计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：多模态知识图谱表示学习综述
作者：王春雷，王肖，刘凯
收稿日期：2023-05-15
网络首发日期：2023-07-31
引用格式：王春雷，王肖，刘凯．多模态知识图谱表示学习综述[J/OL]．计算机应用．
https://kns.cnki.net/kcms2/detail/51.1307.tp.20230728.1508.010.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

doi:10.11772/j.issn.1001-9081.2023050583

多模态知识图谱表示学习综述

王春雷[1,2]，王肖[1*]，刘凯[3]

(1.上海大学 人工智能研究院，上海 200444； 2.上海人工智能实验室，上海 200232；3.上海大学 计算机工程与科学学院，上海 200444)

(*通信作者电子邮箱 1046252251@qq.com)

摘 要: 综合对比了传统知识图谱表示学习方法，包括优缺点以及适用任务，分析得出传统的单一模态知识图谱无法很好地表示知识。因此，如何利用文本、图片、视频、音频等多模态数据进行知识图谱表示学习成为一个重要的研究方向。同时，详细分析了常用的多模态知识图谱数据集，为相关研究人员提供数据支持。在此基础上，进一步讨论了文本、图片、视频、音频等多模态融合下的知识图谱表示学习模型，并对其中各种模型进行了总结和比较。最后，总结提出了多模态知识图谱表示学习如何增强经典应用，包括知识图谱补全、问答系统、多模态生成和推荐系统在实际应用中的效果，并对未来的研究工作进行了总结展望。

关键词: 多模态知识图谱；表示学习；多模态融合；知识图谱补全；多模态生成

中图分类号:TP182 **文献标志码：** A

Multimodal knowledge graph representation learning: A review

WANG Chunlei[1,2], WANG Xiao[1*], LIU Kai[3],

(1.Institute of Artificial Intelligence, Shanghai University, Shanghai 200444, China;

2.Shanghai AI Laboratory, Shanghai 200232, China;

3.School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: By comprehensively comparing the learning methods of traditional knowledge map representation, including the advantages and disadvantages and applicable tasks, the analysis shows that the traditional single-modal knowledge map cannot represent knowledge well. Therefore, how to use multimodal data such as text, pictures, video, and audio for knowledge graph representation learning has become an important research direction. At the same time, the commonly used multimodal knowledge graph dataset is analyzed in detail to provide data support for relevant researchers. On this basis, the knowledge graph representation learning model under multimodal fusion of text, picture, video, and audio is further discussed, and various models are summarized and compared. Finally, the effect of multimodal knowledge graph representation on how to enhance classical applications, including knowledge graph completion, question answering system, multimodal generation and recommendation system, in practical applications is proposed, and the future research work is summarized and prospected.

Keywords: multimodal knowledge graph; representation learning; multimodal fusion; knowledge graph completion; multimodal generation

0 引言

“模态”(Modality)是德国理学家赫尔姆霍茨提出的生物学概念，指的是将多种感官信息进行融合，其中包括嗅觉、味觉、视觉、听觉和触觉等。随着互联网的普及和大数据的发展，不同模态的数据不断涌现，多模态知识图谱表示学习的发展在人工智能领域中引起了广泛的关注。目前大多数研究都是对单一模态的文本进行研究，然而，只有对不同模态数据的研究相互辅佐才能使得知识的表示更加完善。在人工智能领域中，多模态往往指感知信息，如图像、文本、语音和音频等，通过对这些多模态信息的综合理解，可以帮助人工智能更准确地理解外部世界。

知识图谱是由Google公司于2012年提出的概念[1]，是知识表示的一种方法。最初，它是以文本形式表示实体关系属性的三元组，但在机器描述世界和理解世界的能力方面存在缺陷。随着机器视觉和多模态学习的研究推进，研究人员发现结合视觉可以更好地为图谱中的实体进行相关表示学习工作，而多模态研究的最早例子之一是视听语音识别(Audio-Visual Speech Recognition, AVSR)[2]。

随着多模态研究的逐渐深入，从单一模态的表示学习逐渐发展为多模态的表示学习。早期的知识图谱表示学习方法

收稿日期: 2023-05-15； **修回日期**: 2023-06-23； **录用日期**: 2023-06-28。

基金项目: 国家杰出青年科学基金资助项目(62225308)。

作者简介: 王春雷(1977—)男，江苏盐城人，研究员，博士，CCF 会员，主要研究方向：知识图谱与认知智能、情感计算与情绪识别；王肖(1998—)，男，安徽亳州人，硕士研究生，主要研究方向：多模态知识图谱；刘凯(1996—)，男，江西抚州人，硕士研究生，主要研究方向：无人艇领域的知识图谱构建。

主要学习基于实体和关系的结构信息，忽略了其他模态数据类型的实体知识。近年来的相关研究工作表明，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[3]。同样，在关系提取任务中，附加图像通常会大幅提高属性和关系的性能。因此，融合各类模态的数据可以更好地对知识进行表示学习，以推动更多相关的典型任务。多模态知识图谱表示学习与应用的研究已经成为必然的趋势。

当前针对多模态知识图谱表示学习与应用的综述大多停留在传统的单一模态下，缺乏系统性总结该领域的发展历程、研究进展、典型应用和未来挑战等方面。因此，本文将从传统的知识图谱表示学习方法入手，对单一模态方法进行总结分析，并得出结论：目前的知识图谱表示学习模型仅在单一模态上进行，没有充分利用涌现的多模态数据。本文调研分析了近年来多模态数据在知识图谱表示学习中的应用方法，发现综合使用多种模态可以弥补单一模态知识表示的不足。此外，本文还分析了多模态知识图谱表示学习在知识图谱补全、问答系统、多模态生成和推荐系统场景中的应用，并从大规模数据处理、数据多样性和数据质量、数据缺失与任务联合、非监督学习、可解释性、评价体系以及人工智能生成内容(Artificial Intelligence Generated Content, AIGC)与知识图谱等方面展望了该领域的研究工作。

本文旨在对多模态知识图谱表示学习领域的研究进展进行系统的回顾与总结。文章将多模态知识图谱表示学习分为四个方面：1)文本信息用于知识图谱表示学习；2)图片信息用于知识图谱表示学习；3)音视频信息用于知识图谱表示学习；4)多模态信息用于知识图谱表示学习。

1 相关研究

1.1 知识图谱表示学习

将表示学习应用于知识表示中是传统知识图谱主流的研究方向，这样的表示学习方法又称之为知识表示学习或知识图嵌入(embedding)。由于深度学习的高速发展，传统知识图谱表示学习主要目的是将知识图谱中的实体和关系嵌入到连续的向量空间中[4]，以便能够输入到深度学习模型之中更好的处理下游任务，同时保持知识图谱的结构信息不变。本章主要介绍在传统知识图谱表示学习中使用的各类模型，通过区分不同的编码方法，简单地将知识图谱嵌入方法分为以下三种类型：基于翻译的模型、因子分解模型和神经网络模型。

基于翻译的模型是在知识图谱嵌入中使用较多的方法，不同的翻译模型象征着不同的嵌入空间和翻译操作。2013 年提出的 TransE[5]是经典的欧氏空间表示学习模型，受 word2vec 的影响，希望 $\boldsymbol{h}+\boldsymbol{l}\approx\boldsymbol{t}$，它的能量函数和损失函数分别如式(1)、(2)所示：

$$E(\boldsymbol{h},\boldsymbol{l},\boldsymbol{t})=\|\boldsymbol{h}+\boldsymbol{l}-\boldsymbol{t}\| \tag{1}$$

$$\tau=\sum_{(\boldsymbol{h},\boldsymbol{l},\boldsymbol{t})\in S}\sum_{(\boldsymbol{h}',\boldsymbol{l}',\boldsymbol{t}')\in S'_{(\boldsymbol{h},\boldsymbol{l},\boldsymbol{t})}}[\gamma+d(\boldsymbol{h}+\boldsymbol{l},\boldsymbol{t})-d(\boldsymbol{h}'+\boldsymbol{l},\boldsymbol{t}')]_{(2)}$$

其中：$\boldsymbol{h}$，$\boldsymbol{l}$，$\boldsymbol{t}$ 分别表示头实体、关系、尾实体；γ 是大于 0 的超参数。但是该模型不能应对单个空间同时表征实体和关系存在不足的问题。于是在 2015 年出现了 TransR[6]模型，与 TransE 模型相比，TransR 采用了不同的表征空间分别表示实体和关系，将实体投影向量定义为 $\boldsymbol{h}_{\mathrm{r}}=\boldsymbol{h}\boldsymbol{M}_{\mathrm{r}}$，$\boldsymbol{t}_{\mathrm{r}}=\boldsymbol{t}\boldsymbol{M}_{\mathrm{r}}$，能量函数定义为 $\mathrm{E}(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})=\|\boldsymbol{h}_{\mathrm{r}}+\boldsymbol{r}-\boldsymbol{t}\|$。其中：$\boldsymbol{M}_{\mathrm{r}}$ 是关系的投影矩阵。随后出现了很多 Trans 系列模型，包括 TransH[7]、TransD[8]、TransA[9]，其中 TransA 基于 TransE 将原来的损失函数中的距离度量改成了马氏距离。TransE 及其变体模型的原理示意如图 1 所示。

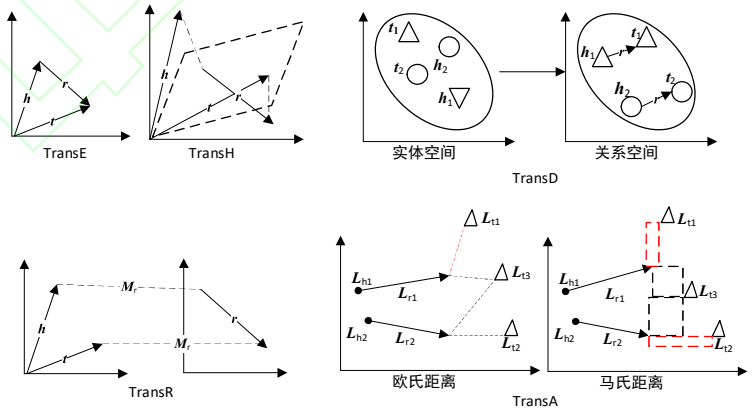

图1 Trans 系列模型原理

Fig. 1 Trans series model schematic

基于因子分解的模型是一类将实体和关系表示为低维因子的模型。文献[10]中提出一种基于三向张量因式分解的关系学习方法 RESCAL，结构如图 2(a)所示。与其他张量方法不同，该方法能够通过模型的潜在组件执行集体学习，并提供计算因子分解的有效算法，即 $\boldsymbol{X}_{\mathrm{k}}\approx\boldsymbol{A}\boldsymbol{R}_{\mathrm{k}}\boldsymbol{A}^{\mathrm{T}}$，其中 $\boldsymbol{A}$ 是矩阵表示全局实体潜在空间，$\boldsymbol{R}_{\mathrm{k}}$ 表示每个潜在成分的关系作用，最后化简为一个优化任务如式(3)所示：

$$\min_{\boldsymbol{A},\boldsymbol{R}_k} loss(\boldsymbol{A},\boldsymbol{R}_k)+\mathrm{reg}(\boldsymbol{A},\boldsymbol{R}_k) \tag{3}$$

其中损失函数如式(4)所示：

$$\mathrm{loss}(\boldsymbol{A},\boldsymbol{R}_k)=\frac{1}{2}\sum_k||\boldsymbol{\chi}_k-\boldsymbol{A}\boldsymbol{R}_k\boldsymbol{A}^{\mathrm{T}}||_{\mathrm{F}}^2 \tag{4}$$

在这基础之上，文献[11]中又提出双线性结构的隐因子模型(Latent Factor Model, LFM)，可以在不同的关系中共享稀疏的潜在因素。TuckER[12]是一个相对简单但强大的线性模型(如图 2(b))，基于知识图三元组的二元张量表示的 TuckER 分解，该模型对知识图谱的张量表示进行链接预测，其中 $\boldsymbol{E}=\boldsymbol{A}=\boldsymbol{C}\in\boldsymbol{R}^{n_e\times d_e}$，$\boldsymbol{R}=\boldsymbol{B}\in\boldsymbol{R}^{n_r\times d_r}$，其中 $\boldsymbol{E}$ 为实体嵌入矩阵，$\boldsymbol{R}$ 为关系嵌入矩阵，n_e 和 n_r 表示实体和关系数量，d_e 和 d_e 表示实体和关系嵌入维数。

