多模态实体识别

多模态实体识别任务（Multimodal Entity Recognition）是指从多种模态的数据源（如文本、图像、视频、音频）中识别和提取实体的过程。这些实体可以是人名、地名、组织名、事件等各种有意义的信息单元。与传统单模态实体识别任务主要依赖于文本数据不同，多模态实体识别任务需要综合利用来自不同模态的信息，以提高识别的准确性和覆盖范围。

核心挑战

实体识别任务本身面临一系列问题：

1. **新实体的不断涌现**：命名实体不断更新，新的命名实体如新的人名、地名、组织等不断出现，增加了识别的难度。这种动态性要求识别系统具有高度的适应性和灵活性。解决思路：
    - 采用更加灵活和动态更新的模型：利用预训练语言模型（如BERT）和动态词典更新机制，使模型能够迅速适应新的命名实体。
    - 实时数据更新：结合最新的新闻、社交媒体和其他实时数据源，定期更新模型的知识库，确保对新实体的快速响应。
2. **命名实体的歧义**：同一命名实体可能对应多种不同的含义。例如，"Washington"可以指地名（华盛顿州或华盛顿特区）也可以指人名（乔治·华盛顿）。这种歧义性对准确识别提出了挑战。解决思路：
    - 利用上下文信息进行消歧：通过上下文语境分析，利用上下文编码模型（如BERT、RoBERTa）来理解实体在特定语境中的含义。
    - 语义相似度计算：结合知识图谱和语义相似度计算方法，提高对歧义实体的区分能力。
3. **复杂的实体结构**：命名实体的结构复杂多样，可能包括别名、缩略词、音译名称和包含数值的实体（如1996年10月4日）等。这种复杂性增加了识别的难度。解决思路：
    - 多层次编码：采用字符级和词级的多层次编码方法（如Char BiLSTM），捕捉复杂的实体结构信息。
    - 序列标注模型：利用BiLSTM-CRF等序列标注模型，准确识别和处理复杂实体结构。
4. **实体多样性**：同一实体可能有多种不同的称呼，例如，"John Biden"和"Sleepy Joe"实际上是指同一个人。这种多样性增加了实体识别的复杂性。解决思路：
    - 共指消解技术：使用共指消解（Coreference Resolution）技术，识别和处理同一实体的不同称呼，确保实体识别的一致性。
    - 融合上下文信息：结合上下文信息和实体关系，进一步提高识别的准确性。
5. **文本信息不足**：在某些情况下，单纯依靠文本信息可能不足以准确识别实体，需要更多的上下文或外部知识的补充。解决思路：
    - 引入外部知识库：结合外部知识库（如Wikipedia、DBpedia）提供的补充信息，增强对实体的理解和识别能力。
    - 预训练语言模型：利用预训练语言模型（如GPT-3、T5）在大量语料库上预训练的知识，提高对上下文的理解和补充信息的能力。

在多模态场景下，实体识别任务又面临一系列新涌现的挑战：

1. **模态对齐**：在多模态情景中，需要在不同模态之间进行有效对齐，使得图像和文本信息能够相互补充。然而，不同模态的数据在表达形式和语义结构上可能存在差异，如何实现高效的模态对齐是一个挑战。解决思路：

- 共享线性映射：通过共享线性映射，将不同模态的数据映射到统一的语义空间，确保各模态之间的语义一致性。
- 表征约束机制：引入表征约束机制，在对齐过程中保持各模态信息的一致性和完整性。

2. **视觉噪音**：图像中可能包含大量不相关的视觉信息，这些视觉噪音可能会干扰实体识别过程，导致识别错误。例如，在一张包含多个对象的图像中，如何筛选出与文本相关的对象是一大难点。解决思路：
   - 视觉注意力机制：引入视觉注意力机制，筛选出与文本语义相关的图像区域信息，减少噪音干扰，提高识别准确性。
   - 多任务学习：结合视觉目标检测任务，提高对相关视觉信息的筛选能力。
3. **图文语义鸿沟**：图像和文本之间存在语义鸿沟，需要构建统一的语义空间进行对齐，实现跨模态的语义理解和融合。解决思路：
   - 多模态图卷积网络：构建多模态图卷积网络，通过图卷积捕捉图像和文本之间的语义关联，实现跨模态的语义对齐。
   - 共享Transformer架构：通过共享Transformer架构，将图像和文本的语义信息进行统一编码，消除图文语义鸿沟。
4. **模态内信息不足**：单一模态的信息可能不足以全面理解实体，需要多模态信息的互补。这在一些图文共现但文本描述不充分的情景中特别明显。解决思路：
   - 多粒度融合策略：利用图像和文本的多粒度融合策略，结合区域、目标等多种粒度的语义信息，提高实体识别的准确性。
   - 引入外部数据：当模态内信息仍然不足时，通过外部数据库引入更多的补充信息，提高对实体的全面理解。
5. **多模态信息融合的复杂性**：不同模态的信息具有不同的表达形式和特点，如何有效融合这些信息是一个复杂的挑战。解决思路：
   - 拼接、注意力机制和图卷积：采用拼接、注意力机制和图卷积等多种信息融合方法，根据具体任务需求选择最优的融合策略，确保多模态信息的有效整合。
   - 自适应融合策略：开发自适应融合策略，根据不同模态的信息特点和任务需求动态调整融合方式，提高信息融合的灵活性和效果。
6. **视觉对象与文本实体的对齐问题**：图像中的视觉对象与文本中的命名实体不总是一一对应，这种不一致可能会引入偏差，影响识别性能。解决思路：
   - 去偏对比学习：使用去偏对比学习，缓解视觉对象与文本实体在类型或数量上的不一致带来的偏置问题，提高对齐准确性。
   - 强化学习：通过强化学习方法，动态调整视觉对象与文本实体的对齐策略，优化对齐效果。
7. **外部数据的引入和利用**：在多模态情景中，模态内信息有时仍不足以满足识别需求，通过引入外部数据库可以提供更多的补充信息。解决思路：
   - 基于MRC框架：使用基于MRC（机器阅读理解）的框架，将实体类型信息作为问句，引导模型进行实体识别。
   - 多模态知识库：结合多模态知识库，利用外部数据提供的补充信息，提高对实体的全面理解和识别能力。

