Knowledge graphs as tools for explainable machine learning: A survey

- abstract
  - 本文对知识图在可解释机器学习中的应用进行了广泛的概述。到目前为止，可解释人工智能已经成为一个非常活跃的研究领域，它解决了最新机器学习解决方案的局限性，这些解决方案通常提供**高度准确但几乎不可检查和解释的决策**。
  - 人们对将知识表示技术集成到机器学习应用中也越来越感兴趣，这主要是由于互补的优势和劣势可能导致新一代混合智能系统。根据这一想法，我们**假设知识图（以机器可读的格式自然提供领域背景知识）可以集成到可解释的机器学习方法中，以帮助它们提供更有意义、有见地和可信的解释。**
  - 使用系统的文献综述方法，我们设计了一个分析框架来探索可解释机器学习的现状。我们特别关注**大规模与结构化知识的集成**，并使用我们的框架分析各种机器学习领域，从不同的角度确定此类基于知识的可解释系统的主要特征。然后，我们总结了这种混合系统的优势，例如提高了**可理解性、反应性和准确性**，以及它们的**局限性**，例如在有效处理噪声或提取知识方面。最后，我们讨论了一系列有待未来研究的开放挑战。
- Introduction
  - 这项工作的目标是研究知识图谱在可解释机器学习背景下的集成和作用。解释长期以来一直是各个领域的研究主题，但由于人工智能 (AI) 的最新进展，包括机器和深度学习系统，现在正被广泛用于决策。制作。然而，**一个主要缺点是他们无法以人类可以轻松理解的方式解释他们的决定，因此需要从用户的角度提高可解释性和可信度**——这是大规模采用他们的关键方面。作为回应，可解释机器学习已迅速成为一个活跃的研究领域，大量的贡献集中在使用各种技术（例如视觉线索、锚点、显着图或反事实）来引发可理解和不可理解的决策（黑色框）方法。
  - 从更广泛的 AI 角度来看，**不透明性**只是现代子符号系统众所周知的限制之一——以及需要**大量训练数据（数据饥饿）、跨任务泛化能力差（脆弱性）、缺乏因果关系或类比推理（反应性）**。近年来，人们对文献的兴趣越来越大，其**目标是促进符号、知识驱动的人工智能（即知识表示）和亚符号、数据驱动的人工智能（机器学习）的整合**，主要是由互补的优势和劣势推动的这可能会导致设计混合的、更智能的系统。神经符号 AI 的想法是符号方法，允许以类似语言的结构化命题的形式编码知识，可以无休止地重新组合以允许跨任务和领域的高级推理，可以与子符号方法及其它们相结合处理大量数据、处理噪声和捕获丰富的感知数据的能力。从这个意义上说，很自然地假设神经符号集成也可以支持可解释系统更加可解释、透明和值得信赖。
  - 在这项工作中，我们特别关注**知识图在可解释机器学习中的作用**。知识表示在操纵、创建、标准化和发布结构化知识方面有着悠久的传统。在过去的二十年里，人们一直致力于扩大技术以应对 Web 的普遍性。语义技术允许轻松访问 Web 上的知识源，而本体、**知识库和图形数据库形式的符号表示允许形式化和捕获有关特定领域的知识和数据以及一般的、百科全书的知识。这种形式化的知识（我们将其称为知识图谱）是机器可读的，主要是可公开访问的，更重要的是，它可以跨领域链接——允许机器以结构化但偶然的方**

式发现知识。本文的主要目标是研究如何将知识图谱集成到可解释机器学习中，以提供更有意义、有见地和值得信赖的解释。

- 为了实现这一目标，我们探索了可解释机器学习的前景，其中子符号系统大规模集成了结构化知识，以确定这种混合集成的特征、优势和局限性。 使用基于系统文献综述的方法，我们使用图形形式的结构化知识来分析不同的机器学习应用程序以生成解释（我们称之为基于知识的解释），范围从早期的规则挖掘任务到图像中的经典任务 识别、项目推荐、自然语言处理和预测任务。
- 本研究并非旨在对可解释人工智能和知识表示的整个领域进行调查，而是特别关注使用知识图作为可解释系统的支持和背景知识的优势和局限性。 特别是，我们提出以下贡献：
  - 我们提出了一个分析框架来系统地分类可解释的、基于知识的子符号系统；
  - 我们根据上述内容塑造和组织基于知识的可解释机器学习的格局，特别是检测每个领域的优势和局限性；
  - 我们讨论了在可解释机器学习的背景下使用知识图谱作为背景知识的优缺点；
  - 最终，我们提供了要解决的开放挑战，以促进下一波可解释系统的设计，充分整合符号和亚符号推理。

- Preliminaries
  - 为了为我们的分析研究奠定相关基础，初步步骤包括建立可解释性的工作定义并提供有关知识图谱的主要概念。 我们通过总结围绕解释的主要理论和机器学习之前的历史概述来实现这一点，然后简要介绍什么是大规模知识图以及存在哪些技术来操纵它们。
    - AI之前的解释概述
      - “解释这个词经常出现，在哲学中占有非常重要的地位，因此花一点时间来确定它的含义将是有益的。”（约翰·斯图尔特·密尔，1884年）
      - 或许在 Mill 的愿望之后的两个世纪，**精确但高深莫测的模型的成功推动了认知科学、法律和社会科学等领域的研究人员与机器学习社区联手，努力为解释的概念提供统一的观点。** 事实上，我们的观点与那些认为可解释的人工智能确实需要来自那些随着时间的推移广泛讨论解释的学科的见解一致。 我们向读者推荐引用的工作，以对该主题进行深入讨论； 在这里，我们通过强调不同学科如何跨时间理解解释的概念来限制自己为我们的研究确定可解释性的工作定义。
      - **纵观历史，哲学家一直将解释视为演绎或归纳的情况，其中一组初始元素（一个事件和一些条件）需要根据一组经验/形而上学定律与随后的现象建立关系（参见亚里士多德的解释的四个原因，Mill的科学解释，Hempel的演绎法学模型）。** 心理学家一直专注于将解释定义为认知-社会过程，见人心理学、信念-欲望-意图模型、脚本理论），而语言学家将解释视为传递知识的过程（解释为对话、论证理论、格莱斯格言） 最后，随着人工智能的出现，解释大多被视为一些初始事实和先验知识可以在特定约束下映射到新知识的过程（参见基于规则的模型，例如专家系统和归纳逻辑）。
      - 虽然很明显，从未就定义达成共同协议，但这里可以评论的方面是，学科确实共享一个共同的抽象模型来定义解释，由法律、由因果关系联系起来的事件以及限制 事件。 最近的工作已经确定了类似的抽象模型（参见图 1 中的 [1] 和 [23]）。

