

文章编号: 1003-0077(2022)05-0001-20

## 多模态信息处理前沿综述: 应用、融合和预训练

吴友政, 李浩然, 姚 霆, 何晓冬

(京东人工智能研究院, 北京 100101)

**摘 要:** 随着视觉、听觉、语言等单模态人工智能技术的突破, 让计算机拥有更接近人类理解多模态信息的能力受到研究者的广泛关注。另一方面, 随着图文社交、短视频、视频会议、直播和虚拟数字人等应用的涌现, 对多模态信息处理技术提出了更高要求, 同时也给多模态研究提供了海量的数据和丰富的应用场景。该文首先介绍了近期自然语言处理领域关注度较高的多模态应用, 并从单模态的特征表示、多模态的特征融合阶段、融合模型的网络结构、未对齐模态和模态缺失下的多模态融合等角度综述了主流的多模态融合方法, 同时也综合分析了视觉-语言跨模态预训练模型的最新进展。

**关键词:** 多模态信息处理; 多模态融合; 多模态预训练; 自然语言处理

**中图分类号:** TP391

**文献标识码:** A

## A Survey of Multimodal Information Processing Frontiers: Application, Fusion and Pre-training

WU Youzheng, LI Haoran, YAO Ting, HE Xiaodong

(JD AI Research, Beijing 100101, China)

**Abstract:** Over the past decade, there has been a steady momentum of innovation and breakthroughs that convincingly push the limits of modeling single modality, e.g., vision, speech and language. Going beyond such research progresses made in single modality, the rise of multimodal social network, short video applications, video conferencing, live video streaming and digital human highly demands the development of multimodal intelligence and offers a fertile ground for multimodal analysis. This paper reviews recent multimodal applications that have attracted intensive attention in the field of natural language processing, and summarizes the mainstream multimodal fusion approaches from the perspectives of single modal representation, multimodal fusion stage, fusion network, fusion of unaligned modalities, and fusion of missing modalities. In addition, this paper elaborate the latest progresses of the vision-language pre-training.

**Keywords:** multimodal information processing; multimodal fusion; multimodal pre-training; natural language processing

### 0 引言

人工智能研究经过 70 多年的探索, 在视觉、语音与声学、语言理解与生成等单模态<sup>①</sup>人工智能领域已取得了巨大的突破。特别是视觉领域的目标检测与人脸识别技术、语音领域的语音识别与语音合成技术、自然语言处理领域的机器翻译与人机对话

技术在限定场景下已经实现了规模化的应用。然而, 人类对周围环境的感知、对信息的获取和对知识的学习与表达都是多模态(Multimodal)的。近些年, 如何让计算机拥有更接近人类的理解和处理多模态信息的能力, 进而实现高鲁棒性的推理决策成为热点问题, 受到人工智能研究者的广泛关注。另一方面, 随着图文社交(Facebook、Twitter、微信、微博等)、短视频(YouTube、抖音、快手)、音频(Club-

收稿日期: 2021-08-16 定稿日期: 2021-09-28

基金项目: 科技创新 2030-“新一代人工智能”重大项目(2020AAA0108600)

① 模态是指信息的来源或者信息表示形式。文本、图像、视频、声音和种类繁多的传感器信号都可以称为一种模态。

house 等)、视频会议(Zoom、腾讯会议等)、直播(抖音、京东、淘宝等)和数字人(2D、3D、卡通、写实、超写实等)等应用的涌现,对多模态信息处理技术在用户理解、内容理解和场景理解上提出了更高的要求,同时也给多模态技术提供了海量的数据和丰富的应用场景。

多模态信息处理技术打破计算机视觉、语音与声学、自然语言处理等学科间的壁垒,是典型的多学科交叉技术。多模态技术从 20 世纪 70 年代开始发展,Morency 等人<sup>[1]</sup>将多模态技术的发展划分为四个阶段,即 1970—1980 年的行为时代(Behavioral Era)、1980—2000 年的计算时代(Computational Era)、2000—2010 年的交互时代(Interaction Era)和 2010 年起的深度学习时代(Deep Learning Era)。多模态核心技术又分为:多模态表示(Representation)、多模态融合(Fusion)、多模态转换(Translation)、多模态对齐(Alignment)和模态协同学习(Co-learning)类。

近些年,研究者从不同的视角对多模态信息处理技术做了很好的总结回顾。Zhang 等人<sup>[2]</sup>围绕图像描述、视觉-语言生成、视觉问答和视觉推理四个应用,从计算机视觉的角度总结了多模态表示学习和多模态融合的最新进展。Summaira 等人<sup>[3]</sup>的综述覆盖了更多的多模态应用,并根据应用组织了每一个多模态应用的技术进展和局限性。

本文从自然语言处理的视角出发,介绍多模态信息处理技术的最新进展,组织结构如下:第 1 节介绍 NLP 领域关注度较高的多模态应用和相关的数据集。多模态融合是多模态信息处理的核心问题。第 2 节从单模态信息的表示方法、多模态信息的融合阶段、融合模型的网络结构、未对齐模态和模态缺失情况下的多模态融合等角度介绍主流的多模态融合方法。第 3 节介绍多模态预训练技术,并从模型的网络结构、模型的输入、预训练目标、预训练语料和下游任务等维度对比最新提出的多模态预训练模型。第 4 节介绍多模态技术在工业界的应用。最后一节是总结和对未来工作的展望。

## 1 多模态应用

我们分析了最近两年在自然语言处理领域国际学术会议上(ACL、EMNLP、NAACL)发表的多模态信息处理的论文,并从应用的角度对论文进行了分类。关注度较高的多模态应用如图 1 所示。本节

将对这些应用展开介绍。除此之外,多模态应用还包括视听语音识别(Audio-Visual Speech Recognition)、多模态语言分析(Multimodal Language Analysis)和视觉辅助的句法分析<sup>[4]</sup>等。文献[4]还获得 NAACL 2021 的最佳长文奖。

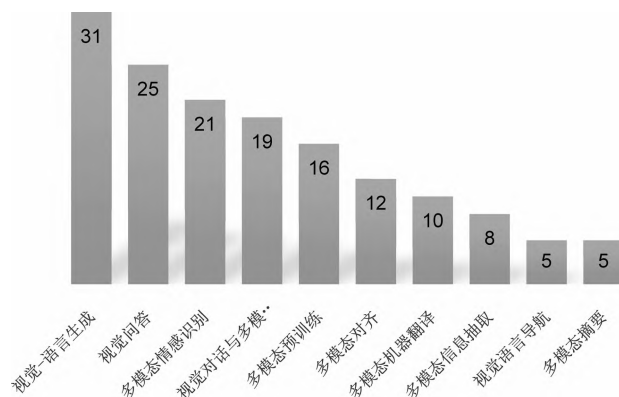


图 1 多模态信息处理论文的分类统计

### 1.1 多模态情感识别

情感是人类区别于机器的一个重要维度,而人的情感往往又是通过语音、语言、手势、动作表情等多个模态表达的。在交互场景下,多模态情感识别研究如何从人的表情和动作手势、语音音调、语言等多模态信息中理解用户细颗粒度的情感表达,进而指导人机交互策略。其主要研究内容有:①基于多模态信息互补性和异步性的动态融合;②高噪声环境下对于模态模糊或模态缺失问题的鲁棒性融合;③客服和营销等自然交互情境下的情感识别等。

多模态情感识别的常用数据集有 IEMOCAP<sup>[5]</sup>、CMU-MOSI<sup>[6]</sup>、CMU-MOSEI<sup>[7]</sup>、CH-SIMS<sup>[8]</sup> 和 IVD<sup>[9]</sup> 等。数据集的多维度比较如表 1 所示。IEMOCAP 数据集收录了 10 位演员的表演数据,包含视频、语音、面部运动捕捉和文本模态,并标注了高兴、悲伤、恐惧和惊讶等共 9 类情感。CMU-MOSI 数据集收录了 89 位讲述者的 2 199 条视频片段,每段视频标注了 7 类情感。CMU-MOSEI 数据集是 CMU-MOSI 的扩展版,收录了 1 000 多名 YouTube 主播的 3 228 条视频,包括 23 453 个句子,每个句子标注了 7 分类的情感浓度(高度负面、负面、弱负面、中性、弱正面、正面、高度正面)和 6 分类的情绪(高兴、悲伤、生气、恐惧、厌恶、惊讶)。CH-SIMS 数据集是一个中文多模态情感分析数据集,该数据集为 2 281 个视频片段标注了细颗粒度

的情感标签。IVD 是从中文语音助手的真实用户对话日志中抽取的语音情感数据集,包括 500 000 条无标注的语音数据和 2 946 条带 6 分类情感标注的语音数据。

随着图文和短视频等新兴社交媒体的迅速发展,人们在社交平台上的表达方式也变得更加丰富。社交场景下的多模态情感识别主要研究基于图文表达的情感倾向<sup>[10]</sup>和方面级的细颗粒度情感<sup>[11]</sup>等。

表 1 常用多模态情感识别数据集对比

| 数据集                      | 语言 | 数据来源                 | 视频片段数  | 说话人数  | 模态类别            | 情感类别数 |
|--------------------------|----|----------------------|--------|-------|-----------------|-------|
| IEMOCAP <sup>[5]</sup>   | 英语 | 实验室录制                | 10 000 | 10    | 视觉、语音、文本和面部运动捕捉 | 9     |
| CMU-MOSI <sup>[6]</sup>  | 英语 | Youtube              | 2 199  | 89    | 视觉、语音和文本        | 7     |
| CMU-MOSEI <sup>[7]</sup> | 英语 | Youtube 上评论、辩论、咨询等视频 | 3 228  | 1 000 | 视觉、语音和文本        | 7     |
| CH-SIMS <sup>[8]</sup>   | 中文 | 电影、电视剧和综艺节目          | 2 281  | 474   | 视觉、语音和文本        | 5     |
| IVD <sup>[9]</sup>       | 中文 | 语音助手的真实用户对话日志        | 2 946  | —     | 语音和文本           | 6     |

## 1.2 视觉-语言生成

视觉(图像或视频)到语言的生成和语言到视觉(图像或视频)的生成打破了计算机视觉和自然语言处理两个领域的边界,成为多模态交叉学科中最热门的研究课题。2021 年初,OpenAI 推出的基于 GPT-3 的语言到视觉的生成模型 DALL-E<sup>①</sup> 可以根据自然语言的描述生成逼真的图像,产生了较大的反响。本节主要介绍视觉到语言生成的相关应用。

### 1.2.1 图像描述

图像描述(Image Captioning)是对给定的一幅自然图像生成一句自然语言描述的任务。2015 年以前,图像描述的主流方法是基于模板的方法。其基本思想是检测图像中的物体、动作,并将这些词作为主语、动词和宾语等填写到预定义的模板中。从 2015 年开始,基于视觉编码器(CNN 等)和语言解码器(RNN/LSTM 等)的序列到序列(Sequence-to-Sequence, Seq2Seq)框架广泛应用于这一任务。通过从视觉图像中解析出属性(Attribute)、关系(Relation)和结构(Hierarchy)等高层语义信息,并将这些语义信息融入视觉编码和语言解码中,提高了图像描述的生成效果。

