多模态知识图谱构建与应用研究综述

陈 烨，周 刚†，卢记仓
（信息工程大学 数据与目标工程学院，郑州 450001）

摘 要： 为了总结前人工作，给相关研究者提供思路，首先讨论了当前多模态知识图谱的基本概念，然后从图数据库和知识图谱这两个角度介绍了多模态知识图谱的构建工作，并总结了两种主要方法的思路。还分析了多模态知识图谱的构建和应用中的关键技术和相关工作，如多模态信息提取、表示学习和实体链接。此外，列举了多模态知识图谱在四种场景中的应用，包括推荐系统、跨模态检索、人机交互和跨模态数据管理。最后，从四个方面展望了多模态知识图谱的发展前景。

关键词： 多模态；知识图谱；构建；应用
中图分类号： TP391.1 **文献标志码：** A **文章编号：** 1001-3695(2021)12-003-3535-09
doi:10.19734/j.issn.1001-3695.2021.05.0156

Survey on construction and application research for multi-modal knowledge graphs

Chen Ye, Zhou Gang†, Lu Jicang
(*Academy of Data & Target Engineering, Information Engineering University, Zhengzhou* 450001, *China*)

Abstract: In order to summarize previous works and provide ideas for related researchers. This paper first discussed the current basic concepts of the multi-modal knowledge graph. Then it introduced the construction of multi-modal knowledge graphs from two perspectives of graph database and knowledge graph, and summarized the ideas of two main methods. It also analyzed the key technologies and related works in the construction and application, such as multi-modal information extraction, representation learning, entity linking. In addition, it enumerated the application of multi-modal knowledge graphs in four scenarios, including recommender systems, cross-modal retrieval, human-computer interaction and cross-modal data management. Finally, it looked forward to the development prospects of multi-modal knowledge graphs from four aspects.

Key words: multi-modal; knowledge graph; construction; application

模态是一种生物学概念，指感官条件下事物发生或存在的方式。互联网的高速发展催生了急剧积累的海量数据，包括文本、音频、图像、视频等多种模态，如何高效地利用这些丰富的信息成为一个关键而具有挑战性的问题。Google 公司于 2012 年提出知识图谱（knowledge graph，KG）的概念[1]。知识图谱本质上是一种基于图模型的关联网络知识表达，旨在采用图的结构来建模和记录世界万物之间的关联关系和知识，以实现更加精准的对象级搜索。相对于传统的语义网络而言，知识图谱的实体覆盖率更高，语义关系也更加全面。近年来，知识图谱技术被广泛用于发现和组织文本知识，对于视觉数据等非结构化的多模态数据的关注度则较低，并且由于多模态数据之间的语义差异和异质性，一度缺乏有效的技术手段来从这些数据中提取结构化知识。随着计算机视觉（computer vision，CV）[2]和多模态学习（multi-modal learning，ML）[3]研究的深入，发现视觉数据作为多模态数据的一个重要部分，能够为知识图谱中实体提供充分的视觉信息；通过单模态的表示学习，可以将文本或图像数据表示为计算机能够理解的数值向量或者进一步抽象为更高层的特征向量，多模态表示学习在保持模态特定语义的完整下，有助于缩小异质性差距；多模态实体链接技术可以帮助相同实体的跨模态的信息进行对齐。现有工作已经证明，加入视觉模态信息能在知识图谱补全和三元组分类等工作中发挥重要作用[4,5]，并且多源信息已显示出在知识图谱上进行推理的潜力[6]。多模态知识图谱在传统知识图谱的基础上，构建了多种模态的实体，以及多模态实体间的语义关系。在多模态数据环境下，跨模态数据之间既拥有模态特性，也拥有语义共性，如何构建和应用多模态知识图谱是当下学术界和工业界的热点之一。现有部分构建工作主要依赖多媒体数据的元数据，而并非其本身的视觉或音频特征，这具有较大局限性。显然，如果可以结合文本和视觉等数据库资源，增加噪声处理能力，扩大实体对齐和链接预测的范围，并开展实体关系挖掘，可以帮助现有的模型在综合考虑文本和视觉特征时获得更好的性能，为上层智能应用提供更接近于现实场景的数据基础，在推荐系统、信息检索、视觉问答以及人机交互中有极大的应用潜力，这也是研究多模态知识图谱的意义所在。

本文是第一个较为全面的多模态知识图谱研究综述，对当前多模态知识图谱的构建方法与应用研究进行了系统的分析与论述。对模态知识图谱概念的基本认识进行了讨论，分析归纳了现有基于图数据库和知识图谱等不同角度下的多模态知识图谱构建工作的方法和技术手段。对标注数据缺失等多模态知识图谱在构建和未来应用中的核心问题与挑战进行了分析和展望。

1 相关研究现状

1.1 知识图谱

知识图谱从语义角度出发，以事实三元组的形式描述客观世界中概念、实体及其关系，将实体和概念抽象为节点，将关系抽象为边，通过结构化的形式对知识进行建模。三元组由头实体、尾实体和描述它们之间的关系组成，如〈张三，国籍，中国〉构建一个三元组实例。知识图谱用本体（ontology）对概念和关系进行形式化描述，知识的本体框架和三元组实例共同构成完

收稿日期：2021-05-05；修回日期：2021-06-21

作者简介：陈烨（1993-），男，浙江诸暨人，硕士研究生，主要研究方向为知识图谱；周刚（1974-），男（通信作者），河南郑州，教授，博导，博士，主要研究方向为大数据分析、数据挖掘（gzhougzhou@126.com）；卢记仓（1985-），男，河南郑州，副教授，硕导，博士，主要研究方向为数据挖掘.

整的知识图谱，并用资源描述框架(resource description framework，RDF) [7] 进行统一表示。

文献 [8] 指出，知识图谱构建一般采用自动或半自动技术从结构化、半结构化以及非结构化数据资源中抽取知识，并存入基于逻辑划分的数据层和模式层，是一个迭代更新的过程，主要包含信息抽取、知识融合、知识加工三个阶段 [9] 。利用自然语言处理、机器学习等技术从多源异构的数据资源中自动构建知识图谱的技术取得长足进展，例如华盛顿大学的 TestRunner [10] 、OLLIE [11] 以及卡耐基梅隆大学的 NELL [12] 。

当前知识图谱被广泛用于处理文本数据，对于半结构化或非结构化的图像、音频、视频等多模态数据的关注度则较低。作为一种知识表示、存储的手段，其可推理、可解释性在图像识别、图像分类中有较好表现 [13,14] 。一方面，知识图谱可以提高视觉识别未知类的性能; 另一方面，视觉信息可以用来扩展知识图谱，两者相辅相成。多模态数据的涌现使跨模态语义理解与知识表示需求变得更加迫切，作为承载底层海量知识并支持上层智能应用的重要载体，知识图谱也急需多模态化。

1.2 多模态学习

从语义感知角度理解，一个客观实体可以被视觉、听觉、触觉等不同模态感知; 从数据层面理解，同一实体可以有图片、文本、语音等数据记录。多模态起源于计算机人机交互领域信息表示方式的研究，模态信息存储在多模态数据中 [15] 。本文讨论的多模态数据可理解为描述同一对象的多媒体数据，多模态数据虽然在底层表征上是异构的，但是相同实体的不同模态数据在高层语义上是一致的。让人工智能更贴近人类对客观世界的认知，实现对多模态数据环境的理解，需要其具备解释多模态数据的能力。多模态数据之间由于其本身结构特点，其技术研究主要面临两大挑战: a) 语义鸿沟(semantic gap) ，指计算机表示系统与人类认知系统对同一个概念形成不同描述的差异，例如对于图像的像素信息、颜色、形状等人类认知中直观的语义表现在计算机视觉的表达中就需要借助复杂的数学形式化方法，利用参数组合对颜色、形状的概念进行编码表示; b) 异构鸿沟(heterogeneity gap) ，指图像、文本等不同媒体的数据具有不同的特征表示形式，它们的相似性难以直接度量 [16] 。

对多源异构数据的挖掘分析可被理解为多模态学习，其任务是通过学习多个模态数据中的信息，实现各个模态信息的转换和交流。不同模态之间的数据在进行综合建模时就会面临语义鸿沟和异构鸿沟带来的问题。随着深度学习的方法在获取自然语言、视觉、听觉等单模态表示上已经取得了较优的效果，多模态学习也已经进入多模态深度学习阶段 [17] 。把不同媒体的数据从各自独立的空间映射到一个第三方的公共空间中进行相似性度量是一种直观的方法。现有的研究大多仅考虑两个模态数据，当同时面临三个或更多模态时，对于公共空间的寻找将面临一定的困难 [18~20] 。

