知识在视觉语言学习中的作用：任务与挑战综述

玛丽亚林佩莱乌[1]，乔治斯塔穆[1]

[1]雅典国立技术大学电气与计算机工程学院 *AILS* 实验室

摘要

近年来，视觉语言（VL）学习取得的进展推动了多种模型和技术的发展，这些模型和技术实现了多项令人瞩目的应用，现已能够解决各种需要视觉与语言协同处理的任务。目前用于视觉语言预训练的数据集仅包含有限的视觉与语言知识，因而大幅限制了许多视觉语言模型的泛化能力。知识图谱（KG）和大语言模型（LLM）等外部知识源能够通过补充缺失的知识来弥合这些泛化差距，进而催生出混合架构。本综述分析了从此类混合方法中受益的任务。此外，我们还对现有知识源及其类型进行分类，并进一步讨论知识图谱与大语言模型之争，以及其对未来混合方法可能产生的影响。

关键词

视觉语言学习、Transformer、知识图谱、大语言模型、混合架构

arXiv:2303.02411v1 [cs.CL] 4 Mar 2023

1. 引言

视觉语言（VL）学习是人工智能领域发展最快的方向之一，尤其是在 Transformer [1], 问世之后，它催生了多种强大的架构。视觉问答（VQA） [2], 视觉推理（VR） [3], 视觉常识推理（VCR） [4], 视觉蕴含（VE） [5], 图像描述（IC） [6], 图文检索（ITR）及其反向任务文图检索（TIR） [7],视觉语言导航（VLN） [8], 视觉叙事（VIST）和视觉对话（VD） [9]等热门视觉语言任务，受益于近期基于 Transformer 的进展，这些进展遵循预训练微调学习框架。预训练负责融合有关视觉和语言模式的一般信息，以及这两种模态如何相互作用的信息，这些信息基于大规模数据集。预训练期间采用自监督目标函数，帮助视觉语言模型学习视觉与语言之间的依赖关系。例如，遮蔽图像描述中的词语，能够促使模型学习如何填补

根据视觉线索识别它们；反过来，也可以遮蔽图像区域，并由语言引导

收录于 *A. Martin*、*K. Hinkelmann*、*H.-G.Fill*、*A.Gerber*、*D. Lenat*、*R. Stolle*、*F. van Harmelen*（编），《*AAAI 2023* 春季研讨会论文集：需要结合机器学习与知识工程的挑战》（*AAAI-MAKE 2023*），凯悦酒店，旧金山机场，加利福尼亚州，美国，*2023*年*3*月*27*日至*29*日。
marialymp@islab.ntua.gr (M. Lymperaiou); gstam@cs.ntua.gr (G. Stamou) 0000-0001-9442-4186 (M. Lymperaiou); 0000-0003-1210-9874 (G.
©2023 年本文作者版权所有。允许依据知识共享署名 4.0 国际许可协议（CC BY 4.0）使用。CEUR 研讨会论文集（CEUR-WS.org）

重建这些区域。基于对视觉与语言的基本理解，针对特定任务进行微调，通过调整训练模型的神经网络权重，使其一次适应一个特定任务，随后进行最终评估。

尽管在这一过程中获得了丰富的视觉语言知识，当前基于 Transformer 的视觉语言模型 [10, 11, 12, 13, 14, 15, 16, 17, 18] 仍无法泛化到若干需要 **常识知识**或关于 **抽象实体**、**事实**和**现实世界事件**的概念与场景。当然，这在某种程度上是意料之中的，因为视觉语言数据集的预训练和微调都不包含、也不要求感知视觉描述之外的概念。图1 展示了几个例子来说明这一点：关于图像（*I*）的问题（*Q*）需要运用视觉领域之外的知识，才能推断出正确的答案（*A*）。



问：我通常会在哪些日子
我通常去
这栋建筑？答：
星期日。



问：在哪个大洲——
nent是谁
出生于图像中？答：
北美洲。



问：图像中的
人物中
谁最年长？答：Per-
左侧的儿子。



问：这个物体叫什么名字
用来
吃这种食物？ A:
筷子。

图1： 回答这些视觉问题需要外部知识 [19, 20, 21]。

图1中的第一幅图像需要结合人类文化和历史知识 [19], 与视觉信息：图像中的物体是一座教堂，人们通常在星期日前往教堂。第二幅图像 [21] 需要多进行一步推理，因为不仅需要检测出这是一张包含某人照片的邮票（视觉信息），还需要知道这个人是谁。关于命名实体的知识识别出此人为亚历山大·汉密尔顿。进一步的事实知识表明，亚历山大·汉密尔顿出生于如今的圣基茨和尼维斯，而圣基茨和尼维斯位于北美洲。结合这两个事实可推导出最终答案：亚历山大·汉密尔顿出生于北美洲。第三幅图像 [20] 需要提取其中出现的两个人。随后，命名实体知识将这两个人分别确定为塞雷娜·威廉姆斯和维纳斯·威廉姆斯。通过结合命名实体知识和事实知识，可以得出她们的年龄，从而得到比较知识事实：塞雷娜·威廉姆斯比维纳斯·威廉姆斯年长。最后，空间知识推断出塞雷娜·威廉姆斯是左边的人。综合命名实体知识、比较知识和空间知识，最终答案是：左边的人。显然，要回答这些问题，需要从外部来源获取更多知识，并将其提取、结合起来以推断出答案。

因此，有必要在预训练/微调过程的早期或后期阶段融入外部知识，以增强视觉语言模型的能力，使其能够应对更多现实世界中的场景。这类知识通常通过实体、关系和语义描述来表示 [22]，并存储在结构化的知识图谱（KG）中。

[23, 24, 25, 26]。BERT 等语言模型（LM） [27] 已被证明能够存储预训练过程中从语言数据中学到的关系知识，由此引出了语言模型即知识库场景 [28]。随后，可以通过构造填空式语句作为查询来检索这些知识，并由语言模型完成填空。后续研究验证了语言模型存储与检索世界知识的能力，同时展示了这种能力随参数数量增加而扩展的特性 [29]。语言模型要成功充当知识库，仍需满足一些前提：以类似知识图谱查询的方式访问数据、更新过时事实同时避免灾难性遗忘的风险、释放其较为隐晦的推理能力，以及衡量其可解释性，这些仍是尚待解决的挑战 [30]。近年来，大语言模型（LLM） [31, 32, 33, 34, 35] 在各种语言任务中取得了令人瞩目的成果，这极大地启发了人们探索将其用作丰富且规模庞大的知识库，以助力视觉语言学习。

以往关于视觉语言学习的综述 [36, 37, 38, 39, 40, 41, 42]并未关注知识与深度学习视觉语言模型之间的协作。 [43] 首次全面介绍了知识增强视觉语言（KVL）这一主题。在本文中，我们聚焦于利用 Transformer 模型进行视觉语言表征的最新研究；当其与外部知识相结合时，便形成了混合式 方法。最后，我们探讨外部知识辅助视觉语言模型的潜在发展趋势，以及这种辅助预计将如何影响该领域未来的应用。

2. 知识与推理

2.1. 外部知识的类型

外部知识源分为两大类：显式 和隐式 [43]。两者都能够提供预训练视觉语言模型所缺少的**事实性知识、常识、时间知识、词汇知识** 或其他知识信息 [44] 。所使用的外部知识源类型，在很大程度上决定了为视觉语言模型检索和利用知识的方式。

显式知识是指以结构化格式存储在知识图谱中的知识。这类知识以符号形式表示为三元组 *(h, r, t)*，其中包含实体 *(h,t)* 及其相互之间的关系 *r*。从知识库中提取答案的过程完全透明，且可以确定性地还原所遵循的路径。这一点尤其对于评估多步推理和组合推理至关重要，从而确保所遵循的推理路径具有事实准确性。然而，构建和维护知识图谱需要人工投入或监督，因此阻碍了此类知识库的自动扩展。

为视觉语言学习做出贡献的热门开源知识图谱包括 ConceptNet[24], DBpedia [26], 维基数据 [25], YAGO [45] 等。知识图谱事实的检索主要基于 SPARQL 查询；SPARQL 是一种为此目的设计的清晰且确定性的查询语言。许多开源知识图谱还提供应用程序接口，以支持频繁查询。由于 SPARQL 查询所施加的限制，弥合自然语言用户查询与 SPARQL 之间的差距是一项有益的工作 [46]。