> 在这些之后，我们将松散地将解释称为为什么问题的答案，其形式为**“当 X 发生时，由于给定的一组环境 C，Y 将由于给定的定律 L 而发生”**。 一旦确定了一个统一的工作模型，我们就可以研究如何将所有解释组件（事件、情况、法律、因果关系）构建为知识，以进一步用于生成解释。

- 以大规模图的形式构建知识
  - 尽管自上世纪 80 年代以来，文献中已经出现了对现实世界中的“事物，而不是字符串”及其（相互）关系进行编码的智能模型的想法，但自从谷歌在2012.然而，迄今为止似乎还没有对术语知识图谱的精确定义。一般的协议是将知识图视为一种数据结构，通过有向边标记图来描述实体及其关系，通常将它们组织在本体模式中，并且还涵盖多个主题。将它们之间的一组概念和属性称为术语框（TBox），将属于这些概念的个人陈述集称为断言框（ABox）。随着用于扩展 Web 的技术数量不断增加，知识图现在是大规模集成、提取和操作来自不同来源的数据的结果。如果遵循语义网标准，可以将知识图称为关联数据。
  - 以图的形式而不是典型的关系设置来构建知识的主要好处之一是对模式的灵活性，维护者可以在稍后阶段定义，并随着时间的推移而改变。 这为数据演化以及捕获不完整知识提供了更大的灵活性。 知识图的推理可以通过标准的知识表示形式（RDF、RDFS、OWL）来执行，允许描述和标记实体以及它们之间的关系。 SPARQL、Cypher、Gremlin 等查询语言允许标准关系操作和导航操作符，允许在实体之间找到任意长度的路径，支持更高级的知识发现。 此外，可以通过能够大规模处理知识图的框架来执行更高级的图操作，例如分析、总结、完成。
  - 由于数据表示、发布和交换的各种标准和实践，过去几年已经在 Web 上提供了许多知识图谱。 文献中采用最多的 KG 如下所示，并在表 1 中进行了总结，并附有一些统计数据。 我们还根据三个类别对它们进行了分类，即包含有关日常生活世界的知识的“常识知识库”，包含有关事实和事件的知识的“事实知识库”，以及编码来自特定领域（语言学、 生物医学、地理等）。
    - OpenCyc，一个从 Cyc 知识库实现的常识性知识图谱；
    - Freebase，一个使用来自包括维基百科计划和贡献在内的多个来源的结构化数据作为众包努力构建的 KG；
    - WikiData，也是一个建立在 wiki 内容之上的免费协作项目，另外还提供有关数据来源的元数据；
    - DBpedia，一个知识图谱，通过从维基百科信息框中自动提取键值对，然后通过众包映射到 DBpedia 本体；
    - YAGO，一个大型 KG，将来自 WikiData、GeoNames 和其他数据源的事实映射到通过结合 WordNet 和 Wikipedia 类别构建的分类；
    - ConceptNet，一个免费的众包语言知识库，包括来自 WordNet、Wikipedia、DBpedia 和 OpenCyc 的信息；
    - WordNet，一个精心策划的概念之间关系的词汇数据库（上位词、相关词、分词等）。 我们参考其 2014 年的 RDF 版本，WordNet3.1RDF。
  - 请注意，上述一些资源已停止使用，而存在许多其他免费可用的资源（例如 NELL、KBpedia、FrameNet），它们的使用迄今为止在文献中受到限制。 此外，

最近几年创建了许多专有的 KG，有时称为企业知识图 (EKG)——参见 Google 和 Facebook 的知识图、亚马逊和 Ebay 的产品图。最后，我们还考虑了没有断言的 KG 域本体，例如 schema.org，这可能是为了支持特定任务而临时构建的。

- 可解释人工智能需要结构化知识：实例
  - 在剑桥分析丑闻和 2016 年美国大选中断等事件发生后，现代机器学习方法在决策中变得更加透明的需求变得显而易见。从那时起，已经启动了许多举措，例如。 DARPA 的 eXplainable AI 计划和 EU Ethical Guidelines for Trustworthy AI 旨在鼓励设计人类能够适当理解、管理和信任的道德系统。
  - 然而，大多数可解释机器学习方法仍然专注于解释黑盒模型的内部功能，例如使用视觉线索、锚点、显着图或反事实来识别与不同输出最相关的输入特征。虽然这只是对复杂决策功能的近似，但这些模型既不能解释任何上下文，也不能解释用户可能拥有的背景知识。让这种方法来为医疗保健等生命攸关的问题做出决策是有问题的，因此需要一种新的智能系统范式，以提供可理解的结果以及透明、可靠的解释。知识图和结构化 Web 代表了一种有价值的特定领域、机器可读的知识形式，可用的连接或集中数据集可以作为 AI 系统的背景知识，以便更好地向用户解释其决策。下面我们展示了一些医学领域的场景，机器学习系统可以从外部知识中受益，以支持领域专家理解为什么算法会得出某些结果。
  - 违反直觉的预测：模型报告的工作违反直觉地预测哮喘患者死于肺炎的风险较低。 为了解释这些决定，需要医生的医学专业知识来揭示这些患者被直接送入重症监护室，接受了积极的护理，确实降低了他们的死亡风险，但也导致了错误的机器驱动结论。 如果该模型还以在外部机器可读知识源（例如，提供患者病史的医院数据库或 DisGeNET 等药物-疾病交互数据集）中发现的解释形式提供证据，则此类决策可能更容易理解。
  - 临床试验推荐：让我们想象一家医疗保健公司使用人工智能系统来支持癌症诊断患者寻找实验性治疗（早期访问计划或 EAP）。患者向系统提供他的病史描述（相关文件、症状、诊断等），然后系统通过文本分析提取显着信息，然后使用搜索引擎组件。然后，公司的医学专家必须根据他的专业知识验证确定的试验，并向患者提交一份完整的报告。将患者的数据与结构化知识（如医学本体、叙词表、PubMed 以往研究的证据）相结合，可以大大减少专家的时间和精力，并且不仅可以为专家提供相关临床试验的列表，还可以解释原因这些被选中。此过程将保证 (a) 专家的知识不会被替代，而是在整个过程中得到补充和整合，以及 (b) 通过证明结果是使用由自动化系统增强的可靠领域专业知识获得的结果来增加信任。
  - 治疗诊断：[34] 的工作展示了以用户为中心的诊断推荐 AI 系统如何要求临床医生用他们自己对患者病例的解释来补充智能代理。 进行用户研究以识别自动推理的不同步骤所需的不同类型的解释，即用于诊断的“日常解释”，用于计划治疗的“基于轨迹的解释”，用于提供科学的“科学解释” 来自现有研究的证据，以及“反事实解释”，以允许临床医生添加/编辑信息以查看推荐的变化。 本体用于对 AI 系统所需的组件（基元）进行建模，以自动组合解释以揭示不同形式的知识并解决代理执行的不同任务。