图像描述任务的常用数据集有 MSCOCO<sup>[12]</sup>、Conceptual Captions<sup>[13]</sup>、Flickr30K<sup>[14]</sup>、Visual Genome<sup>[15]</sup>和 SBU Captions<sup>[16]</sup>。MSCOCO 数据集是微软发布的可用于目标检测(Object Detection)、人体姿势识别(DensePose)、关键点检测(Keypoint Detection)、实例分割(Stuff Segmentation)、全景分割(Panoptic Segmentation)、图片标注(Category Labelling)和图像描述(Image Captioning)的数据

集。该数据集有 91 类物体(人、猫和卡车等),共计 32.8 万幅图像,每幅图像包含 5 个英文描述。Conceptual Captions 数据集收录了 330 万幅“图像,描述”对,是目前最大的多模态数据集,其中的图像有自然图像、产品图像、专业照片、卡通和绘图等类型,描述取自 HTML 中的 Alt-text 属性字段值。Flickr30K 收录了来自 Flickr 的共计 31 783 幅日常活动、事件和场景的图像,每幅图像通过众包方式标注了 5 个图像描述。Visual Genome 是基于 10.8 万幅图像的大规模多模态数据集,该数据集标注了 380 万个对象、280 万个属性、230 万个关系、170 万个“图像、问题、答案”三元组和 540 万个区域描述。图像中的对象、属性、关系、区域描述和视觉问答中的名词与短语还被归一化到相应的 WordNet 同义词集。

### 1.2.2 视频描述

视频描述(Video Captioning)是对给定的一段视频(通常是几十秒的短视频)生成一句准确、细致描述的任务。视频除了图像信息外,还包括时序和声音等信息。视频描述可提取的特征更多,技术挑战也更大。

视频描述任务的常用数据集有 MSR-VTT<sup>[17]</sup>、ActivityNet-Captions<sup>[18]</sup>、YouCook2<sup>[19]</sup>和 ACTIONS<sup>[20]</sup>等。MSR-VTT 数据集由 1 万个网络视频剪辑、20 万“视频,描述”对组成。MSR-VTT 数据集涵盖了音乐、游戏、体育、教育等 20 多个类别的视觉内容,每个视频剪辑时长 10~20 秒,人工为每个视频剪辑标注了 20 个描述句子。YouCook2 数据集是一个烹饪教学视频数据集,包括 89 个食谱的 2 000 个未经剪辑的教学视频

① <https://openai.com/blog/dall-e/>

(最长 10 分钟,平均 5 分钟)。ACTIONS 是首个无需人工标注、从数以亿计的网页内容中自动提炼“视频,描述”对的视频描述数据集,总共包含了 163 183 个 GIF 视频。

### 1.2.3 视觉叙事

视觉叙事(Visual Storytelling)要求模型对于给定的图像序列,在深度理解图像序列的基础上生成连贯的叙事故事。相比于图像描述和视频描述,视觉叙事更具挑战性。在视觉理解上,视觉叙事的输入是有时序关联的图像序列,需要模型具备根据历史视觉事件推测当前的视觉事件的能力。在语言生成上,对比图像描述和视频描述中的客观文字描述,视觉叙事的输出由更多评价性、会话性和抽象性语言组成。SIND<sup>[21]</sup>是一个视觉叙事数据集,该数据集收集了 81 743 幅图片,以及排列成符合文字描述和故事情节的 20 211 个序列。

## 1.3 视觉问答和多模态对话

### 1.3.1 视觉问答


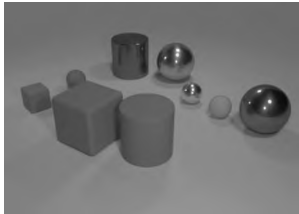

视觉问答(Visual Question Answering, VQA)<sup>[22-27]</sup>是 2015 年新提出的任务,简单来说就是图像问答。给定一幅图像和一个关于该图像的开放式自然语言

问题,要求模型准确回答该问题。视觉问答是一个典型的多模态问题,需要模型具备物体定位、属性检测、事件分类、场景理解和推理及数学计算等能力。根据图片类型的不同,VQA 又分为自然图像理解 VQA<sup>[22-23]</sup>、合成图像推理 VQA<sup>[24]</sup>和自然图像推理 VQA<sup>[25]</sup>。表 2 列举了这 3 种 VQA 的示例。

VQA 常用数据集有 VQAv1/v2<sup>[22-23]</sup>、CLEVR<sup>[24]</sup>和 GQA<sup>[25]</sup>。VQAv1/v2 是自然图像理解 VQA 数据集,VQAv2 解决了 VQAv1 中明显的语言先验(Language Priors)问题。CLEVR<sup>[24]</sup>是合成图像推理问答数据集。CLEVR 中的图像由简单的几何形状的物体组成,旨在测试模型对组合式语言的理解能力和对视觉场景的推理能力。CLEVR 数据集中的图像是程序合成的,其场景的复杂度与自然场景相去甚远。对此,Hudson 等人<sup>[25]</sup>发布了基于自然图像的组合格式问题视觉问答数据集 GQA,该数据集包括关于 11.3 万幅图像的超过 2 000 万的问题。每幅图像都标注了一个场景图(Scene Graph),表示图像中的对象、属性和关系。每个问题都对应一个功能性程序(Functional Program),列出了获得答案所需执行的一系列推理步骤。每个答案都有与之对应的验证信息,指向图片中的相关区域。

表 2 三类视觉问答的示例

(示例引自文献[24-25])

| 自然图像理解 VQA  | 合成图像推理 VQA   | 自然图像推理 VQA  |
|---|--|---|
|  |                   |  |
| Q: What color are her shoes?<br>A: white  | Q: What shape is the small rubber object that is the same color as the large rubber cube?<br>A: cube | Q: Is there any fruit to the left of the tray the cup is on top of?<br>A: yes         |

### 1.3.2 视觉对话

视觉对话(Visual Dialog)<sup>[28-32]</sup>是给定一幅图像(或视频等视觉内容)和一个上下文相关的问题,要求模型根据图片(或视频)内容回答该问题。与视觉问答相比,视觉对话还要解决对话中特有的挑战,如共指(Co-references)和省略(Ellipsis)等。视觉对话也被认为是视觉图灵测试。视觉对话常用数据集有 VisDial<sup>[28]</sup>、IGC<sup>[29]</sup>、GuessWhat<sup>[30]</sup>、Image-Chat<sup>[31]</sup>和 AVSD<sup>[32]</sup>。VisDial 中的问题和答案都是形式自

由的。GuessWhat 是通过一系列“是/否”问题发现图像中的物体。IGC 是一个闲聊型的视觉对话数据集,但闲聊的话题受限于给定的图像。Image-Chat 也是一个闲聊型视觉对话数据集。与 IGC 不同的是,Image-Chat 数据集还限定了对话参与者 A 和 B 的风格特征。AVSD 定义了一个视听场景的多轮对话任务,要求机器在理解问题、对话历史和视频中的场景等语义信息的基础上回答用户问题。

视觉对话中的用户问题只与单个图像(视频)相

关,且用户问题和模型回答都是文字的。

1.3.3 多模态对话

多模态对话(Multimodal Dialog)关注更接近人类自然对话的多模态人机对话技术的研究。它与上一节介绍的视觉对话的主要差异有:①多模态对话给定的输入图像可能是多幅的;②随着对话的推进,图像是不断更新的;③用户问题和模型的回答可以是文本的、图像的或者图文结合的;④模型可能需要查询外部领域知识库才能回答用户的问题(如购物者希望看到更多与特定商品相似的商品,或者要求提供满足某些特征的商品,或者查询特定商品的属性等);⑤模型可能需要通过反问等对话策略澄清用户需求。零售和旅游等限定领域的多模态对话最近受到了越来越多的关注。

常用的面向购物场景的多模态对话数据集有MMD<sup>[33]</sup>、SIMMC<sup>[34]</sup>和JDDC<sup>[35]</sup>。MMD是在服饰专家的指导下通过模拟扮演(Wizard-of-Oz, WoZ)的方式收集的时尚购物场景的数据集。SIMMC 2.0是时尚和家具购物场景的数据集。其中,时尚和家具杂乱的购物场景是通过逼真的VR场景生成器(VR Scene Generator)生成的。与MMD和SIMMC不同,JDDC 2.0是从电商平台客服和消费者之间的真实对话数据中采样的(图2)。JDDC 2.0包括多模态对话24.6万,其中,图片50.7万张,平均对话轮数14轮。此外,JDDC 2.0还提供了30 205个商品

的759种商品属性关系,共计21.9万的<商品ID、属性、属性值>三元组。

视觉对话和多模态对话常用数据集的详细对比如表3所示。

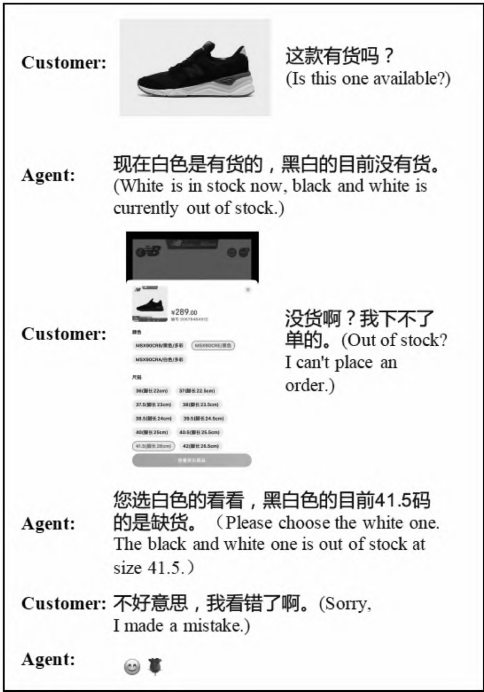


图2 JDDC 2.0中的多模态对话示例  
图片引自文献[35]

表3 视觉对话和多模态对话常用数据集的对比表

| 数据集                        | 对话场景         | 构建方式 | 模态    | 对话数量  | 平均对话轮数 | 特点                    |
|----------------------------|--------------|------|-------|-------|--------|-----------------------|
| VisDial <sup>[28]</sup>    | 日常场景         | 众包   | 图像与文本 | 123K  | 10     | 关于图像内容的多轮问答           |
| IGC <sup>[29]</sup>        | 事件场景         | 众包   | 图像与文本 | 4K    | 6      | 以图像为基础的多轮闲聊           |
| GuessWhat <sup>[30]</sup>  | 开放域          | 众包   | 图像与文本 | 155K  | 5.2    | 通过多轮“是/否”问答发现图像中的物体   |
| Image-Chat <sup>[31]</sup> | 开放域          | 众包   | 图像与文本 | 202K  | 1.5    | 以图像为基础的限定风格的多轮闲聊      |
| AVSD <sup>[32]</sup>       | 日常室内活动场景     | 众包   | 视频片段  | 11K   | 10     | 关于视频内容的多轮问答           |
| MMD <sup>[33]</sup>        | 服饰购物场景       | 众包   | 图像与文本 | 150K  | 40     | 商品选购多轮对话              |
| SIMMC <sup>[34]</sup>      | 服饰家具购物场景     | 众包   | 图像与文本 | 6.6K  | 10.78  | 商品选购多轮对话              |
| JDDC 2.0 <sup>[35]</sup>   | 服饰小家电客服和购物场景 | 真实场景 | 图像与文本 | 24.6K | 14.06  | 选自真实场景客服和消费者之间的真实对话日志 |