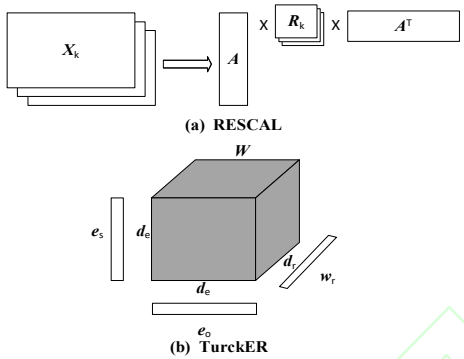

图2 两种模型架构

Fig. 2 Two model architecture diagrams

随着深度学习的发展，基于神经网络的模型也逐渐涌现，多层感知机(Multilayer Perceptron, MLP)将实体和关系一起通过全连接网络进行编码，然后使用第二层带有 sigmoid 的神经元打分。早在 2011 年 Bordes 等[13]将线性神经网络用于知识图谱表示学习。随后 Yang 等[14]提出 DistMult 简化版的双线性公式用于处理大型知识图谱，其将实体表示和关系表示分别用不同的形式表示，用 $\boldsymbol{X}_{e_1}$ 和 $\boldsymbol{X}_{e_2}$ 表示实体 e_1 和 e_2 的输入向量，则学习到的实体表示如式(5)所示：

$$\boldsymbol{y}_{e_1}=\mathrm{F}(\boldsymbol{W}\boldsymbol{x}_{e_1}),\boldsymbol{y}_{e_2}=\mathrm{F}(\boldsymbol{W}\boldsymbol{x}_{e_2}) \tag{5}$$

W 表示第一层投影矩阵，关系表示则用评分函数的形式表示出来，即(6)和(7)所示：

$$\boldsymbol{g}_{\mathrm{r}}^{\alpha}(\boldsymbol{y}_{e_1},\boldsymbol{y}_{e_2})=\boldsymbol{A}_{\mathrm{r}}^{\mathrm{T}}\begin{pmatrix}\boldsymbol{y}_{e_1}\\ \boldsymbol{y}_{e_2}\end{pmatrix} \tag{6}$$

$$\boldsymbol{g}_{\mathrm{r}}^{\mathrm{b}}(\boldsymbol{y}_{e_1},\boldsymbol{y}_{e_2})=\boldsymbol{y}_{e_1}^{\mathrm{T}}\boldsymbol{B}_{\mathrm{r}}\boldsymbol{y}_{e_2} \tag{7}$$

其中 $\boldsymbol{A}_{\mathrm{r}}$ 和 $\boldsymbol{B}_{\mathrm{r}}$ 是特定关系参数，公式(6)为基本线性变换，(6)与(7)结合即为双线性变换，基于语义模型的 DistMult 的评分函数如图 3 所示。

基于卷积神经网络的知识图谱嵌入，文献[15]中提出了 ConvE 模型，使用嵌入的 2D 卷积来预测知识图中缺失链接的模型，其将实体和关系输入卷积层和全连接层最后由 2D 卷积来进行特征分数评分，评分函数如式(8)所示：

$$\psi_{\mathrm{r}}(\boldsymbol{e}_{\mathrm{s}},\boldsymbol{e}_{\mathrm{o}})=\mathrm{F}(\mathrm{vec}(\mathrm{F}([\overline{\boldsymbol{e}}_{\mathrm{s}};\overline{\boldsymbol{r}}_{\mathrm{r}}]*\boldsymbol{w})\boldsymbol{W})\boldsymbol{e}_{\mathrm{o}} \tag{8}$$

其中：$\boldsymbol{r}_{\mathrm{r}}$ 是关系参数，$\boldsymbol{e}_{\mathrm{s}}$、$\boldsymbol{e}_{\mathrm{o}}$ 分别表示主体和对象的嵌入，$\boldsymbol{w}$ 代表滤波器。ConvE 与 DistMult 和关系图卷积网络(Relational Graph Convolutional Network, R-GCN)具有相同的性能，参数分别减少了 8 倍和 17 倍。ConvKB[16]对 ConvE 进行了优化操作，将 ConvE 中的大小调整操作更改为串联，保留了转移特征。最终，在两个数据集上，ConvKB 获得了比 ConvE 更好的分数。在 2020 年，文献[17]中提出了一种用于知识图谱嵌入的具有联合局部全局结构信息的关系感知网络(Relation-aware Inception Network With Joint Local-global structural Information For Knowledge Graph Embedding ,ReInceptionE)，模型将 ConvE 和 KBGAT 的优点结合在一起，形成了一个具有联合局部全局结构信息的关系感知启始网络，可用于知识图嵌入。ReInceptionE 的一般模型框架如图 4 所示。

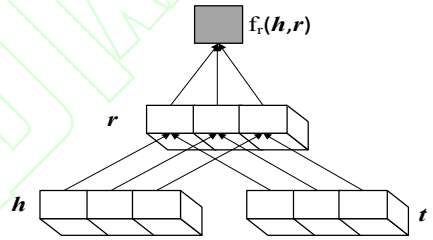

图3 DistMult 评分函数原理

Fig. 3 DistMult scoring function schematic

卷积神经网络大多用于处理图片格式信息，缺少时序信息。2019 年提出的递归跳过网络(Recurrent Skipping Network, RSN)[18]模型，它利用跳过机制来弥合实体之间的差距。RSN 将循环神经网络(Recurrent Neural Network, RNN)与残差学习相结合，以有效地捕获知识图谱(Knowledge Graph,KG)内部和 KG 之间的长期关系依赖性。RSN 的基础模型框架如图 5 所示。

Haithem 等[19]基于一阶和子图感知接近的原理，提出了一种上下文感知知识图嵌入方法(Dilated Recurrent Neural Network, DRNN)，并定义了一种推荐算法，以根据目标用户的上下文提供最高评级的服务。相较于 RNN，RSN 多了残差链接，有效地解决了深度神经网络退化问题。而 DRNN 在深度方面超过了 RNN，在处理更大数据时更占优势。然而，通过对模型框架的分析，我们发现基于 RNN 的模型都有一个通病，即不能处理长序列问题。

Schlichtkrull 等[20]提出了关系图卷积网络(Relation Graph Convolution Network, R-GCN)模型，该模型使用特定关系转换来建模图谱的方向特征，之后学术界一直在不断的优化 GNN 的性能。2021 年提出的关系感知图注意力网络(Relation Aware Graph Attention Network,RAGAT)模型[22]为不同的关系构建了独立的消息函数，旨在利用知识图的异构特性。同

年，Li 等[22]提出了一种基于注意力机制的异构图神经网络(Graph Neural Network, GNN)框架，该方法不仅聚合了来自不同语义方面的实体特征，还为它们分配了适当的权重。不论是 R-GCN 还是后续优化的模型，都是基于 GNN 进行扩展。通过观察图 6 中 R-GCN 模型原理可以看出,GNN 的优点在于能够任意深度表示其附近的信息，但缺点也显而易见，即难以收敛并且扩展性比较差。

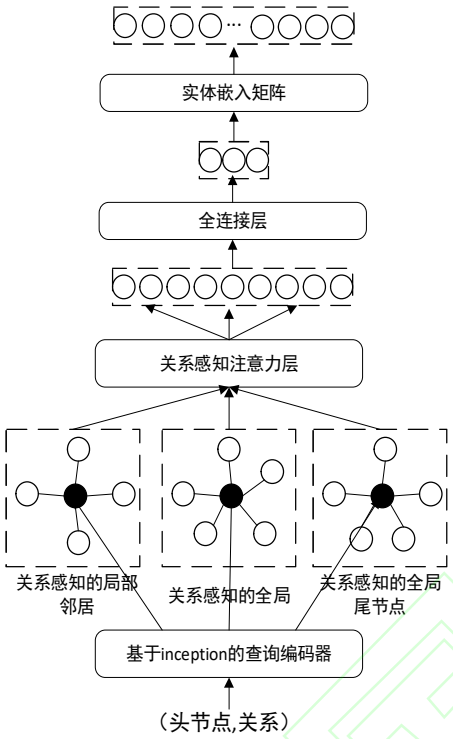

图4 RelnceptionE 模型原理

Fig. 4 RelnceptionE model schematic

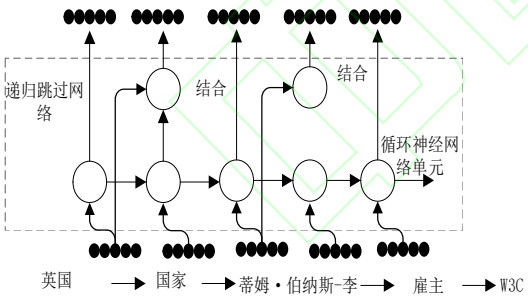

图5 RSN 模型原理

Fig. 5 RSN model schematic

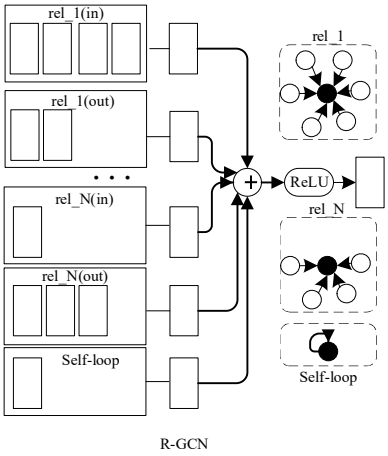

图6 R-GCN 模型原理

Fig. 6 R-GCN model schematic

2017 年 Google 在论文中提出了 Transformer[23]模型，其中使用的自注意力机制取代了自然语言处理(Natural Language Processing,NLP)任务中常用的 RNN 网络结构。随之，BERT[24]模型的出现彻底改变了 NLP 的现状。2019 年，上下文知识图谱嵌入(Contextualized Knowledge Graph Embedding,CoKE)模型[25]将序列作为输入，并利用 Transformer 编码器获得上下文化表示，这些表示自然地适应于输入，捕捉了实体和关系的上下文含义。同年 liang 等[26]提出了一个名为 Knowledge Graph Bidirection Encoder Representations from Transformer(KG-BERT)的新框架对这些三元组进行建模，并采用了预训练模型的思想，其中 Bidirection Encoder Representation from Transformers(BERT)被用作实体和关系的编码器。随后提出的基于源实体邻域的层次转换器(Hierarchical Transformer Model Based on A Source Entity`s neighborhood,HittER)模型[27]采用两个不同的 Transformer 块组：底部块从源实体的局部领域中提取每个实体关系对的特征，而顶部块则从底部块的输出中聚合关系信息。随着预训练模型的流行，现在越来越多的预训练模型被用于图谱表示学习。2022 年，Alam 等[28]提出了一个观点，认为仅仅限于实体之间的结构和联系可能会对模型产生负面影响，他们提出了利用预训练的语言模型，在嵌入模型的学习过程中包含文本知识，以弥补这一差距。表 1 对上述模型进行了总体总结，包括优点、缺点以及适用任务。

表1 知识图谱表示学习模型

Tab. 1 Summary of Knowledge Graph Representation Learning Models

模型基础	具体模型	优点	缺点	适用任务
基于翻译	TransE	简单直观，计算复杂度不高	不适用于复杂关系建模	链接预测
	TransH	解决了 TransE 不适用于复杂关系建模，让实体在多关系的情况下仍能获得合适的表示	实体关系在同一个空间，某些场景不适用	链接预测，三元组分类，事实提取