结合文献 [18,19] 中对多模态深度学习的综述，多模态深度表示学习是多模态深度学习的一个重要研究方向，主要作用是用深度学习的方法将多模态数据在同一高层语义表示空间进行对齐，以便进行对齐、比较和融合。如图 1 为多模态深度表示学习一般框架 [21] ，通常用合适的神经网络学习文本、图像、音频、视频等多模态数据在相应特征空间的表示，随后将各模态的表示作为输入，继续构建更深层的神经网络结构，利用深度跨模态表示学习构建的神经网络融合各模态的语义信息得到在共同表示空间中各模态的高层语义表示。

多模态数据还有数据量大、数据分布稀疏等特点。相较于单模态的研究，多模态数据集的构建更为困难，指代同一实体的不同模态数据通常需要昂贵的人工标注，大规模的多模态研究面临着训练数据缺失的难题，已有研究通过跨模态知识迁移 [22] 和预训练 [23] 来避免对标注数据的过高依赖。

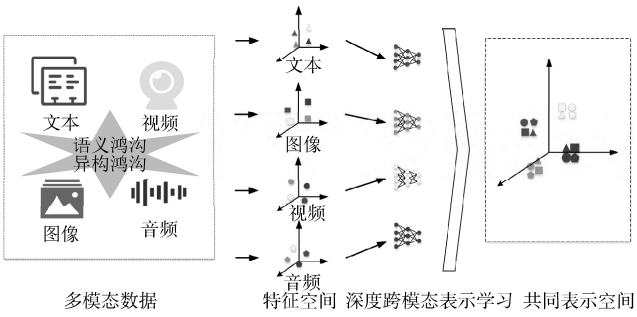


图 1 多模态深度表示学习一般框架

Fig. 1 A general framework of multi-modal deep representation learning

2 多模态知识图谱

认知科学研究表明，个体感知外界进而形成知识的过程是多感官信息的融合处理。让机器具备理解和解释能力以有效地处理来自不同模态的信息，离不开大规模、结构化的背景知识，多模态知识图谱正是知识图谱与多模态学习的有机结合。从早期图数据库发展而来，当前多模态知识图谱研究的重心主要围绕在图像和文本两种模态上 [24~26] 。文献 [24] 是从链接预测和实体匹配任务出发定义多模态知识图谱，是包含了所有实体的数值特征和图像，以及多个知识图谱之间的实体对齐的知识图谱。文献 [25] 中认为，多模态知识图谱在传统知识图谱的基础上，构建了多种模态下的实体，以及多种模态实体间的语义关系。文献 [27] 中将多模态知识图谱定义是包含文本和图像等多种数据类型的知识图谱，将视觉或文本信息引入到知识图谱中，将图像或文本作为实体或实体的属性。

基于现有研究，多模态知识图谱的一般认识比较简单，可以将其看做是具有多模态化的实体和属性的知识图谱，其中视觉等多模态数据具有实物演示、消除歧义、补充细节的作用，可以促进语义理解和可解释的推理 [24,28] 。由于加入了多种模态数据，除文本外其他模态信息以何种形式表示在多模态知识图谱中是一个难点。现有的技术手段主要通过利用文本知识图谱实体链接对应的其他模态信息的 URL(uniform resource locator) 链接，这具有一定局限性。如图 2 是一个广义的多模态知识图谱的示例。参照 Richpedia 中的形式化定义，遵循 RDF 框架，扩展图像模态作为实体在知识图谱中的存在，并强调图像模态实体之间的关系发现，可以将多模态知识图谱视为多模态知识图谱三元组的集合: 在多模态三元组 T_M 中，实体集合 E 包括文本知识图谱实体 E_{KG} 和图像实体 E_{IM}，R 表示一系列关系的集合，L 是描述文本的集合，B 表示为一系列的空白节点，T_M 表示格式为〈主语，谓语，宾语〉= $(E \cup B) \times R \times (E \cup L \cup B)$ [25] ，其中 E 和 R 利用统一资源标志符(uniform resource identifier，URI) 表示。

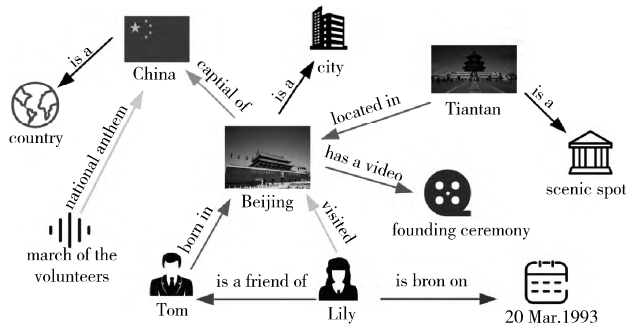


图 2 一个广义的多模态知识图谱示例

Fig. 2 Example of a generalized multi-modal knowledge graph

2.1 多模态知识图谱的构建

知识图谱经历了早期由人工和群体智慧构建到利用机器学习和信息抽取等技术自动获取的过程，并逐渐从单一的文本

模态扩展到庞大的多模态共存。多模态知识图谱的构建在传统知识图谱构建基础上,经历了早期的图数据库时代和近期关系更加复杂、数据更加全面的图谱化时代。如表 1 为部分现有可视化数据资源和多模态知识图谱。

表 1　部分现有可视化数据资源和多模态知识图谱

Tab. 1　A part of existing visual data resources and multi-modal knowledge graphs

图谱	发布时间	数据资源	数据规模	网址
ImageNet [29]	2009	WordNet [30], Web	21 841 个实体,18 种关系,14 197 122 张图片	http://www.image-net.org/
Wikidata [31]	2014	Wikipedia, Freebase [32]	超过 1 400 万个包含 URI 的实体	https://www.wikidata.org/wiki
DBpedia [33]	2015	Wikipedia	超过 260 万个包含 URI 的实体	https://wiki.dbpedia.org/develop/datasets
Visual Genome [34]	2016	Wordnet Synsets, Web	75 729 个实体,40 480 种关系,1 531 448 个三元组,108 077 张图片	https://visualgenome.org
MS-Celeb-1M [35]	2016	Facebook, Google	100 万个名人的图片	https://cn.bing.com/images
IMGpedia [36]	2017	Wikimedia, DBpedia Commons	1 500 万张图片的视觉内容描述,4.5 亿个视觉相似性关系	https://dx.doi.org/10.6084/m9.figshare.4991099.v2
CN-DBpedia [37]	2017	百度百科,互动百科,Wikipedia(中文)	超过 900 万个实体,超过 6 700 万个三元组	http://kw.fudan.edu.cn/cndbpedia/search/#
ImageGraph [28]	2017	Freebase	14 870 个实体,1 330 种关系,829 931 张图片	https://github.com/nle-ml/mmkb
MMKG [24]	2019	Freebase15K, Yogo15K, DBpedia15K	15 000 余个实体,1 300 余种关系,13 000 余张图片	https://zenodo.org/record/1245698
Richpedia [25]	2019	Wikipedia, Web	30 638 个实体,2 883 162 张图片,119 669 570 个三元组	http://rich.wangmengsd.com
DCC [38]	2020	科学出版物及源代码	539 个出版物,7 999 个 DL 实体,174 张 DL 插图,256 个 DL 资源库	https://brat.nlplab.org/

2.1.1　基于图数据库视角的构建

在早期描述多模态数据的数据集中,如 Wikidata [31]、DBpedia [33],重点是捕获多媒体文件的作者、创建日期、大小等高级元数据,而非多媒体内容本身的音频或视觉特征,并不能很准确地支撑多模态知识图谱的构建。2004 年的 ESG 项目,首先以网页游戏的形式,将图像标注任务人工外包,进行视觉模态数据的收集。ImageNet [29]、Visipedia [39] 等图像数据集主要应用在计算机视觉领域。ImageNet 是根据 WordNet [30] 层次结构组织的图像数据集,其目标是收集大量带有标注信息的图片数据供计算机视觉模型训练,每个概念图像都是经过质量控制和人为标注的,这种收集方式在多样性和伸缩性方面受限,并且消耗极大的成本 [40]。ImageNet 包含 1 500 万张已标注的高清图片,涉及 22 000 个类,并对近百万张图片中主要物体的定位边框进行了标注,其实质是一个图片版的 WordNet。2013 年,Chen 等人 [41] 提出 NEIL,大规模、自动化的图像抽取开始成为主流; Vijayanarasimhan 等人 [42] 使用亚马逊土耳其机器人进行抽取任务,但检索精度不高是一个弊端。Wikidata 中存在大量多模态资源,是一个免费的协作知识库,提供了可靠而强大的数据共享查询服务。DBpedia 是早期的语义网项目,采用了严格的本体形式组织从维基百科中抽取的知识条目; 对于每个实体,DBpedia 定义了一个唯一的全局标志符,可以将其引用为网络上描述的一个实体,目前涵养了超过 260 万个实体,对于构建多模态知识图谱提供了很大的便利。Visual Genome [34] 也是一个映射到 WordNet 同义词集的大规模图片语义理解数据集。Guo 等人 [35] 识别了 Facebook 图像中的 100 万个名人,并将其链接到 Freebase 中的相应信息构建了名人图谱 MS-Celeb-1M。2016 年,哥伦比亚大学 Li 等人 [43] 提出一种多模态模式挖掘方法,用于挖掘和命名来自高级新闻事件图像标题对的语料库中的多模态视觉模式,从计算机视觉角度推动了图数据库的构建。