1.4 多模态摘要

多模态摘要是基于对多模态输入(文本、语音、图像和视频等)的理解,归纳并生成单模态或者多模态的概括性总结(摘要)的任务。根据具体任务类型,多

模态摘要又可细分为视频会议摘要<sup>[36]</sup>、教学视频摘要<sup>[37]</sup>、多模态新闻摘要<sup>[38-42]</sup>和多模态商品摘要<sup>[43]</sup>。

视频会议摘要方面,Li等人<sup>[36]</sup>提出了一个从音视频会议输入中提取会议文本摘要的方法,并在AMI数据集上验证了方法的有效性。AMI数据

集<sup>[44]</sup>包含 137 场视频会议。每场会议持续 30 分钟,包含 4 名参与者和约 300 字的文本摘要。

教学视频摘要方面,Palaskar 等人<sup>[37]</sup>提出一种融合视觉信息和文本信息(用户生成的和语音识别系统输出的)的生成式文本摘要方法,同时在开放域教学视频数据集 How2<sup>[45]</sup>上验证了方法的有效性。

多模态新闻摘要方面,Li 等人<sup>[38]</sup>提出一种从异步的多模态(文本、图像、音频和视频)输入中抽取文本摘要的方法,并发布了中文和英文数据集 MMS。Li 等人<sup>[39]</sup>提出一种为“文本,图像”对生成多模态摘要的模型,同时发布了英文数据集 MMSS。Zhu 等人<sup>[41]</sup>提出了一种从异步的多模态(文本和多张图片)输入中生成多模态(一段短文和一张图片)摘要的方法,同时发布了英文数据集 MSMO。

多模态商品摘要方面,Li 等人<sup>[43]</sup>提出了一种从异构的多模态输入(文本、图像、商品属性表)中生成商品摘要的方法,同时发布了数据集 CEPSUM<sup>①</sup>。CEPSUM 数据集由 140 万“商品文本介绍,商品图片,文本摘要”三元组组成,涉及 3 个商品大类。

### 1.5 多模态对齐

多模态对齐研究多个模态不同颗粒度元素间的对齐关系,具体又分为显式对齐和隐式对齐。视觉-语言跨模态的显式对齐任务研究图像和句子<sup>[46-47]</sup>、图像和词<sup>[48]</sup>、图像中的目标和句子中的短语<sup>[49-50]</sup>间的对齐关系。多模态对齐方法可直接应用于多模态检索等应用,也可作为图像描述、VQA、多模态预训练的训练语料,尤其是在缺乏大规模多模态人工标注语料的场景。

图像和句子(或文档内其他文本单元)间的显式对齐通常是不存在的。对此,Hessel 等人<sup>[46]</sup>提出了一种将同一网页内的图像和句子对齐的无监督方法。该方法在 7 个难度不同的数据集上获得了不错的性能。Suhr 等人<sup>[47]</sup>定义了一个视觉推理任务 NLVR2,对于给定的两幅图像和一段自然语言的描述,要求模型判断它们是否存在语义上的对齐关系。

文本预训练语言模型已经取得了巨大的成功,但该方法仅使用文本上下文信息作为监督信号,导致词的上下文表示学习严重依赖词的共现关系(Co-occurrence),缺乏外部物理世界的背景知识。为了给预训练语言模型提供视觉监督信号,Tan 等人<sup>[48]</sup>提出了 Vokenization 技术(图 3),其通过给文本中的每一个词打上一幅图像的标签,实现在大规模文本语料上自动构建多模态对齐语料库。在大规

模图像-词汇对齐的多模态语料库上训练的预训练语言模型可增强其对自然语言的理解能力。实验证明,该模型在多个纯文本的任务上(如 GLUE、SQuAD 和 SWAG 等)均获得了显著的性能提高。

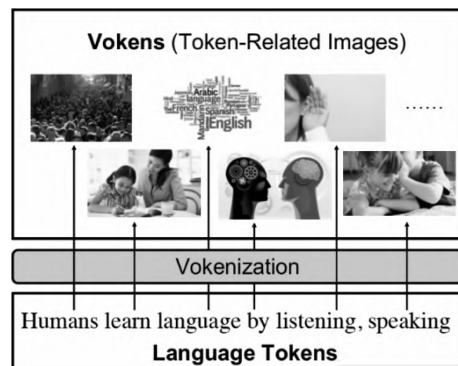


图 3 Vokenization 技术示例

图片引自文献<sup>[48]</sup>

图像中的目标和文本中的短语对齐也被称为图像短语定位(Phrase Grounding),可用于提高图像描述、VQA、视觉导航等视觉-语言下游任务的性能。Plummer 等人<sup>[49]</sup>发布了一个大规模的短语定位数据集 Flickr30k Entities,如图 4 所示。Wang 等人<sup>[50]</sup>提出了一种基于细粒度视觉和文本表示的多模态对齐框架,在 Flickr30k Entities 数据集上显著提高了短语定位的性能。

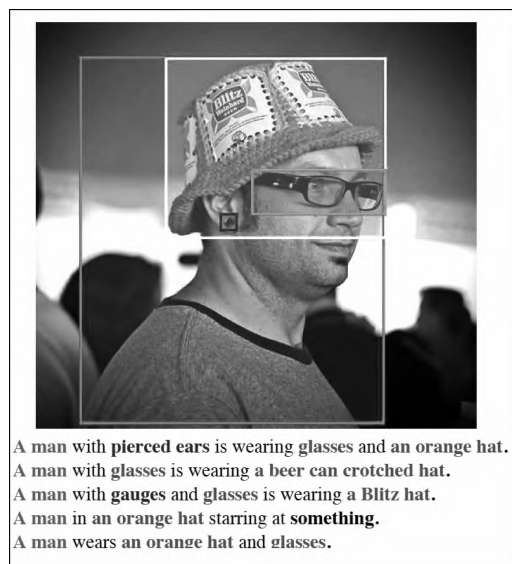


图 4 Flickr30k Entities 标注示例

对齐的图像中的目标和文本中的短语用相同的颜色标记。

图片引自文献<sup>[49]</sup>

① <http://jd-nlg-rhino.github.io/>



视频定位(Video Grounding)<sup>[51]</sup>是多模态对齐中另一项重要且具有挑战性的任务。给定一个查询(Query),它要求模型从视频中定位出与查询语言对应的一个目标视频片段。该技术可应用于视频理解、视频检索和人机交互等场景。常用数据集有 Charades-STA<sup>[52]</sup>、ActivityNet-Captions<sup>[53]</sup>和 TACoS<sup>[54]</sup>。Charades-STA 数据集是基于 Charades 数据集<sup>[55]</sup>构建的,包括 6 672 个视频和 16 128 个“查询,视频片段”对。ActivityNet-Captions 数据集包含两万个视频和 10 万个“查询,视频片段”对,其覆盖的视频类型更多样。TACoS 数据集包含 127 个烹饪视频和 18 818 个“查询,视频片段”。

## 1.6 多模态翻译

多模态翻译是将多模态输入(文本、图像或视频等)中的源语言文本转换为目标语言文本的过程。多模态翻译的目标是在视觉等多模态信息的辅助下,消除语言的歧义,提高传统文本机器翻译系统的性能。

Elliott 等人<sup>[56]</sup>于 2015 年首次提出多模态翻译任务。随后,在 2016 年举办的第一届机器翻译会议上成功组织了第一届多模态机器翻译比赛,并于接下来的两年连续举办了两届比赛,引发了研究者们对多模态机器翻译的关注热潮。目前的工作主要集中在 Multi30k 数据集<sup>[57]</sup>上。该数据集是英语图像描述数据集 Flickr30k<sup>[14]</sup>的多语言扩展,每幅图像配有一个英语描述和一个德语描述,任务定义为给定图像和英语描述,生成德语描述。

模型方面,Huang 等人<sup>[58]</sup>首先从图像中提取视觉全局表示(参见 2.1.1 节的介绍)和视觉目标表示(参见 2.1.3 节的介绍),提取的视觉表示被视为源语言中特殊的单词与文本拼接,再融入编码器-解码器神经网络翻译模型中的编码器中。在 Calixto 等人<sup>[59]</sup>提出的模型中,视觉特征被视为源语言中特殊的单词,或者融入编码器中,或者融入解码器中。Calixto 等人的模型显著提高了模型的翻译效果。文献[58-59]中的模型依赖大量的多模态翻译对齐语料(源语言、图像、目标语言)。对此,Elliott 等人<sup>[60]</sup>将多模态机器翻译分解为两个子任务:文本翻译和基于视觉的文本表示(Visually Grounded Representations)。该模型不依赖昂贵的(源语言、图像、目标语言)对齐语料。模型可以分别在文本翻译语料(源语言,目标语言)和图像描述(图像,源语言)语料上训练。受文献[60]的启发,Zhou 等人<sup>[61]</sup>

提出了一种机器翻译任务和视觉-文本共享空间(Vision-Text Shared Space)表示学习任务相结合的多任务多模态机器翻译框架(VAG-NMT)。VAG-NMT 首先把文献[60]中的基于视觉的文本表示(即从文本表示重建图像)修改为视觉-文本共享空间表示学习。其次,VAG-NMT 还提出了一种视觉文本注意机制,可以捕获与图像语义强相关的源语言中单词。多模态机器翻译中的视觉信息只在非常特殊的情况下(如文本上下文不足以消除歧义词的歧义)对翻译模型有帮助。对此,Ive 等人<sup>[62]</sup>提出了一种翻译-优化(Translate-and-refine)的两段式翻译方法。该方法先翻译源语言中的文本,再使用视觉目标表示对第一阶段的翻译文本进行调整。大多数的多模态机器翻译模型没有考虑不同模态的相对重要性,但同等对待文本和视觉信息可能会引入一些不必要的噪声。Yao 等人<sup>[63]</sup>基于 Transformer,提出了一种多模态自注意机制,探索了如何消除视觉特征中的噪音信号。一方面,单层多模态注意力模型难以有效提取视觉上下文信息,另一方面,多层多模态注意力模型容易导致过拟合,尤其是对训练数据少的多模态翻译。对此,Lin 等人<sup>[64]</sup>提出一种基于动态上下文指导的胶囊网络(Dynamic Context-guided Capsule Network,DCCN)提取和利用两种不同颗粒度(视觉全局表示和视觉区域表示)的视觉信息。也有研究者对多模态翻译的可解释性进行了探索。Wu 等人<sup>[65]</sup>的研究表明,视觉特征对多模态翻译的帮助来自于正则化,视觉特征的合理选取对模型性能至关重要。

