计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200086

跨媒体分析与推理技术研究综述

王树徽　闫　旭　黄庆明
中国科学院计算技术研究所　北京 100190
(wangshuhui@ict.ac.cn)

摘　要　当前,以网络数据为代表的跨媒体数据呈现爆炸式增长的趋势,呈现出了跨模态、跨数据源的复杂关联及动态演化特性,跨媒体分析与推理技术针对多模态信息理解、交互、内容管理等需求,通过构建跨模态、跨平台的语义贯通与统一表征机制,进一步实现分析和推理以及对复杂认知目标的不断逼近,建立语义层级的逻辑推理机制,最终实现跨媒体类人智能推理。文中对跨媒体分析推理技术的研究背景和发展历史进行概述,归纳总结视觉-语言关联等任务的关键技术,并对研究应用进行举例。基于已有结论,分析目前跨媒体分析领域所面临的关键问题,最后探讨未来的发展趋势。

关键词:跨媒体分析与推理;深度学习;多模态融合;视觉-语言分析
中图法分类号　TP181

Overview of Research on Cross-media Analysis and Reasoning Technology

WANG Shu-hui, YAN Xu and HUANG Qing-ming
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract　Cross-media presents complex correlation characteristics across modalities and data sources. Cross-media analysis and reasoning technology is aimed at multimodal information understanding and interaction tasks. Through the construction of cross-modal and cross-platform semantic transformation mechanisms, as well as further question-and-answer interactions, it is constantly approaching complex cognitive goals and modeling high-level cross the logical reasoning process of modal information, finally multimodal artificial intelligence is realized. This paper summarizes the research background and development history of cross-media analysis and reasoning technology, and summarizes the key technologies of cross-modal tasks involving vision and language. Based on the existing research, this paper analyzes the existing problems in the field of multimedia analysis, and finally discusses the future development trend.

Keywords　Cross-media analysis and reasoning, Deep learning, Multi-modal fusion, Visual-and-language analysis

1　引言

随着互联网及媒体技术的不断普及,以网络内容为代表的媒体内容数据逐渐呈现跨模态、跨数据源的复杂关联与协同动态演化特性。如图1所示,以“新冠疫情”主题为例,不同平台、不同来源的文本、图像、视频、音频等信息共同刻画相同或相关的主题内容,呈现复杂、多层级的语义关联关系。在物理空间中,信息技术与传统行业的不断融合也促成了不同模态、不同来源但具有复杂相关性的多源异构数据和信息的爆炸式增长。例如,在城市环境下,各种各样的摄像头及环境传感器,对物理世界中同一个体或场景进行协同感知和记录。网络空间与物理空间的不同来源、不同模态的数据,以多个角度共同刻画了相同或相关的主题和事件,形成了“跨媒体”信息。

图1　跨媒体信息示意图

Fig. 1　Schematic diagram of cross-media information

与传统多媒体[1]数据相比,跨媒体信息呈现出了迥然不同的特点。首先,包含不同模态的多媒体数据之间呈现出内

到稿日期:2021-01-20　返修日期:2021-02-09
基金项目:科技部重点研究发展计划项目(2018AAA0102003);国家自然科学基金项目(62022083,61672497);中国科学院前沿科学重点研究项目(QYZDJ-SSW-SYS013)
This work was supported by the National Key R&D Program of China(2018AAA0102003), National Natural Science Foundation of China(62022083,61672497) and Key Research Program of Frontier Sciences of CAS (QYZDJ-SSW-SYS013).
通信作者:黄庆明(qmhuang@ucas.ac.cn)

蕴同步的语义关联，而跨媒体的不同来源、不同模态的信息呈现动态、复杂、多层次的时空、语义关联。其次，跨媒体形式异构、内容多样、分布复杂，传统的分析处理方法大多基于独立同分布等假设，难以对海量复杂的跨媒体信息进行有效利用和模型学习。最后，跨媒体涉及的应用场景比多媒体更加广泛，如有害网络内容监测与管理、跨媒体内容搜索、推荐、问答等[2]。跨媒体呈现的上述特点对跨媒体分析与推理技术提出了迫切的需求。

借助强大的脑功能，人类对不同模态的信息进行符号化转换和统一表征，进而在符号表示的基础上实现推理与决策，具有天然的跨媒体综合处理能力。类似于人类大脑，实现海量、复杂、异构的跨媒体语义贯通与统一表征是人工智能系统能够有效处理跨媒体信息的先决条件。首先，不同媒体信息的统一表征与关联度量，是实现跨媒体分析与推理的基础。在统一表征与度量的基础上，实现跨媒体内容的理解与转换，是提升跨媒体语义贯通水平的重要方式。其次，在跨媒体内容理解的基础上实现跨媒体推理与决策，是跨媒体类人智能发展必须解决的关键技术问题。跨媒体分析推理技术的发展，对实际应用中的问题也提供了更多的关键技术支撑。

本文第2节详细介绍了跨媒体分析领域的关键技术，包括跨媒体统一表征、跨媒体理解与内容转换生成、跨媒体推理与决策等；第3节介绍了跨媒体深度学习技术的应用示例，包括视觉语言导航、跨模态检索和基于知识图谱的视觉问答系统；第4节总结全文，分析目前跨媒体分析领域存在的主要挑战，并对跨媒体分析与推理技术的未来发展趋势进行总结与展望。