技术路线

粗粒度融合

粗粒度融合方法主要关注获取整个图像的全局语义信息，并将其与文本信息进行融合。通过这种方式，可以获得图像全局语义感知的文本表征，提升实体识别的准确性。

核心工作

1. Multimodal Named Entity Recognition for Short Medical Posts-2018-NAACL
    - **挑战**：社交媒体中的推文往往存在语法不规范、信息不完整的问题，这给传统的基于纯文本的实体识别方法带来严峻挑战。同时，社交媒体中的推文逐渐以图文共现形式为主，配带的图像可以为文本中的实体提供辅助信息。
    - 解决方案：
        - **文本编码器**：采用GLoVE进行词编码，使用Char BiLSTM进行字符编码，捕捉文本中的语义信息。
        - **图像编码器**：使用CNN进行全局图像编码，获取图像的全局语义表征。
        - **模型骨架**：基于BiLSTM+CRF的序列标注模型，通过BiLSTM捕捉上下文信息，CRF学习标签分布约束。
        - **图文特征融合**：采用Modality Attention机制，在每个时间步将字符特征、词特征和图像特征进行融合，提升实体识别性能。

细粒度融合

技术路线 细粒度融合方法注重获取图像的局部语义信息，如区域、目标等，并将这些细粒度信息与文本信息进行融合。这种方法能够更准确地捕捉图像中的局部细节，从而提升文本表征的准确性。

核心工作

1. **Visual Attention Model for Name Tagging in Multimodal Social Media-2018-ACL**
    - **挑战**：文本语义存在不确定性，当文本信息较短时，难以准确理解实体的真实含义。附带的图像信息具有消除文本中实体歧义的作用，然而如何有效利用这些信息是一个难题。
    - 解决方案：
        - **文本编码器**：采用GLoVE进行词编码，使用Char BiLSTM进行字符编码，捕捉文本中的语义信息。
        - **图像编码器**：使用ResNet-152进行图像区域编码，获取图像的区域表征。
        - **视觉注意力**：引入文本引导的视觉注意力机制，获取与文本整体语义相关的视觉区域特征。
        - **图文特征融合**：采用基于门控机制的方法，有选择地融合文本特征与视觉区域特征，提高实体识别的准确性。
2. **Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts-2020-MM**
    - **挑战**：全局图像信息可能会引入大量视觉噪音，影响模型的实体识别性能。图像中的目标信息可为实体识别提供更精细化的信息，然而如何准确对齐图像目标和文本实体是一个难点。
    - 解决方案：
        - **文本编码器**：采用GLoVE进行词编码，使用Char BiLSTM进行字符编码，捕捉文本中的语义信息。
        - **图像编码器**：使用Mask RCNN进行图像目标编码，获取图像中的目标表征。

        - **图文特征融合**：提出Dense Co-Attention Layer，通过Self-Attention学习模态内目标或实体间的关联性，通过Guided-Attention学习模态间目标与实体的关联性。
        - **多头注意力机制**：利用多头注意力机制，进一步增强图文特征融合的效果。
3. **Multi-Modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance-2021-AAAI**
    - **挑战**：图像与文本之间细粒度语义单元之间（图像目标与文本实体）的对应关系没有被充分挖掘。如何构建一个统一的语义空间，消除图像与文本之间的异构鸿沟与语义鸿沟，是一个挑战。
    - 解决方案：
        - **文本编码器**：采用BERT Encoder进行词编码，捕捉文本中的深层语义信息。
        - **图像编码器**：使用ResNet Encoder进行图像编码，获取图像的全局语义表征。
        - **多模态图构建**：根据图像中的目标与文本中实体之间的对应关系，构建一个统一的多模态图，通过多模态图卷积网络（G(V,E)）捕捉模态内和模态间的语义关联。
        - **模态交互**：堆叠多层基于图的多模态融合层，每个融合层都包含模态内与模态间的交互，增强图文语义融合的效果。
4. **Reducing the Bias of Visual Objects in Multimodal Named Entity Recognition-2023-WSDM**
    - **挑战**：图像中目标与文本中实体并不总是一一对应，可能会引入视觉偏置，影响模型性能。图像中目标类型与文本中实体类型不一致，图像中目标个数与文本中实体个数不一致，如何处理这些问题是关键。
    - 解决方案：
        - **文本编码器**：采用BERT Encoder进行词编码，捕捉文本中的深层语义信息。
        - **图像编码器**：使用ResNet Encoder进行图像编码，获取图像的全局语义表征。
        - **模态交互**：堆叠多层模态内和模态间的交互层，基于Multi-Head Self-Attention（SA）和Multi-Head Cross-Attention（CA），增强图文特征的融合。
        - **去偏对比学习**：通过对比学习有效拉近正样本之间的距离，并推远负样本之间的距离，采用硬样本挖掘策略和去偏见对比学习损失，缓解目标与实体数量不一致带来的偏置问题。

引入外部知识的多模态实体识别

引入外部知识的方法通过利用外部数据库或知识库，为多模态实体识别提供更多的补充信息，特别是在模态内信息不足时，这种方法尤为重要。

核心工作

1. **Query Prior Matters: A MRC Framework for Multimodal Named Entity Recognition-2022-MM**
    - **挑战**：现有的数据集中缺少细粒度图像标注，难以获取图像中目标与文本中实体之间更加精确的对应。实体类型可以作为图像目标与文本实体之间的连接桥梁，通过这种方式，可以更精细化地进行实体识别。
    - 解决方案：
        - **问题构建**：根据实体类型的标注说明构建Query，将Query与原始输入文本拼接，得到新的文本序列，并基于BERT获得词表征序列。
        - **视觉定位**：基于构建的问句以及训练好的视觉定位工具，获取图像中top k个目标图像，基于ResNet获取其编码表征。