## 1.7 多模态信息抽取

命名实体识别(NER)是指识别自由文本中的具体特定意义的实体(如人名、地名和组织机构名等)。命名实体识别虽然取得了较大的成功,但对于社交媒体中大量的用户生成内容(User-Generated Content,UGC),仅根据文本模态的信息来定位和分类其中的实体仍然存在一些挑战。多模态命名实体识别(MNER)通过引入视觉、语音等其他模态作为文本模态的补充,识别社交媒体中高噪声短文本中的实体,最近几年受到了比较多的关注。

模型方面,Moon 等人<sup>[66]</sup>首次提出了融合图像和文本模态信息的通用多模态注意力模型。文献[66]还发布了 SnapCaptions 数据集,该数据集由 1 万张“图像,短文本标题”对构成,并标注了短文本标题中的四类命名实体(实体类型:PER、LOC、

ORG、MISC)。一方面,文献[66]中的方法提取的是图像的视觉全局表示,这可能把图像中的噪声信息也引入到模型中。另一方面,视觉和文本模态的特征融合较简单。对此,Zhang 等人<sup>[67]</sup>提出了一种自适应的协同注意力网络(Adaptive Co-attention Network, ACN)。ACN 首先提取图像的视觉区域表示(参见 2.1.2 节的介绍),再通过文本到视觉和视觉到文本的协同注意力剔除图像中的噪声信息,以提高 MNER 的性能。文献[67]在内部数据集上验证了该方法的有效性。基于类似的出发点,Lu 等人<sup>[68]</sup>提出了一种注意力机制与门控机制相结合的模型提取视觉图像中与文本最相关的区域的特征。该模型可忽略不相关的视觉信息。文献[68]基于注意力机制获取了单词感知(word-aware)的视觉表示,却忽略了图像感知(image-aware)的单词表示。对此,Yu 等人<sup>[69]</sup>首次将 Transformer 应用于多模态 NER 任务中,并提出了实体片段检测辅助任务,进一步消除视觉偏差,提升了模型效果。

Sui 等人<sup>[70]</sup>提出了融合语音和文本信息的多模态 NER,并在自建的中文数据集 CNERTA 上验证了方法的有效性。

多模态信息抽取领域中另一个受到较多关注的研究方向是多模态商品属性抽取。多模态商品属性抽取是指从给定商品文本描述和商品图片中抽取商品的属性信息,例如商品的“颜色”“材料”等属性值。为了推动多模态商品属性抽取的研究,IV 等人<sup>[71]</sup>发布了首个大规模多模态属性提取英文数据集 MAE。MAE 包含 400 万图片和 760 万“属性-属性值”对。文献[71]提出的多模态属性抽取模型需要对每一个属性识别其对应的属性值,且无法滤除视觉噪声。为了提高模型的效率,Zhu 等人<sup>[72]</sup>将属性预测和属性值抽取建模为一个层叠化的多任务学习过程,实现了多个属性及其对应属性值的一次性识别,且视觉全局表示和视觉区域表示通过门控机制和文本信息融合,可有效过滤视觉噪声。Zhu 等人还发布了一个包含 9 万“属性-属性值”对的多模态商品属性抽取中文数据集 MEPAVE。

## 2 多模态融合

多模态融合将多个单模态表征整合成为一个多模态信息表征,它是多模态信息处理的核心问题。多模态融合的示例如图 5 所示,其中, $N_i \{i=1, \dots, K\}$  表示单模态表示学习模型的模型深度, $M$  表示  $K$  个

多模态表示的融合模型深度。多模态融合的研究方向有:基于多模态互补性的全模态融合问题、模态模糊或者模态缺失下的鲁棒性融合问题、非对齐的多模态融合问题等。目前,大部分工作是关于模态对齐且无模态缺失情况下的多模态融合算法研究,这也是多模态融合中最基础的挑战。本节根据单模态的特征表示、多模态融合的阶段、多模态融合模型结构等对多模态融合方法进行分类介绍。

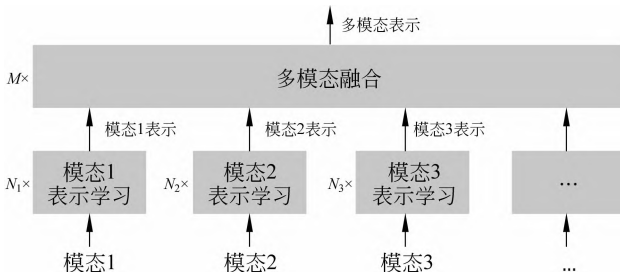


图 5 多模态融合示例

### 2.1 根据单模态表示进行分类

单模态的特征表示是多模态融合的基石。这一类方法重点研究如何在多模态融合之前提取更好的单模态特征表示。以视觉-语言-音频多模态应用为例,如何从视觉内容中解析出高层语义信息以增强视觉特征表达是这一类方法的主要研究内容。例如,从视觉内容中识别目标(Object)、属性(Attribute)、动作(Action)、关系(Relation)、场景图(Scene Graph)<sup>[73-75]</sup>和树形语义结构(Hierarchy)<sup>[76]</sup>等,进而实现对视觉内容的全局(Global)、区域(Regional)、目标(Object)和关系(Relation)等颗粒度的视觉语义建模。语言表示通常使用词的独热编码表示、词的上下文表示(Contextual Representation)<sup>[77-78]</sup>、句子表示<sup>[79-80]</sup>、句法依存关系(Syntactic Dependency)表示<sup>[81]</sup>、场景图表示<sup>[82]</sup>等。音频表示可使用基于 COVAREP<sup>[83]</sup>提取底层声学特征表示<sup>[85]</sup>、基于预训练模型 wav2vec<sup>[84]</sup>提取低维特征向量表示<sup>[85]</sup>等。本节侧重介绍多模态融合中的视觉特征表示方法。

#### 2.1.1 视觉全局表示

视觉全局表示(Global Representation)是从图像编码器的高层网络提取一个  $D$  维静态向量  $\mathbf{v}$  表示一幅图像。相关工作<sup>[43,72]</sup>通常使用预训练的 ResNet<sup>[86]</sup>对图像编码,再提取 ResNet 的最后一个池化层作为视觉全局表示(ResNet152 池化层输出是  $1 \times 2\,048$  维向量,即  $D=2\,048$ )。视觉全局表示



可用来初始化多模态自动摘要模型的解码器<sup>[43]</sup>,或作为一个特殊的字符与文本字符拼接,再用递归神经网络对拼接的字符序列编码<sup>[58]</sup>,或通过注意力机制学习与其他模态特征的联合表示<sup>[72]</sup>等。由于视觉全局表示将图像信息压缩到一个静态的向量中,这可能会导致大量图像细节信息的丢失。

### 2.1.2 视觉区域表示

视觉区域表示(Regional Representation)是从图像编码器的高层网络中提取一组  $D$  维向量表示一幅图像。每个  $D$  维向量表示图像中特定的大小相同的区域<sup>[87]</sup>。具体的,预训练 ResNet 先编码输入的图像,再提取 Conv5\_x 层的输出作为视觉区域表示  $v = \{v_1, \dots, v_K\}$  (ResNet152 的 Conv5\_x 层输出是  $7 \times 7 \times 2048$  的张量,即  $K = 49$ ,  $v_i$  的维度是 2048)。视觉区域表示与注意力机制相结合,通过在每一步解码过程中关注不同的图像区域可生成内容丰富的图像描述<sup>[87]</sup>。视觉区域表示实现了图像的细颗粒度表示,但是每个特征的感受野大小和形状相同,同一个目标(Object)可能被切分到多个区域中,它无法表达视觉上完整的语义信息。

### 2.1.3 视觉目标表示

视觉目标表示(Object Representation)也是用一组  $D$  维向量表示一幅图像,但每个  $D$  维向量表示图像中的一个目标(Object)。具体的,预训练 Faster R-CNN<sup>[88]</sup> 通常被用来检测目标所在的区域,再使用目标所在区域的视觉特征和边界框(Bounding-box)特征作为该视觉目标表示<sup>[79,81,89-90]</sup>。视觉目标表示与注意力机制等多模态融合方法相结合,可进一步提高视觉-语言任务的性能。例如,受人类视觉系统的启发,Anderson 等人<sup>[78]</sup>首次提出了一种“自底向上”和“自顶向下”相结合的注意力机制(BUTD)。BUTD 在 2017 年 CVPR 视觉问答比赛中获得冠军。视觉目标表示通过目标定位与分类实现视觉图像的浅层语义理解,但它无法刻画图像中多个目标间的语义关系。

### 2.1.4 视觉场景图表示

视觉场景图表示(Scene Graph Representation)是用场景图  $G = (V, R)$  表示一幅图像。场景图中的节点  $V = \{v_1, \dots, v_K\}$  是图像中的目标集合,关系  $R = \{r_1, \dots, r_R\}$  是图像中目标和目标间的显式语义关系(如 Wearing、Eating)、空间位置关系(如 Cover、Intersect、In)和隐式语义关系的集合,如图 6 所示。视觉场景图表示可实现模型对视觉内容的深度理解。Yao 等人<sup>[75]</sup>提出了基于 GCN-LSTM 的网络结

构,将视觉场景图中的显式语义关系和空间位置关系集成到图像编码器中。GCN-LSTM 网络显著提高了图像描述任务的性能。Li 等人<sup>[79]</sup>提出了一种关系感知的图注意力网络(ReGAT),它通过图注意力机制对图像目标间的显式关系(语义关系和空间关系)和隐式关系进行建模,学习问题自适应的多模态联合表示,ReGAT 可提高 VQA 的性能。文献[75、79]使用 Faster R-CNN 识别图像中目标,并提取目标的视觉特征表示  $v_i$ 。

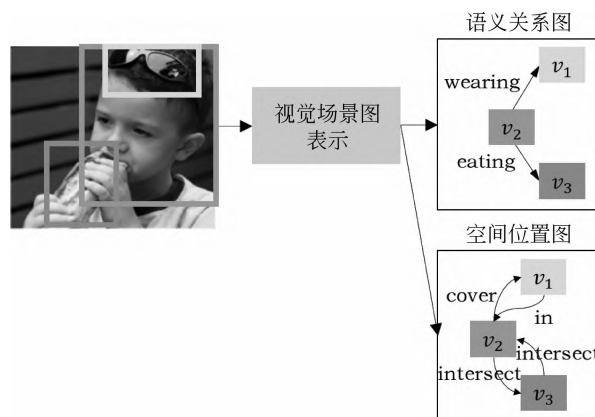


图 6 视觉场景图表示

除了场景图表示,Yao 等人<sup>[76]</sup>把视觉内容解析成一个树状结构,其根节点是整个图像,中间节点为一组图像物体,叶子节点则是在图像目标的基础上应用图像分割技术得到的图像 Instance 级的区域。