	TransR	表示空间进行了改进，考虑了实体和关系投影在不同空间中的情况	复杂度较高，参数急剧增加。并且会使得投影矩阵仅仅和关系有关	链接预测，三元组分类，关系事实提取
	TransD	将投影矩阵分解成两个解决了在同一关系下头，尾实体是相同的投影矩阵的问题	模型训练过程复杂，可能不如其他模型，需要较大的计算资源	三元组分类，链接预测
因子分解	RESCAL	该方法能够通过模型的潜在组件执行集体学习，并提供计算因子分解的有效算法	模型参数较多，复杂性较高，计算需要很大的资源。并行性不高	链接预测，三元组分类
	TuckER	模型表达能力强大，可以将正例与反例完全区分开，模型是线性的相对简单	不适用于一般预测任务，模型方程和优化算法是针对单个任务单独导出，适用性不强	链接预测
CNN	ConvE	表现力强、参数效率高	不足以捕获输入实体和关系的交互，文章中模型仅在输入实体和关系的邻接矩阵中建模交互	链接预测，实习关系预测
	ConvKB	在 ConvE 上做了优化操作，将 ConvE 中的大小调整操作更改为串联，保留了转移特征	仅独立考虑三元组，无法覆盖三元组周围的局部邻居中固有的复杂和隐藏信息	链接预测
	ReInceptionE	综合考虑 ConvE 交互次数受限提出 Inception 增强交互，基于 KBGAT 的缺点，提出充分利用局部和全局结构信息的嵌入模型	超参数太多，对于不同的数据集需要重新进行调参	链接预测
RNN	RSN	通过跳过机制弥合实体之间差距，增加残差学习以有效的捕获 KGs 内部和 KGs 之间的长期关系依赖性	需要大量的计算资源和高质量的数据训练，并且在实际应用中，模型的缺点会比它的优点更为突出	实体对齐，知识图谱补全
	DRNN	基于一阶和子图感知邻近的原理，提出了一种上下文感知知识图嵌入，提高了准确性和可扩展性	虽然减少了大量参数，但是也带来了信息损失，在较长的序列中会造成梯度逐渐消失或者爆炸	上下文感知推荐
GNN	R-GCN	增加了聚合关系的维度，使节点的聚合操作变成一个双重聚合的过程，以此增加表征知识的能力	对于学习参数和学习关系没有轻重之分，忽略了不同关系之间的区别	链接预测，实体分类
GNN+ attention	RAGAT	优化了不同关系之间的区别，增加注意力机制，充分利用知识图的异构特性	参数量多，训练方式没有利用高阶邻居，容易发生过度平滑	知识图谱补全
Transformer	CoKE	CoKE 允许每个实体或关系使用单个静态表示，提出使用 Transformer 编码器获得上下文化表示。CoKE 学习动态适应每个输入序列的 KG 嵌入，捕捉实体和其中关系的上下文含义	模型太大太深导致学习参数过多，需要更多的计算资源	链接预测，路径查询
BERT	KG-BERT	通过增加上下文信息的嵌入来捕获丰富的语义知识	模型复杂性变高,虽然模型效果提高,但是可解释性大大下降	三元组分类
Transformer	HittER	底部块提取源实体的本地邻域中每个实体关系对的特征，顶部块从底部块的输出中聚合关系信息,更好的提取丰富的语义知识	可解释性差,模型通过引入丰富的上下文信息因为其中包含虚假信息,会降低原实体的信息,并且可能导致过拟合	链接预测

1.2 多模态知识与多模态知识图谱

根据表 1 可以看出，知识图谱表示学习主要应用于链接预测、三元组分类和知识图谱补全等场景。实际上，除了文

本和结构化数据，视觉和听觉数据，如图片、视频和音频也可以作为数据源[29]。这些多种模态数据形成的知识被称为多模态知识。目前，对于多模态知识的需求越来越大，这为突破传统知识图谱应用的瓶颈提供了机会[30-31]。

传统的知识图谱可以定义为$G=\{\boldsymbol{E},\boldsymbol{R},\boldsymbol{F}\}$，其中$\boldsymbol{E}$、$\boldsymbol{R}$、$\boldsymbol{F}$分别表示实体关系事实的集合。客观世界的事实三元组可以表示为$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$，其中$\boldsymbol{h}$、$\boldsymbol{r}$、$\boldsymbol{t}$分别表示头实体，关系和尾实体。根据文献[32]中知识图谱的定义，多模态知识图谱可以看成是将实体与非文本形式的数据(如图像、视频等)相关联的知识图谱。现有的多模态知识图谱主要采用两种不同方式来表示多模态信息：一种是将多模态信息作为实体的属性值，另一种是将多模态信息表示为实体，并通过特定类型的关系进行关联。简而言之，多模态知识图谱就是包含多种模态知识的知识图谱。

相较于传统的知识图谱，多模态知识图谱能够更全面地描述现实世界中的概念和事物，同时能够更好地支持音频、图像、视频等非文本形式的信息处理和理解。这些不同的数据源在描述同一对象时相互补充和增强，从而提高知识图谱表示学习相关任务相对于单一模态模型的性能，并推动机器对真实数据场景的感知能力[33]。

1.3 多模态知识图谱数据集

在具体介绍多模态知识图谱表示学习之前，先简要介绍目前存在的多模态知识图谱数据集以及相关系统。从仿生学的角度来看，人类从外界获取信息并且形成知识的过程是一个多方面信息融合的过程。因此，机器在形成属于自己的知识过程种，也必然需要进行多模态数据融合。为了使机器能够像人一样具备理解和解释能力，以有效地处理不同模态的信息，我们需要建立基于超大规模结构化知识基础的多模态知识图谱。多模态知识图谱可以被理解为机器提供的抓门的超大规模结构化知识基础。

在过去的很长一段时间里，涌现了各个领域的知识图谱。早期的知识图谱数据集大多是单一模态数据。随着需求的变化，越来越多的多模态知识图谱数据集出现，以便于学者进行后续的多模态任务研究。因此，知识图谱的多模态化及其应用正在蓬勃发展[34-35]。

DBpedia[36]最早是作为语义网项目提出，主要从维基百科中提取结构化信息，它允许对于维基百科的数据集进行复杂的查询，并且将 Web 上的其他数据连接到维基百科数据。DBpedia 包含的数据包括但是不限于文本、图谱、文本描述等，是目前知识图谱研究领域的核心数据集，包含超过 260 万个实体，提供了约 47 亿条数据。

Never Ending Image Learner(NEIL)[37]是一个通过每天 24 h、每周 7 d 运行的计算机程序从互联网数据中自动提取视觉知识的数据集。NEIL 使用半监督学习算法，共同发现常识关系，并标记给定视觉类别的实例。这项工作旨在通过最小化人为标注的努力来创建世界上最大的视觉结构化知识库。该知识库包含 1152 个对象类别、1034 个场景类别和 87 个属性的本体。在这个过程中，NEIL 发现了 1700 多个关系，并标记了超过 400000 个可视实体。

Wikidata[38]是一个自由开放的数据库，可供人类和机器同时阅读。它是结构化数据的集中存储，为 Wikimedia 项目提供支撑，包括 Wikipedia、Wikivoyage、Wikionary、Wikisource 等，其中包含超过 1400 万个包含 URI 的实体。

IMGPedia[39]与前面提到的 DBpedia 和 Wikidata 有本质上的差别，该数据集更强调在原有的知识图谱文本信息中添加非结构化的图片信息。其中包含 1500 万张图片的视觉内容描述，4.5 亿个视觉相似性关系，可以称之为多模态知识图谱的先例。

ImageGraph[40]提出的目的是通过从 Web 提取的知识图中回答视觉关系查询的新的机器学习方法。为了实现这个目的，创建了该图谱，其中包含 1330 个关系类型、14870 个实体和 829931 个从 web 上抓取的图像。使用像 ImageGraph 这样的可视化关系知识图谱，可以引入新的概率查询类型。

多模态知识图谱(Multimodal Knowledge Graphs, MMKG)[41]以传统的文字知识图谱 Freebase 为基础，使用常用的子集 Freebase15k 进行构建。该多模态知识图谱是三个知识图谱的集合，包含了所有实体的数字特征和图像链接，并且通过 SameAs 进行实体对齐，其中包含了 15000 多个实体，1300 多种关系，13000 多张图片。实现了三个知识图谱之间的链接。

GAIA[42]是第一个全面的开源多媒体知识提取系统。该系统将来自各种来源和语言的大量非结构化、异构多媒体数据流作为输入，并创建了一个连贯且结构化的知识库。它基于丰富而细粒度的本体对实体、关系和事件进行索引。其中包含 457000 个实体，67000 个三元组，38000 个事件信息。

DCC (Deep Code Curator)[43]提出的原因是深度学习模型的表达和实现非常复杂。对于希望开发新方法的研究人员或希望使用现有方法解决现实问题的从业者来说，跟上所有最新出版物及其附带的源代码都是一个巨大的挑战。该研究的主要目标是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，以解决这个问题。该数据集中包含 539 个出版物，7999 个 DL 实体，174 张 DL 图，256 个 DL 资源库。

Richpedia[44]的目标是通过向 Wikidata 中的文本实体分配多个不同图像，为用户提供全面的多模态知识图谱。该数据集基于 Wikipedia 中的超链接和描述，通过资源描述框架链接(视觉语义关系)在图像实体之间建立链接。该数据集中包含 30638 个实体，2883162 张图片，119669570 个三元组。

VisualSem[45]是一个包含 90 万个节点，130 万关系和 938k 张图像的视觉和语言高质量知识图谱。VisualSem 的节点表示概念和命名实体，包括多个高质量的说明性图像，以及多

达 14 种不同语言的注释。VisualSem 中的节点链接到维基百科文章、WordNet 同义词集和 ImageNet 中的高质量图像。

KgBench[46]的提出是针对图形神经网络和其他机器学习模型，为基于关然而，该方面数据集有限。因此，他们在 RDF 编码的知识图上引入了一组新的基准任务。该数据集主要关注节点分类，因为这种设置无法仅通过节点嵌入模型来解决。对于每个数据集，作者提供了 1000 个实例的测试和验证集，并且所有的数据集都以 CSV 的格式打包。

知识图谱的发展历程从单一模态转变到多模态，离不开多模态数据抽取的快速发展。从最初只使用文本构建知识图谱，到现在将图片，视频，音频等用于知识图谱构建，经历了翻天覆地的变化。如今，一些超大规模的多模态知识图谱和多模态数据抽取系统也在不断的增加。

2 多模态知识图谱表示学习

知识图谱在人工智能领域有着举足轻重的地位。知识表示学习是研究知识图谱的基本任务之一，也被称为知识图嵌入。在过去的几十年里，大多数图嵌入都集中于研究实体和关系之间的结构知识，并且仅限于使用单一的模态知识进行学习，这些方法仅仅在嵌入层面上表现较为简单，并忽略了三元组之间的多模态信息。随着多模态数据的兴起，越来越多的工作开始利用多模态数据进行增强表示学习。实现证明，实体可以拥有多种模态的知识，如图片、文本描述、音频、视频等[47]。然而，传统方法忽略了这些多模态信息。例如，图 7(a)、(b)是文献[47]中一些多模态知识图的示例，说明了图像可以在知识图中编码显性知识，如(椅子，有部分，腿)，也可以包含隐性知识，如“蝴蝶与花朵高度相关”。

(a) 显性知识 (b) 隐性知识

图7 显性与隐性知识

Fig. 7 Explicit and implicit knowledge

多模态知识图谱中的实体通常包含文本和图片信息，而某些特殊实体还可能包括视频或者音频等信息。相较于单一模态的知识图谱，多模态知识图谱更加生动，信息更加细致。本小结主要介绍在知识图谱表示学习中引入不同模态信息的研究。

2.1 文本信息用于知识图谱表示学习

在众多融合多模信息进行知识图谱表示学习的工作中大多数以文本，图片，视频模态为主。早期对于知识图谱的拓展从增加单词信息开始，Wang 等[48]先从实体与单词入手，提出一种将实体和单词联合嵌入同一连续向量空间的新方法。嵌入过程试图保持知识图中实体与文本语料库中单词的一致性之间的关系，实体名称和维基百科锚被用来对齐实体和单词在同一空间中的嵌入。单纯的单词辅助映射过于简单，后续 Toutanova 等[49]在 Riedel 等[50]的工作基础上提出了一个模型改进方法(CONV)，该方法在先前模型基础上捕获文本关系的组成结构，并联合优化实体，知识库和文本关系表示，共同学习知识库和文本关系的连续表示以此提升链接预测任务的性能。2016 年 Wang 等[51]指出 TransE、TransR、TransH 系列模型无法很好的解决非一对一关系，提出利用外部文本上下文信息辅助知识图谱表示学习，文本增强方法如图 8 所示。