知识图谱表示学习提供低维分布式向量，遵循语言嵌入的常见策略 [22]。这种方法使得能够将

深度学习技术应用于知识图谱，同时也建立了知识图谱与视觉语言模型之间更合适的通信方式，因为这三种模态都可以视为数值向量。

隐式知识涵盖存储于神经网络权重中的非结构化知识，描述模型预训练期间学习到的事实和关系。这一过程能够整合多个大规模数据源，而无需人工监督。通过重新训练或微调现有神经网络（LM/LLM），可以扩展和更新现有知识。然而，对于大多数研究机构而言，这一过程的计算成本过高；同时，由于学习到的知识和推理在很大程度上取决于所使用训练数据的质量，其事实准确性、公平性和可信度也存在疑问。

目前已有一系列语言模型作为知识库，因为任何在自监督学习目标下完成预训练的语言模型都有可能充当知识库。由于语言模型的结构不透明，从语言模型中检索知识并不像从知识图谱中检索那样直接。因此，访问语言模型知识主要有两种方式：间接访问（微调）和直接访问（提示）。

微调一直是利用语言模型知识最具体的方式，尽管它实际上并不会检索现有知识。与视觉语言模型的微调方式类似（第1节），语言模型微调是指：在适合目标任务的小型带标签语言数据集上训练若干轮，使神经网络权重适应下游语言任务。

提示遵循预训练、提示、预测流程 [47],，只要设计出合适的提示，就能直接访问预训练语言模型中存储的信息。这正是该方法的难点所在：如何构建最优提示仍是一个开放的研究课题，而次优的提示模板可能仅仅反映了语言模型所包含知识的下限 [48, 49]。文本形式的提示称为离散提示，对人类而言相当直观，因此，可以依据人类的对话行为来寻找编写提示的线索。为此，研究者提出了从大型语料库中挖掘模板 [49], 、对现有提示进行释义改写 [49, 50],、通过语言生成实现填空 [51, 52] 等可行方向。另一方面，软提示通过直接访问语言模型的嵌入空间，以牺牲可解释性为代价换取效率。利用冻结语言模型中的连续任务专用向量进行前缀调优 [53], 、基于离散提示初始化的软提示 [54] 等方法均展现出令人鼓舞的成果。尽管基于提示的知识检索需要进行不小的探索，但其回报是实现少样本乃至零样本推理，从而有望彻底改变知识视觉语言模型。

2.2. 知识图谱与大语言模型中的推理

推理一直是衡量大语言模型能否充分发挥知识库作用的重要里程碑之一。它被视为根据给定信息得出结论并做出决策的能力，可分为形式 推理和非形式推理。形式 推理指以合乎逻辑且确定的方式遵循一套规则。另一方面，非形式推理主要基于对世界的一般经验和直觉，因此容易出现错误，但也更加灵活 [55]。

知识图谱倾向于采用形式化推理路径，这源于其结构化特性以及决策所伴随的确定性。非形式化推理大多与非结构化知识库相关联，这类知识库对世界的认知具有更强的概率性。只要规模足够大，语言模型就展现出一定的推理能力，因此

属于大语言模型 [56]。然而，大语言模型推理的全部潜力尚未得到充分探索。提示已被用于发掘大语言模型的推理能力，鼓励它们展现思维链（CoT），而不只是提供最终答案 [57, 58]。事实证明，思维链（CoT）有助于揭示隐藏的推理能力：在少样本设置下 [58], ，模型会收到若干所需推理过程的示例以及一条指导性提示语；在零样本设置下 [57], ，模型则只会收到指导性提示语，而没有示例。另一类研究提出，利用精心构建的数据集评估大语言模型在下游任务上的表现，每个数据集都专门针对不同类型的推理。为此，各种测试考察了大语言模型在算术 [59], 、符号 [58], 、常识 [60], 及其他类型推理方面的能力。总体而言，前述研究得出的发现表明，大语言模型确实展现出涌现推理能力，能够模拟人类思维模式，但仍无法应对复杂的推理挑战。然而，这些证据无法确定大语言模型是否具备真正的推理能力，还是仅仅能够对其接触到的大量信息完美过拟合 [55]。迄今为止，知识增强型视觉语言研究在多个下游任务中采用语言模型即知识库范式，并取得了成功。

3. 结合知识的视觉语言任务

3.1. 视觉问答（VQA）

在视觉问答（VQA）中，模型接收一幅图像I和一个指向该图像的文本问题Q，并预测一个文本答案A。答案A 可以从预定义的候选项中选出，此时将 VQA 视为一个分类问题；也可以由模型生成，因此将 VQA（自由文本 VQA）归入语言生成领域。知识引导有助于处理仅凭视觉信息不足以作答的情形，如图1所示。研究者已开发出多种用于知识驱动型视觉问答（K-VQA）的知识需求型数据集 [19, 20, 61, 62, 63, 64, 65], 为相关模型实现奠定了良好的起点。早期的 K-VQA 尝试严重依赖通过 SPARQL查询在视觉或文本概念与知识图谱节点之间进行精确匹配 [66, 61, 67], ，当不存在这种显式概念映射时便会产生错误。嵌入表示提供了一种更灵活的解决方案，可通过检索与视觉和文本概念相似的知识图谱事实 [62, 68, 20]。然而，经典词嵌入方法不考虑上下文 [69, 70] ，因此只能适用于有限情形，难以推广到需要上下文化的场景。Transformer 在语言侧的应用进一步改进了 K-VQA 模型 [71, 72] ，为后续端到端视觉语言方法铺平了道路。

ConceptBERT [73] 是迈向统一的知识增强型视觉语言Transformer架构的首个突破。它通过在联合表示中考虑全部三种模态，从而能够在推理过程中融入**常识**知识。遵循统一的知识增强型视觉语言策略进行**事实性知识**注入也得到了探索，只需微调即可利用外部知识图谱 [21]的贡献。统一的知识增强视觉语言架构能够融合多种知识类型，例如**事实性**知识和**视觉**知识，以交叉核验预测答案的有效性 [74]。无论外部知识的类型如何，动态地融入外部知识都会将知识注入转化为段落检索问题，从而为相关架构提供更强的

适应性 [75]。 [76], 也采用从维基百科进行段落检索的方法，将不同粒度的视觉线索（全局描述、图像标签和场景文本）结合起来检索事实；随后，所有语言信息都会输入 T5 变换器模型 [77], ，由其生成最终答案 A。类似地， [78] 也从外部来源进行段落检索，但与先前工作不同，它提出联合训练检索模块和 T5答案生成器。通过段落检索从外部知识源获取的通用信息被认为不足以回答定向视觉问题。为此，知识获取经过改进，重点关注查询、检索段落和图像中共同出现的实体 [79]。将多个知识图谱 [24, 26, 80, 81] 整合到一个统一的大型知识图谱中，有助于检索知识；检索依据是经过上下文增强的问题（以视觉描述和场景文本作为上下文）与知识图谱事实之间的语言相似度。增强后的输入会被送入 T5变换器，最终生成自由文本答案A [82]。针对多模态段落检索，研究提出了多模态逆向完形填空任务作为预训练目标，用于学习维基百科条目中视觉信息与文本信息之间的对齐。该技术可用于涉及命名实体识别的 K-VQA [83]。

不同于将知识图谱用作现有知识来源的做法，最早利用语言模型即知识库范式的研究是针对 K-VQA 提出的。具体而言，GPT-3 [31] 可通过少样本方式提供事实，并接收视觉描述作为提示 [84], ，这与传统知识图谱接收 SPARQL 查询的方式类似。通过将多个图像描述用作提示，输入多种预训练大语言模型，可以提升性能，实现零样本推理 [85]。再次借助语言图像描述作为模态中介， [86] 利用冻结的大语言模型解决零样本视觉问答。大语言模型的思维链（CoT）提示是另一条有趣的研究方向，它通过揭示中间推理步骤[87]，提升答案推导流程的可解释性。其他方法则不再借助语言模态来获取单模态大语言模型提示，而是选择将视觉编码器与大语言模型联合微调，从而获得对齐的大语言模型-视觉语言表示 [88]。

