及与主题无关的讨论等, 给细粒度视频链接带来了极大挑战. 过往的研究经验[17]表明, 实体的解释信息可以提升实体链接的准确率. 而大模型具有很强的理解总结能力, 能够从大量语料中把握关键主旨信息. 本文利用大模型总结语音文本, 期望大模型抽取出讲解视频中的实体和属性, 从而更加准确地链接实体. 实验表明, 我们的方法在数据集上达到最好的效果.

综上, 本文的主要贡献如下:

1) 提出视频上的细粒度实体链接任务, 并立足于直播场景, 构建了细粒度视频实体链接数据集.
2) 针对视频细粒度实体识别任务难点, 提出利用大模型摘要抽取出实体以及实体的多种属性, 结合对比学习获取相似视频、实体的更好表示, 实现了基于对比学习的视频链接模型.
3) 实验结果表明, 视频链接模型能够有效处理细粒度视频实体链接任务.

本文第 1 节介绍多模态视频实体链接以及视频检索的相关方法和研究现状. 第 2 节介绍本文所需的基础知识, 包括对比学习、大语言模型. 第 3 节介绍细粒度视频实体链接数据集构建方式. 第 4 节介绍本文构建的基于对比学习、大语言模型的直播商品识别模型. 第 5 节通过实验分析验证所提模型的有效性. 第 6 节总结全文.

1 相关工作

1.1 多模态及视频实体链接

近年来, 多模态实体链接在学界受到了广泛的关注, Moon 等人[14]首次提出在社交媒体中构建出多模态命名实体识别数据集, 数据集主要来源于推特的文本, 期望利用视觉领域的信息补充文本信息, 在推特的图文中识别并链接到社会名人上; Adjali 等人[5,6]以及 Zhang 等人[8]提出了在社交媒体上构建多模态实体链接的数据集方法, 并提出了解决多模态实体链接任务的方法; Gan 等人[13]在电影上构建了 M3EL 数据集, 并用图文融合编码分步骤解决多模态实体链接问题. 除此之外, 由于多模态数据的多样性, Zheng 等人[10]提出零样本多模态实体链接, Zhou 等人[9]将纯文本的实体链接数据集扩充到多模态领域, Wang 等人[11]在此基础上增加了实体的主体数量和类别的多样性. 然而, 这些多模态实体链接方法更多针对的是图文对数据, 并没有应用到视频领域.

Li 等人[2]首次提及视频链接, 并利用 YouTube 选取了名人以及动物领域关键词构建了视频实体链接的数据集, 例如, “科比职业生涯集锦”构建出的视频链接到科比·布莱恩特这一名人上. 然而, 链接的粒度过于粗略, 同时, 数据集中还带有视频的标题, 大大降低了实体链接任务的难度. Venkitasubramanian 等人[3]利用动物视频和视频描述, 期望视频和描述之间能够互为信息补充, 识别出视频或视频描述文字中的主体, 并链接到知识库中. 然而, 这种识别的粒度过于粗略, 同时, 视频通常采用纪录片形式, 纪录片为了清晰表达主题, 语音中包含的噪声较少, 图像中主体较为突出.

1.2 视频检索

视频检索[18]可以分为两种任务: 一种是根据文本检索最相似的视频, 另一种是根据视频检索最相似的文本. 视频检索任务在学术界受到广泛关注, Chen 等人[19]以及 Gabeur 等人[20]主要从“专家”角度将知识注入模型中, 期望提升模型的表现. 随后, 越来越多的研究专注于端到端的模型, Bain 等人[21]、Miech 等人[22,23]提出了端到端模型, 通过视频-文本联合训练来提高下游任务的性能. Wu 等人[24]利用预训练模型并注入领域知识, 提

升了视觉预训练表示能力以及领域检索的水平. 近期, Zhao 等人[25]、Bain 等人[26]试图将 CLIP 的训练权重中学习到的知识转移到视频领域, 将视频看作一个具有时间序列的关键帧序列, 从而提高视频检索任务的性能. 在视频检索任务中, 视频的相关文本的形式大多是视频的描述, 例如“一个小男孩在雪地行走”, 而实体链接任务则是识别并链接视频中的主体, 例如识别出“小男孩”; 同时, 与视频匹配的只有文本信息, 实体链接任务是匹配到多模态知识库中的实体中. 这两种任务存在一定的区别, 但在形式上又具有一定的相似性, 并且可以通过 prompt 技术弥补两者之间的任务差距, 因此, 本文采用视频检索作为对比实验.

2 基础知识

本文所提方法主要基于大规模语言模型、对比学习, 下面就相关概念和基本知识予以介绍.

2.1 大规模语言模型

大规模语言模型是一种基于深度学习的语言模型, 通过处理大量的自然语言数据来学习语言的概率分布, 并能够对新的文本进行自然语言理解、生成和推理等任务. 其中, 基于 Transformer[27]解码器和编码器架构的模型, 如 BERT[28]、GPT 系列[29−31]、T5[32]等模型已经在自然语言处理任务中大放异彩. 基于 Transformer 编码器架构的模型在自然语言理解任务中效果极佳, 而基于 Transformer 解码器架构的模型则更适合自然语言生成任务. 这些模型的成功离不开 Transformer 架构, 如 Self-Attention 机制, 它们通过这些机制来学习语言的概率分布, 并解决自然语言处理中的各种任务.

除了自然语言理解和生成任务, 大规模语言模型在其他领域也有着广泛的应用. 例如: 在推荐系统中, 可以通过大规模语言模型来学习用户的历史行为和兴趣, 以推荐更符合用户兴趣的商品或内容; 在智能客服中, 可以通过大规模语言模型来理解用户的问题并给出答案或建议; 在信息检索中, 可以通过大规模语言模型来进行文本匹配和相似性计算, 以提高搜索结果的准确性.