2 跨媒体分析推理技术研究框架

跨媒体信息包含不同的模态(Modality)信息，如图像、视频、文本、语音等。多模态深度学习(Multimodal Deep Learning)[3]通过深度学习实现对多个模态信息的统一表征、转换及深层理解，是跨媒体分析推理任务涉及到的基础技术。人工智能的目的是让机器实现类人智能，因此让机器具有像人一样处理跨媒体信息的能力，是人工智能领域中重要的发展方向之一。其中，涉及到图像、视频和文本的图文理解任务是跨媒体分析领域主要的研究方向，旨在用文字辅助对视觉内容的理解，或以视觉内容刻画文字所表达的语义。跨媒体分析推理任务层次分析如图2所示。

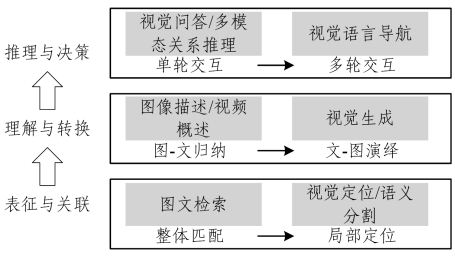

图2 多模态任务层次分析

Fig. 2 Hierarchical analysis of multimedia tasks

跨媒体分析与推理包含表征、理解与推理3个科学问题。表征即对相互关联但存在语义鸿沟的多模态信息进行统一(可度量)表征，可分为整体匹配与局部定位，对应的任务有图文检索、视觉定位、语义分割等；理解指在统一表征的基础上，对具有互补性的多个模态信息进行高层次相互转换，实现跨模态信息的融合与理解，代表性技术包括图像/视频描述、文本-图像/视频生成等；在理解的基础上，进一步完成跨媒体推理与决策，代表性研究任务与技术包括视觉问答、视觉语言导航等。

2.1 跨媒体统一表征

跨媒体统一表征任务的信息由两个及以上模态组成，旨在将多模态信息表示为计算机可以处理的数值向量或更高维的特征向量，利用多模态信息间的语义一致性，剔除模态间的冗余信息，通过跨模态相互转化来实现多模态融合的统一表征，以学习更全面的特征表示。使用该特征表示可解决不同模态间的语义鸿沟问题，为完成跨模态检索、图像语义分割等任务奠定基础。具体地，跨模态检索任务指将一种模态的数据作为查询去检索另一模态对应数据，主流的方法有实值表示学习和二值表示学习。实值表示学习直接学习不同模态之间的信息，二值表示学习将多模态信息映射到汉明二值空间后再拟合。图像分割也是一个典型的多模态信息表征任务，其将文本分配给图像中某个区域的局部定位任务。最典型的版本是语义分割，即每个像素都被分类到唯一语义实体。相应地，可以将文本中提到的多个语义实例分类到一起，对场景中最重要的对象进行分割定位，如显著性检测、文本视觉定位等。

跨媒体统一表征任务主要解决的问题是模态异构性，主要包含两大研究方向，即联合表征和协同表征，如图3所示。联合表征的主要思想是将多模态信息分别嵌入公共空间后，计算任意一对样本距离；而协同表征的核心是分别对多模态信息做嵌入操作，学习特征之间的相似度度量。其中，特征表示具有多个基元层级，可根据尺度大小分为单词、短语及句子级别。图像中的物体类别、区域及对应属性可由单词进行描述，细粒度物体、视频中的行为可由词语描述，而视频中的某个具体事件或整幅图像的内容则需要用句子表征。

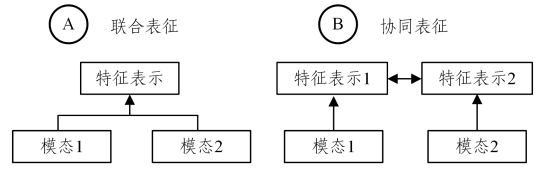

图3 跨媒体表示学习

Fig. 3 Cross-media representation learning

典型相关性分析(Canonical Correlation Analysis, CCA)[4]是联合表征的基本算法，分别学习图片表示和文字表示映射的子空间，通过最大化两种模态数据的投影相关性调节关联关系，以期学习同构空间。核典型相关性分析(Kernel Canonical Correlation Analysis, KCCA)[5]在CCA的基础上，使用较低阶的核函数提高模型的描述能力。而多视角分析方法(Generalized Multiview Analysis, GMA)[6]则通过拓展线性判别分析和边缘Fisher分析，引入语义信息丰富的标签。此外，还可以通过构造非参数映射[7]和线性交叉模态投影[8]拓展CCA。随着深度学习的发展，视觉语义嵌入的方法成为主流。Wang等使用双分支神经网络[9]学习跨模态信息联合嵌

入；Liu 等构建了基于循环残差融合模块（Recurrent Residual Fusion，RRF）[10]的公共空间；Andrew 等将传统 CCA 拓展为深度典型相关分析（Deep Canonical Correlation Analysis，DCCA）[11]，利用深度神经网络学习更具判别性的语义表征，并用不同尺度网络完成相关性分析。