        - **模态交互**：采用图文交互模块和文本内交互模块，提高不同模态信息的融合效果。
        - **多任务联合预测**：结合多任务学习方法，分别进行句子中实体类型的判断、目标区域置信度估计和实体span预测，提高实体识别的综合性能。
2. **ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition-2022-NAACL**
    - **挑战**：基于不同的编码器获得图像表征和文本表征，由于训练所使用的数据集和模型架构不同，导致他们的语义表征并不在相同的语义空间。
    - 解决方案：
        - **图像内容获取**：将图像中包含的信息转为文本形式，获取图像中的目标标签词、图像文本描述和图像中的文本字符。
        - **图像与文本特征交互**：将获得的图像内容与输入的文本进行拼接，并送入到Transformer编码器中，基于Transformer中的多头注意力模块可以学习模态内与模态间的语义交互。
        - **跨模态对齐**：考虑两种情形（图像中含有噪音和图像信息缺失），提出基于输出概率分布的隐式对齐，消除图像和文本之间的语义鸿沟。
3. **Chain-of-Thought Prompt Distillation for Multimodal Named Entity Recognition and Multimodal Relation Extraction-2024**
    - **挑战**：多模态实体识别需要基本的推理能力来理解复杂的语言和多模态信息。基于数据库的检索增强可能会导致检索的信息与原始信息之间存在语义偏差，进而对模型的引起误导。
    - 解决方案：
        - **思维链提示蒸馏**：提出一种思维链（CoT）提示蒸馏方法，基于知识蒸馏技术将大模型（教师模型）的推理能力转移到小模型（学生模型）。
        - **多样化思维链提示**：提出基于多样化的思维链提示引导LLMs从多粒度和数据增强的角度解释每个样本，将不同CoT提示的解释组合在一起，称为CoT知识。
        - **知识蒸馏**：利用知识蒸馏技术，将CoT知识和图像以特定任务的形式蒸馏到学生模型中，提高学生模型在复杂任务中的推理性能。

多模态属性抽取

多模态属性抽取任务是指从文本、图像和音频等多种模态数据中提取实体的相关属性及其属性值。它包括属性预测和值提取两个子任务，即从非结构化数据中识别并提取出与特定实体相关的属性及其具体值。

核心挑战

任务自身面临的困难

1. **标签缩放问题**：
    - 多分类式的属性抽取方法将任何目标属性值视为类标签，当属性值的数量巨大时（数千或更多），模型的训练和预测过程会变得非常复杂和困难。
    - 需要大量标注数据来支持大规模的标签集，增加了数据标注的成本和难度。
    - **解决方案**：采用更加灵活的标签表示方法，如序列标注、问答式或生成式的方法，减少对大规模标签集的依赖。
2. **封闭世界假设**：
    - 传统多分类式方法假设所有可能的属性值都已知，无法发现训练数据中标签集之外的新值。
    - **解决方案**：引入开放世界假设，通过问答式或生成式的属性抽取方法，允许模型识别和生成未知的属性值。

3. **标签独立性假设**：
    - 将每个属性值视为彼此独立，忽略了属性值之间的依赖关系，影响了模型的整体性能。
    - **解决方案**：利用属性值之间的依赖关系，通过联合建模和多任务学习方法提高模型的识别能力。
4. **属性多样性与复杂性**：
    - 实体属性多样且复杂，可能包含别名、缩略词、音译名称和包含数值的实体等。
    - **解决方案**：采用多层次编码方法（如字符级和词级编码），利用先进的序列标注模型（如BiLSTM-CRF）进行准确的属性识别。
5. **属性体系庞大且分布不均**：
    - 产品属性体系庞大，可能跨领域重叠，属性值的分布通常严重倾斜，大多数属性集中在少数产品中。
    - **解决方案**：使用主动学习框架，优先标注置信度最低的样本，并结合多任务学习和数据增强技术，提升对少见属性的识别能力。
6. **有限的标注数据**：
    - 标注数据通常有限，无法提供大量标注数据，同时模型需要具有可解释性。
    - **解决方案**：引入主动学习和迁移学习技术，减少对大规模标注数据的依赖，同时提高模型的解释性。
7. **新属性的引入**：
    - 新属性不断出现，现有模型难以及时应对和适应新的属性需求。
    - **解决方案**：利用问答式和生成式方法，通过动态更新和开放世界假设，增强模型对新属性的适应能力。

在多模态场景下额外暴露的问题

1. **模态对齐**：
    - 多模态属性抽取需要将文本、图像和音频等不同模态的数据进行有效对齐，实现模态间信息的互补和增强。
    - **解决方案**：采用共享线性映射和表征约束机制，实现不同模态数据的语义对齐。
2. **视觉噪音**：
    - 图像中可能包含大量不相关的视觉信息，这些噪音可能干扰属性的正确识别。
    - **解决方案**：引入视觉注意力机制，筛选出与属性相关的图像区域，减少噪音干扰，提高识别准确性。
3. **图文语义鸿沟**：
    - 图像和文本之间存在语义鸿沟，需要构建统一的语义空间进行对齐，确保不同模态信息的一致性。
    - **解决方案**：采用多模态图卷积网络，通过图卷积捕捉图像和文本之间的语义关联，消除图文语义鸿沟。
4. **模态内信息不足**：
    - 单一模态的信息可能不足以全面描述属性，需要结合多模态信息进行补充和增强。
    - **解决方案**：利用多粒度融合策略，将图像和文本的多种粒度信息进行融合，提高属性抽取的全面性和准确性。
5. **多模态信息融合的复杂性**：
    - 不同模态的信息具有不同的表达形式和特点，如何有效融合这些信息是一个复杂的挑战。

- **解决方案**：采用拼接、注意力机制和图卷积等多种信息融合方法，根据具体任务需求选择最优的融合策略。

6. **外部知识的引入和利用**：
   - 当模态内信息不足时，可以通过引入外部数据库或知识库，提供更多的补充信息。
   - **解决方案**：使用基于MRC框架的方法，将实体类型信息作为问句，结合外部数据进行属性抽取。

核心工作

1. **OpenTag: Open Attribute Value Extraction from Product Profiles. KDD 2018**
   - **挑战**：产品简介中存在开放世界假设，属性值词汇表不再是有限且预定义的。此外，属性的堆叠和不规则的结构使信息高度紧凑，增加了抽取难度。
   - 解决方案：
     - **序列标注**：借鉴命名实体识别的思路，采用BiLSTM+CRF模型，结合注意力机制，用于解释模型的标记决策。
     - **主动学习**：引入主动学习框架，以减轻人工注释的负担，提高标注效率。
     - **置信度和标签翻转**：通过最低置信度（Least Confidence）和标签翻转（Tag Flip）策略，优化模型性能。
2. **Scaling up Open Tagging from Tens to Thousands: Comprehension Empowered Attribute Value Extraction from Product Title. ACL 2019**
   - **挑战**：在开放标签集下，属性值数量庞大，标签缩放问题严重，模型需要处理成千上万的标签。
   - 解决方案：
     - **问答式属性抽取**：以属性作为问题，在文本中标注答案（属性值），允许属性无限扩展。
     - **蒸馏掩码语言模型**：改进对完全不可见的属性和值的推广能力，增强模型的泛化性。
     - **无答案分类器**：引入无答案分类器，提高模型预测属性值不存在的能力。
3. **Learning to Extract Attribute Value from Product via Question Answering: A Multi-task Approach. KDD 2020**
   - **挑战**：属性的开放性和不规则性使得传统方法难以应对，同时需要在有限的标注数据下实现高效抽取。
   - 解决方案：
     - **多任务学习框架**：将属性预测和值提取任务合并到一个统一的学习框架中，提高模型的综合能力。
     - **AE-pub数据集**：使用包含110K样本、2.7K唯一属性、10K唯一属性值的AE-pub数据集进行训练，提高模型的泛化能力。
     - **问答式抽取**：利用问答方式提取属性，允许属性无限扩展，并通过蒸馏掩码语言模型改进推广能力。
4. **Multimodal Attribute Extraction. AKBC 2017**
   - **挑战**：在电商产品中，属性数据存在于文本和图像中，如何有效地从多模态数据中抽取属性是一个挑战。
   - 解决方案：