ChatGPT 是一种大规模语言模型, 由 OpenAI 开发完成, 是 GPT-3[31]的进一步升级. ChatGPT 的主要目标是, 理解和生成人类的对话. 这种模型通过从大量的对话数据上对齐, 以理解人类的对话模式和结构. 与 ChatGPT 类似的大规模语言模型在大量的语料上训练, 通过生成的方式统一了自然语言处理的各种任务, 并通过 ICL (in-context-learning)等技术, 在很多任务上超过了现有最佳模型. 此外, 截至日前, 国内多家高校开源了自产模型, 例如清华大学的 ChatGLM[33,34]、复旦大学的 MOSS 等.

大规模语言模型在大量的语料上训练, 对文本信息具有很强的理解能力. 首先, 给定一段文本, 能够理解文本主旨; 其次, 通过合适的提示可以让大规模语言模型适应各种任务, 正确的提示可以提升语言模型的任务表现; 此外, 大规模语言模型凭借其丰富的背景知识, 是目前学界处理文本零样本学习的最好方式; 最后, 其本身在训练的过程中可能接受有部分包含有噪声的语料, 对噪声具有一定的适应和纠正能力. 鉴于大规模语言模型各方面的优势, 本文主要利用大规模语言模型摘要语音文本, 抽取实体以及对应的属性, 达到消除部分文本噪音的目的.

2.2 对比学习

对比学习是一种机器学习方法, 主要用于学习更好的特征表示和相似性度量. 在对比学习中, 模型通过比较两个或更多样本之间的相似性和差异性来学习特征表示, 以便更好地区分不同样本之间的关系. 对比学习在计算机视觉和自然语言处理等领域中得到了广泛应用, 如人脸识别、图像检索、文本分类和推荐系统等.

对比学习的核心思想是: 通过比较不同样本之间的相似性和差异性, 来学习更好的特征表示. 在对比学习中, 通常会使用一对或多对样本, 其中每个样本都有一个对应的标签. 模型通过比较不同样本之间的相似性和差异性来学习特征表示, 并使用这些特征来计算样本之间的相似性得分. 在训练过程中, 模型通过最大化同类样本之间的相似性得分和最小化异类样本之间的相似性得分来优化模型.

对比学习的另一个重要组成部分是相似性度量方法, 相似性度量方法可以衡量两个样本之间的相似性得分, 例如欧氏距离、余弦相似度和曼哈顿距离等. 在对比学习中, 相似性度量方法通常与特征表示一起使用,

以计算每个样本之间的相似性得分. 相似性度量方法的选择, 对于对比学习的性能至关重要.

OpenAI 将对比学习与跨模态表示结合起来, 利用对比学习将匹配的图文对投影到相似的向量空间, 将不匹配的图文对投影到距离较远的向量空间中. 其在大规模数据集训练出来的 CLIP[35], 提升了图文检索, 图文问答等多种下游任务上表现.

本文主要利用对比学习得到视频和知识库中实体在领域内更好的多模态特征表示.

3 直播商品数据集构建

为了推进视频上实体链接的前沿研究, 本文构建了 VMEL (video multimodal entity linking)数据集. 本节首先介绍数据集构建的流程, 接着展示并分析数据集的属性与特点.

3.1 数据集构建

数据集构建的主要流程如图 2 所示, 主要分为 3 个步骤: 第 1 步是视频爬取, 第 2 步是视频分割和语音转写, 第 3 步是视频标注. 除此之外, 为了更好地推进视频实体链接任务, 本文还构建了商品对应的知识库.

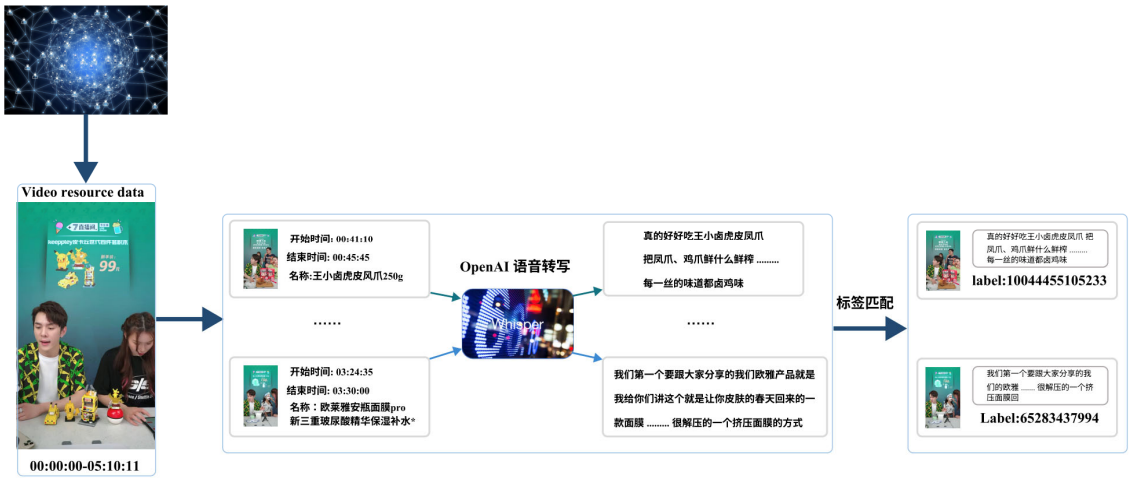

图 2 数据集构建的主要流程

如图 2 所示: 我们首先从淘宝的直播回放页面获取到页面的 m3u8 地址, m3u8 地址常用于直播场景的视频传输; 获取到直播的 m3u8 地址后, 可以解析到视频的“.ts”片段, 将视频的 ts 片段拼接起来, 即可获得直播回放的完整视频. 我们首先获取到 3 个月内 42 场视频的原文件, 每场直播视频时长平均 5.6 h, 每场直播平均包含 51.3 个直播商品. 在爬取到视频的同时, 我们同时获得到每场直播商品的名称列表.