协同表征使用成对匹配方法，利用多模态数据之间的相似性度量反映匹配程度。Wu 等使用图像-文本和文本-图像双向语义相似性学习双线性相似度[12]；Karpathy 等提出深度视觉语义对齐算法（Deep Visual-Semantic Alignments，DVSA）[13]，使用 R-CNN（Region Convolutional Neural Networks）检测和编码图像区域，计算所有可能的图像区域-单词对相似性分数；Ma 等提出多模态卷积神经网络（Multimodal-CNN，m-CNN）[14]，使用不同的句子片段与多层次图像特征交互，削弱了单词对句子的依赖；Huang 等提出的 SCO 模型[15]首次使用多区域多标签的卷积神经网络预测图像语义特征，同时使用自注意力机制的长短期记忆网络显式地学习语义顺序，实现了匹配和生成的联合学习；而 Wang 等用注意力机制学习单词和词语间的语义关系，提出了联合全局共注意力表示学习方法[16]；Wu 等通过引入生成对抗网络的思想，使用对抗训练方式实现语义一致的跨模态数据对齐[17]。

2.2 图文转换

图文转换也可以称为图文映射，负责将一个模态的信息转换至另一模态，常见的应用包括图像视频概述（基于输入图像或视频，输出描述该视觉内容的文本）、文本生成图像（基于文本内容生成对应语义的图像）等。图-文总结任务的主要框架为自注意力机制的编码器-解码器模型，编码器将变长输入序列转换为背景向量后输入到解码器中，解码器在不同时间步参考不同隐变量的自注意力权重生成另一模态向量。文-图演绎任务的常用模型为对抗生成网络，由生成器和判别器组成，生成器生成合理样本，判别器判断该样本是否合理。

模态转换主要有两大难点：1）未知结束位，如实时的机器翻译任务在翻译时的句尾信息未知性；2）主观评判性，即对于目前的模态转换任务，暂时没有客观且全面的评价标准，因此进行模型之间的性能比较具有较大困难。

传统的图像视频描述技术是基于人工设计的语言模板。其首先使用物体检测方法识别视觉物体和动作，然后对复杂词汇进行统计，确定与视觉内容对应的文本词汇后，将其填入预先设计好的语言模板中，生成自然语言文本。与此类方法相比，基于深度神经网络的视频描述任务通用模型（见图 4）可实现端到端学习。2015 年，Vinyals 等使用基于卷积神经网络和循环神经网络的模型[18]来解决图像描述任务；2015 年，Venugopalan 等使用等间隔采样方法[19]，将多个基于帧的特征向量转化为单个基于视频的聚合向量，以完成视频描述任务；Li 等在预训练的活动识别数据集的基础上提出了三维卷积神经网络结构[20]，对各种复杂的动态特征进行建模，实现变长输入（视频）至变长输出（单词或句子）的映射；随后，Cornia 等提出针对图像中某个感兴趣的局部区域进行标题生成任务[21]；Yin 等使用三路长短期记忆网络[22]，同时关注目标特征、全局特征和区域特征，以获取更多先验信息；Zheng 等针对图像中某个未知的特定物体[23]，使用物体检测模型和两层长短期记忆网络生成详细描述该物体的文本。

然而，一句话并不能对视频内容进行详尽的描述，因此生成多句文本描述任务应运而生。2017 年，斯坦福大学李飞飞团队定义了密集视频字幕任务[24]，即对一个视频生成段落式描述文本。其基础模型首先使用提案网络对长视频中的多个事件进行检测[25]，然后按时序逐个生成描述文本。2018 年，第一个使用 Transformer 代替 LSTM 的模型被提出[26]，解决了长时间依赖问题，它是后续几个相关变体模型的基础。2019 年，使用强化学习解决密集视频字幕任务的 SDVC 模型[27]被提出，该模型通过加强生成单词与视觉内容和上下文单词之间的关联，解决了生成字幕冗余问题；Yu 等提出了基于 GAN 框架的 SeqGAN 模型[28]，构建生成式对抗网络框架来解决序列生成问题，通过引入判别器提高了标题生成质量；Chen 等采用监督学习和强化学习相结合的方法训练网络参数[29]，采用梯度策略算法设计视觉一致性与文本多样性的奖励机制，提高了描述文本的独特性。

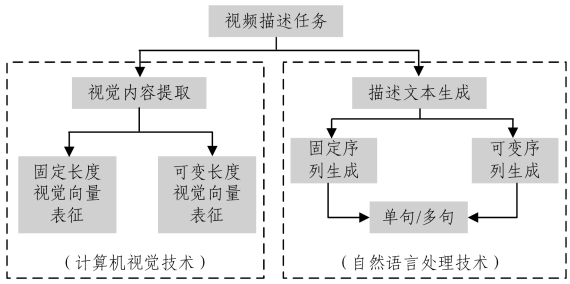

图 4 基于深度神经网络的视频描述任务通用模型

Fig. 4 Deep neural network-based general video captioning models

近年来，图像标题生成领域也出现了一些新颖的工作。Guo 等提出了生成浪漫、幽默、消极等多风格描述文本任务[30]；Shuster 等融合了情感任务，试图让生成的图像标题更具有不可重复性[31]；Xu 等对图像字幕模型进行了对抗攻击研究[32]，通过在输入图像上添加少量噪声，让模型生成指定文本，他们还研究了新闻图像的标题生成任务，解决了生成标题中包含未在训练集中出现过的命名实体的文本不准确的问题。另外，为了解决图文转换中评价指标的主观性问题，Dogin 等提出了 SemanticScore 指标[33]，使用图像与文本向量分别经过 CCA 和 HKSE 空间映射后计算所得的余弦相似度来拟合人类评判标准。