2.1.2　基于知识图谱视角的构建

部分大规模的知识图谱 [32,44] 尝试合并视觉信息,但仅是通过超链接将它们链接到文本,使用时才缓存视觉内容。IMGpedia [36] 是一个多模态知识图谱的先例。从 Wikimedia Commons 数据集中的图像中收集大量的可视化信息,构建并生成 1 500 万个视觉内容描述符,图像之间有 4.5 亿个视觉相似关系。此外,IMGpedia 中单个图像与 DBpedia、Wikidata 之间存在链接。IMGpedia 使用四种图像描述符进行基准测试,并且这些描述符的引用和实现是公开的。IMGpedia 相较于 DBpedia 提供了一个更好的可视化语义查询平台,但存在关系类型稀疏、关系数量少、图像分类不清晰等缺陷,并且没有对图像内可能存在的实体间关系进行深入挖掘。CN-DBpedia [37] 是一个由复旦大学知识工场实验室研发并维护的中文大规模通用领域结构化百科知识库。

2017 年,文献 [28] 探索了一种新的机器学习方法来回答网络提取的知识图谱中的视觉相关查询,由此创建了一个包含 1 330 种关系类型、14 870 个实体和 829 931 张从网络中爬取的图像的知识图谱 ImageGraph,其中图像被视为一等公民(first-class citizen),并引入了卷积神经网络(convolutional neural networks,CNN)和知识图谱嵌入(knowledge graph embedding,KGE)的组合来回答视觉相关的查询。2019 年,Liu 等人 [24] 在之前 Rubio 的工作基础上构建了一个包含所有实体的数字特征和图像的三个知识图谱的集合 MMKG,基于 N-Triples 格式,选择在知识图谱补全文献中广泛使用的数据集 FB15K 作为基准与其他知识图谱中的实体进行对齐。与之前在同一图像中执行视觉推理任务不同,MMKG 主要用于联合不同知识图谱中的不同实体和图像执行关系推理,多关系链接预测和实体匹配任务可以从该资源中受益。但 MMKG 主要针对小数据集,并且图像没有作为单独的图像实体存在,依赖于相应的传统文本实体,同时也没有考虑图像的多样性。

同一阶段,在 Li 等人 [43] 的工作扩展上,建立在复杂的场景图谱生成(scene graph generation,SGG) [45,46] 基础之上的多模态知识图谱概念被提出,结合了计算机视觉领域的部分工作。SGG 任务是建立一个图谱来表示某个场景或事件中多个实体之间的关系,包括位置关系及一些简单的谓词。文献 [47,48] 是哥伦比亚大学 Chang 团队在这类构建思路下的代表性工作,在分别构建文本知识图谱和场景图谱的基础之上,通过动态消息传递和桥接算法融合成一个图,生成一个新的多模态知识图谱支撑下游应用任务。常识知识图谱(common sense knowledge graph)是丰富的知识存储库,编码了客观世界的结构,以及一般概念如何相互作用,场景图谱则可视为常识知识图谱的一个具有图像条件的实例化 [48]。场景图谱通过图像识别技术将图像解析为抽象的语义元素,即对象及其交互,其中实体对应的是图像中的一个区域。

2019 年,Wang 等人 [25] 通过网络链接和图片搜索,向 Wikidata 中的文本实体补充足量和多样的图像,构建了一个较全面的多模态知识图谱 Richpedia,构建工作主要分为数据获取、图像处理、关系发现三个阶段。通过借助外部图像搜索引擎获取图像以缓解单个资源中信息稀疏导致的长尾问题(long

tail)[49]；利用搜索引擎中图像的相关性排序获得与文本实体最相关的图像模态数据；通过去噪操作和 RGB 多样性检测对图像进行预处理和筛选，确保图像实体的相关性和多样性；为所有实体添加了 URI，并利用基于规则的关系抽取模板构建多模态语义关系。Richpedia 中借鉴了场景图的思想，考虑单张图像包含多个客观实体的情况，由于技术限制，根据不同图像的像素特征直接检测实体及语义关系困难较大，在图像实体之间的语义关系检测工作还是借助图像的文本描述完成的。

2020 年，Kannan 等人[38] 在 DCC 项目中通过从科学出版物和附带的源代码中提取信息并将其表示为统一的知识图谱，主要使用了知识图谱对齐方法，是一个集成多媒体文档中文本、结构化图像和源代码的研究。Li 等人[50] 提出了一个全面的、开源的多媒体知识提取系统 GAIA，系统由文本知识提取分支和可视化知识提取分支组成，每个分支采用相同的文档集作为输入，使用在相同语义空间中定义的类型，分别编码各自模态信息创建单独的知识库，通过共享空间使两个知识库可以融合成一个单一、一致的多媒体知识库。

2.2 小结

当前对于多模态知识图谱的构建工作，根据研究方向侧重点的不同，典型思路有两个，如图 3 所示。

a) 自然语言处理角度出发的多模态知识图谱构建相对主流，但是还没有摆脱对传统文本知识图谱的依赖。主要在文本知识图谱构建的基础上，给实体补充视觉信息，工作实质是知识图谱补全，并在图像间做视觉关系的发现和跨模态实体链接，对于图像实体的扩充以及链接关系的确定主要依靠多模态数据的元数据。这种方法对视觉特征的提取及多模态关系挖掘还是粗粒度的，并且在真实场景中，一些实体如概念并不包含图像信息。

b) 计算机视觉角度出发的多模态知识图谱构建工作建立在场景图谱生成基础之上，是一种分布式的构建方法，将视觉知识与外部文本知识图谱相桥接。相较于常识知识图谱中一个节点即表示一个实体或谓词类，场景图谱的图节点表示特定图像中的一个实体或谓词实例，需要链接到文本知识图谱中的相应实体或谓词类完成构建。由于计算机视觉对于图像识别效果偏向于抽象的概念层面识别，并且视觉关系比较单一，存在较多噪声影响。该种构建思路下要达到细粒度的实体对齐与视觉关系发现仍需要更多新技术的辅助。

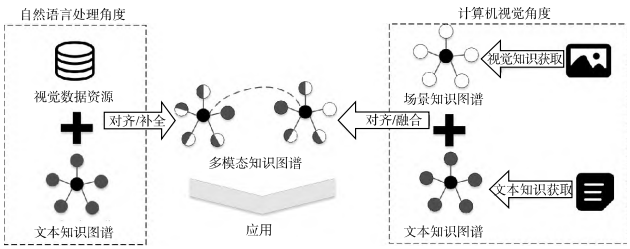


图 3 多模态知识图谱构建的两种典型方法

Fig. 3 Two typical methods of multi-modal knowledge graph construction

3 多模态知识图谱关键技术

图像等视觉信息当前可以通过链接或者以具有数据类型的二进制字符串形式包含在知识图谱中。如果不能将异构的跨模态信息有效关联，单纯将文本知识图谱进行图像扩充，以各模态信息源实体为核心所构建的图谱将是一个个孤立的抽取图谱，无法真正反映跨模态数据中的知识。本文主要分析多模态信息抽取、多模态表示学习和多模态实体链接三种多模态知识图谱关键技术。

3.1 多模态信息抽取

信息抽取的主要目标是从无结构的生文本中抽取出结构化、半结构化或非结构化的数据。多模态信息抽取是多模态学习与信息抽取技术的结合，是当前多模态知识图谱研究的主要工作之一。单模态研究的进步为多模态的研究提供了扎实的基础，对于不同模态的信息可采取先进行单模态抽取，再进行多模态处理的方法[40]。本节主要讨论图像模态的信息抽取并简要介绍声音模态的信息抽取。