有一些结合显式知识源和隐式知识源的实现，可以兼得两者的优势。KRISP [89] 利用了多个外部知识图谱 [24, 26, 81], 、来自 Visual Genome 的视觉知识 [90], 以及来自 BERT 的隐式知识 [27]。REVIVE[91] 利用多种视觉特征，从维基数据和 GPT-3 等不同知识源中检索知识。事实证明，视觉特征引导对于改进知识检索过程至关重要。在视觉语言推理过程中融合隐式知识和显式知识，也是 KAT 所采用的方法；该方法使用一个改进的框架，从维基数据和 GPT-3 中获取信息，并在此基础上进行联合推理。变换器解码器接收推理模块的输出，以生成最终答案 [92]。

3.2. 视觉常识推理（VCR）

视觉常识推理（VCR）是一项与视觉问答（VQA）密切相关的任务。给定关于图像Q I 的一个具有挑战性的问题，VCR 模型需要预测答案A，并附带理由R [4]。**常识**知识可以来自大规模知识图谱（KGs），例如 ConceptNet [24] 和 ATOMIC [93], ，也可以来自专门的数据集，例如 SWAG [94], ，其中包含关于事件序列的描述。

基于 Transformer 的知识辅助视觉常识推理（K-VCR）研究通常利用 BERT[27] 作为骨干架构，以构建端到端的知识视觉语言模型（KVL models）。在 KVL-BERT [95], 中，输入Q与候选答案A共同引导相关常识事实的检索[24], ，从而生成知识增强的语言输入。随后，将视觉特征与该增强输入一同送入类似 BERT 的视觉语言模型（VL model，即 VL-BERT [96]），以选出正确的A。因此，推断R需要将预测得到的A、候选理由R和视觉特征输入 VL-BERT。文献 [97]提出在单一多模态嵌入中对齐彼此独立的模态表示。同一研究还提出了视觉语言预训练目标的扩展 [43]，以将来自 [24] 的常识知识作为额外模态纳入，从而促进知识视觉语言关系的学习。动态增强图文训练数据是一项值得探索的方向，同时还可以学习基于从常识知识库（commonsense KBs）中检索到的知识事实来重建隐藏的视觉标签 [98]。

隐式知识源近年来在 K-VCR 文献中日益受到关注。GPT-2[99] 已协助对图像进行动态推理，推断关于所描绘情境之前可能发生什么以及之后可能发生什么的**时间知识**假设 [100]。思维链（CoT）推理与 VCR 有着内在联系，因为推理路径与选择理由高度相关R。思维链技术在语言任务中的日益普及，与大语言模型的发展高度相互关联；事实证明，大语言模型能够揭示中间推理步骤 [57]。尽管引入带有有依据的答案理由的新型适当数据集凸显了此类方法的前景，但视觉语言方向目前尚未有许多相关工作 [101]。具体而言， [101] 通过对图像进行图像描述来处理 VCR，然后将该图像描述与现有的语言输入一同馈送给大语言模型。该方向的另一项有前景的工作引入了多模态思维链，不使用语言作为中介模态，而是提出一个两阶段过程，分别推断答案 A 和理由 R，同时指出，参数量少于 10 亿的语言模型已足以达到当前最先进的性能 [102]。预计大语言模型在复杂语言问答推理 [103] 中的快速普及，很快将催生更多大语言模型增强的视觉常识推理方法，从而涵盖推理的更多方面。

3.3. 图像描述（IC）

图像描述（IC）是一项广泛应用的视觉语言（VL）任务，要求模型为图像I生成图像描述c。一些知识增强图像描述（K-IC）技术采用循环神经网络（RNN）或长短期记忆（LSTM）网络等相关变体，同时利用知识源来增强图像描述中的**常识** [104, 105, 106, 107]。

使用 Transformer 将外部知识整合到图像描述中的方法，最早在 [108],中进行了探索，其中**事件**知识和**命名实体**知识与文本及视觉数据融合到 Transformer编码器 [1] 中，以生成实体／事件感知描述。由 ConceptNet [24] 和 ATOMIC [93] 衍生的**常识**描述能够辅助视觉常识生成（VCG）——这是一项具有挑战性的任务，要求推断**意图**和事件的**时间知识**序列 [109]。该方法通过将**常识**描述融入 BART（一种强大的语言生成模型） [110]来实现。利用**地理**信息引导**事实性知识**检索以辅助图像描述的研究，最早在 [111], 中进行了探索，其中视觉特征与提取出的事实一同输入 Transformer编码器-解码器结构，最终生成图像描述c。

除外部知识方面的考量外，图像描述作为一项语言生成任务还面临其他挑战：视觉语言变换器并不适合生成式任务，尽管它们在理解任务中表现出色；在理解任务中，答案必须从一组预定义选项中选出。XGPT 通过针对视觉语言任务 [112],调整生成式预训练 [99, 31] 来应对这一挑战，具体做法是引入新颖的生成式预训练目标。GPT-2 [99] 与 CLIP [15] 的协作被认为极具潜力，因为这两个模型都基于大量网络数据进行训练，从而以隐式知识的方式融入了大量知识。ClipCap [113]利用了这种协作，而无需重新训练 CLIP 或 GPT-2；相反，它训练了一个轻量级的基于变换器的映射模块，将 CLIP 表示与 GPT-2 对齐，最终生成图像描述c。VC-GPT [114]也沿用了 CLIP-GPT-2 的组合方式。另一种结合预训练的 CLIP 视觉编码器与冻结的 GPT-2 文本解码器的轻量级改进方法，进一步提升了低资源方法的性能 [115]。文献 [116], 提出了一种跨模态过滤器，用于选取最相关的视觉信息，从而减少图像描述错误，同时仍遵循冻结的 CLIP-GPT-2 框架。

制约知识增强图像描述能力的一个重要因素，是缺乏用于测试的专门数据集。目前，图像描述模型是在包含图像和图像描述的经典数据集上进行评估的，例如 COCO [117] 和 Flickr [118], 。然而，这些数据集在所需外部知识方面并不具有挑战性。构建符合知识需求型视觉问答数据集范式的适当数据集，例如 OK-VQA [19], 、K-VQA [20], 、FVQA[62], 、KB-VQA [61], 或视觉常识推理数据集 [4]，将有助于凸显知识增强图像描述模型的能力。

3.4. 序列生成

序列生成涉及两种任务，具体取决于每次生成的模态。视觉叙事（VIST）是指为由一系列画面 I_1、I_2..... I_N组成的视觉故事生成图像描述 c_1、c_2..... c_N，可视为图像描述任务在序列上的扩展。反向任务——故事可视化（SV）——则是根据文本描述 c_1、c_2..... c_N合成视觉画面 I_1、I_2..... I_N。这两种任务都要求故事内容前后一致，并且两种模态彼此相关。

将**常识**知识用于 VIST 的尝试始于 [119], ，随后是[120, 121], ，但这些方法都采用了用于文本生成的 RNN 架构。在 [122], 中，研究者探索了基于变换器的模型，并通过 ConceptNet 丰富视觉概念。所有相关的丰富概念都会提供给 BART，最终生成恰当的图像描述。至于 SV，则采用了另一条架构路线：[123, 124, 125,126] 基于生成对抗网络（GANs）[127]。[125], 主要探讨了**常识**和**空间**知识，结果令人鼓舞，表明使用外部知识具有优势。不过，近期的一些方法采用了融入隐式知识的基于变换器的方法。具体而言，Story-DALL-E[128] 甚至能够以零样本方式合成未见过的故事。它利用 DALL-E [129]作为多模态非结构化知识库，从文本生成高质量的视觉合成内容。

与图像描述类似，序列生成任务也面临缺乏适当数据集的问题；有了这类数据集，外部知识的贡献才会更加显著且有意义。

3.5. 多任务模型

无知识视觉语言模型已经能够在同一预训练主体下整合多种视觉语言任务，只需在规模较小的带标签文本图像数据集上进行微调。这样一来，就无需为每个独立任务单独设计并实现架构，而是可以利用用于预训练的大规模数据集中已有的视觉语言关系，例如 COCO [117], Visual Genome [90], Conceptual Captions [130] 和 SBU [131]。