        - **多模态数据结合**：从文本和图像中提取属性值，并使用Diffbot产品API爬取商业网站数据进行属性值抽取。
        - **基准模型**：提出一个多模态数据抽取的基准模型，通过Concat融合（拼接+全连接）和GMU（Gated Multimodal Unit）融合机制，决定何时看图、何时读字，增强模型性能。
5. **Multimodal Joint Attribute Prediction and Value Extraction for E-commerce Product. EMNLP 2020**
    - **挑战**：电商产品属性预测和值提取需要同时进行，如何有效地结合多模态信息进行联合建模是一个难题。
    - 解决方案：
        - **多任务学习框架**：应用多任务学习框架进行属性预测和值提取的联合建模。
        - **全局门控跨模态注意层**：引导属性预测任务，通过KL损失惩罚产品属性预测的分布与值提取的分布之间的不一致，提高模型一致性。
        - **区域门控跨模态注意层**：引导值提取任务，从区域层面进行细粒度的属性值抽取。
6. **Large Scale Generative Multimodal Attribute Extraction for E-commerce Attributes. ACL Industry Track 2023**
    - **挑战**：电商产品的属性值数量庞大且分布不均，生成式多模态属性抽取需要处理多个属性并进行高效扩展。
    - 解决方案：
        - **生成式多模态属性抽取**：以属性作为问题，基于文本直接生成答案（属性值）或属性+答案。
        - **可扩展性**：处理多个属性，不需要为每个属性组合单独训练一个模型。
        - **多模态信息提取**：从文本、图像、视频等多种模态中提取属性值，支持零射推理和无值推断，提升模型的广泛适用性。

多模态关系抽取

多模态关系抽取任务旨在从多模态数据（如文本和图像）中识别并提取实体间的关系。具体分为两种形式：

- **给定实体对的关系抽取**：给定输入文本以及文本中的实体对，预测实体之间的语言关系。
- **实体关系联合抽取**：只给定输入文本，同时完成实体识别与关系抽取。

核心挑战

关系抽取任务本身存在的挑战

1. **文本信息不足**：
    - 文本中信息量有限，可能无法全面反映实体间的关系。
    - **解决方案**：通过增加数据源或采用外部知识库增强信息量，提高模型对实体关系的理解。
2. **上下文信息捕捉困难**：
    - 实体关系的识别依赖于文本上下文的信息捕捉，尤其是长距离依赖关系。
    - **解决方案**：采用先进的编码器（如BERT）和上下文捕捉机制（如BiLSTM-CRF），提高上下文信息的捕捉能力。
3. **关系类型多样**：
    - 实体间的关系类型多样且复杂，可能存在多种潜在关系。

- **解决方案**：利用多任务学习模式和多层次分类器，提高对多种关系类型的识别能力。

4. **数据标注困难**：
   - 大规模关系抽取任务需要大量标注数据，获取这些数据成本高且难度大。
   - **解决方案**：引入主动学习和迁移学习技术，减少对大规模标注数据的依赖。

多模态场景下额外暴露的挑战

1. **模态对齐**：
   - 不同模态（如文本和图像）之间的信息需要进行有效对齐，确保模态间信息的互补性。
   - **解决方案**：采用基于图对齐的多模态关系抽取模型（如MEGA），结合视觉对象和文本实体之间的结构相似性和语义一致性，实现更好的文本和视觉关系对齐。
2. **视觉噪音**：
   - 图像中可能包含大量不相关的视觉信息，这些噪音可能会干扰关系识别过程。
   - **解决方案**：引入层次化的视觉前缀融合网络，将视觉表征以可插拔的前缀提示加入到文本编码器的注意力层，减少视觉噪音的干扰。
3. **语义一致性**：
   - 确保检索到的文本和视觉线索的语义与原始输入文本和图像的语义一致，以避免引入无关信息。
   - **解决方案**：通过跨模态检索增强的方法，从不同粒度（如视觉对象、句子、图像等）进行检索，获取更多多模态线索，并基于注意力机制捕获模态间的交互。
4. **模态内信息不足**：
   - 单一模态的信息可能不足以全面理解实体关系，需要多模态信息的互补。
   - **解决方案**：引入基于外部知识库检索的方法，通过外部知识增强信息，缓解数据内部信息不足的问题。
5. **预训练方法的局限性**：
   - 现有的多模态预训练方法在模态对齐时，大多只考虑粗粒度的文本与图像对齐，忽略了细粒度的实体与关系对齐。
   - **解决方案**：提出新颖的多模态预训练方法，充分利用大量无标注的图像-文本描述对语料，学习图像和文本之间的目标-实体与图像-关系的对应。

技术路线

基于数据内部信息的多模态关系抽取

充分利用数据集内部的多模态信息，通过融合语义和结构信息，实现实体关系的统一抽取或联合抽取。

1. **Multimodal Relation Extraction with Efficient Graph Alignment-2021-MM**
   - **挑战**：图像中目标之间的关系（如场景依赖关系、位置关系等）对判断文本中实体之间的语义关系有帮助，然而如何有效结合这些信息是一个难题。
   - 解决方案：
     - **图对齐模型**：提出一种基于图对齐的多模态关系抽取模型（MEGA），结合了图像中的视觉对象和句子中的文本实体之间的结构相似性和语义一致性。
     - **多模态对齐**：同时考虑模态间的语义与结构对齐，通过结构和语义特征找到两个图之间最相似的节点，实现更好的文本和视觉关系对齐。
     - **文本结构表征**：基于依存树表示文本结构，使用BERT编码器捕捉文本语义。