3.1.1 图像模态信息抽取

ImageNet[29]、IMGpedia[36] 等视觉数据库通常是图像或视频数据的丰富来源，能够为知识图谱中的实体提供充分的视觉信息。基于实体属性补充角度的图像模态信息抽取方法比较简便，通常基于文本知识图谱，利用关键词，通过搜索引擎从视觉数据库中检索与文本实体对应的图像模态数据进行补充，图像信息的判别主要依赖图像元数据对图像内容的描述，具有一定的局限性。基于图像模态实体构建与视觉关系发现角度的图像模态信息抽取中，对于元数据的作用同样不可忽视，但更多关注点在内容本身的特征上。直接通过人为理解从图像中提取出有效而丰富的特征较难，早期主要考虑图像的色域、对比度和纹理等，在深度学习出现之前，图像识别主要借助尺度不变特征转换(scale-invariant feature transform，SIFT)[51]、方向梯度直方图(histogram of oriented gradient，HOG)等算法提取具有较好区分性的特征，再集合 SVM[52] 等机器学习算法进行图像识别，HOG + SVM 在行人检测中有着优异的效果。CNN 可以直接将图像的原始像素作为输入，避免了先使用 SIFT 等算法提取特征，减轻了大量数据预处理工作。

在图像特征提取的角度下，CNN 通过卷积和池化操作，产生图像模态的矩阵表示，图像特征的向量表示工作由全连接层或全局均值池化层的输入完成。代表性的 LeNet-5[53] 在手写体数字和字母的识别中得到了非常高的精度，在车牌识别等场景中得到实际应用。在 LeNet-5 的基础上，通过增加网络的深度或对卷积和池化操作进行变形，后续研究提出了更加复杂和高效的卷积神经网络，陆续有 AlexNet[54]、VGG[55]、NIN[56]、GoogLeNet[57]、ResNet[58] 和 CapNet[59] 等，对图像模态的特征提取能力进行了拓展，极大地提高了图像识别的精度。借助经过预训练的 CNN 模型，可以在图像实体的类识别上取得较优效果。表 2 是对经典 CNN 的一个统计。

表 2 部分经典卷积神经网络统计

Tab. 2 Survey of several classical convolutional neural networks

CNN 网络	方法	效果
LeNet-5[53]	利用卷积和池化提取图像特征，将结果输入到全连接层生成图像特征向量	在手写体数字和字母识别中得到极高精度，是现代卷积网络的基础
AlexNet[54]	使用 ReLU/normalization/dropout 增加网络深度，得到包含图像深度语义信息的特征表示	ILSVRC 2012 冠军，top-5 错误率 16.4%，8 层神经网络
VGGNet[55]	利用小卷积核和多个卷积层来捕获精细的图像特征	ILSVRC 2014 年度亚军，top-5 错误率 7.3%，19 层神经网络
NIN[56]	使用 MLPconv 算法改进卷积层，使用全局平均池代替全连接层	获得高度非线性图像的矩阵表示，提高了网络的泛化能力
Google InceptionNet[57]	使用不同大小的卷积内核的 inception 模块	ILSVRC 2014 年度冠军，top-5 错误率 6.7%，22 层神经网络
ResNet[58]	融合身份映射和残差映射	2015 年 ILSVRC 冠军，top-5 错误率 3.57%，152 层神经网络
CapsNet[59]	输入和输出均由胶囊神经元向量形式构成，每个元素是图像中实体特征的参数表示，通过动态路由算法将相邻的胶囊层连接起来	避免了图像中实例丢失方向和空间信息

诸如 Faster-RCNN[60] ,Mask-RCNN[61] 等目标检测模型在区域特征提取上有着较好表现,以 Faster-RCNN 为例,其输入为一张图像,输出为一个序列的兴趣区域 RoI(region of interest) ,其中每个 RoI 都附带其在原始图像中区域范围左上角和右下角的坐标向量、一个 RoI 的表示向量和一个表示类型的标签。Faster-RCNN 在细粒度的图像目标识别上有着广泛的应用,在基于实体的多模态知识图谱构建工作中,可以帮助识别同一张图像中多个实体的类别并标注位置信息。

3.1.2　声音模态信息抽取

在现有多模态知识图谱研究中,声音模态主要还是以实体的属性形式存在,代表描述实体的一段音频。从构建角度除了利用元数据,语音数据也可通过语言解析为文本进而用传统的文本的方法进行语义特征提取。

3.1.3　小结

现有多模态知识图谱构建工作中,对于文本模态之外的模态信息的抽取大多属于被动抽取,是在依托现有文本实体的基础上在有限范围内定向地进行模态数据补充,在视觉关系发现中也仍需要文本描述等额外知识的介入[23] 。未来对于没有元数据描述的未知的多模态数据中的信息抽取仍然存在较大考验。

3.2　多模态表示学习

为了方便对抽取到的多模态信息进行处理,需要对输入的数据进行表示,在深度学习时代,采用的表示方法是将输入的数据表示成向量,进而通过深度神经网络的强大建模能力,自动地对输入数据中的特征进行提取。在知识图谱的研究和应用中,知识表示学习是其基本任务之一。信息可以由单模态的表示学习转换为能被机器处理的数值向量或者进一步抽象为更高层的特征向量,多模态表示学习旨在减小模态信息在联合语义子空间中的分布差距,同时保持模态特定语义的完整。

3.2.1　文本模态表示

文本表示的核心是对语言基本单元进行表示,然后用神经网络学习语言模型提取文本特征,最后用神经网络的某个输出向量作为文本表示。早期使用独热表示(one-hot) 单词,每个词的表示为词典中该词的索引,然而这种形式空间损耗较大,并且不能建模词之间的语义相似性,同时存在数据稀疏问题。后续一般使用 Mikolov 等人[62] 用神经网络模型得到的向量作为词向量。提取文本特征的神经网络主要包括简单的前馈神经网络,以及擅长序列建模的循环神经网络(recurrent neural network,RNN) ,例如长短时记忆网络(long short term memory network,LSTM) [63] 及其变体。近年来,基于自注意力机制的 Transformer[64] 及其变体 BERT[65] 具有比 RNN 更好的文本建模效果,逐渐取代 RNN 成为主流的文本特征提取方法。

3.2.2　图像模态表示

卷积神经网络是在多层神经网络的基础上发展起来的针对图像而特别设计的一种深度学习方法,在图像处理上取得了优异的效果[19] 。近年来由于深度学习技术的发展以及计算机处理能力的提高,图片等多媒体数据可以和文本采用相同的深度学习框架分析,这为多模态研究提供了便利。

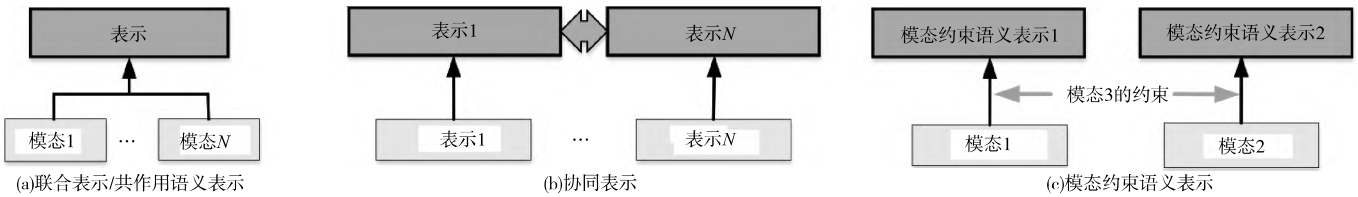
3.2.3　声音模态表示

声音模态数据通常以模拟信号的形式存在,声音的时域波形只代表声压随时间变化的关系,不能很好地体现声音的特征,一般需要进行数字化处理以获得数字信号序列,然后通过生理学、语言学相关的先验知识对离散化的数字信号序列进行声学特征向量提取。当前对声音信号的处理技术主要有傅里叶变换、线性预测和倒谱分析等[19] 。考虑同一实体的多模态数据之间存在富模态信息和缺失模态信息的不平等情况,文献[3] 将机器学习中模态表示分为联合表示(joint representation) 和协同表示(coordinated representation) 。联合表示将多个模态的信息一起映射到一个统一的多模态向量空间; 协同表示负责将多模态中的每个模态分别映射到各自的表示空间,映射后的向量之间满足一定的相关性约束。文献[19] 将多模态表示分为模态共作用语义表示和模态约束语义表示。与联合表示的定义类似,模态区作用语义表示指融合各单模态的特征表示,以获得包含各模态语义信息的多模态表示; 模态约束语义表示指用一个单模态表示结果去约束其他模态的表示,以使其他模态的表示能够包含该模态的语义信息。它并非融合各输入的信息,而是通过将输入模态的表示映射到目标模态的语义空间中,使得映射结果与语义相同的目标模态的相似性大于语义不同的目标模态,此类代表性的神经网络包括前馈神经网络和递归神经网络。