- Research methodology

- 我们的研究方法主要体现在定性研究的三个常见步骤中，即确定包含要分析的主要变量的分析框架，通过文献探索评估这些变量，以及文献综合和讨论。
- 研究问题和分析框架
  - 如第 1 节所述，我们的主要研究问题是如何将大规模知识图谱集成到可解释机器学习系统中以提供更可信的解释？ 为了回答这个问题，我们重点分析以下子问题的文献：
    - **哪些特征具有由生成基于知识的解释的子符号系统使用的知识图？ 它们代表哪种类型的知识（领域知识、事实知识、常识知识），它们的表现力如何（ABox、TBox，两者都有）？** 生成解释的知识是**自动提取**的（例如，通过点击图中的链接）还是**手动提取**（例如，使用专家选择图表的相关部分）？ **解释是通过重用现有的知识图谱构建的，还是临时构建的结构？** 考虑了多少张图表？ 所有这些问题都被归类为知识 (KG) 变量。
    - **哪种类型的子符号方法能够生成基于知识的解释？** 特别是，**模型处理哪种类型的输入数据（表格数据、图像、文本数据），模型用于哪个任务，以及使用哪种方法（顺序神经网络、卷积、基于树等）？** 此外，**该模型如何在内部（即嵌入全局系统行为中的语句）或外部（即使用后验）整合结构化知识？** 我们将这些方面称为模型 (ML) 变量。
    - 系统处理哪种类型的解释？ **它们以何种形式进行交流（文本/自然语言、视觉图像等）**，是**分类解释（解释结果的属性）**、**机械解释（导致结果的机制）**还是**功能解释（解释某事的行为和最终目标）**？最后，**解释内容是否涉及模型的输出、其行为或两者的组合（事后或集成可解释性，或混合）？** 这些被定义为解释 (XP) 变量。
  - 所有这些变量都可以组织在表 2 所示的分析工具包中。在接下来的部分中，我们将使用该工具包首先分析和讨论一组相关研究，最后确定一些指导方针以改进 可解释机器学习中的知识图谱。
- 文献检索与选择
  - 所有研究均通过以下两种方式之一进行搜索：从主要学术数据库、IEEExplore、ACM 数字图书馆、谷歌学术和 ScienceDirect 中广泛搜索知识表示论文的全面自上而下方法。 这伴随着一种自下而上的方法来检查研讨会和人工智能会议，以获取可解释的 AI 领域新发表的研究成果，此外还通过引用的文献滚雪球。
  - 我们对包含表 3 中总结的任何术语的组合的作品进行关键字搜索，从而选择了 10,000 多篇文章。 大约。 根据以下标准对 150 个进行了彻底扫描，并直接从相关工作部分确定了大约 100 个。 只要有可能，我们将同行评审的出版物和主要期刊/会议优先于白皮书或未经审查的提交。 只有在呈现子符号系统（包括从数据中学习的某种形式，并使用图表形式的背景知识产生任何类型的解释）时，才会选择研究。 这意味着，例如，专家系统、基于案例的推理和其他基于规则的方法以及规划和决策应用程序不在本工作的范围内。 通过这种方法，我们最终确定了一组 53 篇论文。
  - 分析工具包以及下一节中介绍和分析的研究都作为开放研究知识图的一部分公开提供。这有助于以结构化方式查阅现有文献并进行相关比较，从而为同一主题的未来研究做出贡献（ cfr. 图 2 中的快照）。 它还将以词汇表的形式推广分析工具包以供重用。