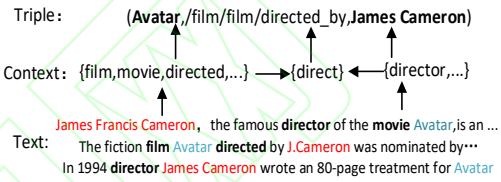

图8 文本增强简单说明

Fig. 8 Simple illustration of text-enhanced , method

基于此网络，定义实体与关系的文本上下文，并将其融入到知识图谱中；最后利用翻译模型对实体与关系的表示进行学习，可以更好的处理 1 对 N，N 对 1 和 N 对 N 的关系。同年 Xie 等[52]扩展了 TransE 模型提出 Description-Embodied Knowledge Representation Learning(DKRL)模型，直接使用卷积编码器从实体描述中学习表征，为了利用事实三元组和实体描述并能够处理零样本场景，提出了基于结构和基于描述的表示。定义 $\boldsymbol{h}$，$\boldsymbol{t}$，$\boldsymbol{r}$ 分为别实体集合、实体、关系，在对结构表示时使用的仍为 TransE 模型的能量函数：$E_S(\boldsymbol{h},\boldsymbol{t},\boldsymbol{r})=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|$，在对文本描述进行表示时使用两种编码来进行这种表示。DKRL 模型使用的能量函数如式(9)，(10)所示：

$$E = E_S + E_D \tag{9}$$

$$E_D = E_{DD} + E_{DS} + E_{SD} \tag{10}$$

其中：$E_{DD}=\|\boldsymbol{h}_d+\boldsymbol{r}-\boldsymbol{t}_d\|$，$E_{DS}=\|\boldsymbol{h}_d+\boldsymbol{r}-\boldsymbol{t}_s\|$，$E_{SD}=\|\boldsymbol{h}_s+\boldsymbol{r}-\boldsymbol{t}_d\|$，$h_d$ 和 t_d 分别表示基于文本描述构建的头和尾，E_S 表示基于结构的能量函数，E_D 表示基于描述的能量函数。该模型使用两个编码器来构建文本的表示，分别是词袋编码器和卷积神经网络。两部分架构图如图 9(a)、(b)所示，左边为连续词袋(Continuous Bag-of-Words,CBOW)编码器，右边为卷积神经网络编码器。

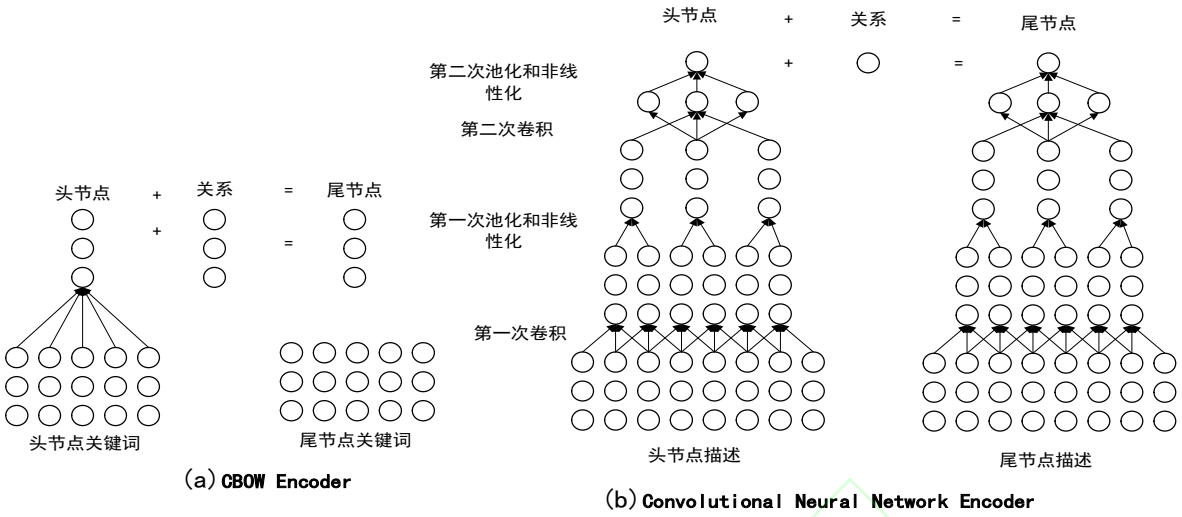

图9 DKRL 模型原理图

Fig. 9 DKRL model schematic

同一时期 Xie 等[53]提出 Type-embodied Knowledge Representation Learning(TKRL)模型，该模型将实体分成具有不同类型的多个表示以此增加了实体类型信息来增强表示。随着研究的深入 Xiao 等[54]提出语义空间投影模型(Semantic Space Projection,SSP)将三元组和文本描述映射到一个空间以此来建模他们之间的强相关性，该模型从符号三元组和文本描述中联合学习，使用文本描述来发现语义相关性并提供精确的语义嵌入。An 等[55]在前人的基础之上引入关系提及和实体描述之间的相互关注机制，以学习更准确的文本表示，从而进一步改进知识图表示。这是一种精确的文本增强知识图表示学习方法，该方法可以通过利用额外的文本描述信息来表示不同三元组中关系及实体的不同表示。与 DKRL 相同，Chen 等[56]提出 KDCoe 模型都旨在为了让新的实体有更好的表现，该方法利用弱对齐的多语言 KG 进行半监督的跨语言学习，使用实体描述，通过创建新的语言间链接在多语言知识图谱的实体之间建立对齐。随着语言模型的崛起，KG-BERT[26]模型使用 BERT 对文本模态信息进行编码，将三元组实体和文本关系描述输入模型进行知识表示学习以此来提高三元组分类，链接预测等任务的精准度。

文本信息用于知识图谱表示学习总结如表 2 所示。

表2 文本信息用于知识图谱示学习

Tab. 2 Text information is used for knowledge graph learning

模型	基础数据集	适用任务	效果说明	改进分析
pTransE	Freebase[57], Wikipedia[58],NY Times	三元组分类,改进关系抽取,类比推理任务	与 TransE 与 word2vec(Skip Gram)相当或稍好，主要为了解决推理新关系事实	基于 TransE 进行相应改变，实体和单词联合嵌入并对同一空间中实体和单词进行对齐以达到更到的嵌入效果
CONV	FB15K-237[15]	文本推理	对具有文本提及的实体对有更大的改进，提高了链接预测性能	在 Riedel 等[50]的子结构上进行建模并在相关依赖路径之间共享参数，使用统一的损失函数学习实体和关系表示，与基本模型的区别在于文本提及的参数化
TEKE	FB13[59],WN18[60],FB15K[5],WN11[8],	链接预测,三元组分类	链接预测任务中的 Hits@10 指标明显且一致地优于其他基线，TEKE_H 比 TEKE_R 表现略好。三元组分类任中，TEKE_E 和 TEKE_H 始终优于其他基线，其中上述三个模型是基于 TransE，TransR，TransH 实现的不同 TEKE 方法	基于 TransE、TransH 和 TransR 但在不同的优化目标上实现，该方法构建了一个基于实体注释文本语料库的共现网络用于将知识和文本信息连接在一起
DKRL	FB15K[5],FB20K(基于FB15K)	知识图谱补全,实体类型分类	DKRL 在现有的基于翻译的模型中 zero-shot 场景下达到最好效果。	基于 TransE 进行改变，实体的嵌入既对相应的事实三元组进行建模，也对其描述进行建

			在实体分类任务中 zero-shot 场景下 DKRL 也有很大优势	模，同时利用事实三元组和实体描述
TKRL	FB15K[5],FB15K+(基于 FB15K)	知识图谱补全,三元组分类	在知识图谱补全任务和三元组分类中比 TransE 和 TransR 效果更好，增强相同类型实体之间的差异对三元组分类任务有明显提高	基于 DKRL 增加多层实体类型进行模型的改进，模型设计了两个类型编码器来建模分层结构
SSP	FB15K[5],FB20K(基于 FB15K),WordNet[61],Freebase[62]	知识图谱补全,实体分类	在知识图谱补全任务中 SSP 优于其他基线，在实体分类任务中 SSP 达到最优。与 TransE 和 DKRL 相比有着更高的精度表现	基于 TransE 在构建三元组和文本描述之间交互过程中增加因子平衡两个部分，同时进行主题模型和嵌入模型以此来共同学习语义和嵌入
AATE, ATE	WN11[8],WN18[63],FB13[59],FB18K,Wikidata[58]	链接预测,三元组分类	链接预测任务中 AATE 和 ATE 完全优于所有基线。AATE 模型比 ATE 取得更好的结果。三元组分类任务中 AATE 模型同样优于所有基线，AATE 比 ATE 更加提高了所有数据集的准确性	基于 Bilstm 对关系和实体进行编码，提出相互关注机制学习关系和实体的更准确文本表示，模型有嵌入层、Bilstm 层和相互关注层
KDCoe	WK3160K(基于 DBpedia[36])	跨语言实体对齐,跨语言知识图谱补全	跨语言实体对齐中 KDCoE 的最后阶段超过了所有基线，在跨语言知识图谱补全中 KDCoE-mono 的表现至少与 TransE 相当，这表明 KDCoE 很好地保留了单语 KG 结构的特征	对多语种 KG 嵌入模型(KGEM)和多语种字面描述嵌入模型(DEM)进行联合训练，KGEM 采用 TransE，DEM 使用门控递归单元编码器(AGRU)来编码多语言实体描述
KG-BERT	WN11[5],FB13[59],FB15K[5],WN18RR[15],FB15K-237[15],UMLS[15]	三元组分类,链接预测,关系预测	三个任务中 KG-BERT 的效果基本优于所有基线，但链路预测任务非常耗时，在该任务中几乎需要所有实体都要替换头部或尾部实体	模型初始化选 BERT Base 在此基础上在进行微调，从而对三元组进行建模表示

表 2 中文本信息用于知识图谱表示学习模型中的基本任务包括三元组分类、实体分类、链接预测、关系预测等，所用的公开数据集包括 Freebase、WN18、FB15K 等。其中，在三元组分类任务上，pTrans 模型在 Freebase 上准确率最高，为 93.4%；在 FB13K 与 WN11 数据集上，KG-BERT 模型准确率都最高，分别达到了 90.4%和 93.5%，AATE 次之，TEKE 最差；在实体分类任务中，SSP 在 FB15K 和 FB20K 上的表现都优于 DKRL，准确率分别达到了 94.4%和 67.4%；在链接预测任务中，AATE 在 WN18 和 FB15K 上的表现都优于 TEKE，Hits@10 分别达到了 94.4%和 88.0%。

2.2 图片信息用于知识图谱表示学习

基于文本描述的知识图谱表示学习仅考虑了三元组结合文本信息，但是图片信息作为多模态数据的一种也可以用来强化知识图谱表示学习。早期对于图片信息用于知识图谱表示学习主要考虑的是与实体相关联的图像数据特征进行融合。

Image-embodied Knowledge Representation Learning(IKRL)模型[64]首先将视觉模式引入知识图，使用神经图像编码器为实体的所有图像构造表示。然后，通过基于注意力的方法将这些图像表示集成到聚合的基于图像的表示中。该方法同样是基于 TransE 来进行设计的，IKRL 的能量函数定如式(11)所示：

$$E(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}) = E_{SS} + E_{SI} + E_{IS} + E_{II} \tag{11}$$

整体的能量函数由结构和图像表示共同决定，E_{SS} 和 TransE 的能量函数相同都取决于结构，E_{II} 代表基于图像的表示，同时设计 E_{IS} 和 E_{SI} 用于确保结构与图像都能映射到同一空间中。IKRL 的整体模型架构图如图 10 所示。