3.2.4　多模态知识图谱表示学习相关工作

早期的表示学习方法主要学习基于实体和关系的结构信息,而忽略了其他数据类型的实体知识,近年来的相关工作证明,从实体的图像和文本描述中可以获得丰富的补充知识,在知识图谱补全和三元组分类工作中发挥重要作用[4,66,67] 。从多模态知识图谱构建角度,多模态知识图谱表示学习工作可以分为基于特征的方法和基于实体的方法两种类型。基于特征的方法将多模态信息作为实体的辅助特征来处理,基于实体的方法将不同的模态信息作为结构化知识的关系三元组,而不是预定的特征[27] 。2013 年,Frome 等人[68] 提出了 DeViSE,首次开展了将图像嵌入到语义空间中的相关工作。DeViSE 利用图像和未注释的文本提供了一个通用的表示学习,利用文本数据显式地将图像映射到一个丰富的语义嵌入空间中,以允许基于文本标签预测未知的视觉类别。

联合表示、协同表示和共作用表示如图 4 所示。



(a)联合表示/共作用语义表示　(b)协同表示　(c)模态约束语义表示

图 4　联合表示、协同表示和共作用表示

Fig. 4　Joint representations, coordinated representations and con-representations

1) 基于特征的方法　基于翻译的模型 TransE[69] 将实体和关系投影到相同连续的低维向量空间,并且将关系解释为头和尾实体之间的翻译操作。在此思想之上,2016 年的 IKRL[4] 是第一个将实体的图片信息集成到现有模型以帮助知识图谱学习表示的工作,并利用了注意力的方法对信息进行聚合。IKRL 为每个实体学习两种独立的表示,一种基于结构化知识,另一种基于视觉知识。然后使用 TransE 的能量函数对三元组进行打分,最后通过排序来预测实体之间的关系。Xie 等人在 DKRL[66] 中提出将 Freebase 中的实体的描述信息引入到知识图谱中,使用连续词袋模型(continuous bag-of-words model,CBOW) 或 CNN 从实体描述来构建实体表示。这两种模型都受到模态数量的限制,很难整合多个模态的知识,因为模态数量的增加会很大程度上加剧模型复杂性。随后文献[5] 改进了 IKRL,将知识图谱三元组能量函数定义为利用结构、视觉和语言知识表示的子能量函数的总和,为实体丰富了文本模态的表示,然后将实体对应的文本表示和图片表示进行融合,得到实体的多模态表示,相较 TranE 和 IKRL 两种基线方法,取得了更好的三元组预测效果。2019 年,MMKG[24] 中通过提取关系、

潜在的数值和视觉特征来进行多模态链路预测和实体匹配。结合多模态特征，并通过匹配实体的底层语义和挖掘嵌入空间中所包含的关系来衡量事实的可信度。在计算每种模态下的事实分数时，它学习了实体对齐的实体嵌入。

2) 基于实体的方法　2017 年，Lonij 等人[13]在神经张量层的基础上设计了一个相互语义嵌入空间，除了考虑线性混合外，还考虑实体向量的乘性混合。在此基础上，提出一种基于该模型的新目标函数来增强评分函数的局部平滑性。Zhang 等人[70]从图像目标关系检测角度出发提出 VTransE，基于翻译模型，VTransE 考虑的是一张图像内的实体的视觉关系检测，作为第一个端到端和全卷积架构，它通过目标检测模块、可微特征提取层和视觉翻译嵌入层来进行谓词分类。VTransE 在关系空间中学习一致的转换向量，不考虑谓词关系中涉及的主语和宾语的不同外观。VTransE 简单易实现，但无法避免 TransE 在应对一对多、多对多等关系预测时的不足。CVPR2019 的改进工作[47]通过学习动态树结构的组合来刻画视觉上下文，并基于此来预测可视关系，从而一定程度上缓解了长尾关系检测难的问题。2018 年，Pezeshkpour 等人[71]基于实体的方法提出 MKBE，思路与 IKRL 相仿，它采用神经编码器和解码器，分别嵌入多模态证据类型和生成多模态属性。研究指出多模态研究在可伸缩性方面需要扩展，尽管多模态证据提供了更多的信息，但这些额外数据的哪些部分对于预测知识库的关系结构是有用的并不明显，并且这些模型容易过拟合。2019 年 Wang 等人[67]提出了一种新的表示学习方法 TransAE，将多模态自动编码器与 TransE 模型相结合，自动编码器的隐含层作为 TransE 模型中实体的表示，它不仅将结构知识编码，还将视觉和纹理等知识编码到最终的表示中。TransAE 在链路预测和三元组分类方面可以显著提高算法的性能，但只考虑了每张图像中的一个实体。2020 年，Chen 等人[72]针对多模态知识图中的实体对齐问题，提出了一种基于知识嵌入的多类型知识表示方法。分别生成关系知识、视觉知识和数字知识的实体表示。通过设计一个基于公共空间学习的多模态知识融合模块，将不同类型知识的多种表示形式整合在一起，从而迁移不同知识空间下的特征。

3.2.5　小结

单模态研究的进步使得多模态的研究有了更扎实的基础，文本、图像、声音模态的深度学习的发展以及算力的支持为多模态表示学习寻找合适的共享空间提供了便利[18,19]。基于特征的多模态表示学习方法根据知识图的结构和实体的视觉表示来定义三元组的评分函数，这意味着每个实体必须包含图像属性。然而，在实际场景中，一些实体不包含多模态信息，所以这种方法不能被广泛使用。基于实体的多模态表示学习方法将多模态信息认为是知识图谱中的一等公民，使用基于 CNN 的方法来训练知识图的嵌入，在部分补全问题上效果优于基于特征的方法，但这种独立的处理方法容易忽略多模态信息的融合。鉴于多模态知识图谱中存在多种模态下的实体及语义关系，未来在其表示学习研究上，或应该结合以上两种角度的方法。

此外，多模态学习过程中哪些额外部分的数据能够为知识抽取和关系发现提供支援并不明确，并且存在的信息冗余和噪声是需要解决的一个关键问题，而注意力机制的引入为缓解噪声、增加任务细粒度效果提供了一定的支持。

3.3　多模态实体链接

实体链接(entity linking，EL)任务是指将从给定资源中抽取的实体对象链接到知识图谱中对应的实体中[73]。其基本思想是从知识图谱中为给定的实体指称选择一组候选实体对象，然后根据相似度比较将实体链接到正确的实体对象。在基于实体的多模态知识图谱的构建工作中，存在同一文本实体对应多张图像的情况，也存在同一图像对应多个文本实体或描述的情况[25]，需要进行多模态实体链接。

3.3.1　多模态实体链接相关工作

Moon 等人[74]提出了一种针对 Snapchat 社交平台的融合图片信息的面向短社交文本的实体链接方法。分别利用 RCNN 和 Bi-LSTM 提取视觉特征和文本特征并获得对应表示，同时预训练了一个基于编辑距离的实体相似性子网络来判断两个实体提及是否是相同，从而获取实体的词汇级别的表示，通过使用注意力机制对三种表示进行融合，最后计算提及的实体与知识库中实体的相似性来得到实体链接的结果。作者从 Snapchat 中抽取了图片和对应标题，然后对标题中的实体及指向知识库的链接进行标注，从而获得数据集。该方法的效果优于传统只利用文本的实体链接方法，在社交媒体数据上取得了很大的提升。Zhang 等人[75]提出了 CAN(adaptive co-attention network)，选择推特作为数据源，爬取并标注了包含配图的数据集，主要包括人物、地点、机构和杂项共四种类型的实体。CAN 拓展了传统的 Bi-LSTM + CRF(conditional random fields)模型，在 CRF 层之前对文本和图片的表示进行了互注意力，从而使得每个词获取了多模态的表示，加入门控机制与过滤器机制来控制每个词对图片和文本的偏好程度。2020 年，文献[76]中提出一种基于联合知识表示学习的多模态实体对齐方法 ITMEA，联合图像和文本数据，采用 TransE 与 TransD 相结合的知识表示学习模型，在低维语义空间中迭代地学习已对齐多模态实体之间的关系，从而实现多模态的实体对齐。Wei 等人[77]提出了一种新的多模态交叉注意网络 MMCA，通过在一个统一的深度模型中联合建模图像区域和句子单词的内部模态和中间模态关系来进行图像和句子匹配；作者设计了一种新的交叉注意机制，利用每个模态内部的模态关系加上图像区域和句子单词之间的模间关系来互补和增强图像和句子匹配。