推理任务，包括视觉问答、视觉文本蕴含和视觉常识推理，由于这些任务需要进行的推理高度相似，因此很容易整合到同一个模型中。VisualCOMET [100]提供的高层认知增强了视觉线索，并与文本输入一起馈入 GPT- 2 模型，为这三项任务生成自由文本解释，从而使答案具备可解释性 [132]。 [133], 中也探索了相同的推理任务，并在知识需求型数据集 [19, 62] 上测试了结果。表示外部知识源 [24, 25]中事实的知识嵌入与文本描述对齐，并与知识特征一起输入多模态变换器。KB-VLP [134] 是另一个多任务模型，旨在处理视觉问答和视觉推理，并通过**常识**和**逻辑**能力加以增强。模型从图像和文本中执行实体抽取，将视觉语言概念映射到维基数据 [25] 条目，并基于知识图谱嵌入，最终协助检索最相关的维基数据事实。

在低资源场景中，通过大语言模型进行少样本学习和零样本学习对于生成式视觉语言任务而言具有很高的实用价值。在 [88], 中，编码器-解码器变换器模型是进行文本生成的最合适选择，可用于处理开放式视觉问答和图像描述；同时，在推理阶段利用各种提示，以确保生成合适的文本预测。PromptCap[135]旨在基于 GPT-3 提示，准确生成细粒度且可控的图像描述。生成的图像描述可用于为视觉问答提供上下文。苏格拉底模型框架 [136] 是一个新颖且前景广阔的方向，因为它证明了借助多模态提示，可以将视觉语言和语言模型在预训练期间获得的互补知识结合起来，以完成多种多模态任务。

围绕外部知识的使用，在构建多任务视觉语言模型的过程中得出了一些重要观察结果。例如，在 [133],中，融入维基数据未能像预期那样增强推理能力。作者将性能下降归因于大型维基数据知识库中存在的歧义和噪声。此外，与视觉问答、图像描述等任务的特定任务实现相比，多任务实现明显更少。这表明，对于多个任务而言，融入外部知识并不像针对特定任务的模型那样直接。

3.6. 视觉语言领域知识的未来

近期知识增强视觉语言研究的文献揭示了若干亟待弥补的空白。在我们看来，最突出的问题是大多数任务都缺乏适当的知识需求型数据集，这限制了对知识视觉语言模型的评估范围。除视觉问答[19, 20, 61, 62, 63, 64, 65] 和 VCR [4], 外，其他视觉语言任务均在经典基准数据集上进行测试；因此，知识对下游性能的贡献并不像在

明确要求运用各种知识的情境中那样显著。

与此同时，我们观察到，许多研究致力于构建语言 基准，用于考察语言模型的推理能力，包括数学 [59],符号推理 [58], 基于策略的隐式推理 [137], 常识理解 [138], 时间知识、因果关系、语言理解等 [139]。我们认为，这类尝试有助于创建适当的视觉语言数据集，将视觉和语言方面的挑战纳入其中，从而使知识的贡献更加具体。

随着模型规模不断增大，更广泛的人工智能社区开始关注可解释性问题 [140], ，尤其是在黑盒视觉语言模型与黑盒非结构化知识库结合使用时。较早的知识增强视觉语言架构通常会将可解释性视为引入知识图谱进行答案预测所带来的直接结果 [67], ，因为可以检索通向最终答案的路径。后来的研究虽然广泛利用知识图谱，但主要致力于提升下游性能，而非增强对推导出这些结果的推理过程的可解释性。目前，知识增强视觉语言方法已完全偏离了追求可解释性的方向，尤其是自从不透明的大语言模型开始充当视觉语言模型的知识库以来。

另一个新兴问题是针对将大语言模型作为外部知识源的知识增强视觉语言架构进行提示设计。视觉语言场景同样面临与语言提示搜索类似的挑战（尤其是许多模型使用语言作为视觉与语言之间的中介），因此对结果的质量以及生成这些结果时所遵循的推理过程产生了一定程度的模糊性。多模态提示设计仍是一个尚未得到充分探索、但对于释放大语言模型视觉语言模型全部潜力至关重要的领域。

最终，我们认为，知识增强视觉语言研究的未来与大语言模型的当前趋势高度相互关联，这既体现在为知识增强视觉语言任务设计恰当的知识需求型基准方面，也体现在回答知识图谱与大语言模型之间这一持续存在的难题上；该问题将在第4节中分析。

4. 知识图谱还是大语言模型?

在整个分析过程中，我们发现了一些潜在趋势：为了辅助视觉语言模型，知识源的选择正逐渐转向混合方法。尽管知识图谱显然在以往的视觉语言实现中占据主导地位，但鉴于大语言模型的快速发展以及其在近期自然语言处理文献中日益提升的普及度，我们可以推测，研究重点正逐渐转向语言模型作为知识库。例如，ChatGPT[1] 的出现引发了前所未有的热潮，并围绕人工智能的整体未来开启了多场讨论。ChatGPT及相关实现近乎人类水平的能力为自然语言处理文献创造了大量机遇，同时也引发了许多不可避免的批评。由于视觉语言学习往往会跟随自然语言处理的发展（正如基于知识图谱的知识增强 [141, 142, 143, 144],所体现的那样），因此，更多基于大语言模型的视觉语言实现很可能会在不久后出现。

然而，这一趋势也有代价：大语言模型的参数规模已达到数十亿甚至数万亿[34] ，而训练如此庞大的模型需要投入大量资源。早在参数规模大幅扩张之前，人们就已开始担忧语言模型预训练的成本[145]；除了计算预算之外，公平获取相关实现的机会以及环境影响等问题，也

[1]ChatGPT

应促使我们质疑研究界对这些模型的依赖和偏好。

与此同时，完全不透明的学习与决策过程可能弊大于利；尽管大语言模型研究论文中报告了已知的偏见、错误和不准确性[2]，但实际使用这些模型时，人们不禁怀疑它们为最终任务提供的知识是否具备应有的质量和可信度。大语言模型问世之初，有关 GPT-3[31], 在数学推理、逻辑和伦理请求方面失败案例的论文[146] 便广受关注。探究大语言模型如何受骗，有助于了解其推理过程以及误导性输入如何使其陷入混乱 [147, 148], ，这或许是克服逻辑脆弱性的一个有前景的起点。相关研究揭示了大语言模型的演绎推理能力 [149], ，证明穷尽式记忆弥补了大语言模型无法学会推理的缺陷。总的来说，大语言模型的推理能力引发了若干尚待解答的问题 [55], ，例如，启发式方法是否掩盖了其推理能力的不足；又或者，鉴于它们有时会斩钉截铁地给出不一致之处和错误的推理依据，其推理步骤是否值得信赖。更深入的挑战表明，大语言模型的世界知识也存在鲁棒性问题，因为它们会混淆可能与不可能的情形，尽管它们能够识别不可能发生的事件 [150]。

即使上述顾虑在相关大语言模型的后续版本中多少得到了解决，通过提示寻找最佳事实检索过程仍然是一个未解难题[47]。此外，推广某些提示用于大语言模型的论文往往避而不谈发现这些最佳提示背后的过程，只提供一些相似提示语句之间的实验比较 [57]。因此，如果黄金提示 [57] 让我们一步一步地思考 是通过大量实验确定的，也无法保证其最优性和可靠性；另一方面，如果确实存在某种方法论，那么它似乎尚未得到充分揭示（或者至少尚未公开）。针对思维链推理的低性能引导式提示在语义上与黄金提示相关，因此令人怀疑大语言模型生成的答案是否一致，以及它们对轻微或更剧烈的输入变化是否敏感。

总的来说，如果大语言模型最终在某种程度上取代知识图谱，那么上述缺陷预计也会延续到视觉语言的实现中，由此引出一个关键问题：是应该迅速顺应这一趋势，还是最好再等等?