        - **视觉结构表征**：使用Faster R-CNN编码器获取视觉语义表征，基于场景图表示视觉结构。
        - **关系预测**：集成所有对齐后的视觉节点特征，并与实体语义进行拼接，用于预测关系类型。
2. **Good Visual Guidance Makes A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction-2022-NAACL**
    - **挑战**：不相关的显著视觉对象会误导模型对文本实体语义的理解，而完全相关的视觉对象又很难获取。
    - 解决方案：
        - **层次化视觉前缀融合**：提出一种层次化的视觉前缀融合网络，将视觉表征以可插拔的前缀提示加入到文本编码器的注意力层，获得图像感知的文本表征。
        - **模态间不同层次特征的交互**：同时考虑图像和文本中的层次化特征，模态间不同层次特征的交互有助于获得更好的文本语义表征。
        - **视觉前缀引导**：在原始文本编码器（BERT）中的每个注意力块中，将视觉特征以可插拔的前缀形式融合到原始文本的多层自注意力操作中。

引入数据外部信息的多模态关系抽取

通过引入外部知识库或无标注语料，以增强信息，缓解数据内部信息不足的问题。

1. **Multimodal Relation Extraction with Cross-Modal Retrieval and Synthesis-2023-ACL**
    - **挑战**：现有数据集中大部分图文对之间的关联性不强，仅依靠数据内部信息可能会引入很多无关的信息。
    - 解决方案：
        - **跨模态检索增强**：提出一种基于跨模态检索增强的方法，从视觉对象、句子、图像等不同粒度进行跨模态检索，以获得更多的多模态线索。
        - **文本线索检索**：基于Google Vision APIs获取实体词列表和图像文本描述，并与原始文本一起送入文本特征编码器。
        - **视觉线索检索**：基于Google custom search API获得与文本余弦相关的图像，并与原始图像一起送入图像特征编码器。
        - **跨模态交互**：基于注意力机制捕获模态间的交互，确保检索到的文本和视觉线索的语义与原始输入文本和图像的语义一致。
2. **Prompt Me Up: Unleashing the Power of Alignments for Multimodal Entity and Relation Extraction-2023-MM**
    - **挑战**：带标记的多模态数据稀缺，现有多模态预训练方法在模态对齐时多为粗粒度对齐，忽略细粒度的实体与关系对齐。
    - 解决方案：
        - **多模态预训练方法**：提出一种新颖的多模态预训练方法，充分利用大量无标注的图像-文本描述对语料，学习图像和文本之间的目标-实体与图像-关系的对应。
        - **对齐损失**：在预训练任务中引入三种对齐损失：Image-Caption对齐、Entity-Object对齐、Image-Relation对齐，以增强模态间的对齐能力。
        - **基于提示的Entity-Object对齐**：通过生成实体的软伪标签，与目标类型概率分布对齐，提高细粒度的文本实体与视觉目标之间的对齐能力。

基于统一抽取框架的多模态关系抽取

将多模态实体识别和关系抽取任务结合，通过统一的框架进行联合建模，实现实体与关系的统一抽取。

1. **Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion-SIGIR-2022**
    - **挑战**：多模态实体识别和多模态关系抽取通常被当作独立任务进行探究，不同任务的架构不具有通用性。
    - 解决方案：
        - **统一Transformer框架**：提出一种统一的基于Transformer的框架，能够处理多种多模态知识获取任务。
        - **多级融合策略**：提出包含粗粒度前缀引导交互模块（PGI）和细粒度关联感知融合模块（CFM）的多级融合策略，减少模态异质性和不相关的视觉噪声。
        - **统一输入格式**：图像和文本均采用统一的输入格式，前LK（LV或者LT）层用于特定模态编码，后LM层用于跨模态交互编码。
2. **A Unified MRC Framework with Multi-Query for Multi-modal Relation Triplets Extraction-ICME-2023**
    - **挑战**：多模态命名实体识别和多模态关系抽取任务并不独立，需要结合文本和图像模态来抽取关系三元组。
    - 解决方案：
        - **统一的机器阅读理解（MRC）框架**：通过二阶段方式实现关系三元组抽取，先抽取实体再抽取关系，执行多模态实体识别和关系抽取。
        - **多查询机制**：在每个阶段构建三个具有相同含义的问题，将丰富的先验知识（如实体和关系类型）融入到原始文本数据中，增强模型对文本语义的理解。
        - **模态间融合**：将文本与视觉编码拼接，基于相同的Transformer架构进行特征交互与融合，实现实体分类和关系分类。

多模态知识补全

多模态知识补全旨在通过引入多种模态的数据（如图像、文本、视频、音频等），提升知识图谱的完整性和准确性。在多模态知识图谱中，知识三元组的实体或属性值不仅限于文本模态，还可以包括其他模态的数据。这种任务的核心目标是根据已有的知识图谱数据，预测缺失的实体关系三元组，使知识图谱更加完整。

核心挑战

知识补全所面临的挑战

1. **噪声信息**：
    - **挑战**：知识图谱通常从非结构化或半结构化数据中构建，往往存在大量噪声信息，这些噪声信息会干扰知识补全的效果。
    - **解决方案**：需要设计有效的去噪算法和策略，识别和过滤掉噪声信息，确保知识补全的准确性。
2. **知识不完备**：
    - **挑战**：由于知识获取的信息有限，知识图谱往往存在不完备的问题，缺失大量实体关系三元组。
    - **解决方案**：通过知识推理和嵌入表示，预测和补全缺失的三元组。

3. **三元组预测**：
    - **挑战**：基于已有实体关系三元组数据，准确预测缺失的三元组成分，即(h,r,?), (h,?,t), (?,r,t)。
    - **解决方案**：利用知识图谱嵌入(KGE)技术，对三元组进行评分和排序，判断其成立性，进而进行补全。
4. **多模态信息的融合与噪声引入**：
    - **挑战**：在多模态场景下，如何有效融合视觉信息与文本信息，并减少噪声引入，是一大难题。
    - **解决方案**：通过设计多模态融合机制（如拼接、门控、注意力机制等），增强知识表征，同时控制噪声引入。