根据文本生成图像是近几年的热门研究领域，该任务旨在生成与描述文本内容对应的图片，主流方法有变分自编码器（Variational Auto-Encoder，VAE）、聚焦机制（Deep Recurrent Attention Writer，DRAW）、生成对抗网络（Generative Adversarial Networks，GAN）等。在 2016 年以前，常用 VAE 和 DRAW 解决图像生成任务，VAE 使用基于统计的方法对数据先验分布和隐变量表示进行建模；DRAW 使用循环神经网络及自注意力机制，按序关注生成对象后叠加得到最终结果。自 Reed 等提出 GAN-INT-CLS[34]以来，使用 GAN 的思想解决从文本到图像的任务成为主流，Reed 等在此基础上又提出了 GAWWN 模型[35]，添加了标定框与关键点以提升图片精度。StackGAN[36]则使用两个 GAN 实现图像的分步生成，首先生成只关注图像背景、颜色及轮廓等基本信息的低精

度图片，然后补充细节信息得到高精度图像；在后续工作中，StackGAN++[37]将 GAN 扩充为树状结构，多个生成器与鉴别器并行训练得到不同精度的图像，逐步提取更精细的文本信息，以解决非限定性生成任务；之后，Xu 等提出 AttnGAN[38]，增加自注意力机制，将全局约束的文本特征精确到单词级别，提升了图像中文本细节的生成质量。除了基于 GAN 增强的方法外，还有基于迁移学习的方法，旨在提升生成器的生成能力，如 GAN-CLS，MirroGAN 等，以及为了提升语言数据表示能力的基于数据增强的 RiFeGAN，Effcient-NET 等。

2.3 图文推理与交互

人工智能的参照样本始终没有离开过人类本身，而可解释的推理学习是人类最重要的能力之一。目前多模态分析研究仍处于表观特征提取与处理阶段，缺乏对复杂推理特征及高层次理解任务的建模，无法完成跨模态信息间的推理交互，因此需要更丰富的向量、复杂的网络来细化多模态信息关联和决策过程，以解决单轮甚至多轮的交互任务，包括抽象推理、视觉问答、视觉对话、视觉语言导航等。其中，视觉推理任务要求在理解文本的基础上结合图片信息进行推理[39]，可分为两个子任务：$\{Q->A\}$根据问题选择答案[40]，$\{QA->R\}$根据问题和答案进行推理。视觉问答任务在视觉推理的基础上，通过引入外部知识库增加回答准确率和泛化性。2018 年，Wang 等对视觉问答任务进行了拓展，提出了视觉对话任务[41]。该任务需要从视觉和文本上下文中提取隐藏信息，实现智能体与人类之间使用自然语言进行多轮有意义的对话交互。视觉语言导航任务通过与机器人领域的技术相结合，要求智能体按照自然语言文本在环境中进行导航。这不仅需要机器同时理解自然语言指令与视角中可见图像的信息，还需要在环境中完成状态评估与决策，通过一系列决策抵达指定位置。

传统的视觉问答模型基于贝叶斯框架的统计学方法，按照训练集中答案出现的统计频率生成候选答案集，将每个答案作为类别标签，将生成问题转化为分类问题进行处理。深度学习的多模态视觉问答模型使用联合嵌入方法，在多模态双线性框架上，通过多模态知识融合技术实现基于视觉的文本问答。常见的融合方法有多模态特征向量相加、利用特征向量做外积融合的双线性汇合匹配、按元素乘的多模态低秩双线性注意力网络[42]、多模态高阶因式分解池化方法[43]等。Han 等将隐式推理过程建模为潜在空间的序列决策过程，为可视化推理过程提供了可解释性[44]。另外，使用协同注意力机制提供图像-文本的双相关特征信息，保证问题和图像的一致性，也是视觉问答任务的一类重要解决方案。

目前，视觉问答任务主要解决的问题已从多模态匹配上升为视觉推理，典型的基于知识图谱的视觉问答模型如下。Wang 等提出了 Ahab 方法[45]，其将提取出的图像信息整合为三元组，从知识图谱中查询与图像信息相关的所有三元组后整合为推理链，实现对答案的推理；Wu 等通过增加数据源来提升模型的性能[46]；Wang 等提出了 FVQA 数据集[47]，通过引入大规模知识数据库，来回答开放性问题。

3 应用举例

3.1 视觉语言导航

视觉语言导航任务是一个典型的跨媒体分析任务，要求智能体感知环境，理解和落实文本内容，在视觉信息的辅助下到达自然语言指定的目标位置[48]。视觉语言导航任务的基本流程如图 5 所示。

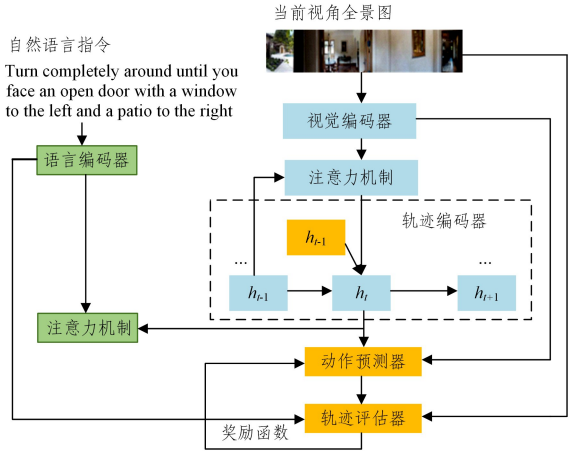

图 5 视觉语言导航任务基本流程图

Fig. 5 Workflow of visual language navigation

在深度学习出现之前，基于图的方法已经得到了广泛研究，这些方法将导航任务明确地分解为一系列子任务，如映射、定位和动作控制，主要从文本的句法、语义信息，图片的空间特征、实体关系推理等角度入手。虽然这些方法获得了一定程度的成功，但分布式设计决定了这类方法易受噪声影响，无论哪个部分被扰动，最终都会传递至控制器，影响整体结果，因此这类算法的鲁棒性较差。同时，这类算法还要求海量数据作为驱动，并且需要人工修正，这极大地限制了模型的泛化能力。