- Using knowledge graphs for explainable machine learning
  - 人工智能文献在利用结构化知识来实现系统可解释性方面有着悠久的历史，从早期的基于知识的系统开始，这些系统是为响应无法提供可信理由的第一代专家系统而设计的。我们在这里专注于基于机器学习的应用程序和根据其通用应用程序领域组织的工作——从规则挖掘方法到图像分类和项目推荐任务，再到自然语言应用程序和预测任务。 该组织纯粹是务实的，旨在最大限度地发挥我们叙述中的逻辑； 然而，我们试图保持从早期、更简单的方法到更复杂、现代架构的历史顺序。
  - **基于规则的机器学习**
    - 知识发现在 80 年代末成为一个成熟的领域，大量工作集中在解释基于规则的机器学习算法的输出，这些算法能够识别大量数据中的有趣模式（所谓的“数据帖子”）。 -处理”或KD管道中的解释步骤）。
    - 对领域本体形式的结构化知识进行了调查，其想法是它可以在这个数据解释步骤中支持（或可能取代）专家——cfr。 [39] 的开创性工作是使用 Horn 子句形式的领域本体将神经网络的输出转换为符号知识。然后，[40,41] 进一步扩展了这个想法，以用手工制作的本体来解释通用数据挖掘模式。这对于分析或过滤生物医学领域中的关联规则特别有用，可以是现有分类法 [42]、元叙词表 [43] 或手工制作的领域本体 [44] 形式的背景知识。类似地，[45,46] 使用领域知识基因本体和 KEGG 本体进行子组发现，其想法是，描述子组的构建规则很好地解释了它们的形成。随着关联数据的兴起，几种方法建议使用跨数据集的链接通过图形探索来解释序列模式 [47-49]。还探索了这一想法，以向非专业观众解释数据，特别是通过使用 DBpedia、Eurostat 和 GADM [50-52] 等多个数据集来增加表格数据和统计分析。
    - Dedalo 还使用了来自多个关联数据源的背景知识，其中执行基于归纳逻辑的图搜索，以通过无监督学习算法（集群、关联规则、时间序列）构建基于路径的数据输出解释。这些方法利用了在 Web 上发布和链接信息的“跟随你的鼻子原则”，因此避免了对特定知识图谱的先验选择来推理。这允许使用通过更深入的图探索（即不限于节点最近的邻居）收集的信息来得出解释，但代价是不使用逻辑推理，并且仅从 ABox 语句构建原子解释。 ILP 还启发了 [54] 的神经前向链接可微规则归纳网络。 TBox 和 ABox 语句使用密集向量表示，这些向量使用梯度下降进行训练，旨在学习给定谓词的最佳表示。为了应对推理的计算成本，作者使用了一种特殊的 is-a、has-a 关系分类法。
    - 表 4 总结了这些系统的一个主要特征，即生成解释的结构化知识是后验集成的，即一旦获得模型的输出。除了少数例外，这些作品的局限性在于它们依赖于手动选择知识图谱，需要专家以陈述的形式从这些有用的背景知识中提取。当依靠重用现有知识图来避免耗时的知识获取步骤时，它们的性能高度依赖于 ABox 断言中陈述的信息的新鲜度，本质上更具动态性并且会随着时间的推移而过期。领域知识图的使用后来随着事实知识的使用而转变，随着解释变得更加直观和基于事实，这表明这些系统已经从针对能够理解清晰解释的领域专家的受众转变为用户需要的系统视觉支持，以更好地理解他们的决定，从而信任他们。此类解释类型的示例，其中使用关联数据的属性和值来解释观察结果，可以在图 3 中看到。
  - **图像识别技术**

- 利用结构化知识进行基于机器学习的视觉解释的早期工作主要集中在图像识别任务上，例如 在 [55] 中，一个手动策划的空间概念、颜色、纹理及其关系的本体被合并到一个多层感知器分类器中，并用于识别图像中的 Brodatz 纹理。 这里的主要见解是使用领域知识来促进过程的透明度，充当分类器和最终用户之间的用户友好中间体。
- 这一工作一直持续到今天，大规模、公开可用的知识图谱被嵌入到更具可扩展性的学习算法中，以直观地解释模型行为（例如，可视化神经网络的隐藏状态）。例如，[56] 展示了如何利用来自 ConceptNet 的背景知识，通过将概念（例如 Dish、Person、Kitchen）和关系（例如 atLocation、 UsedFor, MadeOf) 连接到图像检索关键字（例如 Chef）。在 Microsoft 的 COCO 数据集上运行的成功实验表明，知识图谱的集成是一个值得探索的有价值的研究方向，但也表明需要重点过滤其中表示的数据以识别相关知识。 [57] 也使用集成在卷积神经网络中的视觉检测器，它利用 Wordnet 的细粒度类别（假设更接近人们可能命名的对象）来自动预测图像中的对象类别。实验表明，该模型以更有效的方式模拟了人类观察者的命名选择。
- 随着复杂的深度学习架构的出现，与知识图谱进行神经符号集成的想法也出现在图像识别任务中。 [58] 提出了一种称为面向对象的深度学习的方法，其中深度网络架构的 N 维张量被对象常识知识（具有最小的属性，如位置、姿势、尺度）替换，目标是在所有网络层中获得更可解释的视觉知识。在对象检测的背景下，[59] 中提出了使用来自知识图谱的非命题规则的多类预测。为此，CNN 架构与 ILP 框架相结合，允许基于本体规则形式的背景知识学习 OWL 类表达式。同样，[60] 的作者将通用 DNN 架构与 WordNet 结合起来用于场景分类任务。 WordNet 中的对象类型与 ADE20K 数据集中的对象对齐，然后使用 WordNet 的层次结构来训练对象识别模块，该模块进一步输入线性回归模型，能够自动提供人类可理解的解释。这两种方法都只依赖关系，限制了使用图表的潜力。这篇概念验证论文还依赖于 Suggested Upper Merged Ontology (SUMO) 的类和一个简单的关系（包含（图像，类））。
- 该领域的见解来自表明知识图谱有效地解决了最先进模型的局限性的工作，例如任务设计工作、不准确性、计算成本或数据饥饿。例如，[61] 报告了在后验过程中使用知识图来解释被卷积自动编码器错误分类的卫星图像。现有的地理知识图谱首先用于识别图像的结构（根据其附近的概念和空间关系），然后对可以证明分类器所犯错误的空间约束进行推理。 [62] 的作者将 WordNet 和视觉基因组分类法集成到基于图形搜索的 CNN 中，该 CNN 推理给定图像中的关系和概念，并在表示视觉概念的节点上产生输出（用于对图像中的对象进行分类） -标签图像识别。对分类的解释是通过跟踪图中信息的传播得出的。 [63]利用集成在马尔可夫逻辑网络中的图像和现有元数据源自动构建的自定义知识图，用于零样本可供性预测和对象识别，避免训练单独的分类器（用于对象标记、属性识别、可供性检测）. [64] 使用了一个图卷积网络，它集成了来自 NELL 的词嵌入和显式知识来学习未见类的视觉分类（零样本学习）。 [65] 还研究了在深度网络结构中集成单词级知识和知识图中表达的常识语义的想法，该想法利用 WordNet 在 ImageNet 和 Wikidata 之间创建映射。预训练的 CNN 用于对使用 OpenCV 捕获的图像进行分类，将输出转换为 WordNet 同义词集，并从它们链接到的 Wikidata 项中检索信息。
- 表 5 总结了这些工作，呈现了许多新颖的特征：模型中知识图谱的集成，使用图来解释（和调整）模型如何得出结论，使用常识知识图而不是 比事实的（可能是由于常识知识在图像中表示），以及使用各种不同的模型来生成基于知识的解释，展示了知