由于商品名称和具体的商品在直播中出现的时间是不确定的, 第 2 步需要人工标注出商品在视频中出现位置以及对应时间间隔, 例如, 某视频 1 h 3 min 11 s–1 h 7 min 43 s 讲的是一件名为“Mido 手表女士花淅简约机械手表”. 对每个视频, 两位标注人员同时标注出商品出现的起点和截止点, 每个视频片段对应的商品名称, 当标注结果有偏差时, 还会选择一位有经验标注人员检验确定最终结果. 过滤掉部分时间过短的视频(小于 1 min)之后, 我们获取到了 1 987 个视频片段, 平均每条视频片段 5.3 min. 与此同时, 视频包含文本模态的信息, 为了更好地利用视频丰富的信息, 我们利用 OpenAI 的开源模型 whisper[36], 将视频的语音模态转成了文字, 同时将语音转写的结果进行简单的清洗, 去除其中语气词、字数过少词等. 清洗后的语音文本,作为每条视频数据纯文本模态的信息.

之后, 我们利用规则标注和人工修正的方法对这些视频进行标注. 对视频进行链接的标注需要构建对应的商品库, 由于淘宝商品变动频繁, 部分商品由于时间推移导致页面失效, 本文商品库的构建主要来源于京东商城. 我们在京东商城上搜索每条视频对应名称, 在搜索页面会获得 30 个返回列表, 列表同时也包含商品

名称, 我们利用两方面分数来评判搜索结果与真实商品之间的相似性: 一方面是真实商品名称结果和搜索商品名称分词结果的碰撞匹配, 另外一方面是商品名称与搜索商品名称的语义相似度分数. 按照两方面分数排序搜索到的商品列表, 从上到下选择商品备选, 如果出现重复的商品, 优先选择官方店铺或者自营店铺. 至此, 初步的商品规则标注已经完成. 由于网络原因以及商品图片爬取 url 随时间丢失的问题, 小部分视频无法获取到真实的链接, 最终得到了 1 838 个带有链接的商品视频.

由于是规则匹配的方法, 每条视频对应的商品不一定是真实的商品, 人工检查时发现, 20%左右的数据规则标注是有误的, 主要体现为型号错误或者品牌错误, 例如, 将视频匹配到相似品牌的同类型产品, 或者匹配到同一品牌下的不同型号的产品. 在人工修正的过程中, 我们发现商品库中不同品牌的竞品以及不同型号的商品图片偏向于一致, 但文本有比较大的差异, 因此, 我们利用视频获取到原商品名称搜索结果, 人工对这些错误标注进行修正, 保证每条视频的真实链接标注是准确无误的. 为了进一步评估人工修正的效果, 我们从数据集中抽取了 100 条商品, 所有视频和对应商品名称都是一一对应.

最后, 为了进一步完善数据集, 我们还需要商品库负样例的构建. 为了尽可能地模拟真实场景, 我们利用正样例京东上商品分类的倒数第 2 层节点(最后一层是商品名称), 与规则标注方法一致, 我们搜索倒数第 2 层节点的名称, 得到与其最相关的前 10 位商品, 除正样例之外的所有商品作为负样例. 由于网络原因和商品图片 url 转移, 最终获取到了 9 410 个负样例, 最终得到的商品库正样例与负样例的比例为 1:5.12.

3.2 数据集属性及分布

表 1 展示了 VMEL 数据集的详细信息, 知识库总共包含 11 000 个实体, 共有 1 838 条视频数据, 分别对应知识库中的 1 838 个实体. 视频的平均长度为 5.3 min, 每条视频在知识库中有唯一的实体与其对应, 每条视频的语音文字长度大约是 1 423.47 字, 这部分信息超过了部分文本编码器的输入, 后续方法为了更好地使用这部分信息, 对这部分信息进行了预处理.

表 1 VMEL 数据集的基本信息

数据集	视频条数	视频平均长度	知识库大小	平均语音长度	正负样例比
VMEL	1 838	5.3 min	11 248	1 423.47 字	1:5.12

图 3 也展示了数据集中视频中提到的实体的类型分布情况, 整体数据集包含有 39 个实体类别, 包含市面上大多数生活用商品实体. 其中, 分布最多的几个实体类别分别是美妆护肤、食品饮料、服饰内衣以及家用电器和个人护理, 最大的美妆护肤类别拥有 349 个实体, 最小的类别只有 1 个对应的实体.

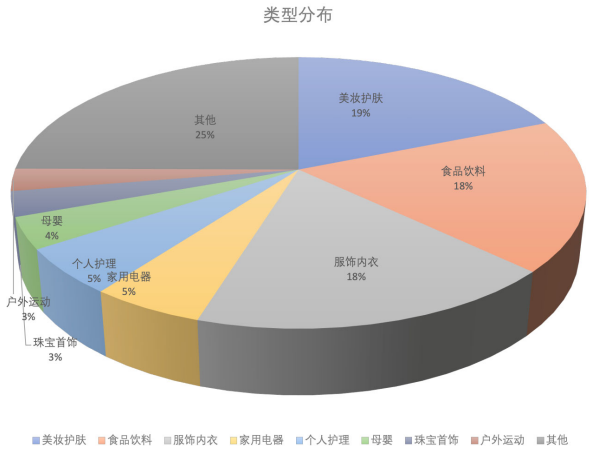

图 3 数据集数据类型分布

4 基于对比学习的实体链接模型

此部分介绍视频上细粒度实体链接的方法，如图 4 所示，主要包括 3 个模块：首先是预处理模块，预处理模块的作用是将视频和文本的信息进行预处理，主要利用大模型对文本信息进行摘要提取；其次是多模态编码模块，这个模块的主要作用是对视频模态和文本信息以及商品库的图文信息进行分别编码和多模态融合成统一编码；最后是基于对比学习的相似度匹配模块，这个模块的主要作用是获取直播领域视频和实体的更好表示.

