**多模态知识图谱(MMKG)（图文）**

**一、多模态知识图谱（MMKG）构建的动机**

1. **传统的知识图谱大多是以文本的形式表示，这削弱了机器描述和理解现实世界的能力。**
2. **构建多模态知识图谱获得更完整的信息，提高传统基于知识图谱模型的性能。**
3. **辅助现有多模态模型，比如高效地引入长尾关系和细粒度关系，对模型进行价值对齐、修正。**

**准确、实时和高效**

**二、MMKG的构建方法及相应挑战**

1. **从图像到符号：文字标注图像---用文本知识图谱中的文字标记图像**

CV 社区开发了许多图像标记解决方案，可用于标记图像。 大多数方案学习从图像内容到各种标签集的映射，包括对象、场景、实体、属性、关系、事件和其他符号。 学习过程由人工标注的数据集监督，这需要众包工作人员绘制边界框并使用给定标签标注图像或图像区域。将图像链接到符号的过程可以分为几个细分的任务：视觉实体/概念提取、视觉关系提取和视觉事件提取。

1. **从符号到图像：符号定位---将图像定位到文本知识图谱中的实体**

符号定位是指找到适当的多模态数据项（例如图像）以表示传统知识图谱中存在的符号知识的过程。与图像标注方式相比，符号定位方式在MMKG构建中应用更为广泛。符号定位主要包括实体定位、概念定位、关系定位。

**a. 实体定位** 旨在将知识图谱中的文本实体定位到其相应的多模态数据，例如图像、视频和音频。将实体定位到图像的**主要挑战如下：**1）如何以低成本为实体找到足够多的高质量图像？ 2）如何从大量噪声中选择最匹配实体的图像？还有很多**未解决的问题：**1）实体被定位成几个图像，每个图像只是实体的一个方面。例如，一个人的图像集合可以是不同年龄的图像、生活照片、事件照片、单人照片和家庭照片。如何确定最典型的图像？2）现实世界的实体是多角度的，在不同的上下文中将一个实体与多个图像相关联是合理的。这促使我们提出一项新的多定位任务，该任务从给定特定上下文的实体中选择最相关的图像。

**b. 概念定位** 概念基础旨在为视觉概念找到具有代表性的、有区别的和多样化的图像。**挑战：**1）并非所有的概念都可以正确可视化。例如，irreligionist（非宗教主义者）不能指定某种具体形式。因此，如何区分可视化概念和非可视化概念成为一个难点。2）如何从一组相关图像中找到一个可视化概念的代表性图像？

**c. 关系定位** 关系指定是从图像数据语料库或互联网中找到可以表示特定关系的图像。输入可以是这个关系的一个或多个三元组，输出应该是这个关系中最具代表性的图像。**挑战：**1）当我们将三元组作为查询来检索关系的图像时，排名靠前的图像通常与三元组的主题和对象更相关，但与关系本身无关。如何找到能够反映输入三元组语义关系的图像？2）有研究主要集中在空间关系和动作关系的基础上，这些关系可以在图像中直观地观察到。但是，大多数其他关系例如isA, Occupation, Team and Spouse在图像中可能并不明显。这些关系通常缺乏训练数据，因此很难用上述两种解决方案训练模型来检索图像。

**三、MMKG的应用**

1. **多模态实体识别与链接**

纯文本的命名实体识别任务（NER）已被广泛研究。多模态 NER (MNER) 旨在检测带有图像的文本中的实体，其中图像可以为实体识别提供必要的补充信息。 MMKG 通过提供视觉特征使得 MNER 附带图像信息用于实体识别。例如，**使用 MMKG 中图像的背景知识来帮助捕获图像的深层特征，以避免来自浅层特征的错误。**

给定带有附加图像的文本，多模式实体链接 (MEL) 使用文本和视觉信息将文本中的模棱两可的提及连接到知识库中的实体。尽管一些早期的尝试基于传统的知识图谱作为知识库进行多模态实体链接，但越来越多的工作更倾向使用多模态知识图谱作为链接的知识库。 **MEL 以两种方式利用 MMKG 中的图像知识：（1）提供实体提及应链接到的目标实体； (2) 使用多模态数据学习每个实体的分布式表示，然后将其用于衡量提及和实体之间的相关性。**图像中视觉信息的使用将有助于捕捉提及和实体之间的关系，但与图像无关的部分也可能成为噪声并对提及和实体的表示学习带来负面影响。为了消除副作用，研究人员提出了一种两阶段的图文关联机制，根据预定义的阈值过滤掉不相关的图像，并且多注意机制还用于通过查询提及的候选实体周围的多跳实体来捕获提及表示和实体表示中的重要信息。

1. **视觉问答**

视觉问答（VQA）是一项具有挑战性的任务，不仅需要对问题进行准确的语义解析，还需要深入理解给定图像中不同对象和场景之间的相关性。新提出的 VQA 任务弥合了人类可以轻松结合各种方式的知识来回答视觉查询的差异。例如，在“哪个美国总统与这里看到的毛绒玩具有关？”的问题中，如果图像中的毛绒玩具被检测为“泰迪熊”，则通过 KG 推断出的答案将是“西奥多·罗斯福”，因为他常常被称为“泰迪罗斯福”，泰迪熊以他的名字命名。