3.3.2　小结

现有多模态实体链接工作主要建立在多模态信息抽取和表示学习的基础上，需要一定的背景知识提供支撑。利用 RCNNs 提取图像的视觉特征有利于缩小候选实体的范围，使用注意力机制能够减小噪声影响，但对于实际多模态知识图谱构建中一对多、多对多的多模态实体关系发现仍面临噪声处理能力不足和先验数据不充分的挑战。

4　多模态知识图谱的典型应用

多模态知识图谱技术可以服务于各种场景，例如多模态实体链接技术可以融合多种模态下的相同实体，能够广泛应用于新闻阅读、商品推荐等场景[20]；通过远程监督可以补全多模态知识图谱，完善现有的多模态知识图谱，利用动态更新技术使其更加的完备，在端到端实体分类[24]、多模态摘要[78]中也有实际应用。

4.1　推荐系统

多模态知识图谱通过将其他模态的信息引入传统的知识图谱，能够提供的丰富的特征和信息[66,79]，将其应用于推荐系统可以有效缓解推荐系统的数据稀疏和冷启动问题，从而使推荐结果更准确，并提供可解释性支撑[80,81]。目标项目及其属性可以映射到知识图谱中以理解项目之间的相互关系[82]。此外，还可以将用户和用户侧信息集成到知识图谱中，以更准确地捕获用户和项目之间的偏好关系。实体的图像和描述可以为知识表示学习提供重要的视觉或文本信息，Sun 等人[27]提出多模态知识图谱注意力网络 MKGAT，作为第一个将多模态知识图引入推荐系统的工作，其包含一个多模态知识图谱嵌入模块和推荐模块，以协作知识图谱作为输入，通过多模态知识图实体编码器和多模态知识图谱注意层为每个实体学习一个新的实体表示，通过聚合实体的邻居的信息，同时保留自身的信息以表示知识推理关系。根据传统的推荐模型生成用户与项目之间的匹配分数，多模态知识图谱在推荐效果上较单模态

知识图谱有更好的表现。

4.2　跨模态检索

跨模态检索研究的基本内容是寻找不同模态样本之间的关系，以一种类型数据作为查询来检索其他类型的数据[16]，例如使用文本去检索相关图片或视频。跨媒体检索能够打破检索结果的媒体限制，从而增强搜索体验和结果的全面性。

跨模态检索在方法上主要分为两大类：a) 实值表示学习；b) 二值表示学习，也称为跨模态哈希方法。实值表示学习直接对从不同模态提取到的特征进行学习；而二值表示学习是对从不同模态提取到的特征先映射到汉明二值空间，然后在此空间中进行学习。如文献[83]中提出将深度玻尔兹曼机(deep Boltzmann machine，DBM)结构扩充到多模态领域，通过多模态 DBM，可以学习到多模态的联合概率分布。针对文本或图片输入，利用条件概率生成相应特征，通过检索出最靠近该特征向量的两个实例，可以得到符合条件的结果。Wei 等人[84]以 ImageNet 为数据源，使用深度语义匹配的方法，将已完成训练的全连接层中的图像特征与文本的语义信息进行配对比较，完成跨模态检索。多模态知识图谱在跨模态检索工作中有较大的应用前景，对于文本检索，输出结果中能够呈现与关键词相关的视觉等其他模态信息，能够有效帮助用户进行实体识别与消歧[73,74]；对于图像检索，通过一个特征提取模块对检索内容进行特征提取和编码，目标识别和视觉问答工作[85,86]也可以从中受益。

4.3　人机交互

利用多模态知识图谱融合不同模态数据的特性，可以推动知识驱动的人机交互(human-computer interaction，HCI)。人类通过人机交互界面与计算机系统进行交流和操作，在实际场景中，通过多传感器的使用，机器能够感知到多模态、数字化的世界，借助多模态知识图谱作为背景知识，有助于机器加强对真实场景的理解，作出更令人舒适、更自然的反馈。例如通过分析人类的语言和面部表情数据对使用者进行情感分析，从而调整环境灯光舒适度等。原来人机交互接入信息更多是从文本、页面中获得，多模态技术会带来新的内容形态，通过听觉和视觉等综合作用，在未来强调沉浸感的人机交互中发挥重要作用。

4.4　跨模态数据管理

诸如 ImageNet 是根据 WordNet 层次结构组织的图像数据集，将大量带有标注信息的图片数据以图数据库的形式对图像数据进行管理。多模态知识图谱形式的跨模态数据管理系统能够将跨模态数据以属性特征或实体的形式表示在知识图谱的结构中，不但能够对跨模态数据进行标准化管理，还能够充分利用跨模态数据之间的结构特点，挖掘隐藏关系。例如在金融证券领域，将不同模态数据进行整合管理后能够在最终控制人识别中发挥重要作用，保障信用风险评估。另外，多模态知识知识图谱的结构也能支撑跨模态数据在推理性问题上的应用。

5　展望

作为一个涉及知识图谱和计算机视觉等多领域交叉的新兴研究，目前多模态知识图谱构建与应用的专门研究还比较有限，相关技术运用还比较初级，大规模多模态知识图谱的构建与应用仍面临较大挑战，本文从以下四个方面对多模态知识图谱的未来前景作下展望：

a) 多模态数据的内容特征提取技术。新技术的崛起很大程度上取决于底层技术的发展和突破。目前计算机视觉领域对目标的分类、识别还停留在偏感知的程度上，对实体的识别限于本体和概念层面，提供的实体信息有限，距达到细粒度、实例化的图像识别还存在一定技术提升空间[50,87,88]。从多模态知识图谱构建的角度出发，如何面向知识图谱本身及下游任务，探究不同模态信息的效果，让模型自适应地从海量的多模态数据中提取对图谱构建和应用有价值的内容特征是首先面临的问题。利用多模态数据的元数据在一个阶段内仍会是构建多模态知识图谱的一个优选方法，但这具有局限性。面向多模态知识图谱构建和应用的多模态数据特征提取的或将是未来发展的主要方向之一。

b) 标注数据缺乏情况下的多模态深度学习技术。由于多模态数据之间存在的语义鸿沟和异构鸿沟，虽然借助深度学习技术已经能够寻找良好的嵌入空间，但要教会机器认识多种跨模态数据代表的是同一个实体，在目前大多数深度学习方法下还依赖于大量有标注的数据，这成为了一个重要瓶颈，对计算和存储能力也是一个严峻的考验。

基于编码器用于自然语言表示建模的 BERT[65]及多模态预训练模型或将为多模态深度学习缓解对有标注数据的依赖提供了一种新选择。预训练模型思路是先利用大规模的廉价标注信息进行无监督的预训练，再使用少量的昂贵的人工标注对模型进行微调，从无标注数据上将更加通用的知识迁移到目标任务上，进而提升任务性能。把为 ImageNet 分类任务预先训练的网络作为初始状态来训练目标检测或语义分割等其他视觉任务是非常流行的做法。图像—文本多模态预训练模型将图像特征嵌入与文本上下文信息经过编码器处理后进行预训练，代表性模型有 ViLBERT[23]、VL-BERT[89]、Image-BERT[90]等，其在掩蔽预测和图像—文本对齐任务上的表现对多模态知识图谱中的补全和匹配工作有一定借鉴性。

c) 多模态知识表示方法。随着构建源数据范围的扩大，更全面的模态层次、更细粒度的知识抽取、更丰富的语义关联将是多模态知识图谱未来发展的方向。在此背景下，传统基于 RDF 的数据模型逐渐不再适用于对多模态知识的建模和表达，需要一种可以有效联合建模逻辑规则、结构化、非结构数据和事件等多模态知识的表示方法，方便知识的增量更新，这也是未来研究的一个重要内容。

d) 多模态数据的冗余分析。当前多模态知识图谱中，从数据多样性讲，一个文本实体可能对应多个其他模态的属性值，例如不同角度拍摄的图片，这些跨模态数据在图谱结构中是平等的，但在数据量上一般是不对等的。在考虑多模态数据的特定应用场景时，如何利用好互补性，减轻富模态数据存在的重复和噪声影响[25]，也是未来多模态知识图谱构建和应用研究中的一个关键内容。迁移学习、图像分割和注意力机制将在这一环节发挥巨大作用。