## 2.2 根据融合阶段进行分类

根据多模态融合的阶段,多模态融合方法可分为早期融合<sup>[79-82,90]</sup>、中期融合<sup>[91]</sup>和晚期融合<sup>[92]</sup>。早期融合的特点是单模态表示学习简单,而多模态融合部分的模型深度大,融合策略复杂。例如,词的独热编码表示和视觉区域表示直接参与多模态融合<sup>[93]</sup>。晚期融合的特点是单模态表示学习模型复杂,多模态融合一般采用拼接、按位乘/求平均等简单策略<sup>[92]</sup>。由于晚期融合抑制了模态之间的交互,目前大部分基于深度学习的模型均使用早期或者中期融合。在第 3 节介绍的多模态预训练模型中,基于单流架构(Single-Stream)的预训练模型把融合操作放在早期阶段,如 VideoBERT<sup>[94]</sup>、Unicoder-VL<sup>[95]</sup>、Oscar<sup>[96]</sup>、VL-BERT<sup>[97]</sup>和 M3P<sup>[98]</sup>等。基于双流架构(Two-Stream)的预训练模型则把融合操作放置在深层模型的中期阶段的多个层中,如 ERNIE-ViL<sup>[82]</sup>、LXMERT<sup>[91]</sup>、ActBERT<sup>[99]</sup>和 ViL-BERT<sup>[100]</sup>等。

Alberti 等人<sup>[90]</sup>通过实验证明在视觉常识推理 (Visual Commonsense Reasoning, VCR) 应用中, 语言与视觉的早期融合是获得高准确率的关键。Shrestha 等人<sup>[80]</sup>也通过实验发现早期融合对他们提出的模型 RAMEN 至关重要, 因为去掉早期融合会导致 VQA 准确率的绝对值在视觉推理数据集 CLEVR 上下降 20%, 在视觉理解数据集 VQAv2 上下降 4%。

### 2.3 根据融合方式进行分类

多模态融合模型的设计是多模态融合的关键研究点。我们将多模态融合模型分为简单融合、门控融合 (Gating)、注意力融合 (Attention)、Transformer 融合、图模型融合 (Graph Fusion) 和双线性注意力 (Bilinear Attention) 融合共六类方法。常见简单融合方法包括编码器、解码器的初始化 (参见 1.6 节和 2.1.1 节)、拼接、按位乘/求和/求平均等操作。本节主要介绍其余的五类较复杂的融合方法。

#### 2.3.1 门控融合

基于自编码 (Auto-encoding)<sup>[101]</sup> 和自回归 (Auto-regression)<sup>[102]</sup> 的大规模预训练语言模型和在下游任务上的微调相结合是自然语言处理研究和应用的新方法。但文本预训练语言模型与下游的多模态任务相结合还是一个尚未充分研究的课题。Rahman 等人<sup>[103]</sup>提出了一种多模态适应门 (Multi-modal Adaptation Gate, MAG) 的网络结构将非语言特征 (视觉和声学特征) 与文本预训练语言模型融合, MAG 与 BERT<sup>[101]</sup> 结合 (MAG-BERT) 以及 MAG 与 XLNet<sup>[104]</sup> 结合 (MAG-XLNET) 都可以有效融合三个模态信息, 并在多模态情感识别数据集 CMU-MOSI 和 CMU-MOSEI 上获得当时最优性能。

#### 2.3.2 注意力融合

Bahdanau 等人<sup>[105]</sup>在 2015 年提出的注意力机制是为了让神经机器翻译模型中的解码器在每一步解码过程中, 有针对性地选择源语言中“对齐”的词来指导目标语言的解码, 包括全局注意力和局部注意力两种方法。2017 年 Vaswani 等人<sup>[106]</sup>提出了由多头注意力和自注意力等模块组成的 Transformer。目前 Transformer 已经成为自然语言处理、计算机视觉和语音领域的标准模型之一。在多模态领域, Yang 等人<sup>[77]</sup>提出了 Stacked Attention Networks (SANs), 通过多层视觉注意力机制逐步过滤掉图像中的噪声区域, 定位到与答案高度相关的图像区域, 从而提高 VQA 准确率。Anderson 等人<sup>[78]</sup>提出

一种“自底向上”和“自顶向下”相结合的注意力机制。具体的, 基于 Faster R-CNN 的“自底向上”的注意力机制提取图像中的兴趣区域, “自顶向下”的注意力机制确定兴趣区域的权重。

上述注意力都是单向的视觉注意力, 即基于文本表示选择性地关注图像中的兴趣区域。Lu 等人<sup>[107]</sup>认为文本注意力和视觉注意力同等重要, 并提出了协同注意力机制 (Co-attention)。协同注意力又根据文本注意力和视觉注意力计算的交替顺序分为平行协同注意力 (Parallel Co-attention) 和交替协同注意力 (Alternating Co-attention) 两种策略。Nam 等人<sup>[108]</sup>基于类似的想法提出了双重注意力网络 (Dual Attention Networks)。受 Transformer 模型的启发, Yu 等人<sup>[109]</sup>提出了一种类 Transformer 结构的协同注意力机制, 可实现文本中的任一词与图像中的任一区域间的完全交互。

#### 2.3.3 Transformer 融合

BERT 凭借着 Transformer 强大的特征学习能力和掩码语言模型 (Masked Language Model) 实现双向编码, 刷新了多个 NLP 任务的最优性能。2019 年 Transformer 开始被应用到多模态领域。基于 Transformer 的多模态融合又分为单流模型<sup>[95-98]</sup>和双流模型<sup>[82, 91, 110-111]</sup>两大类。单流模型使用一个 Transformer 在一开始便对多模态信息进行充分的交互。双流模型则对不同的模态使用独立的 Transformer 编码, 再通过协同注意力机制实现不同模态间的融合, 如图 7 所示。双流模型可以适应不同模态独立的处理需求。ViLBERT<sup>[100]</sup>证明了双流模型的性能优于单流模型, 但目前没有更多的对比实验分析单流模型和双流模型的优点和不足。

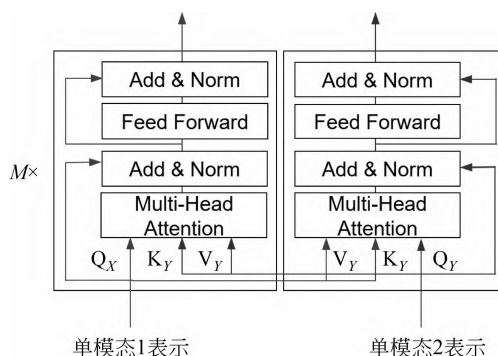


图7 基于Transformer的双流多模态融合

#### 2.3.4 图模型融合

对比 CNN/RNN 等神经网络模型, 图神经网络模型的优势是可处理具有复杂结构的异构数据, 并且具

备一定的关系推理能力和可解释性。图神经网络最近一两年在自然语言处理领域也受到了广泛的关注。

在视觉-语言任务中,将图像进行结构化(参见 2.1.4 节),再与图神经网络结合,有助于对图像的深度理解,进而提高图像描述和视觉问答等视觉-语言任务的性能<sup>[79]</sup>。Huang 等人<sup>[81]</sup>提出一种双通道图卷积网络(DC-GCN)。DC-GCN 通过 I-GCN 模块学习图像中物体间的关系、通过 Q-GCN 模块学习问题中词的依存关系,再通过注意力对齐模块学习多模态的联合表示。Yin 等人<sup>[112]</sup>将基于图的多模态融合编码器应用到多模态神经机器翻译模型中。不同于 DC-GCN 对图像和文本独立建图,Yin 等人<sup>[112]</sup>把源语言中的词和图像中的物体放到了同一个图中,再堆叠多个基于图神经网络的多模态融合层(在每一层顺序执行模态内融合和模态间融合)。该方法可以同时学习模态内和模态间的各种颗粒度的语义关系,进而显著提高了机器翻译的性能。

最近,基于图模型的多模态融合也被广泛应用于多模态情感识别任务。Hu 等人<sup>[113]</sup>提出了一种基于图卷积网络的多模态融合模型(MMGCN),它可以有效地融合多模态信息和学习长距离的依赖关系,还可以通过说话人向量(Speaker Embedding)把说话人的音色特征等信息融入情感识别模型中。

### 2.3.5 双线性注意力融合

协同注意力机制虽然同时引入了文本和视

觉注意力,实现了文本和图像双向交互。但为了减少计算量,协同注意力为每个模态建立了独立的注意力分布。因此,协同注意力忽视了问题和图像之间的两两交互。对此,Kim 等人<sup>[114]</sup>提出了双线性注意力网络(BAN)。双线性注意力网络是低秩双线性池化方法的一般推广。本文不展开介绍 BAN 模型,有兴趣的读者请参考相关文献。

### 2.3.6 多模态融合模型小结

门控融合和注意力融合是早些年提出的基础的多模态融合方法。它们的优点是能方便地与 CNN/LSTM/Transformer 等主流的神经网络结构相结合,也能与 2.1 节介绍的多种单模态表示相结合。图模型融合和 Transformer 融合是近几年提出的新方法,它们的模型结构较复杂,且对单模态的表示要求较高。如图模型融合需要跟视觉场景图表示(和文本的图表示)相结合。双流 Transformer 融合视觉-语言-语音 3 种模态信息,需要多个 Transformer<sup>[110-111]</sup>。图模型融合和 Transformer 融合通常可获得更好的性能,如表 4 所示。门控机制和注意力机制跟预训练模型结合,也能取得不错的性能,如门控机制跟 XLNet 相结合的 MAG-XLNet 模型在情感识别上获得了最佳的性能。

表 4 多模态融合方法的代表性模型在视觉问答、图像描述和情感识别数据集上的性能对比

| 多模态融合          | 典型模型                       | 发表刊物         | 视觉问答        |             | 图像描述        | 情感识别     |
|----------------|----------------------------|--------------|-------------|-------------|-------------|----------|
|                |                            |              | VQA v2.0    |             | MSCOCO      | CMU-MOSI |
|                |                            |              | Test-dev    | Test-std    |             |          |
| 门控融合           | MAG-XLNet <sup>[103]</sup> | ACL 2020     | —           | —           | —           | 87.9     |
| 注意力融合          | DAN <sup>[108]</sup>       | CVPR 2017    | 64.3        | 64.2        | —           | —        |
|                | Bottom-Up <sup>[78]</sup>  | CVPR 2018    | 65.3        | 65.7        | 36.2        | —        |
|                | DCN <sup>[87]</sup>        | CVPR 2018    | 66.9        | 67.0        | —           | —        |
| Transformer 融合 | MuT-XLNET <sup>[110]</sup> | ACL 2019     | —           | —           | —           | 83.1     |
|                | OSCAR <sup>[96]</sup>      | ECCV 2020    | 73.2        | 73.6        | 36.5        | —        |
|                | ERNIE-ViL <sup>[82]</sup>  | AAAI 2021    | 72.6        | 72.8        | —           | —        |
|                | E2E-VLP <sup>[93]</sup>    | ACL 2021     | 73.3        | 73.7        | 36.2        | —        |
|                | UNIMO <sup>[122]</sup>     | ACL 2021     | <b>73.8</b> | <b>74.0</b> | <b>38.8</b> | —        |
| 图模型融合          | MN-GMN <sup>[89]</sup>     | ACL 2020     | 73.2        | 73.5        | —           | —        |
|                | ReGAT <sup>[79]</sup>      | ICCV 2019    | 70.3        | 70.6        | —           | —        |
|                | DC-GCN <sup>[81]</sup>     | ACL 2020     | 71.2        | 71.5        | —           | —        |
| 双线性注意力融合       | BAN <sup>[114]</sup>       | NeurIPS 2018 | 70.0        | 70.4        | —           | —        |