提取视觉概念之间的关系和理解问题中的语义信息是 VQA 的两个关键问题。然而，如果不结合更多各种模态的知识，它就无法通过语义解析和匹配对图像-问题-答案三元组进行推理，并且很难推广到更复杂的情况。 **MMKG 有助于处理问题并增强答案的可解释性。**首先，**MMKG 提供有关命名实体及其在图像中的关系的知识**，从而导致更深入的视觉内容理解。其次，**MMKG 中的结构化符号知识使其成为进行推理过程和预测最终答案的更明确的方式。**

1. **图文匹配**

图像-文本匹配是许多文本和图像相关应用程序中的一项基本任务，例如图像-文本和文本图像检索，其目的是输出输入图像和文本对之间的语义相似度得分。图像-文本匹配通常是通过将文本和图像映射到一个联合语义空间中，然后学习一个统一的多模态表示进行相似度计算来实现的。一种通用的方法是利用多标签检测模块来提取语义概念，然后将这些概念与图像的全局上下文融合。然而，预训练的基于检测的模型很难找到长尾概念，这将模型限制在那些已经被检测到的概念上，导致性能不佳。

**为了克服检索任务训练数据中的偏差，可以利用多模态知识图谱来扩展更多的视觉和语义概念。**此外，MMKG还可以帮助构建场景图，引入视觉概念之间的信息，进一步增强图像表示。例如，可以提取在 MMKG 的多模态三元组中经常同时出现的概念对，例如房屋、窗户和树叶，以增强图像中概念的表示，从而为语义理解提供强大的上下文信息，并提高图像-文本匹配的性能。此外，考虑到图像-文本匹配任务中的一个关键步骤是在不同模式下对齐局部和全局表示，一些工作建议**在 MMKG 中合并关系以表示具有更高级别语义的图像和文本。这种图结构信息更好地增强了多模态数据的推理和推理能力，具有更高的可解释性。** MMKG 还通过学习更统一的多模态表示来帮助跨模态对齐。

1. **多模态生成任务**

**a. 图像标签** 传统的图像标签方法很容易受到标签统计偏差以及嘈杂和不精确的标签的限制。通过**将 MMKG 中的概念知识嵌入到图像中，可以极大地改善图像的表示，从而提高图像标注的性能。**另一项工作提出构建一个名为 Visio-Textual Knowledge Base (VTKB) 的 多模态知识图谱，其中包括文本和视觉信息以及它们之间的关系。基于此 VTKB，提出了一种新颖的图像标记框架，将视觉信息结合到 VTKB 中，以帮助消除概念歧义并将它们与图像更好地联系起来。

**b. 图片说明** 主流的基于统计的图像字幕模型有两个缺点：首先，它们严重依赖目标检测器的性能。具有独立的检测和字幕生成的编码器-解码器框架总是导致预定义的对象/关系和目标文本描述之间的语义不一致。其次，看不见的物体总是给他们带来很大的挑战。在图像-标题并行语料库上训练的模型总是无法描述看不见的对象和概念。幸运的是，MMKG 可以通过以下方式帮助缓解图像字幕的两个障碍：1）一些工作提出**利用 MMKG 进行关系推理，从而产生更准确和合理的字幕。**更具体地说，可以为从候选图像提议中嵌入的视觉和知识向量构建语义图，然后可以对语义图进行编码以生成文本描述。通过这种方式，可以充分利用 MMKGs 中总结的语义约束。2) **来自 MMKG 的符号知识可能有助于理解看不见的物体。**具体来说，符号知识提供了关于未见对象的符号信息，并在符号知识方面建立了已见对象和未见对象之间的语义关系。

意义：

**在多模态大模型的趋势下，多模态知识图谱的价值**

传统的知识图谱大多是以文本的形式表示，引入多模态信息可以进一步丰富模型获取到的信息，提高模型的能力。现有大模型正朝着多模态通用模型方向发展，构建多模态知识图谱可从以下几个方面有效地提高多模态模型的性能：

1. **可解释性:** 知识图谱可以通过可视化工具呈现出来，使我们能够更直观地理解知识结构，可以使大模型的输出更具可解释性，对推理的依据可溯源。
2. **可扩展性：**知识图谱是一种开放性的模型，可以随着需求不断扩展，不必对大模型进微调就可及时地引入新的知识。
3. **准确性：**知识图谱构建的知识准确可靠，可高效、低成本地引入长尾关系和细粒度关系。大模型般需要通过使用大量有较大噪声的数据来训练，以捕捉长尾关系和细粒度关系， 这种方法效率低、成本高。
4. **安全性：**大模型训练由于训练数据量大，质量控制较难，导致所得到的模型容易引发理论和安全问题，可用专门针对此类问题构建知识图谱，对大模型的知识进行补充、对齐和修正。