6　结束语

在人工智能从单一模态逐步向多模态演进、从感知智能向认知智能发展的大背景下，多模态数据学习与知识图谱的交互作用为大数据的价值在应用上的落地提供了极富想象力的可能性。本文对目前多模态知识图谱构建与应用相关的研究现状进行了全面的调研和分析，讨论了当前对多模态知识图谱概念的基本认识，其可看做是具有多模态化实体和属性的知识图谱；介绍了图数据库和知识图谱两个视角出发的构建工作，并归纳了基于属性和基于实体的两种构建方法的主要思路；分析了多模态信息抽取、表示学习、实体链接等多模态知识图谱构建和应用中的关键技术和相关工作，并对技术发展面临的标注数据缺失、噪声影响等挑战进行了讨论；列举了多模态知识图谱在推荐系统、跨模态检索、人机交互和跨模态数据管理四个场景中的应用。最后，从四个方面对多模态知识图谱的发展前景进行了展望。希望能为多模态知识驱动的人工智能相关领域研究者提供研究思路。

参考文献：

[1]　Singhal A. Introducing the knowledge graph: things, not strings [EB/

OL]. (2012). https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.
[2] Szeliski R. Computer vision: algorithms and applications [M]. [S.l.]: Springer Science & Business Media, 2010.
[3] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423-443.
[4] Xie Ruobing, Liu Zhiyuan, Luan Huanbo, *et al*. Image-embodied knowledge representation learning [C]//Proc of the 26th International Joint Conference on Artificial Intelligence. 2017: 3140-3146.
[5] Mousselly-Sergieh H, Botschen T, Gurevych I, *et al*. A multimodal translation-based approach for knowledge graph representation learning [C]//Proc of the 7th Joint Conference on Lexical and Computational Semantics. 2018: 225-234.
[6] Wang Peng, Wu Qi, Shen Chunhua, *et al*. Explicit knowledge-based reasoning for visual question answering [C]//Proc of the 26th International Joint Conference on Artificial Intelligence. 2015: 1290-1296.
[7] Klyne G, Carroll J J, McBride B. Resource description framework (RDF): concepts and abstract syntax [R]. 2004.
[8] 刘峤,李杨,段宏,等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53(3): 582-600. (Liu Qiao, Li Yang, Duan Hong, *et al*. Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53(3): 582-600.)
[9] 徐增林,盛泳潘,贺丽荣,等. 知识图谱技术综述 [J]. 电子科技大学学报, 2016, 45(4): 589-606. (Xu Zenglin, Sheng Yongpan, He Lirong, *et al*. Review on knowledge graph techniques [J]. Journal of University of Electronic Science and Technology, 2016, 45(4): 589-606.)
[10] Etzioni O, Banko M, Soderland S, *et al*. Open information extraction from the web [J]. Communications of the ACM, 2008, 51(12): 68-74.
[11] Schmitz M, Soderland S, Bart R, *et al*. Open language learning for information extraction [C]//Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 523-534.
[12] Carlson A, Betteridge J, Kisiel B, *et al*. Toward an architecture for never-ending language learning [C]//Proc of AAAI Conference on Artificial Intelligence. 2010.
[13] Lonij V P A, Rawat A, Nicolae M I. Extending knowledge bases using images [C]//Proc of the 6th Workshop on Automated Knowledge Base Construction. 2017.
[14] Marino K, Salakhutdinov R, Gupta A. The more you know: using knowledge graphs for image classification [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 20-28.
[15] 陈鹏,李擎,张德政,等. 多模态学习方法综述 [J]. 工程科学学报, 2020, 42(5): 557-569. (Chen Peng, Li Qing, Zhang Dezheng, *et al*. A survey of multimodal machine learning [J]. Chinese Journal of Engineering, 2020, 42(5): 557-569.)
[16] Wang Kaiye, Yin Qiyue, Wang Wei, *et al*. A comprehensive survey on cross-modal retrieval [EB/OL]. (2016). https://arxiv.org/abs/1607.06215v1.
[17] Ngiam J, Khosla A, Kim M, *et al*. Multimodal deep learning [C]//Proc of the 28th International Conference on Machine Learning. 2011.
[18] 孙影影,贾振堂,朱昊宇. 多模态深度学习综述 [J]. 计算机工程与应用, 2020, 56(21): 1-10. (Sun Yingying, Jia Zhentang, Zhu Haoyu. Survey of multimodal deep learning [J]. Computer Engineering and Applications, 2020, 56(21): 1-10.)
[19] 刘建伟,丁熙浩,罗雄麟. 多模态深度学习综述 [J]. 计算机应用研究, 2020, 37(6): 1601-1614. (Liu Jianwei, Din Xihao, Luo Xionglin. Survey of multimodal deep learning [J]. Application Research of Computers, 2020, 37(6): 1601-1614.)
[20] Zhang Chao, Yang Zichao, He Xiaodong, *et al*. Multimodal intelligence: representation learning, information fusion, and applications [J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478-493.
[21] Wang Yang. Survey on deep multi-modal data analytics: collaboration, rivalry and fusion [EB/OL]. (2020). https://arxiv.org/abs/2006.08159.
[22] Huang Xin, Peng Yuxin. Deep cross-media knowledge transfer [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 8837-8846.
[23] Lu Jiasen, Batra D, Parikh D, *et al*. Vilbert: Pretraining task-agnostic visiolinguisticre presentations for vision-and-language tasks [EB/OL]. (2019). https://arxiv.org/abs/1908.02265.
[24] Liu Ye, Li Hui, Garcia-Duran A, *et al*. MMKG: multi-modal knowledge graphs [C]//Proc of European Semantic Web Conference. Cham: Springer, 2019: 459-474.
[25] Wang Meng, Qi Guilin, Wang Haofen, *et al*. Richpedia: a comprehensive multi-modal knowledge graph [C]//Proc of Joint International Semantic Technology Conference. Cham: Springer, 2019: 130-145.
[26] Wilcke W X, Bloem P, De Boer V, *et al*. End-to-end entity classification on multimodal knowledge graphs [EB/OL]. (2020). https://arxiv.org/abs/2003.12383.
[27] Sun Rui, Cao Xuezhi, Zhao Yan, *et al*. Multi-modal knowledge graphs for recommender systems [C]//Proc of the 29th ACM International Conference on Information & Knowledge Management. 2020: 1405-1414.
[28] Oñoro-Rubio D, Niepert M, García-Durán A, *et al*. Answering visual-relational queries in Web-extracted knowledge graphs [C]//Proc of the 1st Conference on Automated Knowledge Based Construction. 2017.
[29] Deng Jia, Dong Wei, Socher R, *et al*. ImageNet: a large-scale hierarchical image database [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 248-255.
[30] Miller G A. WordNet: a lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41.
[31] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase [J]. Communications of the ACM, 2014, 57(10): 78-85.
[32] Bollacker K, Evans C, Paritosh P, *et al*. Freebase: a collaboratively created graph database for structuring human knowledge [C]//Proc of International Conference on Management of Data. 2008: 1247-1250.
[33] Lehmann J, Isele R, Jakob M, *et al*. DBpedia: a large-scale, multilingual knowledge base extracted from Wikipedia [J]. Semantic Web, 2015, 6(2): 167-195.
[34] Krishna R, Zhu Yuke, 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.
[35] Guo Yandong, Zhang Lei, Hu Yuxiao, *et al*. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition [C]//Proc of European Conference on Computer Vision. Cham: Springer, 2016: 87-102.
[36] Ferrada S, Bustos B, Hogan A. IMGpedia: a linked dataset with content-based analysis of Wikimedia images [C]//Proc of International Semantic Web Conference. Cham: Springer, 2017: 84-93.
[37] Xu Bo, Xu Yong, Liang Jiaqing, *et al*. CN-DBpedia: a never-ending Chinese knowledge extraction system [C]//Proc of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Cham: Springer, 2017: 428-438.
[38] Kannan A V, Fradkin D, Akrotirianakis I, *et al*. Multimodal knowledge graph for deep learning papers and code [C]//Proc of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2020: 3417-3420.
[39] Perona P. Vision of a visipedia [J]. Proceedings of the IEEE, 2010, 98(8): 1526-1534.
[40] Gong Dihong, Wang D Z. Extracting visual knowledge from the Web with multimodal learning [C]//Proc of the 26th International Joint Conference on Artificial Intelligence. 2017: 1718-1724.
[41] Chen Xinlei, Abhinav S, Abhinav G. Neil: extracting visual knowledge from Web data [C]//Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2013: 1409-1416.
[42] Vijayanarasimhan S, Grauman K. Large-scale live active learning: training object detectors with crawled data and crowds [J]. International Journal of Computer Vision, 2014, 108(1): 97-114.
[43] Li Hongzhi, Ellis J G, Ji Heng, *et al*. Event specific multimodal pattern mining for knowledge base construction [C]//Proc of the 24th ACM International Conference on Multimedia. New York: ACM Press, 2016: 821-830.
[44] Dong Xin, Gabrilovich E, Heitz G, *et al*. Knowledge vault: a Web-scale approach to probabilistic knowledge fusion [C]//Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 601-610.
[45] Yang Jianwei, Lu Jiasen, Lee S, *et al*. Graph R-CNN for scene graph generation [C]//Proc of European Conference on Computer Vision. Cham: Springer, 2018: 670-685.