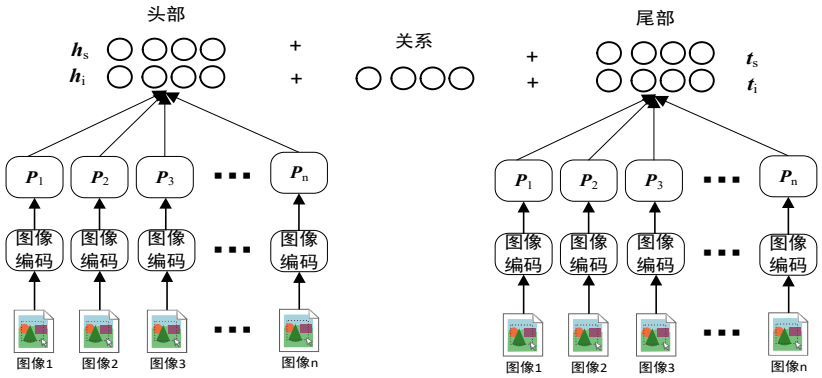

图10 IKRL 模型

Fig. 10 IKRL model

IKRL 是一种传统的基于翻译的方法来学习基于结构的表示。类似的，Mousselly-Sergieh 等[65]提出了一种基于 TransE 的知识图谱表示学习方法 Multimodal Translation-based Knowledge Graph Representation Learning(MTKGRL)。该方法利用两种不同类型的外部多模态表示：一种是通过分析文本语料库中知识图谱实体的使用模式创建的语言表示；另一种是从对应的知识图谱实体的图像获取的视觉表示。MTKGRL 为每种表示及其组合定义了一个能量函数，包括结构能量、多模态能量和结构-多模态能量。结构能量采用 TransE，而头部和尾部的多模态表示分别计算如式(12)和(13)所示。其中 $\oplus$ 可以是一个连接运算符或者是一个函数映射，$\boldsymbol{h}_{\mathrm{w}}$ 为头节点的语言表示，$\boldsymbol{h}_{\mathrm{i}}$ 为头节点视觉表示，$\boldsymbol{t}_{\mathrm{w}}$，$\boldsymbol{t}_{\mathrm{i}}$ 分别表示尾节点的语言表示和视觉表示。

$$\boldsymbol{h}_{\mathrm{m}}=\boldsymbol{h}_{\mathrm{w}}\oplus\boldsymbol{h}_{\mathrm{i}} \tag{12}$$

$$\boldsymbol{t}_{\mathrm{m}}=\boldsymbol{t}_{\mathrm{w}}\oplus\boldsymbol{t}_{\mathrm{i}} \tag{13}$$

除了基于翻译的嵌入方法外，Lonij 等[66]在神经张量网络(Neural Tensor Network,NTN[63])的基础之提出新的目标函数 Text，用于改进模型，从而提出神经张量层(Nerual Tensor Layer,NTL)将图像转换为其内容的结构化语义表示。他们设计了一个相互嵌入的空间将知识图谱和图片联系起来，通过这个设计可以捕捉知识图中图像和已知实体之间的关系，从而使用标记的图像将新概念添加到知识图谱中。与前面提出的方法相比，TransAE[47]是一种新的表示学习方法，它将多模态自动编码器和 TransE 相结合。在 TransAE 中，自编码器的隐藏层被用作 TransE 模型中实体的表示，因此它不仅能够编码结构知识，还可以将视觉和纹理等多种模态的知识编码到最终的表示中。TransAE 整体架构如图 11 所示。

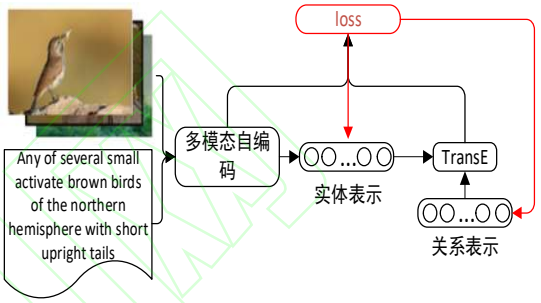

图11 TransAE 模型

Fig. 11 TransAE model

这种方法具有很高的灵活性和可扩展性，能够更好地捕捉数据的多模态特征，从而提高表示学习的质量。随后，Wang 等[67]设计的关系敏感多模态嵌入(Relation Sensitive Multimodal Embedding,RSME)模型。该模型能够在表示学习中自动增强或减弱视觉上下文的影响，进一步证明了在适当情况下利用视觉信息有助于更好的知识图谱嵌入。2021 年，Liu 等[68]通过使用视觉语义表征来对齐异质知识图谱(KGs)中的实体，提出实体视觉对齐(Entity Visual Alignment,EVA)模型。该模型通过基于注意力的模态加权方案把知识图谱的多模态信息融合到联合嵌入中。在 2022 年，Liang 等[69]提出超节点关系图(Hyper-node Relation Graph Attention,HRGAT)网络，可以通过端到端的训练过程有效地捕获多模态知识图的多模态信息和图结构信息。同年，Lu 等[70]提出了多模态表示学习(Multimodal Knowledge Representation Learning,MMKRL)模型，将组件对齐方案与 TransE 方法相结合，实现多模态知识图谱表示学习。不同于之前的模型将多模态知识和结构化数据简单映射到一个空间中，该模型通过对不同的合理性函数求和重构多源知识，然后使用特定的范数约束来对齐源知识。

图片信息用于知识图谱表示学习总结如表 3 所示。

表3 图片信息用于知识图谱表示学习

Tab. 3 Image information is used in the knowledge graph to represent learning

模型	基础数据集	适用任务	效果说明	改进分析
IKRL	WN9-IMG(基于WN18[60], WordNe[61], ImageNet[71])	知识图谱补全, 三元组分类	在知识图谱补全任务和三元组分类中 IKRL 在整体质量上显著优于基线，实验表明图像信息可以提供补充信息，并且基于注意力的方法可以联合考虑多个实例	基于翻译方法的整体架构，该架构结合了结构化知识信息和视觉信息，并利用神经图像编码器和实例级注意力的方式来联合学习图像和结构的表示
MTKGRL	WN9-IMG[64],FB-IMG(基于FB15K[5])	链接预测, 三元组分类	在链接预测任务中优于 IKRL 模型。在三元组分类任务中模型更有效利用了多模态信息，与 TransE 相比平均精度提高超过一个百分点	基于翻译的表示学习方法，通过融合视觉和语言信息，扩展了三元组的定义从而建立新的多模态表示
NTL 改进模型	WN1M(基于WordNet[61]), ILSVRC-2012[72]	类别属性预测	在三个数据集上都明显优于基线 INTL 模型。开放世界(OW)案例中，其性能仅比零样本(ZS)数据集稍差。在 1K 数据集上的性能较低	训练了两种知识图谱嵌入函数，一种基于原始 NTL 架构，另一种基于改进的平滑 SNTL 方法。对于图像嵌入，使用 VGG16 架构，训练了两个嵌入模型，分别以 NTL 和 SNTL 的实体向量为目标
TransAE	WN9-IMG-TXT (基于WN9-IMG[64])	链接预测,三元组分类	链接预测任务中 TransAE 模型优于所有基线模型,与 IKRL 和 DKRL 相比有一定的优势。三元组分类任务中 TransAE 效果最佳,大多数情况下,准确性足够高,可以将知识图中的正三元组与负三元组区分开来	将 Multimodal autoencoder 和 TransE 结合，同时学习多模态知识和结构知识。提取视觉和文本特征向量，将这些向量输入多模态自动编码器，得到联合嵌入作为实体表示
RSME	WN18-IMG(基于WN18[60]), FB15K-IMG(基于FB15K[5]), WN18-UMG-S(基于 WN18-IMG), FB15K-IMG-S (基于 FB-15K-IMG)	链接预测	RSME 的性能优于所有其他模型。RSME(VIT)和 RSME(No Img) 之间的差异比较显著，表明视觉上下文的引入确实有帮助。RSME(VIT)和 RSME(VIT+Forget)的对比也说明 forget gate 在大多数情况下确实有进一步的提升	由一个基本的 KG 嵌入模型和三个门(过滤门、遗忘门和融合门)。使用过滤门来自动过滤不相关的图像，图像通过遗忘门来增强有益的特征。在遗忘门之后，视觉信息和 KG 结构信息在融合门中融合，最后通过最小化损失函数获得实体和关系的嵌入
EVA	DBP15k[73], DWY15k[18]	实体对齐	半监督 EVA 在两个 EA 基准测试中获得了新的 SOTA，大大超过了以前的方法。无监督 EVA 达到了大于 70%的准确率	利用视觉相似性创造初始种子字典，提供了一个完全无监督的解决方案。通过多模态嵌入学习过程和对齐学习过程联合以解决实体对齐任务
HRGAT	FB15K-237[74],WN18RR[15], DB15K[41], YAGO15K[41]	多模态知识图谱补全	HRGAT 与其他基线相比，多数评测指标都为最高。与四个基础模型比较，优于所有传统知识图谱嵌入方法。HRGAT 对于多模态知识图的补全任务达到了高质量水平	HRGAT 主要包括信息融合模块(通过预训练嵌入和低秩多模态融合融合多模态特征)，信息聚合模块(捕获多模态知识图谱中的结构信息)，预测块(用于多模态知识图谱补全)
MMKRL	WN9-IMG[64], FB-IMG[65]	链接预测, 三元组分类	链接预测中 MMKRL 与文中多模态 KRL 模型相比较，除 Raw 指标，其他所有指标均为最高。三元组分类中 MMKRL 明显优于所有模型，在 FB-IMG 数据集上效果最好	MMKRL 主要分成两个模块其中知识重构模块中使用不同预训练编码模型对各种知识进行嵌入以此重构多模态知识图谱，而 AT 模块中使用联合学习框架学习结构化和多模态表示

表 3 中图片信息用于知识图谱表示学习模型中，主要任务集中在三元组分类、链接预测等。所用公开数据集包括 WN9-IMG 和 FB-IMG 等。在三元组分类任务中，MMKRL 在数据集 WN9-IMG 和 FB-IMG 上准确率最高，分别为 97.91%和 69.59%，MTKGRL 次之，但与前者相差不到百分之一，IKRL 表现最差；在链接预测任务中，MTKGRL 在 WN-IMG 数据集上略优于 MMKRL，在 Hits@10 指标上达到了 83.78%，但与 MMKRL 相差不到 0.2%；在 FB-IMG 数据集上，MMKRL 比 MTKGRL 高出两个百分点。

2.3 音视频信息用于知识图谱表示学习

结合音频、视频等进行图谱表示学习的研究较少。早期研究中，Jin 等[75]提出了 TransFusion，这是一种端到端的监督学习框架，它将多模态嵌入(视频、音频)与知识嵌入相融合，是第一个在多关系数据中进行视频标签推理的工作。TransFusion 通过串联融合视频嵌入，如图 12 所示，能够跨不同模态[76]融合视频的特征。

文中将融合视频嵌入的一般形式如式(14)所示：

$$\boldsymbol{h}_{\text{vid}}=[\boldsymbol{z}^{(K)},\{\boldsymbol{z}^{(M_i)}\}] \tag{14}$$

其中：$\boldsymbol{z}^{(K)}$ 表示视频知识嵌入，$\boldsymbol{z}^{(M_i)}$ 表示补充模态的视频嵌入，K 表示知识形态，$\boldsymbol{M}_i$ 表示状态空间(视觉、音频、文本)，作者考虑了最多三中模态，并且不考虑融合实体嵌入中的重复。

紧接着，针对视频理解问题，Deng 等[77]制作了一个包含多模态视频实体和丰富的常识关系的异构数据集。在该数据集上提出了一种基于 Contrastive Language-Image Pre-Training(CLIP)的端到端模型，通过知识图嵌入联合优化视频理解目标。这种方法不仅可以更好的把知识注入视频理解中，还可以帮助模型生成更好的知识图谱嵌入。Li 等[78]提出了一种基于教育词典的微调双向编码器表示，称为 Educational Bidirection Encoder Representation from Transformers(EduBERT)。他们收集了教师的语音数据，用于构建多模态知识图谱。此外，他们创新性地提出了一种语音融合方法，将这些数据作为一类实体链接到图中。音视频信息用于知识图谱表示学习总结如表 4 所示。