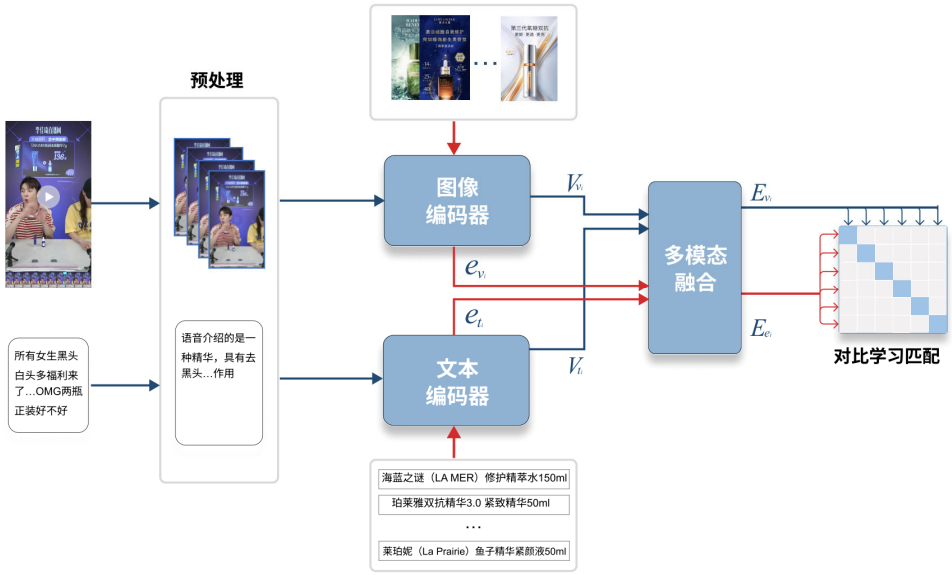

图 4 基于对比学习的视频链接模型

4.1 任务定义

首先给出任务的定义. 给定一个视频 $V_i=\{v_i,t_i,\ldots\}$，视频 V_i 中包含有与之对应的各种模态的信息 $v_i,t_i,\ldots$，其中，v_i 表示图像信息，t_i 表示文本信息. 除此之外，可能还包含例如语音、时序等模态信息. 同时，给定一个多模态知识库 $KB=\{e_1,e_2,\ldots,e_n\}$，包含有多个多模态实体 $e_j=\{v_j,t_j\}$，v_j 与 t_j 分别表示实体的图像和文本信息. 视频上的多模态实体链接的任务，需要将给定视频，链接到知识库中与之对应的实体 e_{v_j} 上. 公式表示如下.

$$V_i = Fusion(v_i,t_i,...),$$
$$e_j = fusion(v_j,t_j),$$
$$e_{v_i} = \arg\max Sim(V_i,e_j).$$

Sim 代表视频和多模态知识库中的相似性分数，$Fusion$ 代表多模态融合的过程，任务需要将多模态视频链接到多模态知识库中的实体.

4.2 预处理模块

预处理模块主要是文本的预处理，语音转成的文本有很多噪声，会对商品的识别造成干扰. 如何从噪声中提取出有效的信息，对最后视频链接的准确率影响非常大. 本文利用大规模语言模型作为预处理模块.

如图 5 所示，对于一段视频的语音文本，本文构建了如下模板，提示 ChatGPT 需要完成任务，模板告诉 ChatGPT 需要总结抽取语音文本的对象，并从中抽取出实体和对应属性.

由于使用的文本是利用语音转成的结果，会带有很多干扰模型做出正确选择的噪声. 例如：首先，过于口语话的表达“但是是什么”，过多的类似表达输入到模型之中会让模型产生理解偏差；其次，语音转写也有一定的错误，有些主体的名称被转写成了其他同音字；此外，主播的一些常用表达也会对主题的向量表示产

生影响, 例如“所有女生”, 在一段文本过多出现, 文本的向量表示可能会出现偏移, 导致不同语音文本之间相似度很高; 最后, 一些少见的品牌转写可能出现错误. 现有方法表明: 实体的上下位信息可以增强实体表示, 商品属性信息也可以达到类似的效果, 利用大模型抽取出商品属性可以帮助修正此类错误. 后续分离实验也证明了这种方法的有效性. 相较于原始语音文本, 总结出的文本少了大量噪声.

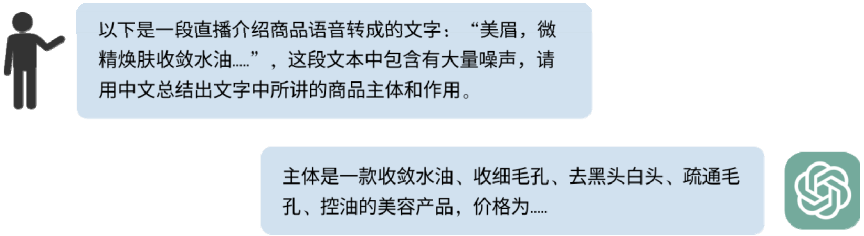

图 5 基于语言模型的摘要去噪

另外一部分是视频的预处理, 由于直播视频场景转换较少, 为了将视频更好地输入模型, 我们对视频每隔固定的时间截取关键帧, 将关键帧序列作为模型的输入.

4.3 基于图文编码的多模态编码模块

获取到文本和关键帧序列之后, 接着对文本和关键帧进行多模态编码. 如图 4 所示: 首先, 分别对文本和图片进行文本和图像编码; 与此同时, 商品库中的图片和商品名称也分别编码进行多模态编码. 文本编码器采用的 Roberta[37], 图像编码器采用的是 Vit[38]. 视频和商品库的文本和图片经过编码器之后, 就获得了不同模态的初始的向量表示, 视频 V_i 经过编码后模态分别记作 V_{v_i}, V_{t_i}, 知识库中的实体 e_j 初始编码分别为 e_{v_i}, e_{t_i}. 接着, 将不同模态的信息送入多模态融合模块.