[46] Gu Jiuxiang, Zhao Handong, Lin Zhe, *et al*. Scene graph generation with external knowledge and image reconstruction [C] //Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1969–1978.

[47] Akbari H, Karaman S, Bhargava S, *et al*. Multi-level multimodal common semantic space for image-phrase grounding [C] //Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12476–12486.

[48] Zareian A, Karaman S, Chang S F. Bridging knowledge graphs to generate scene graphs [EB/OL]. (2020). https://arxiv.org/abs/2001.02314.

[49] Wang Weitao, Wang Meng, Wang Sen, *et al*. One-shot learning for long-tail visual relation detection [C] //Proc of AAAI Conference on Artificial Intelligence. 2020: 12225–12232.

[50] Li Manling, Zareian A, Lin Ying, *et al*. GAIA: a fine-grained multimedia knowledge extraction system [C] //Proc of the 58th Annual Meeting of Association for Computational Linguistics: System Demonstrations. 2020: 77–86.

[51] Lowe G. Sift-the scale invariant feature transform [J]. International Journal of Computer Vision, 2004, 60(2): 91–110.

[52] Joachims T. Making large-scale SVM learning practical [R]. [S.l.]: Technische Universität Dortmund, 1998.

[53] LeCun Y. LeNet-5, convolutional neural networks [EB/OL]. 2015. http://yann.lecun.com/exdb/lenet.

[54] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] //Advances in Neural Information Processing Systems. 2012: 1097–1105.

[55] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014). https://arxiv.org/abs/1409.1556.

[56] Lin Min, Chen Qiang, Yan Shuicheng. Network in network [EB/OL]. (2013). https://arxiv.org/abs/1312.4400.

[57] Szegedy C, Liu Wei, Jia Yangqing, *et al*. Going deeper with convolutions [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 1–9.

[58] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al*. Deep residual learning for image recognition [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 770–778.

[59] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules [C] //Advances in Neural Information Processing Systems. 2017: 3856–3866.

[60] Ren Shaoqing, He Kaiming, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137–1149.

[61] He Kaiming, Gkioxari G, Dollár P, *et al*. Mask R-CNN [C] //Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 2961–2969.

[62] Mikolov T, Chen Kai, Corrado G, *et al*. Efficient estimation of word representations in vector space [EB/OL]. (2013). https://arxiv.org/abs/1301.3781.

[63] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735–1780.

[64] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need [C] //Advances in Neural Information Processing Systems. 2017: 5998–6008.

[65] Devlin J, Chang M W, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2018). https://arxiv.org/abs/1810.04805.

[66] Xie Ruobing, Liu Zhiyuan, Jia Jia, *et al*. Representation learning of knowledge graphs with entity descriptions [C] //Proc of AAAI Conference on Artificial Intelligence. 2016: 2659–2665.

[67] Wang Zikang, Li Linjing, Li Qiudan, *et al*. Multimodal data enhanced representation learning for knowledge graphs [C] //Proc of International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2019: 1–8.

[68] Frome A, Corrado G S, Shlens J, *et al*. DeViSE: a deep visual-semantic embedding model [C] //Proc of Neural Information Processing Systems. 2013: 2121–2129.

[69] Bordes A, Usunier N, Garcia-Duran A, *et al*. Translating embeddings for modeling multi-relational data [C] //Proc of Neural Information Processing Systems. 2013: 1–9.

[70] Zhang Hanwang, Kyaw Z, Chang S F, *et al*. Visual translation embedding network for visual relation detection [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 5532–5540.

[71] Pezeshkpour P, Chen Liyan, Singh S. Embedding multimodal relational data for knowledge base completion [C] //Proc of Conference on Empirical Methods in Natural Language Processing. 2018: 3208–3218.

[72] Chen Liyi, Li Zhi, Wang Yijun, *et al*. MMEA: entity alignment for multi-modal knowledge graph [C] //Proc of International Conference on Knowledge Science, Engineering and Management. Cham: Springer, 2020: 134–147.

[73] Moon S, Neves L, Carvalho V. Multimodal named entity recognition for short social media posts [C] //Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 852–860.

[74] Moon S, Neves L, Carvalho V. Multimodal named entity disambiguation for noisy social media posts [C] //Proc of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 2000–2008.

[75] Zhang Qi, Fu Jinlan, Liu Xiaoyu, *et al*. Adaptive co-attention network for named entity recognition in tweets [C] //Proc of AAAI Conference on Artificial Intelligence. 2018.

[76] 王会勇,论兵,张晓明,等. 基于联合知识表示学习的多模态实体对齐 [J]. 控制与决策, 2020, 35(12): 2855–2864. (Wang Huiyong, Lun Bing, Zhang Xiaoming, *et al*. Multi-modal entity alignment based on joint knowledge representation learning [J]. Control and Decision, 2020, 35(12): 2855–2864.)

[77] Wei Xi, Zhang Tianzhu, Li Yan, *et al*. Multi-modality cross attention network for image and sentence matching [C] //Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10941–10950.

[78] Zhu Junnan, Zhou Yu, Zhang Jiajun, *et al*. Multimodal summarization with guidance of multimodal reference [C] //Proc of AAAI Conference on Artificial Intelligence. 2020: 9749–9756.

[79] Garcia-Duran A, Niepert M. KBLRN: end-to-end learning of knowledge base representations with latent, relational, and numerical features [EB/OL]. (2017). https://arxiv.org/abs/1709.04676.

[80] Palmonari M, Minervini P. Knowledge graph embeddings and explainable AI [M] //Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges. 2020: 49.

[81] Guo Qingyu, Zhuang Fuzhen, Qin Chuan, *et al*. A survey on knowledge graph-based recommender systems [EB/OL]. (2020-03-03). https://arxiv.org/pdf/2003.00911.pdf.

[82] Zhang Fuzheng, Yuan N J, Lian Defu, *et al*. Collaborative knowledge base embedding for recommender systems [C] //Proc of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 353–362.

[83] Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines [J]. The Journal of Machine Learning Research, 2014, 15(1): 2949–2980.

[84] Wei Yunchao, Zhao Yao, Lu Canyi, *et al*. Cross-modal retrieval with CNN visual features: a new baseline [J]. IEEE Trans on Cybernetics, 2017, 47(2): 449–460.

[85] Antol S, Agrawal A, Lu Jiasen, *et al*. VQA: visual question answering [C] //Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 2425–2433.

[86] Zhu Yuke, Zhang Ce, Ré C, *et al*. Building a large-scale multimodal knowledge base system for answering visual queries [EB/OL]. (2015). https://arxiv.org/abs/1507.05670.

[87] Xu Huapeng, Qi Guilin, Li Jingjing, *et al*. Fine-grained image classification by visual-semantic embedding [C] //Proc of the 26th International Joint Conference on Artificial Intelligence. 2018: 1043–1049.

[88] Hao Maoxiang, Li Zhixu, Zhao Yan, *et al*. Mining high-quality fine-grained type information from Chinese online encyclopedias [C] //Proc of International Conference on Web Information Systems Engineering. Cham: Springer, 2018: 345–360.

[89] Su Weijie, Zhu Xizhou, Cao Yue, *et al*. VL-BERT: pre-training of generic visual-linguistic representations [EB/OL]. (2019-08-22). https://arxiv.org/abs/1908.08530.

[90] Qi Di, Su Lin, Song Jia, *et al*. ImageBERT: cross-modal pre-training with large-scale weak-supervised image-text data [EB/OL]. (2020-01-22). https://arxiv.org/abs/2001.07966.