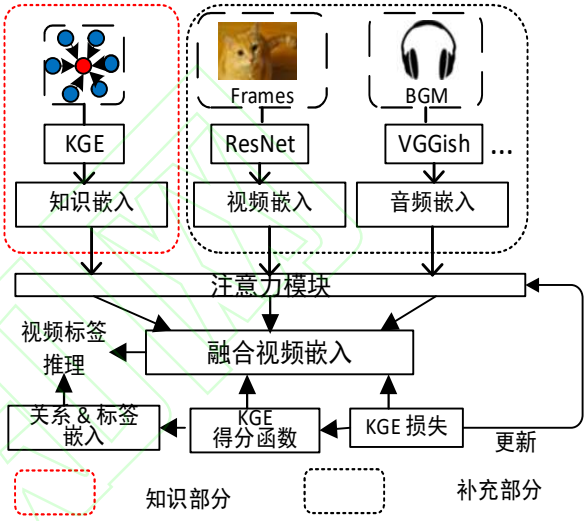

图12 TransFusion 整体架构

Fig. 12 Overall architecture of TransFusion

表4 音视频信息用于知识图谱表示学习

Tab. 4 Audio and video information is used to represent learning in the knowledge graph

模型	基础数据集	适用任务	效果说明	改进分析
TransFusion	FB15K[5], ImageNet[71]	视频标签推理	TransFusion 变体都优于基本模型 TransE，包括没有任何预训练视频嵌入的 TransFusion-0。集成任何额外组合(单一、双重或三重)模态的 TransFusion 的性能明显优于 TransFusion-0 和所有基线	是部分可训练的模型，通过融合 KG 嵌入和多种模态的预训练视频嵌入。结合预定义的评分函数，融合的视频嵌入用于导出语义关系的嵌入，这些嵌入进一步用于推断标签，作为常规链接预测任务
基于 CLIP 的方法[77]	CN-DBpedia 等自制数据集	视频关系标签(Video-Relation-Tag, VRT),视频关系视频(Video-Relation-Video,VRV)任务	该方法在 HITS@10 上分别在 VRV 和 VRT 任务中获得了 42.36%和 17.73%的大幅提升，优于所有基于两阶段 KGE 的方法	模型首先对视频编码器进行视频理解训练，通过基于 CLIP 的方法将视频嵌入投影到相同的标签嵌入空间中，最后在一个模型中联合优化 KGE、CLIP 和视频理解目标

表 4 音视频信息被用于知识图谱表示学习模型中，每个模型针对的任务不同，包括视频关系标签(VRT)、视频关系视频(VRV)任务等，并使用了公开数据集如 FB15K 和 ImageNet。在视频标签推理任务中 TransFusion 优于没经过预训练的基础模型，并且相对于 TransE 及其文献中提及的所有自定义 baseline，在 Hits@10 上至少提高了 9.59%；在 VRT 与 VRV 任务中，基于 CLIP 优化的模型优于传统知识图谱表示学习模型，相对于诸如 TransE、TransH、TransR 等模型，在 Hit@10 指标上至少分别提高了 42.36%和 17.73%。

2.4 多模态信息用于知识图谱表示学习

我们将涉及三种及以上模态数据总结为多种模态信息。此类的研究主要集中在文本、图片、数字、音视频等多种模态信息。

Liu 等[41]提出 Product of Experts(PoE)模型，该模型通过提取关系、潜在的数字和视觉特征来进行多模态知识图谱中的实体对其。实体嵌入通过链接预测来进行训练。Pezeshkpour 等[79]从关系模型角度出发，提出了多模态知识图谱嵌入模型(Multimodal Knowledge Base Embeddings,MKBE)，其中文本、数字和图像字词被一起建模。该模型使用不同的神经编码器处理不同类型的数据，并将它们与现有的关系模型结合起来，以学习实体和多模态数据的嵌入。Chen 等[80]设计一种新的嵌入方式多模态实体对齐(Multimodal Entity Alignment,MMEA)，用于分别生成关系、视觉、数字知识的实体表示，以解决多模态知识图谱中实体对齐问题。该模型使用 TransE 模型嵌入关系数据，利用 Visual Geomentry Group(VGG)模型学习图像嵌入表示视觉信息，将数字信息通过应用径向基函数转换为高维空间中的嵌入表示。多模态信息用于知识图谱表示学习总结如表 5 所示。

表5 多模态信息用于知识图谱表示学习

Tab. 5 Multimodal information is used in knowledge graph to represent learning

模型	基础数据集	适用任务	效果说明	改进分析
PoE	FB15K[5],DB15K,YAGO15k(通过与 FB15k 实体对齐构建而成)	链接预测	通过融合了三个多模态知识图谱验证了不同模态对于 sameAs 链接预测任务是互补的假设。PoE-lrni 的后缀代表嵌入了 l(潜在)、r(关系)、n(数字)、i(图像)PoE-lrni 在链接预测任务中优于其他基线模型。其中 PoE-lrni 和 PoE-rni 效果最佳，并且嵌入专家响应占主导地位	PoE 模型通过合并视觉信息并且进行扩展，在 KG 中，目标是学习一个 PoE，将高概率分配给真实的三元组，并将低概率分配给假定为假的三元组。对于每个关系类型都对于一个 expert
MKBE	YAGO-10[81], MovieLens-100k[82]	链接预测, 评级预测	与链接预测模型 DistMult 和 ConvE 相比，MKBE 在链接预测任务中准确率更高	MKBE 将神经网络编码器和解码器来替换任何基于嵌入的关系模型的初始层，将其应用于 DistMult 和 ConvE。适合数据类型(例如文本、图像、数值和分类值)的知识图谱建模
MMEA	FB15K-DB15K[83],FB15K-YAG15K[83]	多模态知识图谱实体对齐	实体对齐任务中对比 TransE, MTransE, IPTransE, SEA, GCN, IMUSE 等,在两个数据集上 MMEA 是性能最好的模型,并且 MMEA 更能充分利用有限的数据	MMEA 包括多模态知识嵌入和多模态知识融合两个模块。在第一个模块中提取关系，视觉和数字信息用于补充实体特征，第二个模块中将多模态知识融合，并使用交互式训练

表 5 多模态信息用于知识图谱表示学习模型中，各个模型针对任务有所差距。这些任务主要集中在链接预测和多模态知识图谱实体对齐，并且所用的数据集也各有不同但相对较为丰富。在链接预测任务中，MKBE 模型在数据集 MovieLens-100k 上相比传统链接预测模型 DisMult 和 ConvE 表现更好，该模型更好的利用多模态数据提高链接预测的准确性，并相较于现有的方法提高了 5%~7%；在多模态知识图谱实体对齐任务中，MMEA 模型在数据集 FB15K-DB15K 和 FB15K-YAGO16K 上表现最佳。相较于 TransE 和 PoE-I 等模型，在 Hit@10 指标上至少提高 10.58%。

3 多模态知识图谱表示学习典型应用

随着大量异构数据的涌现，多模态知识图谱的研究越来越受到重视。知识图谱表示学习一直以来都是热点研究方向，如今不同模态辅助知识图谱表示学习的研究也越来越多。多种模态综合学习更加提高了知识表示的准确性，并且提高了各类应用的性能和效果。

本章主要介绍多模态知识图谱表示学习在知识图谱补全、问答系统、多模态生成和推荐系统方面的潜在应用场景。在这些应用场景中，多模态知识图谱表示学习能够有效地利用不同模态的信息，提高应用的准确性和效果。

3.1 知识图谱补全

“三元组分类”或“链接预测”任务都可以用于知识图谱的补全。在三元组分类任务中，模型需要预测给定三元组在知识图谱中是否存在。而在链接预测任务中，模型需要预测三元组中缺失的元素，如实体或关系。链接预测任务可以被转化为三元组分类任务，因为缺失元素可以被视为待预测的关系类型。这两种任务都可以用于预测新的知识，即不存在于当前知识图谱中的三元组。

多模态知识图谱表示学习是一种将知识图谱中的实体和关系映射到低维空间的方法。它利用多个模态的信息来推断实体之间的关系以及缺失值的推测。例如，图像、文本和语音数据都可以被视为不同的模态。通过将多种模态的信息融合到表示学习中，可以更准确地预测实体之间的关系和缺失值。由于加入了多种模态的嵌入，多模态知识图谱表示学习应用场景得到了高速发展，成为知识图谱补全中的一个研究热点。2018 年 Mousselly-Sergieh 等[65]提出多模态(视觉和语言)翻译表示学习模型用于链接预测，最终结果显示利用多模态信息会带来显著改善。尽管通过该方法使得结构表示变得不那么具有辨别力，但由于综合考虑了多模态知识表示从而弥补了这种影响，进而提高了预测精度。再如 Xie 等[64]提出基于图像的知识表示模型用于知识图谱补全中的三元组分类，最终结果证明了将三元组中的结构化信息与图像中的视觉信息相结合的模型的有效性和鲁棒性。更具体的说，例如实体分类任务可以被视为一个特殊的链接预测任务，Wilcke 等[84]在实验中通过多种不同模态组合，提出多模态消息传递网络用于学习不同类型的模态中实体和概念的嵌入，在此基础上 Bloem 等[46]引入了一组新的基准任务，用于更加精准的评估多模态知识图谱上的实体分类任务。在实体分类任务之下实体对齐也是一个重要的方向，多模态知识图谱的覆盖率通常比较低下并且不够完成，为了能更好利用现有的多模态知识图谱的知识，Guo 等[85]将欧氏表示扩展到双曲面流形，提出双曲多模态实体对齐(Hyperbolic Multimodal Entity Alignment,HMEA)模型用于实体对齐。后续 Chen 等[80]在此基础上设计了新的损失函数用于增强多种模态的互补性。

3.2 问答系统

知识图谱是构建智能问答系统的重要基础。多模态知识图谱表示学习可以进一步提高问答系统的智能水平，使其更好地理解用户的意图。在旅游领域，多模态知识图谱表示学习能够将来自图片和文本的信息进行融合，从而更好地回答用户提出的旅游相关问题。在视觉问答领域，多模态知识图谱表示学习提供了有关图像中实体及其关系的知识，进而加深了对视觉内容的理解。这使得系统能够以一种更加准确的方式进行推理过程，并预测最终答案。因此，多模态知识图谱表示学习在智能问答系统中扮演着至关重要的角色。其中 Chen 等[86]通过多模态知识图谱提供实体的视觉特征来增强图像或文本的表示来增强多模态实体识别。除此之外在视觉问答中大家越来越发现很多问题需要结合外部知识来进行视觉推理，而多模态知识图谱可以很好的解决这个问题。Yu 等[87]从视觉、语义、事实的角度建立新的表示，在此基础上将视觉问答问题转化为递归推理过程。除了视觉问答领域，在图像文本匹配领域为了减少传统方法导致的检索任务训练数据中的偏差，通过多模态知识图谱扩展更多的视觉语义概念，引入视觉概念相关性知识来增强图像表示成为主流方法。Shi 等[88]提出新的模型通过对 SCG 扩展，将上下文概念的表示与图像嵌入特征融合来改进图像文本匹配。

3.3 多模态生成

多模态生成任务是指根据多模态输入生成对应的多模态输出，如图像描述生成、文字生成等。多模态生成任务在自然语言处理、计算机视觉、语音识别等领域都有广泛应用。但多模态数据集往往规模较小，存在数据不足、数据不一致和模态之间交互等问题，影响任务效果。利用多模态知识图谱可以解决这些问题。多模态知识图谱是基于知识表示的多模态数据结构，包含多个模态，通过实体之间的语义关系进行连接。多模态知识图谱可以将不同模态的信息整合在一起，扩充数据集规模。同时，多模态知识图谱中的关系信息可以对不同模态之间的关系和交互进行建模，更好地处理数据不一致和模态之间的交互问题。多模态知识图谱应用于多模态生成任务可以提供有价值的信息，解决数据不足、数据不一致以及模态之间的交互等问题，提高任务效果。举例来说，Chaudhary 等[89]提出了一种新的标签分配框架(Tag Assignment Using Knowledge Embedding,TAKE)。这个框架使用了来自外部知识库的知识嵌入，将多模态知识图谱中的概念知识嵌入到图像中，从而大大改进了图像的表达能力。作者将视觉信息与 Visio 文本知识库(即多模态知识图谱)相结合，以帮助消除概念歧义，并更好地与图像相关联。