受到深度学习快速发展的影响，基于学习的方法被应用于视觉语言导航任务并取得了较大进展。该类方法自动将输入图片和文本映射为序列数据，不需要进行人工特征选取，在保证效果的前提下极大地提升了效率和泛化能力。目前基于深度学习的视觉语言导航模型可分为监督学习方法和强化学习方法两类。其中，使用监督学习方法的模型，通过最大化预测动作和真实轨迹之间的相似度进行训练；而使用强化学习方法的模型，则通过最大化累积目标奖励进行训练。基于监督学习的方法大多要求大量的训练数据，需要人工生成大量的最优轨迹，同时忽视了环境噪音的可控性影响，而基于强化学习的方法则不需要使用轨迹标签数据，但需要较长的训练时间。

对于跨媒体分析与推理任务，需要对多模态信息使用编码器将其映射至特征空间，使用自注意力机制计算不同时间步的跨媒体加权特征向量，而解码器模块在视觉语言导航中对应动作预测器，实现对前进方向的评估与预测。为保证决策动作与视觉和文本信息的一致性，可以使用强化学习训练轨迹评估器，为前进方向的决策过程提供上下文信息。其中，从状态到动作的映射过程可由基于模型的强化学习方法直接

学习映射关系，也可通过基于模型的强化学习方法对环境进行建模。

最早，RAP 模型[49]提出了一种杂交的强化学习模型，将两种方法预测的动作分布都输入到动作预测器中，提升了预测结果的准确率。Speaker-follower 模型[50]则使用扬声器模块对给定路径生成新的指令，使用数据增强方法解决有标签数据不足的问题。Wang 等提出基于强化学习的跨媒体匹配模型[51]，使用强化学习方法匹配局部与全局信息，同时使用模拟学习通过历史已知信息适应新环境，以解决模型泛化能力差的问题。类似地，EnvDropout 模型[52]通过丢弃环境中的部分物体构造新的环境，并在新的环境中生成路径信息，使用半监督学习方法提高模型的泛化性能。2019 年，较多研究考虑引入大量辅助信息来帮助最终动作决策。Ma 等提出在进行动作选择时，不仅要考虑文本和视觉的上下文环境信息，还需要计算轨迹进度信息[53]，而 Zhu 等在此基础上还引入了角度预测、轨迹重构等丰富的已知信息来辅助训练[54]。综上所述，视觉语言导航任务除了要解决跨媒体的语义鸿沟难题外，在处理复杂的语言指令、不完全可知的视觉信息的基础上，还需要完成复杂导航决策。目前，强化学习方法还存在虚拟环境与真实环境差距较大、模型不稳定、模型泛化能力差等问题，但基于强化学习的视觉语言导航任务仍是一个极具发展潜力的跨媒体分析推理研究任务。

3.2 跨模态检索

图文匹配和检索是多模态分析的基本任务，目标是学习一种多模态的相似性度量，对于给定的查询词，返回另一模态最相似的样本，该任务可分为全局匹配与局部检索两大类。跨模态检索任务的难点主要有不同模态特征具有异构性、底层内容和高层语义之间存在语义鸿沟、模态间信息不对齐等。

目前常用的跨模态检索方法可分为 3 类：典型相关性分析、视觉语义嵌入和基于 BERT(Bidirectional Encoder Representation from Transformers)[55]的预训练模型。前两类方法是从多模态表示学习延伸出的解决方案，基本流程图如图 6 所示。第一类方法的基本思想是将不同模态的特征投影至同一个公共子空间，度量不同模态间的相似性，实现跨模态检索；第二类方法从底层抽取不同模态的有效表示，在高层建立语义关联。随着 BERT 在文本领域的广泛应用，基于 BERT 的预训练模型也被用于解决跨模态检索任务，该类方法包括单流和双流两种模型。代表性的单流模型有 VideoBERT[56]，VisualBERT[57]，VLBERT[58]。VideoBERT[56]将视频信息注入预训练语言模型进行训练，在视频动作分类、视频字幕等任务上都取得了较好结果；VisualBERT[57]将输入文本中的单词与输入图像中的局部区域进行隐式对齐，实现局部匹配；VLBERT[58]将图像中感兴趣区域和文本中单词的嵌入特征同时作为输入，可捕捉更细节的视觉线索。ViLBERT[59]是典型的双流网络模型，使用两个 BERT 流分别预处理视觉、文本输入，并在 Transformer 层中进行交互，实现特征的相互提取与优化；LXMERT[60]则使用关系对象编码器、语言编码器和多模态编码器进行多任务预训练。