5. 结论

在本综述中，我们分析了外部知识与视觉语言学习的协同作用。现有模型和数据集为这一新兴领域带来了诸多挑战，因为知识增强方法的全部潜力尚未得到充分发挥。知识图谱和大语言模型都可以作为视觉语言领域的知识库，各有优缺点。为此，本文深入探讨了视觉语言领域中知识库与大语言模型之间的取舍，并着重指出了与语言模型即知识库范式当前发展状况相关的重大待解问题。总之，我们希望本研究能引导研究人员探索知识增强的视觉语言研究，同时指出知识采纳过程中的挑战。

[2]鸟并非真实存在

致谢

本研究工作得到了希腊研究与创新基金会（HFRI）通过HFRI第三轮博士奖学金计划（奖学金编号5537）提供的支持。

参考文献

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, 《Attention is All You Need》, 载于：I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett（编）, 《神经信息处理系统进展》, 第30卷, Curran Associates, Inc., 2017年。网址：https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf。[2] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, D. Parikh, 《VQA：视觉问答》, 2016年。arXiv:1505.00468.[3] F. He, Y. Wang, X. Miao, X. Sun, 《可解释视觉推理：综述》, 《图像与视觉计算》112（2021年）104194。网址：https://www.sciencedirect.com/science/article/ pii/S0262885621000998。doi:https://doi.org/10.1016/j.imavis.2021.104194.[4] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, 《从识别到认知：视觉常识推理》, 2019年。arXiv:1811.10830.[5] N. Xie, F. Lai, D. Doran, A. Kadav, 《面向视觉 grounding 语言学习的视觉蕴含任务》, arXiv预印本 arXiv:1811.10582（2018年）。[6] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, 《从展示到讲述：基于深度学习的图像描述综述》, 2021年。arXiv:2107.06912.[7] S. R. Dubey, 《基于深度学习的基于内容的图像检索十年综述》, 《IEEE视频技术电路与系统汇刊》（2021年）1–1。网址：http: // dx.doi.org/10.1109/TCSVT.2021.3080920。doi:10.1109/tcsvt.2021.3080920.[8] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, A. van den Hengel, 《视觉与语言导航：解读真实环境中的视觉 grounding 导航指令》, 2018年。arXiv:1711.07280.[9] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, G. W. Taylor, 《讲述、绘制并重复：基于持续语言指令生成和修改图像》, 2019年。arXiv:1811.09845.[10] J. Lu, D. Batra, D. Parikh, S. Lee, 《Vilbert：面向视觉与语言任务的任务无关视觉语言表征预训练》, 2019年。arXiv:1908.02265.[11] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, 《Oscar：面向视觉语言任务的对象语义对齐预训练》, 2020年。arXiv:2004.06165.[12] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, D. Kiela, 《Flava：基础语言与视觉对齐模型》（2021年）。[13] W. Kim, B. Son, I. Kim, 《Vilt：无需卷积或区域监督的视觉与语言Transformer》, 2021年。arXiv:2102.03334.[14] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, Y. Cao, 《Simvlm：简单的视觉语言

模型弱监督预训练》，2021年。网址：https://arxiv.org/abs/2108.10904。doi:10.48550/ARXIV.2108.10904.[15] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, 《通过自然语言监督学习可迁移视觉模型》，2021年。[16] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, J. Fu, 《跳出框架看问题：面向视觉语言表征学习的端到端预训练》，2021年。网址：https://arxiv.org/ abs/2104.03135。doi:10.48550/ARXIV.2104.03135.[17] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, P. Zhang, 《Florence：面向计算机视觉的新型基础模型》，2021年。网址：https://arxiv.org/abs/2111.11432。doi:10.48550/ARXIV.2111.11432.[18] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, T. Duerig, 《通过噪声文本监督扩展视觉与视觉语言表征学习》（2021年）。网址：https://arxiv.org/abs/2102.05918。doi:10.48550/ARXIV.2102.05918.[19] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, 《OK-VQA：需要外部知识的视觉问答基准》，2019 IEEE/CVF计算机视觉与模式识别会议（CVPR）（2019年）3190–3199。[20] S. Shah, A. Mishra, N. Yadati, P. P. Talukdar, 《KVQA：知识感知视觉问答》，载于：AAAI，2019年。[21] D. Garcia-Olano, Y. Onoe, J. Ghosh, 《通过实体增强型知识注入改进并诊断基于知识的视觉问答》，2021年。[22] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, 《知识图谱综述：表示、获取与应用》，《IEEE神经网络与学习系统汇刊》PP（2021年）。[23] C. Fellbaum, 《WordNet：电子词汇数据库》（1998年）。[24] R. Speer, J. Chin, C. Havasi, 《ConceptNet 5.5：常识开放多语言图谱》，载于：AAAI，2017年。[25] D. Vrandeˇci`c, M. Krötzsch, 《维基数据：免费的协作式知识库》，《ACM通讯》57（2014年）78–85。[26] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. G. Ives, 《DBpedia：开放数据网络的核心》，载于：ISWC/ASWC，2007年。[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 《BERT：用于语言理解的深度双向变换器预训练》，arXiv abs/1810.04805（2019年）。[28] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, 《语言模型能否作为知识库？》，2019年。网址：https://arxiv.org/abs/1909.01066。doi:10.48550/ARXIV.1909.01066.[29] C. Wang, X. Liu, D. Song, 《语言模型是开放知识图谱》，2020年。网址：https://arxiv.org/abs/2010.11967。doi:10.48550/ARXIV.2010.11967.[30] B. AlKhamissi, M. Li, A. Celikyilmaz, M. Diab, M. Ghazvininejad, 《关于语言模型作为知识库的综述》，2022年。网址：https://arxiv.org/abs/2204.06031。doi:10.48550/ARXIV.2204.06031.[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan,

R. Child、A. Ramesh、D. Ziegler、J. Wu、C. Winter、C. Hesse、M. Chen、E. Sigler、M. Litwin、S. Gray、B. Chess、J. Clark、C. Berner、S. McCandlish、A. Radford、I. Sutskever、D. Amodei，语言模型是少样本学习器，载于：H. Larochelle、M. Ranzato、R. Hadsell、M. F. Balcan、H. Lin（编），《神经信息处理系统进展》第33卷，Curran Associates, Inc.，2020年，第1877–1901页。URL：https://proceedings.neurips.cc/paper/ 2020年/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf。[32] J. Hoffmann、S. Borgeaud、A. Mensch、E. Buchatskaya、T. Cai、E. Rutherford、D. d. L. Casas、L. A. Hendricks、J. Welbl、A. Clark、T. Hennigan、E. Noland、K. Millican、G. v. d. Driessche、B. Damoc、A. Guy、S. Osindero、K. Simonyan、E. Elsen、J. W. Rae、O. Vinyals、L. Sifre，训练计算最优的大语言模型，2022年。URL：https://arxiv.org/abs/2203. 15556。doi：10.48550/ARXIV.2203.15556。[33] A. Chowdhery、S. Narang、J. Devlin、M. Bosma、G. Mishra、A. Roberts、P. Barham、H. W. Chung、C. Sutton、S. Gehrmann、P. Schuh、K. Shi、S. Tsvyashchenko、J. Maynez、A. Rao、P. Barnes、Y. Tay、N. Shazeer、V. Prabhakaran、E. Reif、N. Du、B. Hutchinson、R. Pope、J. Bradbury、J. Austin、M. Isard、G. Gur-Ari、P. Yin、T. Duke、A. Levskaya、S. Ghemawat、S. Dev、H. Michalewski、X. Garcia、V. Misra、K. Robinson、L. Fedus、D. Zhou、D. Ippolito、D. Luan、H. Lim、B. Zoph、A. Spiridonov、R. Sepassi、D. Dohan、S. Agrawal、M. Omernick、A. M. Dai、T. S. Pillai、M. Pellat、A. Lewkowycz、E. Moreira、R. Child、O. Polozov、K. Lee、Z. Zhou、X. Wang、B. Saeta、M. Diaz、O. Firat、M. Catasta、J. Wei、K. Meier-Hellstern、D. Eck、J. Dean、S. Petrov、N. Fiedel，PaLM：利用路径扩展语言建模，2022年。URL：https://arxiv.org/abs/2204.02311。doi：10.48550/ARXIV.2204.02311。[34] N. Du、Y. Huang、A. M. Dai、S. Tong、D. Lepikhin、Y. Xu、M. Krikun、Y. Zhou、A. W. Yu、O. Firat、B. Zoph、L. Fedus、M. Bosma、Z. Zhou、T. Wang、Y. E. Wang、K. Webster、K. Pellat、K. Robinson、K. Meier-Hellstern、T. Duke、L. Dixon、K. Zhang、Q. V. Le、Y. Wu、Z. Chen、C. Cui，GLaM：利用混合专家高效扩展语言模型，2021年。URL：https://arxiv.org/abs/2112.06905。doi：10.48550/ARXIV.2112.06905。[35] S. Borgeaud、A. Mensch、J. Hoffmann、T. Cai、E. Rutherford、K. Millican、G. v. d. Driessche、J.-B. Lespiau、B. Damoc、A. Clark、D. d. L. Casas、A. Guy、J. Menick、R. Ring、T. Hennigan、S. Huang、L. Maggiore、C. Jones、A. Cassirer、A. Brock、M. Paganini、G. Irving、O. Vinyals、S. Osindero、K. Simonyan、J. W. Rae、E. Elsen、L. Sifre，通过从数万亿个词元中检索来改进语言模型，2021年。URL：https://arxiv.org/abs/2112.04426。doi：10.48550/ARXIV.2112.04426.[36] T. Baltrušaitis、C. Ahuja、L.-P. Morency， 多模态机器学习：综述与分类体系，2017年。arXiv:1705.09406。[37] K. Kafle、R. Shrestha、C. Kanan，视觉与语言研究中的挑战与前景，2019年。arXiv:1904.09317。[38] W. Guo、J. Wang、S. Wang，深度多模态表征学习：综述，《IEEE Access》7（2019年）63373–63394。doi：10.1109/ACCESS.2019.2916887。[39] A. Mogadala、M. Kalimuthu、D. Klakow，视觉与语言研究融合的趋势：任务、数据集与方法综述，《人工智能研究期刊》71（2021年）1183–1317。URL：http://dx.doi.org/10.1613/jair.1.11688。doi：10.1613/jair.1.11688。[40] C. Zhang、Z. Yang、X. He、L. Deng，多模态智能：表征学习、信息融合与应用，《IEEE信号处理选题期刊》