3.4 推荐系统

推荐系统是一种重要的人工智能应用，旨在根据用户历史行为和偏好，向用户推荐符合其需求的商品、服务、内容等。然而，由于用户兴趣和偏好的复杂性以及数据的稀疏性和噪声等问题，传统的推荐系统往往难以实现精准和个性化的推荐。为了解决这些问题，研究人员提出了很多方法，其中多模态知识图谱表示学习是一种非常有前景的方法。

多模态知识图谱表示学习通过将用户的行为和偏好以及商品等物品的信息整合到一个多模态知识图谱中，并通过学习图谱中实体和关系之间的表示，更好地理解用户的行为和偏好，并提供更精准的推荐结果。在电子商务领域，多模态知识图谱可以将用户的购买历史、浏览行为和评论等信息整合到一个图谱中，并通过学习图谱中实体和关系之间的表示，来提高推荐的准确性和个性化程度。

近年来，多模态知识图谱表示学习已经成为推荐系统领域的一个热点研究方向。研究人员们提出了很多方法，其中 Tao 等[90]提出的多模态知识感知强化学习网络(Multimodal Knowledge-aware Reinforcement Learning Netword,MKRLN)是一种比较有代表性的方法。MKRLN 通过在多模态知识图谱中提供实际路径来耦合推荐和可解释性，并且利用多模态知识图谱中的路径内的顺序依赖关系来推断代理多模态知识图谱交互的潜在合理性。同时，MKRLN 还设计了一种新的

分层注意力路径，可以使用户将注意力集中在他们感兴趣的项目上，从而减少了知识图谱中的关系和实体，缩短了到目标实体的路径，并提高了推荐的准确性。

4 研究工作展望

多模态知识图谱表示学习是指在知识图谱中结合多种不同模态数据(如文本、图像、音频、时间、事件等)，对实体和关系进行表示学习的技术。它旨在扩展知识图谱的功能，以更好地捕捉和表示知识，并以此提高推理和检索的准确性。其典型任务极其广泛，并且多模态知识图谱表示学习的质量直接影响后续任务评估的指标。多模态知识图谱表示学习是一个非常具有挑战性的领域，对于未来工作的展望包括以下几个方面：

大规模数据处理：随着知识图谱的不断扩大，多模态知识图谱表示学习需要处理大量的数据，而不影响模型的性能，一方面要从高效的算法入手，多模态数据的数据量通常很大，需要考虑如何实现高效的算法，以支持实际应用。面对此类问题可以采用分布式计算等技术，以加快处理速度，例如，可以使用基于图神经网络(GNN)的方法，这种方法在处理图数据时比传统机器学习算法更为高效。另一方面数据的质量尤为重要，目前由于人工标注成本过高仍然缺少多模态知识图谱相关的高质量数据，尤其是在视频和音频方面。面对此类问题可以采用半监督学习，弱监督学习等技术，通过仅使用少量标注数据和大量非标注数据来训练模型，以此提高数据利用效率。

数据多样性和数据质量：多模态数据具有不同类型和质量，需要考虑如何统一处理和结合各种模态的信息。数据质量和鲁棒性是处理多模态数据时必须关注的问题，其中数据缺失是一个普遍的问题，它会影响模型性能和准确性。为了解决这个问题，可以采用数据增强和合成等方法，采用综合性策略来处理和结合各种模态信息，提高模型泛化性能。同时，跨模态学习的方法可以将来自不同模态的数据映射到统一特征空间中，例如可以采用神经网络中的跨模态对齐技术。值得注意的是，当前研究主要集中在针对图片、文本等多模态数据的研究，面对视频、音频等数据的多模态处理研究还较少。

数据缺失与任务联合：多模态知识图谱表示学习需要开发新的方法以解决缺失数据对模型性能的影响。在补充新的数据的同时，如何设计模型，使其在联合多个任务(如链接预测、知识图谱补全、三元组分类等)的同时学习到实体和关系的表示。为了解决这个问题，可以采用多任务学习的方法，将不同任务的目标函数同时加入导模型的训练中，以共同学习实体和关系表示。此外，集成学习的方法也可以用于任务联合多样性表达，将不同的模型集成在一起，以实现更好的性能。综上所述，处理多模态数据和缺失数据的方法需要考虑到数据质量和鲁棒性，以及如何设计模型实现任务联合的多样性表达。

非监督学习：目前大多数多模态知识图谱表示学习方法都是基于监督学习的，监督学习的特点就是需要大规模高质量数据，而大规模高质量数据正是现在所急缺的。因此开发非监督学习方法以解决数据缺失问题将是一个未来挑战。面对这种困境，可以从非监督学习出发深入研究，例如现在比较流行的比如生成对抗网络(Generative Adversarial Network,GCN)可以进一步深入研究。

可解释性：随着深度学习的发展，越来越多的深度学习模型在各个领域表现得十分出色，但这些模型本质上并没有模拟出人类在决策某些事情的思维过程，可以说深度学习的可解释性非常差。而多模态知识图谱表示学习模型通常是黑盒模型，因此如何解释模型的决策是一个未来的挑战。面对此类问题，我们可以从模型架构设计，可视化方法，解释方法，对抗样本四个方面入手。首先对于模型架构首选可解释性更好的模型，比如决策树，规则等，或者将神经网络进行可解释改造。其次通过可视化的方式使用图像或者网络结构展示出模型对于不同模态的输入数据处理的过程。再者就是通过解释方法提高可解释性，例如，将模型的决策结果进行本体化，用语义化的方式来解释模型的输出。最后就是通过对抗性样本，发现模型漏洞，进而提高模型的可解释性。

评价体系：如何评价多模态知识图谱表示学习的效果，并与其他方法进行比较也是一个比较重要的问题，现在的大都数研究都是基于自己设定的数据集并且在特定的任务上的评估表现，无法更加全面并且多方位的对一个多模态知识图谱表示学习模型进行评估。面对此类问题我们建议综合使用多任务评估并且进行对比实验，在有必要的情况下可以邀请特定领域专家进行评估，以此来确定模型生成的表示是否合理。

AIGC 与知识图谱：随着 ChatGPT 病毒式传播，生成式 AI(AIGC，即人工智能生成内容)因其出色的语言理解、生成、推理能力在学术界引起了巨大的关注[91]。针对 AIGC 任务，ChatGPT 以及各类大语言模型作为目前最重要的工具，擅长推理和问答，这些领域恰巧是知识图谱应用的典型场景。两种不同的技术在同一个任务场景下并非对立竞争关系。对于问答领域而言，AIGC 的最大痛点在于其存在幻视和偏见问题，以及对垂直领域的不够专业。而知识图谱在这些方面具有天然的优势。因其提取客观存在的事实知识，知识图谱的真实性可保证，并且已经通过人为筛选排除了有害信息。它可以指导当前的 AIGC 工具进行学习，以消除幻视和偏见问题。对于垂直领域的问题，垂直领域知识图谱包含了专业的垂直领域知识，结合垂直领域的知识图谱和 AIGC 工具解决其不够专业的问题也是未来面临的一大挑战。在推理领域，目前 AIGC 技术的最大痛点在于推理过程的黑盒性。目前常用的基于思维链进行上下文学习的推理方法虽然效果可观，但业界目前无法确定模型在该过程中的推理过程。而知识图

谱具有天然的可解释性，作为符号主义的延申其在推理领域更具说服力。总的来说，无论是在问答领域还是推理领域，目前的 AIGC 技术还不足以直接取代知识图谱。从另一个角度看，两者应该是互利共生的关系。目前的 AIGC 技术可以看作是数据驱动的，而知识图谱则是知识驱动的。未来如何综合利用知识驱动和数据驱动，实现更加完善的 AIGC 体系，是一个巨大的挑战。

5 结语

知识图谱是感知智能通往认知智能的基石，而多模态知识图谱则是更高级的形式。虽然大型语言模型已经表现出了惊人的能力，但是它们也有显而易见的缺点。大型语言模型只是在粗略地模仿人类语言的规则而并没有真正理解语言的语义。然而，多模态知识图谱可以让机器理解浅层的语义。未来的人工智能是数据驱动和知识驱动两种形式共同推动的。本文讨论了针对传统单一模态知识图谱表示学习所使用的学习方法。先前的模型都只在单一模态上进行知识图谱表示学习，没有利用到现在涌现的多模态数据。因此，调研分析了近年来多模态数据用于知识图谱表示学习的方法，得出结论综合使用多种模态可以在多方面弥补单一模态知识表示的不足。本文还举例分析了多模态知识图谱表示学习在知识图谱补全、问答系统、多模态生成和推荐系统场景中的应用。最后，本文从七个方面总结了多模态知识图谱表示学习未来的挑战。希望这篇文章能为多模态知识图谱表示学习领域的研究者提供研究思路。

参考文献

[1] SINGHAL A. Introducing the knowledge graph : things, not string[EB/OL].(2012-05-16)[2023-03-12].https://www.blog.google/products/search/introducing-knowledge-graph-things-not.

[2] YUHAS B P, GOLDSTEIN M H, SEJNOWSKI T J. Integration of acoustic and visual speech signals using neural networks[J]. IEEE Communications Magazine, 1989, 27(11): 65-71.

[3] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423-443.

[4] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: Representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.

[5] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[J]. Advances in Neural Information Processing Systems, 2013, 26:1-9.

[6] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2015: 2181-2187.

[7] WANG Z, ZHANG J, FENG J, et al. Knowledge graphembedding by translating on hyperplanes[C]// Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2014: 1112-1119.

[8] JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2015: 687-696.

[9] XIAO H, HUANG M, HAO Y, et al. TransA: an adaptive approach for knowledge graph embedding[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2015:1-7.

[10] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. Red Hook, NY: Curran Associates Inc., 2011: 809-816.

[11] JENATTON R, ROUX N, BORDES A, et al. A latent factor model for highly multi-relational data[J]. Advances in Neural Information Processing Systems, 2012, 25: 3176-3184.

[12] BALAŽEVIĆ I, ALLEN C, HOSPEDALES T. TuckER: tensor factorization for knowledge graph completion[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: Association for Computational Linguistics, 2019: 5185-5194.

[13] BORDES A, WESTON J, COLLOBERT R, et al. Learning structured embeddings of knowledge bases[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2011, 25(1): 301-306.

[14] YANG B, YIH S W, HE X, et al. Embedding entities and relations for Learning and Inference in knowledge bases[C]// Proceedings of the International Conference on Learning Representations (ICLR). Ithaca, NY: openreview.net, 2015:1-13.

[15] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2d knowledge graph embeddings[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2018, 32(1):1811-1818.

[16] DAI QUOC NGUYEN T D N, NGUYEN D Q, PHUNG D. A novel embedding model for knowledge base completion based on convolutional neural network[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Stroudsburg, PA: Association for Computational Linguistics, 2018: 327-333.

[17] XIE Z, ZHOU G, LIU J, et al. ReInceptionE: relation-aware inception network with joint local-global structural information for knowledge graph embedding[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 5929-5939.

[18] GUO L, SUN Z, HU W. Learning to exploit long-term relational dependencies in knowledge graphs[C]// Proceedings of the 2019 International Conference on Machine Learning. PMLR.New York: Association for Computing Machiery, 2019: 2505-2514.

[19] MEZNI H, BENSLIMANE D, BELLATRECHE L. Context-aware service recommendation based on knowledge graph embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 34(11): 5225-5238.

[20] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]// Proceedings of the 2018 European Semantic Web Conference. Cham: Springer, 2018: 593-607.

[21] LIU X, TAN H, CHEN Q, et al. RAGAT: Relation aware graph attention network for knowledge graph completion[J]. IEEE Access, 2021, 9: 20840-20849.

[22] LI Z, LIU H, ZHANG Z, et al. Learning knowledge graph embedding with heterogeneous relation attention networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(8): 3961-3973.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30:5998-6008.

[24] KENTON J D M W C, TOUTANOVA L K. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 NAACL-HLT. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171-4186.