多模态场景下知识补全带来的变化

1. **丰富的视觉信息**：
    - **变化**：图片能够提供更多关于实体的视觉细节信息，帮助理解实体之间的关系，进而辅助知识补全任务。
    - **作用**：视觉信息可以更好地反映三元组事实，如“头盔”是“铠甲”的“一部分”。
2. **多模态特征的交互依赖**：
    - **变化**：在多模态知识图谱中，需要建立不同模态特征的交互依赖，以提高知识表征的准确性。
    - **作用**：通过模态编码和模态融合机制，实现多模态特征的有效交互，提升知识补全效果。
3. **噪声控制与信息过滤**：
    - **变化**：引入多模态信息时，可能带来额外的噪声，需要有效的噪声控制和信息过滤机制。
    - **作用**：利用注意力机制和门控机制，过滤不相关的图像信息，确保融合的信息是有用的。

常规知识补全技术路线

- 基于平移距离的方法：这类方法通过定义距离得分函数，使得成立的三元组具有较低的距离得分，利用这种距离进行知识补全。
    - **基本思想**：假设尾实体是由头实体通过关系进行平移或旋转得到的。通过定义距离函数，使得正样本的距离最小化，负样本的距离最大化。
    - **代表方法**：TransE, TransH, TransR, RotatE等。
- 基于语义匹配的方法：这类方法通过多个低维的矩阵或张量的积代替原始的关系矩阵，从而用少量的参数代替稀疏而大量的原始数据。
    - **基本思想**：用相似度的得分函数或双线性函数，通过矩阵或张量分解捕捉实体和关系之间的语义关系。
    - **代表方法**：RESCAL, DistMult, ComplEx, ANALOGY等。
- 基于神经网络的方法：这类方法利用卷积神经网络（CNN）或图神经网络（GNN）等深度学习模型，对实体和关系的嵌入进行特征提取。
    - **基本思想**：通过一维、二维卷积或图卷积对实体和关系的嵌入进行特征提取，捕捉复杂关系。
    - **代表方法**：ConvE, ConvKB, R-GCN, SACN等。
- 基于语言模型的方法：这类方法利用预训练的大规模语言模型（LLM），通过编码或生成的方式进行知识补全。
    - **基本思想**：利用大语言模型的强大文本生成和编码能力，对知识图谱中的三元组进行编码和预测。
    - **代表方法**：KG-BERT, GenKGC等。

多模态知识补全代表性工作

1. IKRL (Image-embodied Knowledge Representation Learning) - IJCAI 2017
    - **核心挑战**：
        - 传统知识图谱方法忽略了从实体图像中获取的视觉信息。
        - 如何在知识表示学习中有效利用实体的图像信息。
    - **核心思路**：
        - **视觉信息增强**：利用实体图像提供的视觉信息，补充传统方法仅从结构化三元组中学习知识表示的不足。
        - **图像编码器**：使用预训练的AlexNet模型提取图像特征，并通过投影矩阵将图像特征映射到实体空间中。
        - **多实例学习**：基于注意力机制，选择最合适的图像信息来增强实体表示。
        - **三元组约束**：保持实体在图像和结构两个方面的三元组约束，确保多模态信息的一致性。
2. RSME (Relation Sensitive Multi-modal KG Embedding) - MM 2022
    - **核心挑战**：
        - 不同关系类型对视觉信息的依赖程度不同，有些关系在视觉上难以反映。
        - 如何自动决定视觉模态信息对不同关系的影响。
    - **核心思路**：
        - **关系敏感性**：根据关系自动鼓励或过滤视觉模态信息的影响，确保视觉信息对特定关系的适用性。
        - **结构和图像编码器**：使用ComplEx作为结构编码器，VGG16或ResNet50作为图像编码器。
        - **过滤门和遗忘门**：通过过滤门选择与其他图像相似度最高的图像，通过遗忘门根据关系的预测分数，设置图像信息的遗忘分数。
        - **融合门**：将图像和结构信息进行线性组合，提升多模态嵌入的效果。
3. OTKGE (Optimal Transport Knowledge Graph Embeddings) - NeurIPS 2022
    - **核心挑战**：
        - 不同模态的嵌入空间是异构的，简单融合会破坏不同模态嵌入的内在空间结构。
        - 如何在多模态嵌入过程中保持不同模态的空间一致性。
    - **核心思路**：
        - **最优传输对齐**：通过最优传输理论，建模多模态分布之间的Wasserstein距离，确保不同模态嵌入空间的一致性。
        - **模态融合**：将文本嵌入和图像嵌入分别转移至结构嵌入空间，然后寻找使得文本、图像、结构总成本最小的融合方式。
        - **损失函数**：基于模态融合嵌入，建立三元组约束关系，利用TransE或RotatE等方法进行关系映射，计算三元组成立的得分函数。
4. LAFA (Link Aware Fusion and Aggregation) - AAAI 2024
    - **核心挑战**：
        - 实体的视觉信息在不同三元组场景中贡献不同，简单融合会引入噪声。
        - 如何根据具体的三元组关系选择和融合最相关的视觉信息。

- **核心思路**：
  - **链路感知融合**：将头尾实体涉及的多个图像进行融合，提出模态交互注意力机制，根据不同的链路评估不同图像的重要性。
  - **链路感知聚合**：对头实体的不同尾实体信息进行聚合，基于初始的实体嵌入映射为一个查询和一个键向量，计算不同尾实体的注意力权重。
  - **多头注意力机制**：引入多头注意力机制，提升稳定性，最终利用ConvE机制作为三元组成立性的得分函数。

多模态实体对齐

实体对齐旨在判断两个或多个来自不同知识图谱的实体是否指代相同事物。通过对齐不同语言、来源、领域的知识图谱中的实体，能够提升知识图谱的知识质量和覆盖范围，为下游任务提供更为丰富和全面的知识支撑。

在多模态知识图谱中，实体不仅包含文本信息，还包括图像作为属性信息。这些图像提供了丰富的视觉描述信息，通过挖掘这些多模态描述信息，可以为实体对齐提供有力的辅助特征，进而帮助更好地识别候选实体之间的等价性。

重要问题

- **候选实体的一致性信息挖掘**：如何从不同图谱中挖掘出候选实体之间的一致性信息是实体对齐任务的重要问题。由于不同知识图谱在语言、来源和领域上的差异，实体的表示方式也会有所不同。有效挖掘这些实体在不同图谱中的一致性信息，是识别它们是否等价的关键。
- **多模态信息的利用**：如何有效利用多模态信息（如图像）来辅助实体对齐。在多模态知识图谱中，实体的图像信息可以提供更丰富的描述，有助于弥补单一文本信息的不足。利用这些多模态信息，可以增强实体的描述，从而提高实体对齐的准确性。