14（2020年）478–493。URL：http://dx.doi.org/10.1109/JSTSP.2020.2987728。doi：10.1109/jstsp.2020.2987728。[41] S. Uppal、S. Bhagat、D. Hazarika、N. Majumdar、S. Poria、R. Zimmermann、A. Zadeh，视觉与语言中的多模态研究：当前趋势与新兴趋势综述，2020年。arXiv:2010.09522。[42] J. Cao、Z. Gan、Y. Cheng、L. Yu、Y.-C. Chen、J. Liu，幕后：揭示预训练视觉语言模型的秘密，载于：ECCV，2020年。[43] M. Lymperaiou、G. Stamou，知识增强型多模态学习综述，2022年。URL：https://arxiv.org/abs/2211.12328。doi：10.48550/ARXIV.2211.12328。[44] F. Ilievski、A. Oltramari、K. Ma、B. Zhang、D. L. McGuinness、P. Szekely，常识知识的维度（2021年）。URL：https://arxiv.org/abs/2101.04640。doi：10.48550/ARXIV.2101.04640。[45] T. Tanon、G. Weikum、F. Suchanek，Yago 4：一个合理的知识库（2020年）583–596。doi：10.1007/978-3-030-49461-2_34。[46] W. Zheng、M. Zhang，面向知识图谱查询的自动查询图生成，载于：《第30届ACM信息与知识管理国际会议论文集》，CIKM '21，美国计算机协会，美国纽约州纽约，2021年，第2698–2707页。URL：https://doi.org/10.1145/3459637.3482235。doi：10.1145/3459637.3482235。[47] P. Liu、W. Yuan、J. Fu、Z. Jiang、H. Hayashi、G. Neubig，预训练、提示与预测：自然语言处理领域提示方法的系统综述，《ACM计算评论》55（2023年）。URL：https://doi.org/10.1145/3560815。doi：10.1145/3560815。[48] L. Adolphs、S. Dhuliawala、T. Hofmann，如何查询语言模型?，2021年。URL：https://arxiv.org/abs/2108.01928。doi：10.48550/ARXIV.2108.01928。[49] Z. Jiang、F. F. Xu、J. Araki、G. Neubig，我们如何知道语言模型知道什么?，2019年。URL：https://arxiv.org/abs/1911.12543。doi：10.48550/ARXIV.1911.12543。[50] A. Haviv、J. Berant、A. Globerson，Bertese：学习与BERT对话，2021年。URL：https: //arxiv.org/abs/2103.05327。doi：10.48550/ARXIV.2103.05327。[51] T. Gao、A. Fisch、D. Chen，使预训练语言模型成为更好的少样本学习器，载于：《第59届计算语言学协会年会暨第11届自然语言处理国际联合会议论文集》（第1卷：长篇论文），计算语言学协会，线上，2021年，第3816–3830页。URL：https://aclanthology.org/2021.acl-long.295。doi：10.18653/v1/2021.acl-long.295。[52] H. Guo、B. Tan、Z. Liu、E. P. Xing、Z. Hu，利用有限的优质数据实现高效的（软）Q学习以进行文本生成，2021年。URL：https://arxiv.org/abs/2106.07704。doi：10.48550/ARXIV.2106.07704。[53] X. L. Li、P. Liang，前缀调优：优化用于生成的连续提示，载于：《第59届计算语言学协会年会暨第11届自然语言处理国际联合会议论文集》（第1卷：长篇论文），计算语言学协会，线上，2021年，第4582–4597页。URL：https://aclanthology.org/2021.acl-long.353。doi：10.18653/v1/2021.acl-long.353。[54] Z. Zhong、D. Friedman、D. Chen，事实探测是 [mask]：学习与学习如何

召回率，2021年。网址：https://arxiv.org/abs/2104.05240。doi:10.48550/ARXIV.2104.05240.[55] J. Huang, K. C.-C. Chang, 《大语言模型中的推理：综述》，arXiv abs/2212.10403（2022）。[56] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, 《大语言模型的涌现能力》，2022年。网址：https://arxiv.org/abs/2206.07682。doi:10.48550/ARXIV.2206.07682.[57] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, 《大语言模型是零样本推理器》，2022年。网址：https://arxiv.org/abs/2205.11916。doi:10.48550/ARXIV.2205.11916.[58] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. hsin Chi, Q. Le, D. Zhou, 《思维链提示可激发大语言模型的推理能力》，arXiv abs/2201.11903（2022）。[59] A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, H. Hajishirzi, 《MathQA：借助基于运算的形式体系实现可解释的数学文字题求解》，载于：2019年计算语言学协会北美分会人类语言技术会议论文集，第1卷（长短论文），计算语言学协会，明尼阿波利斯，明尼苏达州，2019年，第2357–2367页。网址：https://aclanthology.org/N19-1245。doi:10.18653/v1/N19-1245.[60] P. Bhargava, V. Ng, 《利用预训练语言模型进行常识知识推理与生成：综述》，2022年。网址：https://arxiv.org/abs/2201.12438。doi:10.48550/ARXIV.2201.12438.[61] Q. Wu, P. Wang, C. Shen, A. R. Dick, A. van den Hengel, 《有问必答：基于外部来源知识的自由形式视觉问答》，2016年IEEE计算机视觉与模式识别会议（CVPR）（2016），第4622–4630页。[62] P. Wang, Q. Wu, C. Shen, A. R. Dick, A. van den Hengel, 《FVQA：基于事实的视觉问答》，《IEEE模式分析与机器智能汇刊》40（2018），第2413–2427页。[63] A. K. Singh, A. Mishra, S. Shekhar, A. Chakraborty, 《从字符串到事物：能够读取和推理的知识增强型VQA模型》，2019年IEEE/CVF国际计算机视觉会议（ICCV）（2019），第4601–4611页。[64] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, J. Tan, 《面向基于知识的视觉问答的跨模态知识推理》，arXiv abs/2009.00145（2020）。[65] Z. Chen, J. Chen, Y. Geng, J. Z. Pan, Z. Yuan, H. Chen, 《利用知识图谱进行零样本视觉问答》，载于：SEMWEB，2021年。[66] Y. Zhu, C. Zhang, C. Ré, L. Fei-Fei, 《构建用于回答视觉查询的大规模多模态知识库系统》，2015年。arXiv:1507.05670.[67] P. Wang, Q. Wu, C. Shen, A. R. Dick, A. van den Hengel, 《用于视觉问答的显式知识推理》，载于：IJCAI，2017年。[68] M. Narasimhan, A. G. Schwing, 《直奔事实：学习用于事实型视觉问答的知识库检索》，arXiv abs/1809.01124（2018）。[69] J. Pennington, R. Socher, C. Manning, 《GloVe：词表示全局向量》，载于：2014年自然语言处理实证方法会议（EMNLP）论文集，计算语言学协会，多哈，卡塔尔，2014年，第