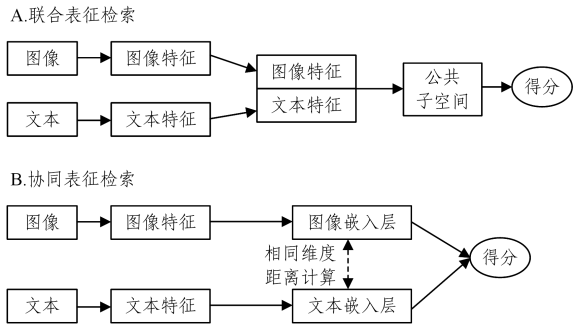

图 6 两种跨模态检索的流程图

Fig. 6 Workflow of two types of cross-modal retrieval

注意力是一个常见但很容易被忽视的事实，人的视觉皮层在接受新的图像输入时，很自然地只关注重点区域而忽略无关信息[61]，因此在多模态的信息匹配中，局部定位比全局匹配具有更广泛的应用范围。同时，人们常使用指示性表达，如“红色的车”“桌子上的电脑”等，因此准确地理解指示性表达的内容定位技术对多模态分析的发展至关重要。内容定位技术旨在给定一张图像和一个细粒度文本描述，准确定位到图像中的对应对象或区域，实现局部匹配[62]。

内容定位技术的基本流程图如图 7 所示。首先对于输入的搜索图像使用预训练的区域生成网络，从而获取一系列候选区域。其次，对于每一个图像区域，使用预训练的卷积神经网络获取视觉表观特征，同时，根据区域坐标计算区域位置特征，与视觉表观特征组合构成区域的视觉特征[63]。通过预训练的词向量模型将输入的查询文本映射为向量表示。在训练阶段，使用双向长短期记忆网络对语言特征进行学习和高维特征表示，并将其映射至视觉特征空间以计算语义距离，最后使用特征匹配及排序算法获得搜索结果。

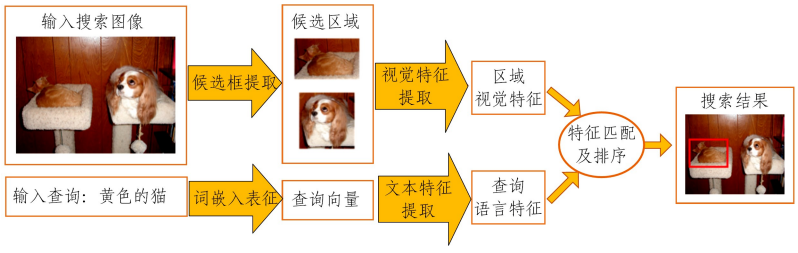

图 7 内容定位的流程图

Fig. 7 Workflow of grounded video description

3.3 基于知识图谱的视觉问答系统

视觉问答系统是让计算机根据视觉信息回答用户所提出的问题，是跨媒体内容服务的一种高级形式。不同于现有的搜索引擎，问答系统返回的不再是基于关键词匹配的相关排

序，而是精确的自然语言形式的答案信息。基于深度学习的视觉问答模型大多为多模态双线性框架，此类方法的主要处理过程有：提取图像特征、处理自然语言文本、融合两模态知识特征向量、使用分类网络得到最终答案。但这类方法不能准确地回答推理性问题，随着知识图谱的不断开放，基于知识图谱的视觉问答系统逐步发展起来。

知识图谱本质上是一种语义网络，由结点和边组成，其结点代表实体或概念，边代表实体之间的各种语义关系，是基于现有数据再加工的一种表示形式[64]。多模态知识图谱利用知识库中已有的知识储备进行多模态信息融合及推理，可用于支撑下一代搜索、推荐和在线广告业务。

大规模知识图谱数据中存储着大量多模态、多层次语义内涵的知识信息。对于非结构化输入（如图像特征），神经网络模型通过多模态表示学习进行概念的结构化构建，将知识图谱中的实体、类别及关系等内容嵌入为数值向量。对于自然语言文本输入，神经网络模型将文本解析映射为相应的低维向量，与知识图谱中的实体、概念、关系等对应连接，使用向量间的相似度计算生成知识图谱中节点和边构成的事实。通过知识嵌入向量和概念元组组合的形式完成跨媒体知识图谱构建，一方面可对单一模态的信息进行准确表征，另一方面实现了多模态相同语义信息的统一嵌入表征。另外，由于用户问题的复杂性，以及跨媒体知识图谱内容的丰富性，在推理过程中可能出现调用多个知识库的情况，此时可使用异构知识关联与对齐技术来解决[65]。

同时，通过问答系统不断产生人机协同交互，模仿人类的自然学习模式，经过自主生成问题[66]-交互获得答案-推理补全关系的方式不断互动，可实现对知识关系的推理补全，从而实现增量式的基于知识图谱的视觉问答人机交互系统（见图 8）。

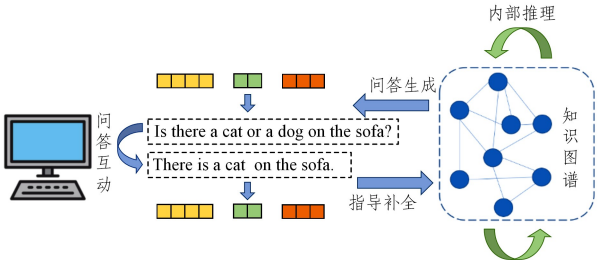

图 8 基于知识图谱的视觉问答人机交互示意图

Fig. 8 Diagram of human-computer interaction for visual question answering based on knowledge graph