[25] WANG Q, HUANG P, WANG H, et al. CoKE: contextualized knowledge graph embedding[EB/OL]. (2020-04-04) [2023-03-25]. https://arxiv.org/pdf/1911.02168.pdf.

[26] YAO L, MAO C, LUO Y. KG-BERT: BERT for knowledge graph completion[EB/OL]. (2019-09-11) [2023-04-03]. https://arxiv.org/pdf/1909.03193.pdf.

[27] CHEN S, LIU X, GAO J, et al. HittER: hierarchical transformers for knowledge graph embeddings[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2021: 10395-10407.

[28] ALAM M M, RONY M R A H, NAYYERI M, et al. Language model guided knowledge graph embeddings[J]. IEEE Access, 2022, 10: 76008-76020.

[29] CHEN Y, GE X, YANG S, et al. A survey on multimodal knowledge graphs: construction, completion and applications[J]. Mathematics, 2023, 11(8): 1815.

[30] NIU Y, TANG K, ZHANG H, et al. Counterfactual VQA: a cause-effect look at language bias[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2021: 12700-12710.

[31] ZHAO W, HU Y, WANG H, et al. Boosting entity-aware image captioning with multi-modal knowledge graph[EB/OL]. (2021-07-26) [2023-04-12]. https://arxiv.org/pdf/2107.11970.

[32] LIANG K, MENG L, LIU M, et al. Reasoning over different types of knowledge graphs: static, temporal and multi-modal[EB/OL]. (2023-05-27) [2023-06-16]. https://arxiv.org/pdf/2212.05767.

[33] WANG M, WANG S, YANG H, et al. Is visual context really helpful for knowledge graph? A representation learning perspective[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 2735-2743.

[34] SUN R, CAO X, ZHAO Y, et al. Multi-modal knowledge graphs for recommender systems[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 1405-1414.

[35] XU G, CHEN H, LI F L, et al. Alime mkg: A multi-modal knowledge graph for live-streaming e-commerce[C]// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: ACM, 2021: 4808-4812.

[36] 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.

[37] CHEN X, SHRIVASTAVA A, GUPTA A. Neil: extracting visual knowledge from web data[C]// Proceedings of the IEEE International Conference on Computer Vision. Washington, DC: IEEE Computer Society, 2013: 1409-1416.

[38] VRANDEČIĆ D, KRÖTZSCH M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.

[39] FERRADA S, BUSTOS B, HOGAN A. IMGpedia: a linked dataset with content-based analysis of Wikimedia images[C]// Proceedings of the 2017 Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Part II 16. Cham: Springer, 2017: 84-93.

[40] LIU Z, WANG S, ZHENG L, et al. Robust Imagegraph: rank-level feature fusion for image search[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3128-3141.

[41] LIU Y, LI H, GARCIA-DURAN A, et al. MMKG: multi-modal knowledge graphs[C]// Proceedings of the 2019 Semantic Web: 16th International Conference ESWC 2019. Cham: Springer, 2019: 459-474.

[42] LI M, ZAREIAN A, LIN Y, et al. Gaia: A fine-grained multimedia knowledge extraction system[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Stroudsburg, PA: Association for Computational Linguistics, 2020: 77-86.

[43] KANNAN A V, FRADKIN D, AKROTIRIANAKIS I, et al. Multimodal knowledge graph for deep learning papers and code[C]// Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 3417-3420.

[44] WANG M, WANG H, QI G, et al. Richpedia: a large-scale, comprehensive multi-modal knowledge graph[J]. Big Data Research, 2020, 22: 100159.

[45] ALBERTS H, HUANG N, DESHPANDE Y, et al. VisualSem: a high-quality knowledge graph for vision and language[C]// Proceedings of the 1st Workshop on Multilingual Representation Learning. Stroudsburg, PA: Association for Computational Linguistics, 2021: 138-152.

[46] BLOEM P, WILCKE X, VAN BERKEL L, et al. Kgbench: a collection of knowledge graph datasets for evaluating relational and multimodal machine learning[C]// Proceedings of the 2021 Semantic Web: 18th International Conference ESWC 2021. Cham: Springer, 2021: 614-630.

[47] WANG Z, LI L, LI Q, et al. Multimodal data enhanced representation learning for knowledge graphs[C]// Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN). Washington, DC: IEEE Computer Society, 2019: 1-8.

[48] WANG Z, ZHANG J, FENG J, et al. Knowledge graph and text jointly embedding[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics, 2014: 1591-1601.

[49] TOUTANOVA K, CHEN D, PANTEL P, et al. Representing text for joint embedding of text and knowledge bases[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.Stroudsburg, PA: Association for Computational Linguistics, 2015: 1499-1509.

[50] RIEDEL S, YAO L, MCCALLUM A, et al. Relation extraction with matrix factorization and universal schemas[C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics,2013: 74-84.

[51] WANG Z, LI J, LIU Z, et al. Text-enhanced representation learning for knowledge graph[C]// Proceedings of International Joint Conference on Artificial Intelligent (IJCAI). San Francisco, CA: Morgan Kaufmann Publishers Inc., 2016: 4-17.

[52] XIE R, LIU Z, JIA J, et al. Representation learning of knowledge graphs with entity descriptions[C]// Proceedings of the 2016 AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2016, 30(1).

[53] XIE R, LIU Z, SUN M. Representation learning of knowledge graphs with hierarchical types[C]// Proceedings of International Joint Conference on Artificial Intelligent (IJCAI). San Francisco, CA: Morgan Kaufmann Publishers Inc., 2016: 2965-2971.

[54] XIAO H, HUANG M, MENG L, et al. SSP: semantic space projection for knowledge graph embedding with text descriptions[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2017, 31(1).

[55] AN B, CHEN B, HAN X, et al. Accurate text-enhanced knowledge graph representation learning[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2018: 745-755.

[56] CHEN M, TIAN Y, CHANG K W, et al. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2018: 3998-4004.

[57] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 1247-1250.

[58] Wikipedia. Wikipedia[M]. PediaPress, 2004.

[59] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2014, 28(1):1112-1119.

[60] BORDES A, GLOROT X, WESTON J, et al. A semantic matching energy function for learning with multi-relational data: Application to word-sense disambiguation[J]. Machine Learning, 2014, 94: 233-259.

[61] MILLER D. On nationality[M]. Clarendon Press, 1995.

[62] LI Z, FENG S, SHI J, et al. Future event prediction based on temporal knowledge graph embedding[J]. Computer Systems Science and Engineering, 2023, 44: 2411-2423.

[63] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[J]. Advances in Neural Information Processing Systems, 2013, 26:926-934.

[64] XIE R, LIU Z, LUAN H, et al. Image-embodied knowledge representation learning[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2017: 3140-3146.

[65] MOUSSELLY-SERGIEH H, BOTSCHEN T, GUREVYCH I, et al. A multimodal translation-based approach for knowledge graph representation learning[C]// Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 225-234.

[66] LONIJ V P A, RAWAT A, NICOLAE M I. Extending knowledge bases using images[C]// Proceedings of the 2017 Workshop on Automated Knowledge Base Construction (AKBC)@ NIPS. Cambridge, MA: MIT Press, 2017: 59-65.

[67] WANG M, WANG S, YANG H, et al. Is visual context really helpful for knowledge graph? A representation learning perspective[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 2735-2743.

[68] LIU F, CHEN M, ROTH D, et al. Visual pivoting for (unsupervised) entity alignment[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2021, 35(5): 4257-4266.

[69] LIANG S, ZHU A, ZHANG J, et al. Hyper-node relational graph attention network for multi-modal knowledge graph completion[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(2): 1-21.

[70] LU X, WANG L, JIANG Z, et al. MMKRL: a robust embedding approach for multi-modal knowledge graph representation learning[J]. Applied Intelligence, 2022, 52(7): 7480-7497.

[71] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]// Proceedings of the 2009 IEEE Conference on Computer Vision and PatternRecognition. Washington, DC: IEEE Computer Society, 2009: 248-255.

[72] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[73] SUN Z, HU W, LI C. Cross-lingual entity alignment via joint attribute-preserving embedding[C]// Proceedings of the 2017 Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Part I 16. Cham: Springer, 2017: 628-644.

[74] TOUTANOVA K, CHEN D. Observed versus latent features for knowledge base and text inference[C]// Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality. Stroudsburg, PA: Association for Computational Linguistics, 2015: 57-66.

[75] JIN D, QI Z, LUO Y, et al. TransFusion: multi-modal fusion for video tag inference via translation-based knowledge embedding[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1093-1101.

[76] SHAN Y, HOENS T R, JIAO J, et al. Deep crossing: web-scale modeling without manually crafted combinatorial features[C]// Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2016: 255-262.

[77] DENG J, SHEN D, PAN H, et al. A unified model for video understanding and knowledge embedding with heterogeneous knowledge graph dataset[EB/OL]. (2023-04-02) [2023-04-19]. https://arxiv.org/pdf/2211.10624.

[78] LI N, SHEN Q, SONG R, et al. MEduKG: a deep-learning-based approach for multi-modal educational knowledge graph construction[J]. Information, 2022, 13(2): 91.

[79] PEZESHKPOUR P, CHEN L, SINGH S. Embedding multimodal relational data for knowledge base completion[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics,2018: 3208-3218.

[80] CHEN L, LI Z, WANG Y, et al. MMEA: entity alignment for multi-modal knowledge graph[C]// Proceedings of the 2020 Knowledge Science, Engineering and Management: 13th International Conference, KSEM 2020. Berlin: Springer, 2020: 134-147.

[81] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2d knowledge graph embeddings[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2018, 32(1).

[82] HARPER F M, KONSTAN J A. The movielens datasets: History and context[J]. ACM Transactions on Interactive Intelligent Systems (TIIS), 2015, 5(4): 1-19.

[83] PANDIT H J, DEBRUYNE C, O'SULLIVAN D, et al. Gconsent — a consent ontology based on the GDPR[C]// Proceedings of the 16th International Conference, ESWC 2019. Berlin: Springer, 2019: 270-282.

[84] WILCKE W X, BLOEM P, DE BOER V, et al. End-to-End entity classification on multimodal knowledge graphst[EB/OL].(2020-05-25) [2023-05-02]. https://arxiv.org/pdf/2003.12383.

[85] GUO H, TANG J, ZENG W, et al. Multi-modal entity alignment in hyperbolic space[J]. Neurocomputing, 2021, 461: 598-607.

[86] CHEN D, LI Z, GU B, et al. Multimodal named entity recognition with image attributes and image knowledge[C]// Proceedings of the 2021 Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021. Berlin: Springer, 2021: 186-201.

[87] YU J, ZHU Z, WANG Y, et al. Cross-modal knowledge reasoning for knowledge-based visual question answering[J]. Pattern Recognition, 2020, 108: 107563.

[88] SHI B, JI L, LU P, et al. Knowledge aware semantic concept expansion for image-text matching[C]//Proceedings of Joint

Conference on Artificial Intelligent (IJCAI). San Francisco, CA: Morgan Kaufmann Publishers Inc., 2019:5182-5189.
[89] CHAUDHARY C, GOYAL P, PRASAD D N, et al. Enhancing the quality of image tagging using a visio-textual knowledge base[J]. IEEE Transactions on Multimedia, 2019, 22(4): 897-911.
[90] TAO S, QIU R, PING Y, et al. Multi-modal knowledge-aware reinforcement learning network for explainable recommendation[J]. Knowledge-Based Systems, 2021, 227: 107217.
[91] ZHANG C, ZHANG C, ZHENG S, et al. A complete survey on generative ai (AIGC): is ChatGPT from GPT-4 to GPT-5 all you need? [EB/OL]. [2023-05-23]. https://arxiv.org/pdf/2303.11717.

This work is partially supported by National Science Fund for Distinguished Young Scholars (62225308).

WANG Chunlei, born in 1977, Ph. D., researcher. His research interests include Knowledge graph and cognitive intelligence, affective computing and emotion recognition.
WANG Xiao, born in 1998, M. S. candidate. His research interests include multimodal knowledge graph.
LIU Kai, born in 1996, M. S. candidate. His research interests include knowledge graph construction in the field of unmanned vehicles.