本文采用线性拼接多模态融合的方法. 如公式所示, 经过多模态融合后, 分别得到了视频的向量表示和知识库中实体的向量表示 V.

$$E_{v_i} = Fusion(V_{v_i}, V_{t_i}),$$
$$E_{e_i} = Fusion(e_{v_i}, e_{t_i}),$$
$$e_{v_i} = \arg\max_{e_j \in KB} Sim(V_i, e_j).$$

4.4 基于对比学习的相似度匹配

获取到视频和知识库中的实体表示之后, 为了让模型学习到视频和知识库之中对应实体的链接信息, 本文利用对比学习来让链接实体之间的相似度更高, 非链接实体之间的相似度分数下降. 对比学习的损失函数如下所示.

$$Loss(E_{v_i}, E_{e_i}) = -\log\left[\frac{\exp(\phi(E_{v_i}, E_{e_i}^{+})/\tau)}{\exp(\phi(E_{v_i}, E_{e_i}^{+})/\tau) + \sum \exp(\phi(E_{v_i}, E_{e_i}^{-})/\tau)}\right]$$

其中, e_j^{+} 表示是视频真实链接到的实体, 而 e_j^{-} 表示非视频真实链接的实体, ϕ表示相似度计算. 本文采用余弦相似度计算视频和实体之间的匹配程度. τ表示对比学习中的温度系数, 其意义是对困难样本的关注程度.

5 实验分析

实验使用的数据集是本文所构建的 VMEL 数据集, 由于实验设备影响, 训练阶段和测试阶段的备选链接实体有一定的差异: 训练阶段是在一个固定的批量大小内, 利用对比学习提升链接视频和实体之间的相似度; 而测试过程则是在固定大小的随机抽样的知识库上, 期望检索到每个视频对应的真实商品.

5.1 实验设置

5.1.1 评估指标

指标采用 *R@K* 以及 *MRR@K* 指标. *R@K* 与 *MRR@K* 是信息检索领域的常见指标, 如公式所示, *R@K* 表示检索到的前 *K* 个备选的召回率, 对于单个查询, 召回 *K* 个候选, 若 *K* 个中有对应匹配项, 则为 1; 若无匹配项, 则为 0. *MRR@K* 是计算检索系统返回的前 *K* 个结果中, 正确答案第 1 次出现的平均位置的倒数, 整体的 *MRR* 值是在单个候选中取平均值得到:

$$R@K=\frac{|relevant:documents:retrieved@rank:\leqslant K|}{|relevant:documents|},$$

$$MRR@K=\frac{1}{|K|}\sum_{i=1}^{|K|}\frac{1}{rank_i}.$$

5.1.2 具体实现

由于中英文编码上的差异, 本文文本编码器采用了 RoBERTa-wwm-Base, 在 RoBERTa 的基础上改进了遮盖策略, 与中文更加契合. 图像编码器采用 ViT-B/16, ViT 是近几年在计算机视觉领域被广泛采用的图片编码器. 本文采用预训练权重在中文上进行微调 Chinese-CLIP[39]的权重, 在 4 张 3 090 上用 64 个批量大小进行训练, 采用 SGD 优化器, 学习率初始设置为 0.000 5, 初始化动量为 0.9, 本文方法记为 CVMEL. 测试阶段, 由于显存大小的限制, 我们随机从知识库中选取 1.8K 个备选作为检索对象, 所有对比实验都是在这随机选取后固定的知识库上进行.

5.1.3 基线模型及其设置

由于现有的方案很少有对视频进行链接, 因此, 本文的对比实验选择了 BLINK[17]、V2VTEL[16]以及 AltCLIP[40]. BLINK 是纯文本实体链接上广泛使用的链接基地模型; V2VTEL 是最近一年发表的多模态实体链接的基础模型; 而 AltCLIP 则是中英文双语多模态检索模型, 在多个中文跨模态检索任务上达到最好效果. 此外, 我们还选择了与视频链接任务相关的视频检索方向的方案, CLIP4Clip[41], 视频检索任务包括视频搜索文本和文本搜索视频, 本文主要关注视频搜索文本.

对比实验中, V2VTEL 的做法是基于图片检索关键帧, 原文效果最好的情况是利用 ResNet[42]找出备选实体, 再利用 CLIP 重新排序. ResNet 更多关注的是像素级别的信息, 因为数据集中实体在视频中出现的位置不一, 本文尝试了 ResNet 初步筛选, CLIP 精确匹配, 但此方法在数据集上表现欠佳. 因此, V2VTEL 只展示用 CLIP 作为实体链接的主体模型, 效果是在数据集上微调的结果.

AltCLIP 是基于 CLIP, 文本端采用双语言文本模型, 利用 CLIP 中的文本编码器增强 XLM-R 双语模型, 在损失部分英文能力的基础上, 在中文多模态检索数据集上达到了最佳. 本文采用 AltCLIP 的预训练权重, 与本文方法一致, 采用对比学习方式获取更好的多模态表示, 结果是在数据集上微调的结果.

视频检索现有的方案大多针对英文场景, Zeng 等人[43]虽然提出了大规模数据集, 但是并没有提供中文预训练的权重, 重新训练一个视频检索的中文模型开销较大, 因此, 本文将知识库翻译成对应的英文, 考虑到中文语音的信息与英文语音有一定的差异, 因此选取目前视频检索方案中不依赖于语音的模型 CLIP4Clip, 每个视频每秒截取了 3 个关键帧, 选取平均池化作为最后视频的向量表示, 结果也是在视频链接数据集上微调的结果.

BLINK 的原方案是以实体提及以及实体提及的上下文作为实体提及的编码, 知识库中的实体和实体解释作为实体的编码, 计算相似度. 本文的做法是: 利用整个语音文本作为实体提及的文本信息, 知识库中的实体的分层信息作为实体的信息补充, 整合实体本身名称作为实体的信息, 两部分信息分别采用两个文本编码器, 实验结果也是在数据集上微调的结果.