1532–1543页。网址：https://aclanthology.org/D14-1162。doi: 10.3115/v1/D14-1162.[70] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, 《词语和短语的分布式表示及其组合性》，2013年。 arXiv:1310.4546.[71] M. Ziaeefard, F. Lécué, 《迈向知识增强型视觉问答》，载于：COLING，2020年。[72] A. Salaberria, G. Azkune, O. L. de Lacalle, A. S. Etxabe, E. Agirre, 《用于在基于知识的视觉问答中有效运用语言模型的图像描述》，arXiv abs/2109.08029（2021）。[73] F. Gardères, M. Ziaeefard, B. Abeloos, F. Lécué, 《ConceptBERT：面向视觉问答的概念感知表示》，载于：FINDINGS，2020年。[74] J. Wu, J. Lu, A. Sabharwal, R. Mottaghi, 《用于基于知识的VQA的多模态答案验证》，arXiv abs/2103.12248（2021）。[75] C. Qu, H. Zamani, L. Yang, W. B. Croft, E. G. Learned-Miller, 《面向外部知识视觉问答的段落检索》，第44届国际ACM SIGIR信息检索研究与发展会议论文集（2021）。[76] F. Gao, Q. Ping, G. Thattai, A. N. Reganti, Y. Wu, P. Natarajan, 《变换-检索-生成：以自然语言为中心的外部知识视觉问答》，2022年IEEE/CVF计算机视觉与模式识别会议（CVPR）（2022），第5057–5067页。[77] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, 《探索统一文本到文本转换器的迁移学习极限》，《机器学习研究杂志》21（2020），第1–67页。网址：http://jmlr.org/papers/v21/20-074.html。[78] W. Lin, B. Byrne, 《利用外部知识进行检索增强型视觉问答》，载于：自然语言处理实证方法会议，2022年。[79] J. Wu, R. J. Mooney, 《面向外部知识视觉问答的实体聚焦稠密段落检索》，载于：自然语言处理实证方法会议，2022年。[80] N. Tandon, G. de Melo, G. Weikum, 《从网络获取比较常识知识》，全国人工智能会议论文集1（2014），第166–172页。[81] S. Bhakthavatsalam, K. Richardson, N. Tandon, P. Clark, 《狗有胡须吗？一个新的部件关系知识库》，2020年。 arXiv:2006.07510.[82] Z. Chen, Y. Huang, J. Chen, Y. Geng, Y. Fang, J. Z. Pan, N. Zhang, W. Zhang, 《LAKO：通过后期知识到文本注入实现知识驱动的视觉问答》，第11届国际知识图谱联合会议论文集（2022）。[83] P. Lerner, O. Ferret, C. Guinaudeau, 《面向基于知识的视觉问答的多模态逆向完形填空任务》，arXiv abs/2301.04366（2023）。[84] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, L. Wang, 《GPT-3用于少样本基于知识的VQA的实证研究》，arXiv abs/2109.05014（2021）。[85] A. M. H. Tiong, J. Li, B. Li, S. Savarese, S. C. H. Hoi, 《即插即用VQA：通过结合大型预训练模型实现零训练零样本VQA》，载于：自然语言处理实证方法会议，2022年。[86] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, S. Hoi, 《从图像到文本提示：利用冻结的大语言模型实现零样本VQA》，arXiv abs/2212.10846（2022）。[87] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, H. Zhang, C. Gan, 《观察、思考、确认：交互式

用于基于知识的视觉推理的视觉与语言模型间提示，arXiv abs/2301.05226（2023年）。[88] W. Jin, Y. Cheng, Y. Shen, W. Chen, X. Ren, 一个好的提示价值数百万参数：面向视觉语言模型的低资源提示学习，2021年。URL: https://arxiv.org/abs/2110.08484. doi:10.48550/ARXIV.2110.08484。[89] K. Marino, X. Chen, D. Parikh, A. K. Gupta, M. Rohrbach, KRISP：融合隐式知识与符号知识以进行开放领域基于知识的视觉问答，2021 IEEE/CVF计算机视觉与模式识别会议（CVPR）（2021年）14106–14116。[90] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, F.-F. Li, Visual Genome：利用众包密集图像标注连接语言与视觉，2016年。arXiv:1602.07332。[91] Y. Lin, Y. Xie, D. Chen, Y. Xu, C. Zhu, L. Yuan, REVIVE：区域视觉表征对基于知识的视觉问答至关重要，2022年。URL: https://arxiv.org/abs/2206.01201. doi:10.48550/ARXIV.2206.01201。[92] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, J. Gao, KAT：用于视觉与语言的知识增强Transformer，载于：计算语言学协会北美分会2022年会议论文集：人类语言技术，计算语言学协会，美国西雅图，2022年，第956–968页。URL: https://acla nthology.org/2022.naacl-main.70. doi:10.18653/v1/2022.naacl-main.70。[93] J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, Y. Choi, COMET-ATOMIC 2020：关于符号与神经常识知识图谱，载于：AAAI，2021年。[94] R. Zellers, Y. Bisk, R. Schwartz, Y. Choi, SWAG：用于基于语境的常识推理的大规模对抗数据集，2018年。URL: https://arxiv.org/abs/1808.05326. doi:10.48550/ARXIV.1808.05326。[95] D. Song, S. Ma, Z. Sun, S. Yang, L. Liao, KVL-BERT：用于视觉常识推理的知识增强视觉与语言BERT，Knowl. Based Syst. 230（2021年）107408。[96] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT：通用视觉与语言表示的预训练，2020年。arXiv:1908.08530。[97] J. Lee, I. Kim，用于视觉常识推理的视觉—语言—知识协同嵌入，2020年。[98] S. Ye, Y. Xie, D. Chen, Y. Xu, L. Yuan, C. Zhu, J. Liao, 通过知识图谱谜题提升视觉与语言模型中的常识能力，2022年。URL: https://arxiv.org/abs/2211.16504. doi:10.48550/ARXIV.2211.16504。[99] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever，语言模型是无监督多任务学习器，2019年。[100] J. S. Park, C. Bhagavatula, R. Mottaghi, A. Farhadi, Y. Choi, VisualCOMET：对静态图像的动态语境进行推理，载于：欧洲计算机视觉会议（ECCV）论文集，2020年。[101] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, 学习解释：通过思维链进行多模态推理，用于科学问题解答，2022年。URL: https://arxiv.org/abs/2209.09513. doi:10.48550/ARXIV.2209.09513。[102] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, A. Smola，语言模型中的多模态思维链推理，2023年。URL: https://arxiv.org/abs/2302.00923。

doi:10.48550/ARXIV.2302.00923。[103] X. Daull, P. Bellot, E. Bruno, V. Martin, E. Murisasco，复杂问答与语言模型混合架构：综述，2023年。[104] Y. Zhou, Y. Sun, V. G. Honavar，通过利用知识图谱改进图像描述，2019 IEEE计算机视觉应用冬季会议（WACV）（2019年）283–293。[105] J. Hou, X. Wu, Y. Qi, W. Zhao, J. Luo, Y. Jia，利用先验知识进行视觉描述的关系推理，arXiv abs/1906.01290（2019年）。[106] F. Huang, Z. Li, S. Chen, C. Zhang, H. Ma，结合内部与外部知识的图像描述，第29届ACM信息与知识管理国际会议论文集（2020年）。[107] D. K. Aditya Mogadala, Xiaoyu Shen，将基于规则的实体屏蔽整合到图像描述中，2020年。[108] W. Zhao, Y. Hu, H. Wang, X. Wu, J. Luo，通过多模态知识图谱增强实体感知图像描述，2021年。URL: https://arxiv.org/abs/2107.11970. doi: 10.48550/ARXIV.2107.11970。[109] Y. Xing, Z. Shi, Z. Meng, Y. Ma, R. Wattenhofer，Km-bart：用于视觉常识生成的知识增强多模态BART，载于：ACL/IJCNLP，2021年。[110] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettle- moyer，BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练，2019年。URL: https://arxiv.org/abs/1910.13461. doi:10.48550/ARXIV.1910.13461。[111] S. Nikiforova, T. Deoskar, D. Paperno, Y. Winter，利用外部百科知识生成图像描述，arXiv abs/2210.04806（2022年）。[112] Q. Xia, H. Huang, N. Duan, D. Zhang, L. Ji, Z. Sui, E. Cui, T. Bharti, X. Liu, M. Zhou，XGPT：用于图像描述的跨模态生成式预训练，2020年。URL: https://arxiv.org/ abs/2003.01473. doi: 10.48550/ARXIV.2003.01473。[113] R. Mokady, A. Hertz, A. H. Bermano, ClipCap：用于图像描述的Clip前缀，arXiv abs/2111.09734（2021年）。[114] Z. Luo, Y. Xi, R. Zhang, J. Ma，一种极其简单的端到端图像描述方法，2022年。[115] R. P. Ramos, B. Martins, D. Elliott, Y. Kementchedjhieva，Smallcap：通过检索增强提示的轻量级图像描述，arXiv abs/2209.15323（2022年）。[116] Z. Luo, Z. Hu, Y. Xi, R. Zhang, J. Ma，I-tuning：利用图像微调冻结语言模型以实现轻量级图像描述，2022年。URL: https://arxiv.org/abs/2202.06574. doi:10.48550/ARXIV.2202.06574。[117] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick，Microsoft COCO：上下文中的常见物体，载于：D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars（编），计算机视觉——ECCV 2014，Springer国际出版公司，Cham，2014年，第740–755页。[118] P. Young, A. Lai, M. Hodosh, J. Hockenmaier，从图像描述到视觉指称：针对事件描述的语义推理的新相似度指标，计算语言学协会会刊 2（2014年）67–78。URL: https://aclanthology.org/ Q14-1006. doi:10.1162/tacl_a_00166。