识图谱跨任务泛化的能力。 图 4 中给出了几个示例，我们可以在其中看到如何使用语义限制来解释模型的图像输出。 我们注意到这些方法主要关注蕴涵关系（即子类），限制了知识图谱的潜力，其特点是提取/对齐知识源的手动步骤，使得诸如缺失/错误陈述和实体歧义等问题出现知识图谱。

- 推荐系统
  - 为模型的输出提供更透明结果的知识图最近在推荐系统领域也受到了重视，目的是提高用户在满意度、信任度和忠诚度方面的体验。 大多数方法都是基于内容的，即它们包括使用来自给定知识图谱的实体以图像或自然语言句子的形式解释推荐。
  - 我们在这里发现主要是集成的方法，例如，[66] 提出 DKN 作为一种深度神经网络，将知识图谱整合到新闻推荐系统中。作者处理点击率预测的任务，将一条候选新闻和用户的点击历史作为输入，给出用户点击新闻的概率作为输出。新闻中的单词会自动与图中的实体（及其直接邻居）相关联，并最终嵌入到 CNN 使用的向量中，该向量可以预测用户的点击可能性。还结合了一个基于注意力的层，以便自动将用户的历史与候选的有趣新闻相匹配。同样，[67] 侧重于通过将协同过滤方法与知识图相结合来推荐亚马逊产品。模型中嵌入了一个特别的、自动构建的实体和用户行为图，以及一组最小属性（例如，由生成、类别、也查看）。然后使用知识图谱上的软匹配算法构建生成的个性化推荐及其以自然语言表达的解释。沿着相同的思路但重用现有知识源，[68] 的作者建议用 DBpedia 的结构替换隐藏层和自动编码器神经网络的连接，从而遵循语义感知自动编码器 [69] 的原则。为了定义电影，作者使用一小组谓词（dct:subject、dct:starring、dct:director、dct:writer），然后依靠与用户个人资料中的特征相关的权重来制定人类可以理解的解释。解释以 3 种形式呈现，基于受欢迎程度（“我们建议 X 和 Y，因为它们在与你喜欢相同电影的人中非常受欢迎”），逐点个性化（“我们猜你想看一些东西，因为它们是关于X 和 Y”）或成对个性化（“我们猜你更喜欢看 X 而不是 Y，因为你可能更喜欢 xi”）。一种非常相似的方法是 [70] 中的一种，它使用 Freebase 来增强电影和书籍的顺序推荐系统，该系统集成了循环神经网络和键值记忆网络。
  - [71] 中提出了一种事后方法，建议知识图不仅可以用于生成人类可理解的解释，还可以通过对手头任务的额外知识来增强模型，用于生成自然语言解释以使用以下方法推荐电影DBpedia 在 ExpLOD 框架之上。提出了一种排名算法来对用户喜欢的项目最相关的属性进行排名，然后将其用于构建自然语言解释。 [72] 提出了另一个有趣的论点，认为可解释的推荐系统仅限于它们仅包含来自节点最近邻居的知识，并且缺乏过滤不相关实体的适当解决方案。作者建议使用具有丰富文本内容的项目的非结构化数据来解释旅行推荐，这些文本内容可能对构建解释很有价值（例如书籍、新闻、旅游.....）。 DBpedia 用于过滤不相关的实体（通过 DBpedia 类别），而集成 DBpedia、schema.org 和 YAGO 的大规模知识图用于增强有关实体的信息，并随后用自然语言为推荐构建解释。有关附加摘要，请参见表 6。
  - 这里的可解释推荐系统的文献具有与上一节中介绍的系统相似的方面，例如。主要是在内部集成知识图谱和书面或视觉形式的解释的模型。显着差异主要在于使用大规模的开放领域知识图（主要是 DBpedia）来提取有关输入数据的附加事实，这些事实可以以用户友好的方式解释推荐，并且缺乏使用术语公理，表明在扩展推理技术方面存在困难。开放领域知识图谱还引入了信息溢出等问题，即节点的高出度限制了图中最近事实的选择，通常表现为从图中提取的多边路径的形式（cfr.example of图 5）。这