5.2 实验结果

实验结果见表 2. 本文提出的视频实体链接模型在数据集上达到了最好的链接效果, 其次是基于纯文本的 BLINK 模型. 我们根据数据集的分析, 这两个方案表现较好是因为文本模态在细粒度实体识别时作用更加

明显. 在视频细粒度实体识别中, 语音转写的文本中会提到实体的品牌、作用等, 虽然出现的顺序不一, 分布也可能不均匀, 但是相较于图像模态, 文本还是发挥更加重要的作用.

表 2 视频链接在数据集上整体效果对比分析

方法	*R*@1	*R*@5	*R*@8	*R*@10	*MRR*@3	*MRR*@5	*MRR*@10
CLIP4Clip	1.06	8.02	13.9	17.1	2.14	3.10	4.27
AltCLIP	8.95	20.62	24.5	26.5	12.4	13.3	13.5
V2VTEL	9.09	24.1	30.5	32.6	13.0	14.2	15.4
BLINK	42.2	72.7	80.2	82.4	53.7	54.8	56.1
CVMEL	**57.7**	**82.3**	**86.1**	**87.7**	**66.0**	**66.8**	**67.7**

CVMEL 与 BLNIK 相比获得了不小的提升, 一方面是因为结合了图像模态的信息, 补充了部分信息缺失, 有时会出现同种产品, 但实际上外形差距很大, 例如不同品牌的乳霜, 会有盒装、瓶装、罐装, 此时, 视频模态可以作为一种信息补充, 将相似实体排序更加靠前; 另外一方面, CVMEL 利用大模型抽取实体和实体的描述信息, 相比 BLINK 可能会因为输入文本长度变化, 输入实体出现的位置不一, 更容易捕获到实体的特征. 消融实验也证明, 纯文本的 CVMEL 的效果也优于 BLINK 方法.

CVMEL 相较于 V2VTEL 有比较大的效果领先, 最重要的原因是 V2VTEL 只包含图像模态, 在取较少待选实体的同时, 会出现很多相似的实体. 例如: 当待推理的实体是“口红”时, 所有条装的口红都有可能作为备选实体被选取出来, 缺少视觉层面的品牌和属性信息, 无法对其进行细粒度的实体消歧.

而 CVMEL 与 AltCLIP 相比, 效果也有较大的领先. 实验过程中, 尽管在数据集上微调使 AltCLIP 效果提升明显, 但由于 AltCLIP 本身是双语言模型, 并且模型的参数较多, 很难学到特定领域(如电商)的知识; 而 CVMEL 由于其模型的参数较少, 只有 AltCLIP 的 1/4, 通过零样本推理结果推测训练数据中包含电商的对齐跨模态数据, 并且文本编码器是纯中文的, 在数据集微调可以很好地提升其表现.

CLIP4Clip 在数据集上的效果很差, 我们推测: 预训练权重是英文数据集上预训练好的参数, 运用到翻译的英文上有一定的差距, 在语料上微调不足以弥补此种差距; 此外, 视频检索任务和实体链接任务有一定的本质区别, 视频检索文本的目标更多是视频的描述信息, 例如“一个小女孩在吃冰糖葫芦”, 而实体链接的目标是将冰糖葫芦识别并链接到知识库中, 两种任务的差异导致结果表现的不一致. 而 CVMEL 提出的方案既考虑了文本, 也考虑了图像, 并利用大模型摘要文本, 不仅学习到了更好的文本表示, 还将图像和文本融合利用对比学习, 区分出高相似视频和实体. 结果表明, 本文提出的视觉方案链接方案适用于细粒度视频链接的任务.

5.3 消融实验分析

在消融实验部分, 首先, 我们分析了不同模态的影响; 其次, 我们分析了不同模块对实验结果的影响.

为了分析本文方法不同模态对视频链接效果的影响, 我们分别只利用文本信息和视觉信息进行链接, 结果见表 3. 由表可见: 纯文本的效果已经非常接近多模态融合之后的效果, 纯视觉的链接的效果并不好. 可能有以下几方面的原因: 首先, 本身直播场景中画面较为单一, 直播中的商品只占其中很小的一部分, 例如口红、修复乳等, 同时又有一些体积大的商品, 例如电视、跑步机等, 占据了页面的很大一部分, 视觉编码很难获取到不同场景下商品的充分信息; 其次, 实体之间的相似性很大, 不同品牌的电视、口红从外形上看没有本质的区别, 很多时候, 人类也不一定分辨清楚, 召回实体的同时就会召回很多相似商品. 多模态融合对纯文本有一定的提升, 说明视频中的语音文本对视频实体链接起主要作用的同时, 图像可以作为一种辅助, 进一步提升实体链接的效果.

表 3 不同模态对实验结果影响分析

方法	*R*@1	*R*@5	*R*@8	*R*@10	*MRR*@3	*MRR*@5	*MRR*@10
纯视觉	7.5	21.4	25.1	27.8	11.3	12.4	13.2
纯文本	55.1	82.3	85.5	86.6	65.0	66.0	66.6
CVMEL	**57.7**	**82.3**	**86.1**	**87.7**	**66.0**	**66.8**	**67.7**

为了分析每个模块对视频链接效果的影响，我们依次删除了大模型摘要部分，删除对比学习模块对实验的影响，实验结果见表 4. 针对这一实验结果，分析如下.