## 2.4 其他融合方法

融合语言、视觉和声学序列信息的多模态情感识别,由于每个模态的采样率不同,多模态序列通常表现出“未对齐”特性(也称之为异步性)。早期的多模态情感识别工作是在词对齐的多模态序列上展开的。最近也有工作提出了基于异步的多模态序列的建模方法。然而, Tsai 等人<sup>[110]</sup>提出的多模态 Transformer(MulT)一次只能接收两个模态。为了实现三个模态的融合,作者使用了六个跨模态 Transformer。Yang 等人<sup>[115]</sup>提出了一个可解释的基于图神经网络的异步多模态序列融合算法:模态-时间注意力图(Modal-Temporal Attention Graph, MTAG)算法。MTAG 算法首先将多模态序列转为一个异构图,再从多模态序列中抽取特征作为节点,节点间通过多模态边(Multimodal Edges)和时间边(Temporal Edge)进行连接。最后,在图上进行融合操作,实现每一个模态的节点与其他模态节点的交互。

由于利用了多个模态间的互补性,多模态系统具有较高的预测鲁棒性。然而,在现实应用场景中,我们经常会遇到模态缺失的问题。例如,由于隐私问题关闭了摄像头、由于语音识别错误带来的语言模态缺失等。模态缺失问题通常会导致现有基于全模态的多模态融合模型失效。对此,Zhao 等人<sup>[116]</sup>提出了基于缺失模态想象网络(Missing Modality Imagination Network, MMIN)来处理不确定的模态缺失问题。由于模态缺失现象的普遍性,该问题将会是多模态领域接下来的一个研究热点。

## 3 多模态预训练

通过预训练语言模型从海量无标注数据中学习通用知识,再在下游任务上用少量的标注数据进行微调,已经成为自然语言处理领域成熟的新范式。从 2019 年开始,预训练语言模型(BERT<sup>[101]</sup>、GPT-3<sup>[102]</sup>、BART<sup>[117]</sup>和 T5<sup>[118]</sup>等)相继被扩展到多语言和多模态等场景。

相对于文本预训练语言模型,多模态预训练模型可以更好地对细颗粒度的多模态语义单元(词或者目标)间的相关性进行建模。例如,基于语言上下文,被掩码的词“on top of”可以被预测为符合语法规则的词“under”或“into”等。但这与关联的图片场景“猫在车顶”不符。通过多模态预训练,模型从图像中捕获“汽车”“猫”之间的空间关系,从而可以准确地预测出掩码词是“on top of”<sup>[82]</sup>。大部分的多模态预训练模型是在视觉-语言对齐数据上进行的。例如,使用图像和文本对齐数据集(MSCOCO<sup>[12]</sup>、Conceptual Captions<sup>[13]</sup>、Visual Genome<sup>[15]</sup>和 SBU Captions<sup>[16]</sup>等)训练的跨模态预训练模型 LXMERT<sup>[91]</sup>、Oscar<sup>[96]</sup>、VL-BERT<sup>[97]</sup>和 ViLBERT<sup>[100]</sup>, M3P<sup>[98]</sup>。使用视频和文本对齐数据集训练的 VideoBERT<sup>[94]</sup>和 ActBERT<sup>[99]</sup>等<sup>[119-120]</sup>。Liu 等人<sup>[85]</sup>最近还发布了视觉、文本、语音三模态预训练模型 OPT。

本文表 5 中从网络结构、模型输入、预训练目标、预训练语料和下游任务等维度对比了最新的视觉-语言跨模态预训练模型 ERNIE-VIL<sup>[82]</sup>、LXMERT<sup>[91]</sup>、LightningDOT<sup>[92]</sup>、E2E-VLP<sup>[93]</sup>、Unicoder-VL<sup>[95]</sup>、Oscar<sup>[96]</sup>、VL-BERT<sup>[97]</sup>、M3P<sup>[98]</sup>、ViLBERT<sup>[100]</sup>、TDEN<sup>[121]</sup>、UNIMO<sup>[122]</sup>。表 5 中的  $\langle \mathbf{I}, \mathbf{w} \rangle$  表示“图像,语言”对,  $\mathbf{I}$  表示一幅图像,  $\mathbf{w} = w_1, \dots, w_T$  表示长度为  $T$  的文本表示。 $\mathbf{g} = g_1, \dots, g_G$  是图像区域表示,  $\mathbf{q} = q_1, \dots, q_K$  和  $\mathbf{v} = v_1, \dots, v_K$  分别表示图像中的目标的文本表示和目标的视觉表示。 $\mathbf{g}$  和  $\mathbf{v}$  的提取可参考 2.1 节的介绍。此外, [SEP]、[IMG]、[CLS]等特殊标记用来分割不同模态。MLM(Masked Language Model)是根据未掩码的词和图像区域预测掩码单词。MOC(Masked Object Classification)根据未掩码的图像区域和文本预测掩码区域的目标类别。MOR(Masked Object Regression)根据未掩码的图像区域和文本预测掩码区域的特征表示。MSG(Masked Sentence Generation)根据输入图像逐字生成句子。VQA 根据输入的图像和该图像相关问题预测该问题的答案。CMCL 是跨模态对比学习任务。VLM 是预测图像-文本对是否语义一致。

表 5 视觉-语言预训练模型对比

| 模型     | 发表刊物       | 网络结构 | 模型输入   | 预训练任务  | 预训练语料       | 下游任务  |
|--------|------------|------|--|--|-------------|-------|
| LXMERT | EMNLP 2019 | 双流模型 | 语言流: $[[CLS]] \mathbf{w}$<br>[END]]<br>图像流: $[\mathbf{v}]$ | 语言任务: MLM<br>图像任务: MOC 和 MOR<br>VL 任务: VLM 和 VQA | 9.8M 图像-语言对 | 多模态理解 |



续表

| 模型            | 发表刊物         | 网络结构 | 模型输入   | 预训练任务   | 预训练语料                                      | 下游任务                             |
|---------------|--------------|------|--|---|--|----------------------------------|
| ViL-BERT      | NeurIPS 2019 | 双流模型 | 语 言 流：[[CLS] w [SEP]]<br>图 像 流：[[IMG]v]                | 语言任务：MLM<br>图像任务：MOC<br>VL 任务：VLM                       | 3.1M 图 像-语言对                               | 多模态理解                            |
| Unicoder-VL   | AAAI2020     | 单流模型 | [[CLS] v [SEP] w [SEP]]                                | 语言任务：MLM<br>图像任务：MOC<br>VL 任务：VLM                       | 3.8M 图 像-语言对                               | 多模态理解                            |
| Oscar         | ECCV2020     | 单流模型 | [[CLS] w [SEP] q [SEP]v]                               | 语言任务：MLM<br>图像任务：预测目标标签 o 与图像 I 是否对齐                    | 6.8M 图 像-语言对                               | 多模态理解<br>多模态生成                   |
| VL-BERT       | ICLR2020     | 单流模型 | [[CLS] w [SEP] v [END]]                                | 语言任务：MLM<br>图像任务：MOC                                    | 3.3M 图 像-语言对                               | 多模态理解                            |
| M3P           | CVPR2021     | 单流模型 | [[CLS]w[IMG]v]<br>或[w]                                 | 语言任务：MLM<br>图像任务：MOC 和 MOR<br>VL 任务：VLM                 | 3.3M 图 像-语言对<br>101G 多语言<br>文 本 维 基<br>百科  | 多模态理解                            |
| ERNIE-ViL     | AAAI2021     | 双流模型 | 语 言 流：[[CLS] w [SEP]]<br>图 像 流：[[IMG]v]                | 语言任务：MLM,场景图预测(目标预测、属性预测、关系预测)<br>图像任务：MOC<br>VL 任务：VLM | 3.8M 图 像-语言对                               | 多模态理解                            |
| Light-ningDOT | NAACL 2021   | 双流模型 | 语言流：[[CLS]w]<br>图像流：[[CLS]v]                           | 语言任务：MLM<br>图像任务：MOC 和 MOR<br>VL 任务：VLM                 | 9.5M 图 像-语言对                               | 多模态理解                            |
| TDEN          | AAAI2021     | 双流模型 | 语 言 流：[[CLS] w [SEP]]<br>数据流：[[IMG]v]                  | 语言任务：MLM<br>图像任务：MOC<br>VL 任务：VLM 和 MSG                 | 3.3M 图 像-语言对                               | 多模态理解<br>多模态生成                   |
| E2E-VLP       | ACL2021      | 单流模型 | [[CLS]w[SEP]g]   | 语言任务：MLM<br>图像任务：目标检测<br>VL 任务：VLM 和 MSG                | 6.01M 图 像-语言对                              | 多模态理解<br>多模态生成                   |
| UNIMO         | ACL2021      | 单流模型 | [[CLS]w[SEP]]<br>或[[IMG]v]<br>或[[IMG] v [CLS] w [SEP]] | 语言任务：MLM<br>图像任务：MOC 和 MOR<br>VL 任务：CMCL 和 MSG          | 54G 文 本<br>语料<br>1.7M 图像库<br>9.58M 图 像-语言对 | 多模态理解<br>多模态生成<br>单模态理解<br>单模态生成 |

从表 5 中的 11 个图像-语言跨模态预训练模型的对比,我们发现的跨模态预训练模型的特点如下:  
①单流模型和双流模型均被广泛采用。虽然双流模型可以适应每种模态的不同处理需求,但目前尚无完整的实验证明双流模型优于单流模型。②多模态预训练模型从应用于多模态理解任务或多模态生成任务发展到可兼顾多模态理解和生成两大任务的统一模型。③相对动辄上百 G 甚至 T 级别的单模态数据,多模态对齐数据的规模有限。最新的多模态预训练模型可以利用互联网上的大规模非对齐的文本数据、图像数据、以及文本-图像对齐数据学习更

通用的文本和视觉表示,以提高模型在视觉和语言的理解和生成能力,如 M3P 和 UNIMO。④多模态预训练模型从仅应用于多模态下游任务发展到可同时应用于单模态下游任务和多模态下游任务。  
上述的多模态预训练模型需要在大量图像文本的对齐语料上进行训练。然而,此类数据的收集成本昂贵,很难扩大规模。受无监督机器翻译<sup>[123-124]</sup>的启发,Li 等人<sup>[125]</sup>提出了一种不依赖图像-文本对齐语料的预训练 U-VisualBERT,该预训练模型的输入是一批文本数据,或一批图像数据,并通过图像中物体标签作为锚点(Anchor Points)对齐两种模