不仅阻止了更扩展的知识发现过程，而且通常还需要进行准确的修剪，以获得最终用户可以更好地信任的解释。

- 自然语言应用
  - 从社会科学中汲取灵感，认为解释还涉及从解释者向解释者传达信息的社会过程，可以在自然语言应用程序中识别大量相关工作，例如基于知识的问答（KB-QA），机器阅读理解和一般的对话式人工智能，其中知识图谱主要用作背景知识，以图像、语音和文本的形式回答常识性知识问题。
  - 例如，[75] 将 ConceptNet 和 WordNet 集成到 Knowledgeable Reader 中，这是一种基于注意力的阅读理解模型，使用从这些来源检索到的外部信息从给定文档中推断出答案。[76] 的作者在 MeRaLi 中结合了现有词汇知识库（BabelNet、NASARI 和 ConceptNet）的本体知识和推理能力，允许在概念相似性任务中建立伴随分数的解释。通过量化 COVER 空间中各个实体向量之间共享的信息量来计算两个术语之间的相似性。通过提取两个比较向量中的（显式和人类可读的）匹配属性和值，以自然语言生成分数的解释。然后在用户研究中评估自动生成的解释的质量。类似地，[77] 使用分布式语义模型和来自 WordNet 的知识以自然语言类人的理由的形式解释了文本文档之间的语义关系。
  - ConceptNet 还用作 KB-QA 中的背景知识来回答特定领域的问题，例如通过结合查询重构、结构化背景知识和文本蕴涵来解释科学问题的答案 [78] 或在 QA 模型 [79] 产生的概念之间提供常识链接。[80] 使用开放域知识图来回答转换为结构化 SPARQL 查询的单事实问题。结构为门控循环单元的循环神经网络 (RNN) 用于生成问题的表示，即检测问题中提到的主题和关系，进一步构建为 SPARQL 查询以检索所需的实体作为对象。该方法在 SimpleQuestions 数据集上进行了训练，该数据集由大约。110k 英语问题与来自 Freebase 的主题-关系-对象三元组配对。[81] 的工作是尝试扩展这一想法并使用多个相互关联的知识图谱来改进开放域问答。从自然语言问题中提取候选三元组模式（主题、属性、对象），然后与基于整数线性规划的联合推理模型对齐到 SPARQL 查询中的变量，用于从图中检索答案和附加信息。也为视觉问答 (VQA) 提供了基于知识的解释。在 [82] 中，DBpedia、WebChild 和 ConceptNet 的组合用于提取支持给定视觉问题答案的三重模式。组合的 RNN-LSTM 模型用于从与输入图像相关的图中提取事实。与 VQA 中的当前方法相反，这种方法允许一种形式的显式推理，其中答案被认为是对答案的指示性解释。这项工作扩展了 [83] 中的一个，其中对图像的推理基于从手动定义的 DBpedia 子集中提取的信息，并且答案是从预定义的问题模板构建的。[84] 提出了一种集成方法，其中通过 GCN 学习由来自 DBpedia 的事实组成的嵌入空间，以预测给定图像和有关图像的问题的正确答案。
  - [85] 将 KB-QA 的想法扩展到对话式 AI，将自动构建的领域知识图用于基于对话的 QA 系统。对话系统与用户进行科学对话，了解问题中的概念如何与科学语料库中的事实命题相关联。学习的概念和关系存储在用于解决问题和解释答案的知识图中。由于重点是使用图表来促进知识获取，因此该方法有目的地保持简单的推理机制。[86] 在讲故事的背景下也使用了浅层推理，使用来自 Framenet、Wordnet 和 Open Mind Common-sense 数据集的知识构建了一个基于 QA 的会话代理。作者利用 WordNet 的语言因果关系 cs(Cause, Effect) 和 ent(Action,Consequence) 来生成关于为什么关于故事的问题的答案，以及生成解释性句子。最近，[87] 对比了在最先进的对话系统中使用模式、框架和主题图作为背景知识，使用 DBpedia 语句作为语音识别的支持。给定

输入的语音，该图用于识别实体并从其中包含的现有三元组中得出答案。[88] 专注于对话系统在支持用户日常活动的智能助手的背景下缺乏推理透明度的问题，并展示了这如何影响关键决策情况（例如隐私、健康）。作者建议，知识图的属性（例如节点之间的语义边和路径）可用于通过利用图的拓扑特征来提供有关对话的额外上下文，从而提供对系统推理的解释。识别知识图推理时的可扩展性等问题。、

- 此处介绍的工作总结在表 7 中，并表明知识图确实为解释各种上下文（如图像、文本、语音）提供了背景知识。 图 6 中的示例进一步揭示了多个连接源的主要用法，而不是单个源。 然而，主要限制之一是知识图中信息的使用有限，主要限于节点及其最近的邻域（即“摩托车比汽车小”等单跳关系）。 这可能是由于与更先进的推理机制相关的计算成本。

- 预测和预测任务
  - 我们分析的最后一个主体是使用知识图来解释和解释预测任务，例如贷款申请、市场分析、交通动态等。这些系统依赖于这样的想法，即可以通过将原始输入数据点链接到 图形，允许通过图形导航检索有关它们的附加信息（如图 7 的示例所示）。 表 8 总结了我们在这方面的工作。
  - 在扩展的基于 Trepan 本体的决策树中，用于解释贷款预测上下文中预测模型的结果。 基于用户的评估表明，本体不仅在提高模型性能方面发挥着积极作用，而且在决策的感知可理解性方面也发挥着积极作用。[90] 还使用了域本体，并将其集成到具有受限玻尔兹曼机器模型的神经符号架构中，以解释健康社交网络中的人类行为。本体用于为用户创建符号表示并预测他们的行为，而解释是作为一组三元组的后验步骤生成的，这些三元组最大化用户行为的可能性。
  - [91] 使用现有知识图谱（Freebase 和 Wikidata）的组合来解释使用时间卷积网络对股票趋势的预测。这些图表用作外部背景知识，以获取从金融新闻数据集中提取的事件和价格值的嵌入。事件之间生成的链接用作解释意外价格变化的视觉解释。在他们的工作中，[92] 使用 CrossE 学习的知识图嵌入来搜索知识图完成任务中预测链接的解释。解释被视为预测链接的头部和尾部实体的封闭路径，并且学习的嵌入相似性允许在召回和平均支持方面识别最可靠的路径。[93] 中提出了一种基于知识图谱的迁移学习方法来解释延误航班的预测。这个想法是首先学习基于数据集和本地 OWL 本体的预测，然后用来自 DBpedia 的 TBox 断言补充学习域，以解释从一个域到另一个域的正负迁移。类似地，[94] 建议使用 RDFS 本体来增强二进制分类器的输入数据点，将它们抽象为概念，以用于得出人类可理解的解释。从 DBpedia 和 Microsoft 概念图中提取概念，然后映射到域本体。
  - 这些方法主要是后验集成知识图谱，提供对训练模型结果的解释，部分忽略本体的表达能力及其为复杂模型提供解释的推理能力。 这与这些工作在利用知识图来解释处理高维输入的模型方面的难度一致。 对解释深度网络的强烈关注，而其他模型（内核机器、线性或逻辑回归、决策树）也可能成为可以利用知识图提供更透明结果的困难模型。 最近，[7] 引起了人们的注意，指出可解释的 AI 领域在包含知识图谱方面的困难很可能是因为深度架构难以整合先验知识。