- 对比学习的相似度微调模块删除之后，实验准确率下降非常明显，*R*@1 甚至下降了 33.2，说明对比学习能够帮助模型学习到视频直播领域的链接方式，通用领域的模型表示在某一个领域应用时，向量表示上会有一定程度上的相似，例如直播场景下主播介绍的场景较为固定，视觉表示时可能会受到固定场景的影响，文本段介绍商品的结构大致相同，会有同样的影响. 使用对比学习能够在通用表示基础上，拉大高相似实体表示之间的差距，得到更好的领域内表示.
- 删除预处理中的大模型总结摘要之后，实验结果显示，拥有预处理模块相较于没有预处理获得了很大的提升. 主要原因可能有：一方面，预处理能够处理总结出文本中的关键信息，去除一定的噪音；另外一方面是部分信息在输入模型之后，因为长度过长可能发生了截断，导致部分信息丢失. 基于大模型摘要的去噪方案也可以为包含噪声较多的真实场景落地提供了参考思路.

表 4　不同模块对实验结果影响比较

方法	*R*@1	*R*@5	*R*@8	*R*@10	*MRR*@3	*MRR*@5	*MRR*@10
对比学习	24.5	44.9	52.9	55.6	29.6	31.6	33.1
大模型摘要	40.1	65.2	69.0	70.6	50.1	50.9	51.6
CVMEL	**57.7**	**82.3**	**86.1**	**87.7**	**66.0**	**66.8**	**67.7**

5.4　定性分析

为了更好地分析视频链接任务中模型的性能瓶颈和表现，我们抽取了 3 个测试用例对模型的表现进行了评估. 如表 5 所示，红框表示命中实体. 在 3 个用例中，CVMEL 都捕捉到了最佳的候选实体，BLINK 在前几位能够链接到候选实体，而 V2VTEL 则要求候选中的图片相似度极高才能准确链接实体.

表 5　3 个用例定性分析结果

输入视频		识别结果(前三位)
 我们下一个细节的王饺子 这个还用多说吗 这个我可以吃 不用多说了吧 买饺子就选王饺子 他们家就像这种王炸单品一样 他们家的饺子我跟你们说 有人没吃过吗 你们在嘉宾直播间 你们没有买过王饺子的有吗	V2VTEL	
	BLINK	
	CVMEL	
 下一个是好欢罗 不用我们介绍了吧 我们要快点把这个螺丝粉剥完 要不然整个人会把娘娘臭到 所有女生们 好欢罗螺丝粉来了 不用我们多介绍了 好欢罗我们准备卖爆炸 里面有迷之臭的 所有女生的酸笋 以及酸豆角 还有萝卜干	V2VTEL	
	BLINK	
	CVMEL	
 就是我们的科润的一个乳霜 真的是经典的不能再经典了 跟大家讲过很多次了 是的 因为对于敏敏肌脆弱肌肤 都懂它的好 科润有多好真的不用多说 因为我之前有一个朋友就是从国外回来之后 皮肤特别不稳定 就是别的面霜用了都不管用了	V2VTEL	
	BLINK	
	CVMEL	

从 BLINK 和 CVMEL 的分析来看, 虽然用例 2 中在 top2 时链接到了实体, 但是与此同时也链接到了多种名称相似的实体, 图片特征却不完全一致; 而 CVMEL 不仅选出了名称相似的实体, 还选择出了与链接实体最相似的排在第二位, 一方面可能由于图像信息的补充, 另一方面可能是由于对比学习微调能够得到实体的更好的表示, 让不同品牌的商品区分开来, 让同一品牌的商品又不过于聚集. 而在用例 3 中, 文本信息中“珂润”甚至发生了转写错误成了“科润”, 这种方法对 BLINK 造成了很大的干扰, 链接到了多种乳霜, 却没有链接到准确的品牌型号, 而利用大模型摘要的 CVMEL 则是在转写错误的情况下准确识别出了实体. 我们推测的原因有: 一方面, 大模型摘要可以抽取出各种实体的属性, 不同品牌的产品侧重点可能会略有不同, 利用这部分属性信息能够帮助模型准确识别出待选实体; 另外一方面, 可能是大模型在大规模语料上进行预训练, 见过足够多的数据, 本身具备一定的纠错能力, 在出现错误的语料上, 也能去除干扰, 抽取出较为准确的信息.

而 V2VTEL 只识别出了一个测试用例, 用例一识别成形状包装较为相似的其他产品, 用例 3 甚至识别出与瓶罐较为相似的水壶, 纯视觉模型可能不适合细粒度的视频实体链接任务.

6 总 结

本文提出了细粒度视频上的多模态实体链接任务, 并立足于直播场景, 构建了对应的 VMEL 数据集, 数据集来源于更加接近生活的直播场景. 本文提出的基于大模型的去噪方案对包含噪声的生活场景有极大的参考价值, 并且实验证明, 去噪对整体实验效果有较大提升. 此外, 当预训练运用到特定领域时, 为了获取特定领域的相似度, 本文使用对比学习来获取特定领域数据的更好表示. 在实验中证明, 本文的方法对细粒度视频上的实体链接有较好的效果.

References:

[1] Shen W, Wang J, Han J. Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Trans. on Knowledge and Data Engineering, 2014, 27(2): 443−460.

[2] Li Y, Yang X, Luo J. Semantic video entity linking based on visual content and metadata. In: Proc. of the IEEE Int’l Conf. on Computer Vision. 2015. 4615−4623.

[3] Venkitasubramanian AN, Tuytelaars T, Moens MF. Entity linking across vision and language. Multimedia Tools and Applications, 2017, 76: 22599−22622.

[4] Grams T, Li H, Tong B, *et al*. Semantic video entity linking. In: Proc. of the European Semantic Web Conf. Cham: Springer, 2022. 129−132.

[5] Adjali O, Besançon R, Ferret O, *et al*. Building a multimodal entity linking dataset from Tweets. In: Proc. of the 12th Language Resources and Evaluation Conf. 2020. 4285−4292.

[6] Adjali O, Besançon R, Ferret O, *et al*. Multimodal entity linking for Tweets. In: Proc. of the European Conf. on Information Retrieval. Cham: Springer, 2020. 463−478.

[7] Dost S, Serafini L, Rospocher M, *et al*. VTKEL: A resource for visual-textual-knowledge entity linking. In: Proc. of the 35th Annual ACM Symp. on Applied Computing. 2020. 2021−2028.