态。U-VisualBERT 在四个多模态任务上取得与使用多模态对齐数据训练的预训练模型接近的性能。该方向可能会是接下来的一个研究热点。


#### 4 多模态技术的产业应用

本节介绍多模态信息处理在商品文案生成、智能客服与营销等场景的应用。

多模态商品文案生成是基于商品的文本描述和商品的图片生成卖点突出的商品介绍文案的任务。

为了生成一段简洁凝练、卖点突出、流畅、合规的商品文案, Li 等人<sup>[43]</sup>提出了一种基于商品要素的多模态商品信息自动摘要模型, 其可以根据商品的文本描述、商品图片信息自动生成商品短文。目前文献[43]中的算法已支持 3 000 多个商品品类, 广泛应用于商品导购机器人、搭配购、AI 直播带货等实际场景中。AI 创作的文案人工审核通过率超过 95%, AI 文案曝光点击率高出专业写手平均水平 40%。表 6 对比了文本模型和多模态模型的生成文案效果。

表 6 文本生成模型 vs. 多模态生成模型

|   |       |   |
|---|-------|---|
|  | 文本模型  | 时尚的两件套设计, 穿出不同种类的风格, 选用优质的针织面料, 手感细腻, 具有良好的亲肤效果, 穿着舒适不紧绷, 配以甜美的喇叭袖, 丰富了整体的视觉效果。 |
|   | 多模态模型 | 这款来自欧芮儿很时尚的套装, 以橙色为主色调的条纹设计, 靓丽吸睛, 让你穿出不一样时髦造型。丰富视觉效果, 凸显层次美感, 木耳边的喇叭袖, 增添潮流亮点。 |

智能客服场景中, 超过 16% 的客服与用户的对话包括一张以上的图片(截屏图片和实拍图片)。所以, 客服机器人不仅要理解文字内容, 还要理解图片等多模态内容, 才能准确回答用户咨询。基于多模态技术的用户意图识别已经应用于京东智能情感客服系统。多模态情感识别也应用到语音客服质检<sup>①</sup>、语音外呼机器人等产品中。此外, 融合语音、计算机视觉和自然语言处理的数字人已应用到智能客服、虚拟主播、数字人直播带货等场景。

#### 5 结束语

多模态信息处理是一个典型的多学科交叉领域。最近几年, 多模态信息处理受到自然语言处理、计算机视觉和语音与声学领域研究者的广泛关注。本文从自然语言处理的视角出发, 首先介绍了目前热点的多模态应用, 接着介绍了多模态的三个重要研究方向及其主流方法: 即视觉的单模态表示(视觉全局表示、视觉区域表示、视觉目标表示和视觉场景图表示)、多模态融合(简单融合、门控融合、注意力融合、Transformer 融合、图模型融合和双线性注意力融合)和通用的多模态预训练。最后, 本文对多模态技术在产业界的应用进行了简要的描述。

多模态信息处理还有很多亟待进一步研究的课

题。我们认为, 以下五个方向将是多模态信息处理技术领域未来重要的研究内容: ①非对齐语料上的多模态信息处理。目前, 大多数下游的多模态任务和多模态预训练模型都依赖多模态对齐语料。相对动辄上百 G 甚至 T 级别的单模态语料, 多模态对齐语料的规模还是很有限。探索如何在海量非对齐多模态语料上训练多模态模型具有非常实用的价值, 也是多模态领域需要重点关注的课题之一。此方向已经有了初步的探索。例如, 利用多模态对齐技术将海量的单模态语料与其他模态进行自动对齐<sup>[48, 122]</sup>。②面向单模态和多模态的理解和生成任务的统一模型。当前的主流模型或面向单模态理解(或生成)或面向多模态理解(或生成)的模型, 构建一个既适用于单模态理解与生成任务, 又适用于多模态理解与生成任务的统一模型是未来非常重要的研究方向。多模态模型在文本任务上的性能未来可能会超过单模态模型<sup>[48, 122]</sup>。③高噪声环境下的多模态鲁棒性融合。真实场景常常有较强的背景噪声, 部分模态的数据通常是模糊或缺失的。因此, 探索如何在高噪声情况下获得信息缺失的有效表征, 提高模型预测鲁棒性和准确性是多模态领域重要的研究课题之一。文献[116]提出一种基于缺失模态

① 语音客服质检是根据语音和 ASR 识别结果识别客服和用户的情绪变化, 提高客服服务的质量。

的想象网络(Missing Modality Imagination Network, MMIN)对该方向进行了初步的探索。④多模态与知识的融合。2.1 节介绍的从视觉内容中提取视觉粗粒度特征表示和基于视觉场景图的细颗粒度特征表示,其目的都是增强视觉特征表示。我们认为,如何提取更精细粒度的视觉特征表示是多模态领域重要的基础研究方向之一。引入知识图谱作为图像实体信息的补充,从而进行知识增强的视觉特征表示是该方向一种探索思路<sup>[126-127]</sup>。⑤复杂交互情境下的多模态应用。第 1 节介绍了多模态信息处理技术的多个应用场景。我们认为,数字人、元宇宙(Metaverse)是多模态信息处理技术最佳的应用场景之一,探索复杂交互情境下的多模态信息处理是多模态领域未来最重要的研究方向之一。

## 参考文献

- [1] Morency L P, Baltrusaitis T. Tutorial on multimodal machine learning[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017.
- [2] Zhang C, Yang Z, He X, et al. Multimodal intelligence: representation learning, information fusion and applications[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478-493.
- [3] Summaira J, Li X, Shoib A M, et al. Recent advances and trends in multimodal deep learning: a review[J]. arXiv preprint arXiv:2105.11087, 2015.
- [4] Zhang S, Song L, Jin L, et al. Video-aided unsupervised grammar induction [C]//Proceedings of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies, 2021: 1513-1524.
- [5] Busso C, Bulut M, Lee C, et al. IEMOCAP: interactive emotional dyadic motion capture database [J]. Language Resources and Evaluation, 2008, 42(4): 335-359.
- [6] Zadeh A, Zellers R, Pincus E, et al. Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages[J]. IEEE Intelligent Systems, 2016, 31(6): 82-88.
- [7] Zadeh A B, Liang P P, Poria S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2236-2246.
- [8] Yu W, Xu H, Meng F, et al. CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotations of modality [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3718-3727.
- [9] Jia J, Zhou S, Yin Y, et al. Inferring emotions from large-scale internet voice data[J]. IEEE Transactions on Multimedia, 2019, 21(7): 1853-1866.
- [10] Cai Y, Cai H, Wan X. Multi-modal sarcasm detection in Twitter with hierarchical fusion model [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2506-2515.
- [11] Truong Q T, Lauw H W. VistaNet: visual aspect attention network for multimodal sentiment analysis [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 305-312.
- [12] Lin T, Maire M, Belongie S, et al. Microsoft COCO: common objects in context [C]//Proceedings of the European Conference on Computer Vision, 2014: 740-755.
- [13] Sharma P, Ding N, Goodman S, et al. Conceptual Captions: a cleaned, hypernymed, image ALT-text dataset for automatic image captioning [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2556-2565.
- [14] Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics, 2014(2): 67-78.
- [15] Krishna R, Zhu Y, Groth O, et al. Visual Genome: connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.
- [16] Ordonez V, Kulkarni G, Berg T T. Im2Text: Describing images using 1 million captioned photographs [C]//Proceedings of the 24th International Conference on Neural Information Processing Systems, 2011: 1143-1151.
- [17] Xu J, Mei T, Yao T, et al. MSR-VTT: a large video description dataset for bridging video and language [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016: 5288-5296.
- [18] Krishna R, Hata K, Ren F, et al. Dense-captioning events in videos [C]//Proceedings of the IEEE International Conference on Computer Vision, 2017.
- [19] Zhou L, Xu C, Corso J J. Towards automatic learning of procedures from web instructional videos [C]//Proceedings of the AAAI, 2018: 7590-7598.
- [20] Pan Y, Li Y, Luo J, et al. Auto-captions on GIF: a large-scale video-sentence dataset for vision-language pre-training [J]. arXiv preprint arXiv: 2007.02375,

- 2020.
- [21] Huang T H, Ferraro F, Mostafazadeh N, et al. Visual storytelling[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 1233-1239.
  - [22] Agrawal A, Lu J, Antol S, et al. VQA: visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 2425-2433.
  - [23] Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA Matter: elevating the role of image understanding in visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 6325-6334.
  - [24] Johnson J, Hariharan B, Maaten L, et al. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1988-1997.
  - [25] Hudson D A, Manning C D. GQA: a new dataset for real-world visual reasoning and compositional question answering[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019: 6700-6709.
  - [26] Zhou Y, Ji R, Su J, et al. More than an answer: neural pivot network for visual question answering[C]//Proceedings of the 25th ACM International Conference on Multimedia, 2017: 681-689.
  - [27] Zhou Y, Ji R, Su J, et al. Dynamic capsule attention for visual question answering[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 9324-9331.
  - [28] Das A, Kottur S, Gupta K, et al. Visual dialog[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1080-1089.
  - [29] Mostafazadeh N, Brockett C, Dolan B, et al. Image-grounded conversations: multimodal context for natural question and response generation[C]//Proceedings of the International Joint Conference on Natural Language Processing, 2017: 462-472.
  - [30] Vries H D, Strub F, Chandar S, et al. Guesswhat?! visual object discovery through multi-modal dialogue[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4466-4475.
  - [31] Shuster K, Humeau S, Bordes A, et al. Image-Chat: engaging grounded conversations[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 2414-2429.
  - [32] Alamri H, Cartillier V, Das A, et al. Audio-visual scene-aware dialog[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019: 7558-7567.
  - [33] Saha A, Khapra M, Sankaranarayanan K. Towards building large scale multimodal domain-aware conversation systems[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 696-704.
  - [34] Moon S, Kottur S, Crook P A, et al. Situated and interactive multimodal conversations[C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020: 1103-1121.
  - [35] Zhao N, Li H, Wu Y, et al. The JDDC 2.0 Corpus: a large-scale multimodal multi-turn Chinese dialogue dataset for e-commerce customer service[J]. arXiv preprint arXiv:2109.12913, 2021.
  - [36] Li M, Zhang L, Ji H, et al. Keep meeting summaries on Topic: Abstractive Multi-Modal Meeting Summarization[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2190-2196.
  - [37] Palaskar S, Libovicky J, Gella S, et al. Multimodal Abstractive Summarization for How2 Videos[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6587-6596.
  - [38] Li H, Zhu J, Ma C, et al. Multi-modal summarization for asynchronous collection of text, image, audio and video[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 1092-1102.
  - [39] Li H, Zhu J, Liu T, et al. Multi-modal sentence summarization with modality attention and image filtering[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 4152-4158.
  - [40] Li H, Zhu J, Zhang J, et al. Multimodal Sentence Summarization via Multimodal Selective Encoding[C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020: 5655-5667.
  - [41] Zhu J, Li H, Liu T, et al. MSMO: multimodal summarization with multimodal output[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 4154-4164.
  - [42] Zhu J, Zhou Y, Zhang J, et al. Multimodal Summarization with Guidance of Multimodal Reference[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 9749-9756.
  - [43] Li H, Yuan P, Xu S, et al. Aspect-aware multimodal summarization for Chinese e-commerce products[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 8188-8195.
  - [44] Carletta J, Ashby S, Bourban S, et al. The AMI