[119] P. Yang, F. Luo, P. Chen, L. Li, Z. Yin, X. He, X. Sun，知识型故事讲述器：一种用于视觉叙事的常识驱动生成模型，载于：第二十八届国际人工智能联合会议论文集，IJCAI-19，国际人工智能联合会议组织，2019年，第 5356–5362 页。网址：https://doi.org/10.24963/ijcai.2019/744。doi:10.24963/ijcai.2019/744.[120] C.-C. Hsu, Z.-Y. Chen, C.-Y. Hsu, C.-C. Li, T.-Y. Lin, T.-H. K. Huang, L.-W. Ku，知识增强的视觉叙事，2019年。arXiv:1912.01496.[121] C. Xu, M. Yang, C. Li, Y. Shen, X. Ao, R. Xu，想象、推理与写作：运用图知识和关系推理进行视觉叙事，载于：AAAI，2021年。[122] H. Chen, Y. Huang, H. Takamura, H. Nakayama，面向多样且信息丰富的视觉叙事的常识知识感知概念选择，载于：AAAI，2021年。[123] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, J. Gao，StoryGAN：一种用于故事可视化的序列条件生成对抗网络，2019年，第 6322–6331 页。doi:10.1109/CVPR.2019.00649.[124] A. Maharana, D. Hannan, M. Bansal，通过语义一致性改进视觉故事的生成与评估，arXiv abs/2105.10026（2021年）。[125] A. Maharana, M. Bansal，将视觉空间、语言和常识结构整合到故事可视化中，arXiv abs/2110.10834（2021年）。[126] N. Tsakas, M. Lymperaiou, G. Filandrianos, G. Stamou，一种用于故事可视化的公正Transformer，2023年。网址：https://arxiv.org/abs/2301.03563。doi:10.48550/ARXIV.2301.03563。[127] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio，生成对抗网络，载于：神经信息处理系统进展，第 27 卷，2014年。网址：https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf。[128] A. Maharana, D. Hannan, M. Bansal，StoryDALL-E：调整预训练文本到图像Transformer以进行故事续写，2022年。网址：https://arxiv.org/abs/2209.06192。doi:10.48550/ARXIV.2209.06192.[129] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever，零样本文本到图像生成，2021年。arXiv:2102.12092.[130] P. Sharma, N. Ding, S. Goodman, R. Soricut，Conceptual Captions：一个经过清理、上位词化的图像替代文本数据集，用于自动图像描述，载于：ACL，2018年。[131] V. Ordonez, G. Kulkarni, T. Berg，Im2text：使用 100 万张带说明文字的照片描述图像，载于：J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger（编），神经信息处理系统进展，第 24 卷，Curran Associates, Inc.，2011年。网址：https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf。[132] A. Marasović, C. Bhagavatula, J. S. Park, R. L. Bras, N. A. Smith, Y. Choi，运用全栈视觉推理的自然语言推理依据：从像素到语义框架，再到常识图，载于：FINDINGS，2020年。[133] V. Shevchenko, D. Teney, A. R. Dick, A. van den Hengel，视觉与语言推理：探索补充知识的优势，arXiv abs/2101.06013（2021年）。[134] K. Chen, Q. Huang, Y. Bisk, D. J. McDuff, J. Gao，KB-VLP：基于知识的视觉与

语言预训练，2021年。[135] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, J. Luo, PromptCap：提示引导的任务感知图像描述，arXiv abs/2211.09699（2022年）。[136] A. Zeng, A. S. Wong, S. Welker, K. Choromanski, F. Tombari, A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, P. R. Florence，苏格拉底模型：将零样本多模态推理与语言相结合，arXiv abs/2204.00598（2022年）。[137] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, J. Berant，亚里士多德用过笔记本电脑吗？一项采用隐式推理策略的问答基准，《计算语言学协会会刊》9（2021年）346–361。网址：https://doi.org/10.1162/ tacl_a_00370。doi:10.1162/tacl_a_00370.[138] A. Talmor, J. Herzig, N. Lourie, J. Berant，CommonsenseQA：一项以常识知识为目标的问答挑战，载于：2019年计算语言学协会北美分会会议：人类语言技术论文集，第1卷（长篇与短篇论文），计算语言学协会，明尼阿波利斯，明尼苏达州，2019年，第4149–4158页。网址：https://aclanthology.org/ N19-1421。doi:10.18653/v1/N19-1421.[139] A. Srivastava, A. Rastogi, A. Rao, A. Shoeb, A. Abid, A. Fisch, A. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. Kocurek, A. Safaya, A. Tazarv, Z. Wu，超越模仿游戏：量化并外推语言模型的能力，2022年。doi:10.48550/arXiv.2206.04615.[140] S. R. Islam, W. Eberle, S. K. Ghafoor, M. Ahmed，可解释人工智能方法：综述，2021年。网址：https://arxiv.org/abs/2101.09429。doi:10.48550/ARXIV.2101.09429。[141] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto，Luke：基于实体感知自注意力的深度上下文化实体表示，载于：EMNLP，2020年。[142] Y. Qin, Y. Lin, R. Takanobu, Z. Liu, P. Li, H. Ji, M. Huang, M. Sun, J. Zhou，Erica：通过对比学习改进预训练语言模型的实体与关系理解，载于：ACL/IJCNLP，2021年。[143] N. Poerner, U. Waltinger, H. Schütze，E-BERT：高效且有效的BERT实体嵌入，载于：计算语言学协会论文集：EMNLP 2020，计算语言学协会，线上，2020年，第803–818页。网址：https://aclanthology. org/2020.findings-emnlp.71。doi:10.18653/v1/2020.findings-emnlp.71.[144] L. Bauer, L. Deng, M. Bansal, Ernie-nli：分析领域特定外部知识对自然语言推理增强表示的影响，载于：DEELIO，2021年。[145] O. Sharir, B. Peleg, Y. Shoham，自然语言处理模型的训练成本：简要概述，2020年。网址：https://arxiv.org/abs/2004.08900。doi:10.48550/ARXIV.2004.08900.[146] L. Floridi, M. Chiriatti，GPT-3：其性质、范围、局限与影响，《心智与机器》30（2020年）681–694。[147] N. Kassner, H. Schütze, 针对预训练语言模型的否定式与错误启动探测：鸟会说话，但不会飞，2019年。网址：https://arxiv.org/abs/1911.03343。doi:10.48550/ARXIV.1911.03343.[148] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. hsin Chi, N. Scharli, D. Zhou，大语言模型很容易被无关上下文分散注意力，arXiv abs/2302.00093

(2023).[149] Z. Yuan, S. Hu, I. Vulic, A. Korhonen, Z. Meng, 预训练语言模型（还能）进行演绎推理吗？，arXiv abs/2210.06442（2022年）。[150] C. Kauf, A. A. Ivanova, G. Rambelli, E. Chersoni, J. S. She, Z. Chowdhury, E. Fedorenko, A. Lenci，大语言模型中的事件知识：不可能与不太可能之间的差距，arXiv abs/2212.01488（2022年）。