常规实体对齐技术路线

基于平移嵌入的方法

1. **MTransE**：
   - **核心思路**: 利用TransE模型编码实体和关系表征，并引入对齐损失以保持等价三元组在两个图谱中的语义一致性。
   - 实现方案:
     - **TransE模型**: 假设关系通过向量平移连接头实体和尾实体，公式为 $h+r \approx t$。
     - **对齐损失**: 定义一个对齐损失函数，以度量不同知识图谱中等价实体的嵌入向量之间的距离，优化过程使等价实体的嵌入向量尽可能相似。
     - **优化目标**: 最小化嵌入向量之间的距离，同时保留TransE的结构约束。
2. **IPTransE**：
   - **核心思路**: 扩展MTransE方法，通过引入关系路径的语义信息增强实体表征能力。
   - 实现方案:
     - **路径表征**: 利用PTransE模型考虑关系路径的语义信息，采用向量加法、向量乘法、RNN编码等方法进行路径表征。
     - 三种对齐策略:
       - **平移距离策略**: 引入代表对齐语义的特殊关系。

        - **线性变换策略**: 利用线性变换建立语义空间关联。
        - **参数共享策略**: 等价实体在两个图谱中共用表征向量，最佳实践为参数共享策略。
    - **优化目标**: 优化路径表征与原有关系表征的语义一致性。
3. **MultiKE**:
    - **核心思路**: 将实体的名称、关系、属性视为不同的视图，分别进行编码，并融合这些视图的实体表征以实现实体对齐。
    - 实现方案:
        - **名称视图编码**: 利用词向量表示实体名称。
        - **关系视图编码**: 通过三元组损失进行表征，假设关系通过向量平移连接实体。
        - **属性视图编码**: 将实体的属性名和属性值拼接为特征矩阵，利用卷积网络提取特征进行表征。
        - **融合表征**: 融合不同视图的实体表征，综合考虑名称、关系、属性信息，进行实体对齐。

基于图神经网络的方法

1. **GCN-Align**:
    - **核心思路**: 首个基于图神经网络的实体对齐方法，通过共享参数的GCN编码两个知识图谱中的实体。
    - 实现方案:
        - **图神经网络**: 利用GCN对实体的结构表征和属性表征进行拼接编码。
        - **语义匹配**: 分别计算结构表征和属性表征的相似度，度量不同知识图谱中等价实体的嵌入向量之间的相似性。
        - **优化目标**: 通过最小化对齐损失，优化实体的嵌入表示，使得等价实体在不同图谱中的嵌入向量尽可能相似。
2. **HGCN**:
    - **核心思路**: 利用Highway-GCN捕捉高阶语义关联，增强实体语义匹配。
    - 实现方案:
        - **Highway-GCN**: 在每层GCN输出中引入highway机制，允许网络在不同层次的信息之间进行灵活传递。
        - **全局关系编码**: 对给定关系在知识图谱中的所有头实体、尾实体进行均值池化，并通过线性变换获得全局关系编码。
        - **语义匹配**: 利用L1范数度量实体对的相似度，并通过margin loss进行训练，以优化实体对齐。
3. **NMN**:
    - **核心思路**: 通过对实体进行邻居采样和子图匹配，增强模型对不同知识图谱结构差异的鲁棒性。
    - 实现方案:
        - **结构嵌入**: 利用GCN进行实体编码，捕捉实体的局部结构信息。
        - **邻域采样**: 利用注意力机制，选择与实体最相关的邻居实体进行采样。
        - **邻域匹配**: 对给定实体的每个邻居与候选实体的每个邻居进行相似性匹配，并按相似性聚合邻居差异信息，与原始邻居表征融合。
        - **邻域聚合**: 将考虑差异信息的邻居表征进行聚合，用于最终语义匹配，实现不同知识图谱之间的实体对齐。

多模态实体对齐

多模态实体对齐方法通过引入多模态信息（如图像、文本等）作为实体的辅助信息，从模态融合和对比训练两个方面增强实体对齐任务效果。

1. **MMEA (Multi-modal Entity Alignment)**:
    - **核心思路**: 利用多模态信息（如图像、数值属性）丰富实体表征，增强对齐效果。
    - 实现方案:
        - **多模态知识嵌入**: 编码关系、图像和数值信息，利用TransE模型学习关系嵌入，利用VGG-16进行图像编码，利用CNN进行数值属性编码。
        - **多模态知识融合**: 在统一空间下融合实体的多模态表征，通过欧氏距离度量不同图谱中的实体是否为等价实体。
2. **EVA (Entity Visual Alignment)**:
    - **核心思路**: 利用注意力机制直接获得模态融合表征，通过迭代学习和无监督种子发现缓解数据稀疏性。
    - 实现方案:
        - **模态信息融合表征**: 利用GCN进行结构编码，ResNet进行视觉编码，词袋模型进行关系和属性编码，通过模态融合表征实现实体对齐。
        - **种子数据稀疏缓解策略**: 利用NCA损失计算嵌入相似度，通过迭代训练扩展训练集，实现无监督种子实体发现。
3. **MCLEA (Multi-modal Contrastive Learning based Entity Alignment)**:
    - **核心思路**: 通过对比学习，挖掘模态内和模态间的特征相关性，增强等价实体的相似性特征。
    - 实现方案:
        - **模态内对比损失**: 对每个实体的模态表征进行对比训练，利用对称损失增强相似性。
        - **模态间对齐损失**: 将模态混合表征的对比相似性分布蒸馏给单模态对比相似性分布，增强单模态表征的整体感知。
4. **MSNEA (Multi-modal Siamese Network for Entity Alignment)**:
    - **核心思路**: 优化视觉、关系、属性模态的嵌入表征，通过对比学习辨别模态重要性。
    - 实现方案:
        - **视觉知识嵌入**: 利用预训练的ResNet进行图像编码。
        - **关系知识嵌入**: 利用TransE假设建立三元组关联，建模实体之间关系，优化实体表征。
        - **属性知识嵌入**: 利用BERT编码属性名，并将图像表征作为注意力query，分配权重加权聚合属性编码。
        - **对比学习损失**: 对单模态和混合模态表征分别施加对比损失，增强细粒度模态对齐约束。

多模态实体链指

实体链指(Entity Linking, EL) 任务旨在非结构化文本中识别出实体指称（mention），并将其链接到知识图谱中的实体（entity）。EL在本质上是非结构化数据和结构化数据之间的桥梁。子任务:

- **指称检测(Mention Detection, MD)**: 从文本中识别出所有的实体指称。
- **实体消歧(Entity Disambiguation, ED)**: 将识别出的实体指称链接到知识图谱中的具体实体。

常规实体链指模型主要包含以下四个部分：

1. **候选实体生成 (Candidate Generation)**:

- **目标**: 从知识图谱中召回与指称相关的实体子集。
- 方法:
  - **字面形式匹配**: 使用Levenshtein距离、n-gram等方法进行字面形式的匹配。
  - **基于别名扩展**: 构建别名/同义词字典，根据字典在候选实体生成阶段进行匹配。

2. **指称上下文编码器 (Mention-Context Encoder)**:
   - **目标**: 对给定指称及其上下文进行编码。
   - 方法:
     - 使用神经网络结构（如CNN、RNN、注意力机制）对指称及其上下文进行编码，得到指称的编码表示。
3. **实体编码器 (Entity Encoder)**:
   - **目标**: 对候选实体进行编码，捕捉实体之间的语义相关性。
   - 方法:
     - **基于共现统计的文本嵌入方法**: 通过统计实体在文本中的共现情况进行嵌入。
     - **基于知识表征学习的方法**: 使用知识图谱中的结构信息进行实体嵌入。
     - **基于深度学习的方法**: 利用神经网络对实体进行上下文嵌入。
4. **实体排序 (Entity Ranking)**:
   - **目标**: 根据候选实体与指称上下文信息进行比对，对候选实体进行排序。
   - 方法:
     - 计算每个候选实体的匹配分数，如点积、余弦相似度等，来度量候选实体与指称的匹配度。
     - 使用排序损失函数进行训练，优化实体排序结果。

NIL问题

NIL问题是指所需要的实体不包含在知识图谱中或者候选实体列表中。主要原因有：

1. **实体缺失**: 知识图谱中不存在所有所需的实体信息。
2. **错误指称**: 在指称检测阶段错误地将无意义的词语识别为指称。
3. **候选实体召回缺失**: 在候选实体生成阶段未能召回所需的实体。

常规解决方法：

1. **添加哨兵实体**: 在候选实体列表中加入一个特殊的“NIL”实体，模型可以在实体排序阶段预测出指称是否与“NIL”实体对应。
2. **实体排序后处理**: 在实体排序阶段后，通过一个二分类模型判断实体排序给出的指称-实体对是否匹配。

核心工作

1. 深度零样本多模态实体链指模型 (DZMNED)
   - **研究动机**: 实体存在多语义性，利用图像信息可以帮助识别实体语义，并补充短文本中的上下文信息稀疏问题。
   - **基本思路**: 在指称和实体的多模态嵌入之间进行匹配，通过注意力机制赋予不同模态不同的权重，实现多模态信息融合。
   - **具体实现**:

        - 模态编码:
            - 图像: 使用Inception预训练模型进行编码。
            - 单词: 使用Bi-LSTM进行编码。
            - 字符: 使用Bi-Char-LSTM进行编码，用于近似编辑距离。
        - 模态融合: 利用模态注意力机制，为每个模态分配不同的权重，并将模态嵌入进行融合。
        - 语义匹配: 利用实体的词法嵌入和知识库嵌入，与指称的多模态嵌入进行相似性计算。

2. 层级门控模态融合及对比模型 (GHMFC)
    - **研究动机**:
        - 实体指称可能存在多种含义（如“黑豹”可以指“动物”或“漫威角色”），利用图像可以更好地辨别指称含义。
        - 文本和图像之间存在多种联系，需要进行细粒度的特征关联。
    - **基本思路**: 利用预训练模型编码单模态信息，通过跨模态相互注意力机制进行多模态特征融合，并通过对比学习增强指称和实体之间的相似性联系。
    - **具体实现**:
        - 特征提取: 使用BERT和ResNet提取文本和图像的特征，利用Conv1D卷积进一步精炼词项表征。
        - 多模态相互注意力: 使用Cross-Modal Transformer (CMT)模块，建立文本指导的图像表征和图像指导的文本表征，进而融合多模态特征。
        - 对比训练:
            - 构造文本-图像和图像-文本的对比损失，使相同实体指称的图像表征和文本表征接近，统一模态表征空间。
            - 构造指称-实体的匹配损失，使指称表征和对应实体表征接近，为给定指称进行候选实体排序。

3. 多粒度模态交互网络模型 (MIMIC)
    - **研究动机**: 隐式图像特征能够弥补短文本的信息不足，需要从不同粒度特征中进行过滤和关联。
    - **基本思路**: 通过多粒度的跨模态交互网络，增强图像和文本的语义关联，挖掘实体和指称的匹配特征。
    - **具体实现**:
        - 输入信息:
            - 文本信息: 指称名及上下文文本，实体的名字与平铺开的三元组构成的文本序列。
            - 图像信息: 图像patches。
        - 编码器: 使用CLIP-ViT进行编码。
        - 匹配器: 包含TGLU、VDLU、CMFU交互单元，建立多粒度的模态特征交互：
            - **TGLU**: 从全局-全局、全局-局部角度，建立实体和指称之间的文本多粒度相似性。
            - **VDLU**: 建立局部-全局-目标实体/指称三者之间的关联，评估图像多粒度相似性。
            - **CMFU**: 建立局部-全局的跨模态特征融合，评估实体和指称之间的相似性。

4. 动态关系交互网络模型 (DRIN)
    - **研究动机**:
        - 以往工作大多利用注意力、门控等机制自动学习跨模态特征关联，缺少对文本和图像模态中目标的显式关联结构。

        - 利用先验知识构建特征关联图，并根据具体情况动态调整特征关联权重。
    - **基本思路**:
        - 在指称和所有候选实体之间显式构建文本-文本、图像-文本、图像-图像关联图，并利用图神经网络增强关联交互，实现链指匹配。
    - **具体实现**:
        - 关联图构建:
            - 节点设置及特征初始化: 在指称-候选实体之间构建关联图，将指称和实体的图像和文本分别作为节点，利用BERT和ResNet作为节点的初始特征编码。
            - 连边设置及连边权重: 利用预训练模型评估特征相似性，利用目标检测模型识别图像物体并评估细粒度图像相似度。
        - 关联交互: 利用图卷积神经网络根据连边权重精炼节点表征，动态更新边权重，适配具体任务。
        - 链指匹配: 计算指称和实体节点表征的语义相似性，得到链指匹配分数，利用正负例构建损失进行训练。