- meeting corpus: a pre-announcement[C]//Proceedings of the 2nd International Workshop on Machine Learning for Multimodal Interaction, 2005: 28-39.
- [45] Palaskar S, Caglayan O, Palaskar S, et al. Metze. How2: a large-scale dataset for multimodal language understanding[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems, 2018: 1-12.
- [46] Hessel J, Lee L, Mimno D. Unsupervised discovery of multimodal links in multiimage, multisentence documents[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019: 2034-2045.
- [47] Suhr A, Zhou S, Zhang A, et al. A corpus for reasoning about natural language grounded in photographs[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6418-6428.
- [48] Tan T, Bansal M. Vokenization: improving language understanding with contextualized, visual-grounded supervision[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 2066-2080.
- [49] Plummer B A, Wang L, Cervantes C M, et al. Flickr30k Entities: collecting region-to-phrase correspondences for richer image-to-sentence models[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 2641-2649.
- [50] Wang Q, Tan H, Shen S, et al. MAF: multimodal alignment framework for weakly-supervised phrase grounding[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 2030-2038.
- [51] Zhang H, Sun A, Jing W, et al. Parallel attention network with sequence matching for video grounding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021: 776-790.
- [52] Gao J, Sun C, Yang Z, et al. Tall: temporal activity localization via language query[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 5277-5285.
- [53] Krishna R, Hata K, Ren F, et al. Dense-captioning events in videos[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 706-715.
- [54] Regneri M, Rohrbach M, Wetzel D, et al. Grounding action descriptions in videos[J]. Transactions of the Association for Computational Linguistics, 2013(1): 25-36.
- [55] Sigurdsson G A, Varol G, Wang X, et al. Hollywood in homes: crowdsourcing data collection for activity understanding[C]//Proceedings of the European Conference on Computer Vision, 2016: 510-526.
- [56] Elliott D, Frank S, Hasler E. Multi-language image description with neural sequence models[J]. arXiv preprint arXiv:1510.04709, 2015.
- [57] Elliott D, Frank S, Simaan K, et al. Multi30k: multilingual English-German image descriptions[C]//Proceedings of the 5th Workshop on Vision and Language, 2016: 70-74.
- [58] Huang P, Liu F, Shiang S R, et al. Attention-based multimodal neural machine translation[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 639-645.
- [59] Calixto I, Liu Q. Incorporating global visual features into attention-based neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 992-1003.
- [60] Elliott D, Kadar A. Imagination improves multimodal translation[C]//Proceedings of the International Joint Conference on Natural Language Processing, 2017: 130-141.
- [61] Zhou M, Cheng R, Lee Y J, et al. A visual attention grounding neural model for multimodal machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 3643-3653.
- [62] Ive J, Madhyastha P, Specia L. Distilling translations with visual awareness[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6525-6538.
- [63] Yao S, Wan X. Multimodal transformer for multimodal machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 4346-4350.
- [64] Lin H, Meng F, Su J, et al. Dynamic context-guided capsule network for multimodal machine translation[C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 1320-1329.
- [65] Wu Z, Kong L, Bi W, et al. Good for misconceived reasons: an empirical revisiting on the need for visual context in multimodal machine translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021: 6153-6166.
- [66] Moon S, Neves L, Carvalho V. Multimodal named entity recognition for short social media posts[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 852-860.
- [67] Zhang Q, Fu J, Liu X, et al. Adaptive coattention network for named entity recognition in Tweets[C]//

- Proceedings of the AAAI Conference on Artificial Intelligence, 2018; 5674-5681.
- [68] Lu D, Neves L, Carvalho V, et al. Visual attention model for name tagging in multimodal social media [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018; 1990-1999.
- [69] Yu J, Jiang J, Yang L, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020; 3342-3352.
- [70] Sui D, Tian Z, Chen Y, et al. A large-scale Chinese multimodal NER dataset with speech clues [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021; 2807-2818.
- [71] IV R L, Humeau S, Singh S. Multimodal attribute extraction [C]//Proceedings of the 31st Conference on Neural Information Processing Systems, 2017; 1-7.
- [72] Zhu T, Wang Y, Li H, et al. Multimodal joint attribute prediction and value extraction for e-commerce product [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020; 2129-2139.
- [73] Johnson J, Krishna R, Stark M, et al. Image retrieval using scene graphs [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015; 3668-3678.
- [74] Lu C, Krishna R, Bernstein M, et al. Visual relationship detection with language priors [C]//Proceedings of the European Conference on Computer Vision, 2016; 852-869.
- [75] Yao T, Pan Y, Li T, et al. Exploring visual relationship for image captioning [C]//Proceedings of the European Conference on Computer Vision, 2018; 711-727.
- [76] Yao T, Pan Y, Li Y, et al. Hierarchy parsing for image captioning [C]//Proceedings of the IEEE International Conference on Computer Vision, 2019; 2621-2629.
- [77] Yang Z, He X, Gao J, et al. Stacked attention networks for image question answering [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016; 21-29.
- [78] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018; 6077-6086.
- [79] Li L, Gan Z, Cheng Y, et al. Relation-aware graph attention network for visual question answering [C]//Proceedings of the IEEE International Conference on Computer Vision, 2019; 10312-10321.
- [80] Shrestha R, Kafle K, Kanan C. Answer them all! toward universal visual question answering models [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019; 10472-10481.
- [81] Huang Q, Wei J, Cai Y, et al. Aligned dual channel graph convolutional network for visual question answering [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020; 7166-7176.
- [82] Yu F, Tang J, Yin W, et al. ERNIE-ViL: knowledge enhanced vision-language representations through scene graphs [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021; 3208-3216.
- [83] Degottex G, Kane J, Drugman T, et al. COVAREP: a collaborative voice analysis repository for speech technologies [C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2014; 960-964.
- [84] Baevski A, Zhou Y, Mohamed A, et al. Wav2vec 2.0: a framework for self-supervised learning of speech representations [C]//Proceedings of the Advances in Neural Information Processing Systems, 2020.
- [85] Liu J, Zhu X, Liu F, et al. OPT: omni-perception pre-trainer for cross-modal understanding and generation [J]. arXiv preprint arXiv:2107.00249, 2017.
- [86] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016; 770-778.
- [87] Nguyen D K, Okatani T. Improved fusion of visual and language representations by dense symmetric Co-attention for visual question answering [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018; 6087-6096.
- [88] Ren S, He K, Girshick R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans Pattern Anal Mach Intell. 2017; 39(6): 1137-1149.
- [89] Khademi M. Multimodal neural graph memory networks for visual question answering [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020; 7177-7188.
- [90] Alberti C, Ling J, Collins M, et al. Fusion of detected objects in text for visual question answering [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020; 2131-2140.

- [91] Tan H, Bansal M. LXMERT: learning cross-modality encoder representations from transformers[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019: 5099-5110.
- [92] Sun S, Chen Y, Li L, et al. LightningDOT: pre-training visual-semantic embeddings for real-time image-text retrieval[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 982-997.
- [93] Xu H, Yan M, Li C, et al. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021: 503-513.
- [94] Sun C, Myers A, Vondrick C, et al. VideoBERT: a joint model for video and language representation learning[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 7463-7472.
- [95] Li G, Duan N, Fang Y, et al. Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 11336-11344.
- [96] Li X, Yin X, Li C, et al. Oscar: object-semantics aligned pre-training for vision-language tasks[C]//Proceedings of the European Conference on Computer Vision, 2020: 121-137.
- [97] Su W, Zhu X, Cao Y, et al. VL-BERT: pre-training of generic visual-linguistic representations[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.
- [98] Huang H, Su L, Qi D, et al. M3P: learning universal representations via multitask multilingual multimodal pre-training[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2021: 3977-3986.
- [99] Zhu L, Yang Y. ActBERT: learning global-local video-text representations[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020: 8743-8752.
- [100] Lu J, Batra D, Parikh D, et al. ViLBERT: pre-training task-agnostic visiolinguistic representations for vision-and-language tasks[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019: 13-23.
- [101] Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.
- [102] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proceedings of the 34th Conference on Neural Information Processing Systems, 2020.
- [103] Rahman W, Hasan M K, Lee S, et al. Integrating multimodal information in large pretrained transformers[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 2359-2369.
- [104] Yang Z, Dai Z, Yang Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019: 5754-5764.
- [105] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]//Proceedings of the 3rd International Conference on Learning Representations, 2015.
- [106] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5998-6008.
- [107] Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016: 289-297.
- [108] Nam H, Ha J, Kim J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2156-2164.
- [109] Yu Z, Yu J, Cui Y, et al. Deep modular co-attention networks for visual question answering[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019: 6281-6290.
- [110] Tsai Y H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6558-6569.
- [111] Sahay S, Okur E, Kumar S H, et al. Low rank fusion based transformers for multimodal sequences[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 29-34.
- [112] Yin Y, Meng F, Su J, et al. A novel graph-based multi-modal fusion encoder for neural machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3025-3035.
- [113] Hu J, Liu Y, Zhao J, et al. MMGCN: multimodal

- fusion via deep graph convolution network for emotion recognition in conversation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021; 5666-5675.
- [114] Kim J H, Jun J, Zhang B T. Bilinear attention networks[C]//Proceedings of the 32rd Conference on Neural Information Processing Systems, 2018; 1571-1581.
- [115] Yang J, Wang Y, Yi R, et al. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences [C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021; 1009-1021.
- [116] Zhao J, Li R, Jin Q. Missing modality imagination network for emotion recognition with uncertain missing modalities [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021; 2608-2618.
- [117] Lewis M, Liu Y, Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020; 7871-7880.
- [118] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. arXiv preprint arXiv:1910.10683, 2019.
- [119] Huang P Y, Patrick M, Hu J, et al. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021; 2443-2459.
- [120] Xu H, Ghosh G, Huang P Y, et al. VLM: task-agnostic video-language model pre-training for video understanding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021; 4227-4239.
- [121] Li Y, Pan Y, Yao T, et al. Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network [C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021; 8518-8526.
- [122] Li W, Gao C, Niu G, et al. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021; 2592-2607.
- [123] Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019; 4996-5001.
- [124] Conneau A, Wu S, Li H, et al. Emerging cross-lingual structure in pretrained language models[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020; 6022-6034.
- [125] Li L H, You H, Wang Z, et al. Unsupervised vision-and-language pre-training without parallel images and captions[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021; 5339-5350.
- [126] Wang M, Qi G, Wang H, et al. Richpedia: a comprehensive multi-modal knowledge graph [C]//Proceedings of the Joint International Semantic Technology Conference. Springer, Cham, 2019; 130-145.
- [127] 郑秋硕,漆桂林,王萌. 多模态知识图谱[EB/OL]. <https://zhuanlan.zhihu.com/p/163278672>. [2020-07-26].



吴友政(1976—),博士,主要研究领域为自然语言处理、人机对话和知识图谱等。  
E-mail: wuyouzheng1@jd.com



姚霆(1982—),主要研究领域为图像理解、视频分析和 3D 视觉等。  
E-mail: tingyao.ustc@gmail.com



李浩然(1990—),博士,主要研究领域为自然语言处理、文本生成等。  
E-mail: lihaoran24@jd.com