- Discussion
  - 在第 4 节中，我们展示了不同类型和不同任务的学习系统如何使用知识图谱为最终用户提供更有意义的解释。 乍一看，这表明可解释的 AI 社区，到目前为止似乎与知识表示和符

号 AI 脱节，而是可以从社区的许多技术中受益。 然而，一些挑战仍然存在并且需要调查。 回到我们最初的研究问题，本节总结了现有的基于知识的可解释（KBX-）系统的特征及其优点和局限性。

- 当前基于知识的解释系统有哪些特点?
  - 根据我们的分析，可以得出一些有趣的观察结果：
    - 在项目推荐或图像识别等领域，主要关注点主要是开发 KBX 系统，为模型的行为提供机械解释。 这与数据挖掘上下文相反，主要关注可以为输入数据生成分类解释的 KBX 系统。 在自然语言和预测应用程序中观察到这两种方法的混合，可能是由于要完成的任务种类繁多。
    - 类似地，图像识别和推荐系统主要依赖于 ABox 语句，而其他应用领域倾向于使用 TBox 和 ABox 语句的组合来生成解释。
    - 与用户交互直接相关的领域，例如 推荐器和对话式 AI 系统非常注重将知识图谱直接集成到训练模型中（模型嵌入知识）。 考虑到在深度架构中集成大规模知识的难度，基于深度架构的 KBX 系统的特点是后验符号知识（后嵌入知识）的集成。 图像识别、数据挖掘和知识发现应用程序更加多样化。
    - 根据手头的任务，可以看到 KBX 系统使用的知识图类型的明显区别。 常识知识图谱用于解释基于神经网络的系统的行为，用于图像识别和 QA 等分类任务，而事实知识用于预测和推荐。 领域知识图的使用仅限于早期系统，以激励他们在基于规则的学习中做出决策。
    - 现有知识图谱的重用似乎是最近几年的既定做法，支持共享、开放知识可以促进研究并降低开发成本的假设。 多个数据源的组合在一定程度上是经过实践的，主要用于会话任务。
    - 从现有资源中提取相关背景知识仍然是在所有任务中手动进行的。
  - 上述几点可以总结为图 8。分析的区域组织在两个主轴上，分别表示 KBX 系统嵌入知识图的方式（模型嵌入与后嵌入知识）及其所针对的解释类型 在自动生成（机械与分类解释）。 代表系统使用的知识图类型的第三个轴用于对不同区域进行颜色编码。 图 9 还显示了基于时间的研究概述。这提供了一个初步的系统概述，可以定义基于知识的解释系统领域的当前状态。
  - 作为一般观察，使用知识图生成行为解释作为后验过程仍然需要调查。 此外，在生成模型行为和数据解释的系统之间可以观察到从使用常识知识图到更特定领域的知识图的转变。 类似地，机械解释往往是使用成员断言（ABox）生成的，而分类解释则是通过利用术语知识（TBox）生成的。 这两种趋势都可以通过 KBX 系统必须推理的知识图的大小来解释，即当前的 KBX 系统依赖于复杂的架构，与分类解释的系统相比，这些架构无法对非常大的知识图进行推理，即 可以根据更小的、特定领域的图进行推理，通常计算成本更低。
- 可解释机器学习的知识图：它们有效吗?
  - 在下文中，我们将重点介绍我们在研究中发现的具体优势和局限性。简而言之：
    - KBX 系统更易于理解，因为它们以符号的、人类可读的规则的形式提供解释，但需要在生成的解释的结构和简洁性之间进行权衡，这妨碍了跨任务的泛化；

    - 重用大规模知识图谱可以提高系统的准确性，但必须采用处理噪声和有效知识提取的技术；
    - KBX 系统以计算效率为代价带来反应性。
- 解释可理解性和任务泛化之间的交易
    - KBX 系统以不同的形式表达解释，从原始文本到自然语言解释再到知情的可视化。 虽然对于什么可以被认为是完整和令人满意的解释没有达成一致，但大多数文献都认为解释是对人类需求的回答，并使用知识图作为一种手段，用用户可能不知道的额外知识来增加这些答案 的。 事实上，知识图谱提供了跨领域的背景知识，改进了生成解释的结构，为不同的任务带来了更多的信息； 同时，这种知识形式化的方式更接近于人类知识概念化，转化为人类更高的信任度，从而使知识图谱的应用在以用户为中心的应用程序中特别成功。
    - 然而，仍然不清楚的是，解释应该如何以及在何种程度上被构建，即它们是否应该被组成为强有力的论据（参见图尔明的论证模型，包括主张、理由、保证和反驳），还是紧凑和简洁的概念。 这也高度依赖于 KBX 系统中使用的背景知识类型。 依赖 ABox 断言的系统会为开放域任务生成更长的解释，但会受到信息不完整和低推理能力的影响。 依赖 TBox 的系统允许更结构化的解释，但推理的高昂计算成本限制了它们对特定领域/专家上下文的适用性。 目前，这种结构与简洁的权衡仍然高度依赖于手头的任务。
- 可重用性与大规模管理
    - 重用（通常是多个）现有知识图谱正在成为一种常见做法，这可能是由于更开放、可用的大量资源以及系统更好地扩展到它们的能力。 从这个意义上说，知识图是构建可解释系统的机会，这些系统可以生成更准确的解释，也可以跨领域重复使用。 跨这些来源管理身份是这里的一项基本要求。 数据源之间的错位和实体之间的歧义仍然是需要手动过滤和干预的问题。 这是大规模知识表示中一个众所周知的问题，其中数据发布指南被故意优先考虑到一个集中的权威，代价是在语义 Web 中不正确地使用身份，有时被定义为“身份危机”或“相同的身份” 问题”。
    - 此外，大多数基于知识的解释系统都依赖于从图中手动选择信息。 这种选择主要是出于与知识图维护相关的问题，例如 信息过时、缺失或不正确，导致定性结果丢失和模型性能下降。 这代表了 KBX 系统 w.r.t 的重大限制。 可解释人工智能的要求，其中之一是实现高性能模型，不需要从专家那里引入任何先验知识。 因此，如何在知识图中获取相关知识仍然是一个悬而未决的问题，即找到与生成基于知识的解释相关的信息。 因此，一个开放的挑战仍然是限制知识提取过程中的人工数量。
- 反应性与近似推理
    - KBX 系统利用知识图谱来补充经典和现代（深度）机器学习方法的局限性，例如需要大量训练数据并且无法转移学习任务，这表明灵活的解决方案可以支持跨任务泛化的端到端方法的开发。 此外，通过对概念、关系及其上下文进行编码，知识图谱为系统提供了整合推理和因果推理的机会，从而提高了它们的反应性和决策能力。 这种组合在神经符号方法中非常有名，旨在结合机器学习方法从经验中学习的能力，以及知识表示框架之一来推理所学内容。 从这个意义上说，知识图