[8] Zhang L, Li ZX, Yang Q. Attention-based multimodal entity linking with high-quality images. In: Proc. of the 26th Int’l Conf. on Database Systems for Advanced Applications (DASFAA 2021). Taipei: Springer, 2021. 533−548.

[9] Zhou X, Wang P, Li G, *et al*. Weibo-MEL, Wikidata-MEL and Richpedia-MEL: Multimodal entity linking benchmark datasets. In: Proc. of the 6th China Conf. on Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction (CCKS 2021), Vol. 6. Guangzhou: Springer, 2021. 315−320.

[10] Zheng Q, Wen H, Wang M, *et al*. Faster zero-shot multi-modal entity linking via visual-linguistic representation. Data Intelligence, 2022, 4(3): 493−508.

[11] Wang X, Tian J, Gui M, *et al*. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. arXiv:2204.06347, 2022.

[12] Wang P, Wu J, Chen X. Multimodal entity linking with gated hierarchical fusion and contrastive training. In: Proc. of the 45th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2022. 938−948.

[13] Gan J, Luo J, Wang H, *et al*. Multimodal entity linking: A new dataset and a baseline. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. 2021. 993−1001.

[14] Moon S, Neves L, Carvalho V. Multimodal named entity disambiguation for noisy social media posts. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). 2018. 2000−2008.

[15] Zheng Q, Wen H, Wang M, *et al*. Visual entity linking via multi-modal learning. Data Intelligence, 2022, 4(1): 1−19.

[16] Sun W, Fan Y, Guo J, *et al*. Visual named entity linking: A new dataset and a baseline. arXiv:2211.04872, 2022.

[17] Wu L, Petroni F, Josifoski M, *et al*. Scalable zero-shot entity linking with dense entity retrieval. arXiv:1911.03814, 2019.

[18] Wang Y, Zhan YW, Luo X, Liu M, Xu XS. Survey on video moment retrieval. Ruan Jian Xue Bao/Journal of Software, 2023, 34(2): 985−1006 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6707.htm [doi: 10.13328/j.cnki.jos.006707]

[19] Chen S, Zhao Y, Jin Q, *et al*. Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 10638−10647.

[20] Gabeur V, Sun C, Alahari K, *et al*. Multi-modal transformer for video retrieval. In: Proc. of the 16th European Conf. on Computer Vision (ECCV 2020), Part IV 16. Glasgow: Springer, 2020. 214−229.

[21] Bain M, Nagrani A, Varol G, *et al*. Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. 2021. 1728−1738.

[22] Miech A, Zhukov D, Alayrac JB, *et al*. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. 2019. 2630−2640.

[23] Miech A, Alayrac JB, Smaira L, *et al*. End-to-end learning of visual representations from uncurated instructional videos. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 9879−9889.

[24] Wu W, Sun Z, Ouyang W. Revisiting classifier: Transferring vision-language models for video recognition. Proc. of the AAAI Conf. on Artificial Intelligence, 2023, 37(3): 2847−2855.

[25] Zhao S, Zhu L, Wang X, *et al*. CenterCLIP: Token clustering for efficient text-video retrieval. In: Proc. of the 45th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2022. 970−981.

[26] Bain M, Nagrani A, Varol G, *et al*. A CLIP-Hitchhiker's guide to long video retrieval. arXiv:2205.08508, 2022.

[27] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017. 5998−6008.

[28] Devlin J, Chang MW, Lee K, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2018.

[29] Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training. 2018. https://openai.com/research/language-unsupervised

[30] Radford A, Wu J, Child R, *et al*. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[31] Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020, 33: 1877−1901.

[32] Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 2020, 21(1): 5485−5551.

[33] Du Z, Qian Y, Liu X, *et al*. GLM: General language model pretraining with autoregressive blank infilling. arXiv:2103.10360, 2021.

[34] Zeng A, Liu X, Du Z, *et al*. Glm-130b: An open bilingual pre-trained model. arXiv:2210.02414, 2022.

[35] Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. arXiv:2103.00020, 2021.

[36] Radford A, Kim JW, Xu T, *et al*. Robust speech recognition via large-scale weak supervision. arXiv:2212.04356, 2022.

[37] Liu Y, Ott M, Goyal N, *et al*. Roberta: A robustly optimized Bert pretraining approach. arXiv:1907.11692, 2019.

[38] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.

[39] Yang A, Pan J, Lin J, *et al*. Chinese clip: Contrastive vision-language pretraining in Chinese. arXiv:2211.01335, 2022.

[40] Chen Z, Liu G, Zhang BW, *et al*. AltCLIP: Altering the language encoder in Clip for extended language capabilities. arXiv:2211.06679, 2022.

[41] Luo HS, Ji L, Zhong M, *et al*. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing, 2022, 508: 293–304.

[42] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770–778.

[43] Zeng Z, Luo Y, Liu Z, *et al*. Tencent-MVSE: A large-scale benchmark dataset for multi-modal video similarity evaluation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2022. 3138–3147.

附中文参考文献:

[18] 王妍, 詹雨薇, 罗昕, 刘萌, 许信顺. 视频片段检索研究综述. 软件学报, 2023, 34(2): 985–1006. http://www.jos.org.cn/1000-9825/6707.htm [doi: 10.13328/j.cnki.jos.006707]

赵海全(2000－), 男, 硕士生, CCF 学生会员, 主要研究领域为自然语言处理, 多模态知识图谱, 多模态实体链接.

王续武(1996－), 女, 博士, CCF 学生会员, 主要研究领域为实体识别, 实体链接, 多模态知识获取和应用.

李金亮(1997－), 男, 硕士, 主要研究领域为深度学习, 自然语言处理, 实体链接.

李直旭(1983－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为知识图谱, 知识工程与认知智能, 自然语言处理.

肖仰华(1980－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为大数据管理与挖掘, 图数据库, 知识图谱.