4 挑战与展望

4.1 主要挑战

虽然跨媒体分析推理目前已经取得了一定的进展，但仍存在一定的局限性：

（1）模型的处理准确率较低。虽然计算机视觉、自然语言处理和语音识别等领域的发展促进了跨媒体分析任务准确率的不断提高，但距离实现高水平人工智能还有很大差距。

（2）模型的推理能力较弱。现有模型实现了多模态信息在同一语义空间的映射，但缺乏高层逻辑推理能力，无法实现对未知信息的预测。

（3）深度学习具有不可解释性。尽管深度学习被广泛应用，但其内在机理的难以解释是亟待解决的难题，这一不足也限制了下游任务的鲁棒性、可信度与性能提升。

4.2 未来展望

基于深度学习的跨媒体分析与推理技术虽然取得了一定的进展，但还未达到人类的预期水平，在未来还可从以下几个方面对该任务进行深入探索：

（1）获取跨媒体信息更全面的高维序列表征，对声音、文本、图像特征使用更合理的融合方式进行表征。

（2）进行模型与技术的创新，解决现有方法固有的局限性，尝试添加或替换模块。

（3）对于描述生成任务，重点提升文本信息的语义准确性和视觉一致性，尤其是长视频中多事件的顺序、联系，以进行更详尽的表达。可参考自然语言处理领域的前沿方法，来提升文本质量，从而提升跨媒体分析与推理任务的性能。

参 考 文 献

[1] SRIVASTAVA N, RUSLAN S. Multimodal learning with deep boltzmann machines[J]. The Journal of Machine Learning Research, 2014, 15(1): 2949-2980.

[2] ATREY P K, HOSSAIN M A, SADDIK A E, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia Systems, 2010, 16(6): 345-379.

[3] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[4] HOTELLING H. Relations Between Two Sets of Variates[J]. Biometrika, 1935, 28: 321-377.

[5] SHAWE-TAYLOR J, CRISTIANINI N. Kernel Methods for Pattern Analysis[M]. Taylor & Francis Group, 2004.

[6] SHARMAA, KUMAR A, DAUME H, et al. Generalized multiview analysis: A discriminative latent space[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012: 2160-2167.

[7] SONG G L, WANG S H, HUANG Q M, et al. Multimodal Similarity Gaussian Process Latent Variable Model[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4168-4181.

[8] YAN H, WANG S, LIU S, et al. Cross-modal correlation learning by adaptive hierarchical semantic aggregation[J]. IEEE Transactions on Multimedia, 2016, 18(6): 1201-1216.

[9] WANG L, LI Y, LAZEBNIK S. Learning deep structure-preserving image-text embeddings[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5005-5013.

[10] WANG L, LI Y, SVETLANA L. Learning a recurrent residual fusion network for multimodal matching[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 4107-4116.

[11] ANDREW G,RAMAN A,JEFF B,et al. Deep canonical correlation analysis[C]// International Conference on Machine Learning. 2013:1247-1255.

[12] WU Y L,WANG S H,HUANG Q M. Online asymmetric similarity learning for cross-modal retrieval[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:4269-4278.

[13] KARPATH Y,ANDRE J,FEI-FEI L. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3128-3137.

[14] MA L,LU Z,SHANG L. Multimodal convolutional neural networks for matching image and sentence[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:2623-2631.

[15] HUANG Y,WU Q,WANG W,et al. Image and sentence matching via semantic concepts and order learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(3):636-650.

[16] WANG S H,CHEN Y Y,ZUO J B,et al. Joint global and co-attentive representation learning for image-sentence retrieval [C]// Proceedings of the 26th ACM international conference on Multimedia. 2018:1398-1406.

[17] WU Y,WANG S,SONG G,et al. Augmented Adversarial Training for Cross-modal Retrieval[J]. IEEE Transactions on Multimedia,2021,23:559-571.

[18] VINYALS O,TOSHEV A,BENGIO S,et al. Show and tell:A neural image caption generator[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3156-3164.

[19] VENUGOPALAN S,XU H,DONAHUE J,et al. Translating Videos to Natural Language Using Deep Recurrent Neural Networks[J]. Human Language Technologies,arXiv:1412.4729,2015.

[20] YAO L,TORABI A,CHO K,et al. Describing videos by exploiting temporal structure[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:4507-4515.

[21] CORNIA,MARCELLA,LORENZO B. Show,control and tell:A framework for generating controllable and grounded captions [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:8307-8316.

[22] YIN G,SHENG L,LIU B,et al. Context and attribute grounded dense captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:6241-6250.

[23] ZHENG Y,LI Y,WANG S. Intention oriented image captions with guiding objects[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:8395-8404.

[24] 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.

[25] QI Z B,WANG S H,SU C. Modeling Temporal Concept Receptive Field Dynamically for Untrimmed Video Analysis[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020:3798-3806.

[26] ZHOU L,ZHOU Y,CORSO J,et al. End-to-End Dense Video Captioning with Masked Transformer[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8739-8748.

[27] MUN J,YANG L,REN Z,et al. Streamlined dense video captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:6588-6597.

[28] YU L,ZHANG W,WANG J,et al. Seqgan:Sequence generative adversarial nets with policy gradient[C]// Thirty-first AAAI Conference On Artificial Intelligence. 2017:2852-2858.

[29] CHEN Y,WANG S,ZHANG W,et al. Less is more:Picking informative frames for video captioning[C]// European Conference on Computer Vision. 2018:358-373.

[30] GUO L,LIU J,YAO P,et al. Mscap:Multi-style image captioning with unpaired stylized text[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:4204-4213.