谱允许开发语义互操作的解决方案，其中系统以更有效的方式交换和解释不同流程产生的信息。

- 这里出现了两个主要限制。 一方面，一个明显的问题是非常大的知识图谱的可扩展性，迫使系统在解释完整性和计算效率（在运行时）之间进行近似推理和权衡。 另一方面，KBX 系统仍然无法有效地结合学习和推理任务之间的信息来实现复杂的认知任务（涉及感知、决策和行动），也称为“绑定问题”。 这导致系统将生成解释所需的背景信息与训练数据相关联是一项手动任务，数据工程师负责从图表中查找和重用相关信息，或者在自动执行时限制在节点的最近邻域。

- 当前的挑战（以及未来的想法）
  - 最后，我们讨论了我们为基于知识的可解释系统确定的一系列开放挑战。
  - 知识图维护
    - 可解释的人工智能系统需要完整性和准确性。 这意味着大规模知识表示领域面临的一项重要挑战是增加信息覆盖率并跨领域明确表示更多知识。 此外，大型知识图谱中信息的正确性和新鲜度是必要的，不仅需要研究大规模知识图谱演化的有效方法，还需要在不需要昂贵的人力的情况下保持高质量的跨领域知识图谱的解决方案。 也可能导致资源中断。 正如已经调查的那样，集中式权威中心可能是解决该问题的潜在方法。
  - 身份管理
    - 不同知识图谱的资源之间的差异和错位是当前 KBX 系统中的一个长期问题。 管理身份是基于知识的可解释系统的一项特权，可以有效地使用可用信息并避免不良的广泛影响。 尽管存在许多用于发布和链接资源的原则，但就什么构成相同实体达成共同协议仍然是一个公开的挑战。 这也影响了可解释 AI 中知识图谱的广泛采用，无法容忍数据质量的不确定性。 这个问题的解决方案，经过部分研究，可能是帮助数据建模者和应用程序识别现实世界中相同实体的服务； 正确使用不同类型的身份链接的更好指南（例如 owl:sameAs、owl:equivalentClass、skos:exactMatch）。 此外，应调查监测和识别误用的错误检测和纠正方法。
  - 从图形中自动提取知识
    - 从现有知识图谱中获取知识仍然是一个开放的挑战，值得深入研究。 我们认为，迫切需要研究新的启发式方法，以处理当前知识图谱的规模，从而自动识别其中的正确部分信息。 考虑到知识图谱的快速增长性质，一个已经探索但需要更多努力的想法。 还应该探索应用新的网络分析方法和基于图的策略来理解手头图的性质。 这将对 KBX 系统带来好处，因为计算成本和人力资源分配都将显着降低。
  - 理解人类的角色
    - 一个开放的挑战仍然是理解人类在 KBX 系统中的作用，即他们是否应该以及在多大程度上应该参与生成解释的过程。 一些 KBX 系统表明，当人类用户向系统提供反馈时，可以获得更好的性能。 从这个意义上说，要研究的一个想法是强化学习和人在环在 KBX 系统中的适用性（迄今为止尚未探索），整合混合智能和协作 AI 的原则和方法。 人类评估还应该用于开发目前在该领域缺乏的 KBX 系统的基准，以便更好地从人类的角度理解什么是“好的解释”。 这也将支持根据类型和满意度标准更好地描述有用的解释。、

- 从知识到意义
    - 最后，我们提到的最大挑战是捕捉意义。 我们分析的 KBX 系统利用知识图谱作为事实的孤岛，从中聚合相关的三元组来支持或解释给定的观察，而不遵循任何特定的语义结构。 我们认为，知识图谱捕获的信息远远超出简单事实，因果关系、模态关系或时空关系可用于构建复杂的叙述，如主张、想法、故事、行为和经验。 通过语义模型将简单的事实转化为连贯的叙述，机器将能够像人类一样捕捉某些经验的意义，从而更连贯地解释它们背后的事件。 研究如何操纵和组合来自知识图谱的信息以支持机器包含含义，这将允许开发以人为中心的人工智能，其中机器支持人类能力而不是人类能力。

- Conclusions
    - 在这项工作中，我们研究了在可解释机器学习的背景下使用知识图谱。受到可以通过结合符号（即知识表示）和子符号（即机器学习）方法来开发新一代混合智能系统的想法的启发，我们研究了机器可读的大规模知识图谱的假设。集成在可解释的机器学习系统中，以提供更有意义、有见地和值得信赖的解释。我们分析了可解释的机器学习系统，这些系统在各种机器学习领域大规模集成结构化知识，以确定这种集成的特征、优势和局限性。我们全面介绍了 KBX 系统的现状以及一些开放的研究挑战，这些挑战可以通过将可解释 AI 和知识表示这两个社区结合在一起来解决，促进设计有效结合非常大的知识图谱的新方法和产生解释的现代学习方法。

幕布 - 极简大纲笔记 | 一键生成思维导图