[31] SHUSTER K,HUMEAU S,HU H,et al. Engaging image captioning via personality[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:12516-12526.

[32] XU Y,WU B,SHEN F,et al. Exact adversarial attack to image captioning via structured output learning with latent variables [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:4135-4144.

[33] DOGNIN P,MELNYK I,MROUE H,et al. Adversarial semantic alignment for improved image captions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:10463-10471.

[34] REE D,SCOTT E. Generative adversarial text to image synthesis[C]// International Conference on Machine Learning. 2016:1060-1069.

[35] REED,SCOTT E. Learning what and where to draw[C]// Neural Information Processing Systems. 2016:217-225.

[36] HAN Z,XU T,HONGSHENG L. StackGAN:Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5908-5916.

[37] ZHANG H,XU T,LI H,et al. StackGAN++:Realistic Image Synthesis with Stacked Generative Adversarial Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,41(8):1947-1962.

[38] XU T,ZHANG P,HUANG Q,et al. AttnGAN:Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:1316-1324.

[39] JOHNSON J,HARIHARAN B,VAN DER MAATEN L,et al. Clevr:A diagnostic dataset for compositional language and elementary visual reasoning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2901-2910.

[40] ANTOL S,AGRAWAL A,LU J,et al. Vqa:Visual question an-

swering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:2425-2433.
[41] WU Q, WANG P, SHEN C, et al. Are you talking to me? reasoned visual dialog generation through adversarial learning [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:6106-6115.
[42] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling[C]// International Conference on Learning Representations. 2017:1-13.
[43] YU Z, YU J, XIANG C, et al. Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12):5947-5959.
[44] HAN X, WANG S, SU C, et al. Interpretable Visual Reasoning via Probabilistic Formulation Under Natural Supervision[C]// European Conference on Computer Vision. 2020:553-570.
[45] WANG P, WU Q, SHEN C, et al. Explicit Knowledge-based Reasoning for Visual Question Answering[J]. Computer Science, arXiv:1511.02570, 2015.
[46] ANDERSON P, WU Q, TENEY D, et al. Image captioning and visual question answering based on attributes and external knowledge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6):1367-1381.
[47] NARASIMHAN M, LAZEBNIK S, SCHWING A. Out of the box: Reasoning with graph convolution nets for factual visual question answering[C]// Advances in Neural Information Processing Systems. 2018:2654-2665.
[48] ANDERSON P, WU Q, TENY D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:3674-3683.
[49] WANG X, XIONG W, WANG H, et al. Look before you leap: Bridgingmodel-free and model-based reinforcement learning for planned-ahead vision-and-language navigation[C]// Proceedings of the European Conference on Computer Vision. 2018:37-53.
[50] FRIED D, HU R, CIRIK V, et al. Speaker-follower models for vision-and-language navigation[C]// Advances in Neural Information Processing Systems. 2018:3314-3325.
[51] WANG X, HUANG Q, CELIKYILMAZ A, et al. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:6629-6638.
[52] TAN H, YU L, BANSAL M. Learning to navigate unseen environments: Back translation with environmental dropout[C]// International Conference on Learning Representations. 2019.
[53] MA C Y, LU J, WU Z, et al. Self-monitoring navigation agent via auxiliary progress estimation[C]// International Conference on Learning Representations. 2019.
[54] ZHU F, ZHU Y, CHANG X, et al. Vision-language navigation with self-supervised auxiliary reasoning tasks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020:10012-10022.
[55] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805, 2018.
[56] 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. 2019:7464-7473.
[57] LI L N, YATSKARM, YIN D, et al. Visualbert: A simple and performant baseline for vision and language[J]. arXiv: 1908.03557, 2019.
[58] SU W, ZHU X, CAO Y, et al. Vl-bert: Pre-training of generic visual-linguistic representations[J]. arXiv:1908.08530, 2019.
[59] LU J, BATRA D, PARIKH D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]// Advances in Neural Information Processing Systems. 2019:13-23.
[60] TAN H, MOHIT B. Lxmert: Learning cross-modality encoder representations from transformers[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019.
[61] HOOD B M, ATKINSON J. Disengaging visual attention in the infant and adult[J]. Infant Behavior & Development, 1993, 16(4):405-422.
[62] LIU X J, LI L, WANG S H, et al. Adaptive reconstruction network for weakly supervised referring expression grounding [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019:2611-2620.
[63] LIU CX, MAO J H, SHA F, et al. Attention correctness in neural image captioning[C]// Proceedings of the Conference on Artificial Intelligence. 2017:4176-4182.
[64] JI S, PAN S, CAMBRIA E, et al. A Survey on Knowledge Graphs: Representation, Acquisition and Applications[C]// Proceedings of the Conference on Artificial Inelligence. 2020.
[65] MALINOWSKI M, FRITZ M. A multi-world approach to question answering about real-world scenes based on uncertain input [C]// Advances in Neural Information Processing Systems. 2014:1682-1690.
[66] WU Q, WANG P, SHEN C, et al. Ask me anything: Free-form visualquestion answering based on knowledge from external sources[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:4622-4630.

WANG Shu-hui, born in 1983, Ph. D, professor, Ph. D supervisor. His main research interests include cross-media understanding, multi-modal learning/reasoning and large-scale Web multimedia data mining.

HUANG Qing-ming, born in 1965, Ph.D, professor, Ph. D supervisor. His main research interests include multimedia computing, image/video processing, pattern recognition and computer vision.