跨模态推理

知识推理是一种为某个本体赋予相关属性的过程，常见的推理方法包括归纳推理和演绎推理。归纳推理的目标是从具体实例中总结出一般规则，而演绎推理则是从已知规则出发，推导出具体结论。使用这些推理方法的算法，如FOIL和AMIE，能够帮助我们在知识库中进行自动推理，从而获取新的知识。

在表示学习的背景下，推理涉及如何提取特征并引入各种知识以进行复杂的推理任务。这可以理解为神经网络前向传播过程的一种直观解释：在前向传播的某个环节，模型将来自不同特征提取器的特征进行融合。例如，基于交叉注意力的跨模态对齐模块，利用文本模态的信息对视觉特征进行重加权操作。

跨模态推理不同于单一模态的前向传播过程，它需要在模型架构层面组合不同模态的特征提取器，通过融合信息来进行复杂的推理任务。例如，融合来自ResNet的图像特征和BERT的文本特征，可以引导模型关注ResNet特征图中与文本相关的部分。通过这种跨模态推理模块，视觉问答（VQA）等领域的一系列具体问题得到了有效解决。

特征融合

特征融合通过将不同模态的信息合并，从而在统一的表示空间中进行推理。可以想象在一家餐馆点餐：你不仅依赖视觉（看菜单上的图片），还结合语言（读菜单上的描述）来做出选择。你可能会同时考虑图片中的菜肴外观和文字描述的口味，从而决定点什么。这就像特征融合，视觉和语言信息在你的大脑中结合，帮助你做出决定。

在跨模态推理中，需要综合处理视觉、语言等多种模态的输入，将不同模态的表征融合为一个反映总体信息的表征。从设计的角度来看，特征融合可以被理解为一个多元到一元的运算。常用的无参数特征融合方法包括特征拼接、按位相加、按位点乘等

- 双线性池化：为了增强不同模态间的交互，双线性池化（Bilinear Pooling）方法被应用于跨模态任务中。双线性池化通过对来自同一样本的特征进行细粒度融合，提高了跨模态表征的效果。具体步骤包括特征矩阵的生成、展开、归一化和融合后的特征计算。
- 压缩双线性融合：虽然双线性池化实现了细粒度交互，但其生成的表征维度较高。压缩双线性池化（Compact Bilinear Pooling, CBP）通过随机投影、张量压缩等数学降维的方式减少特征矩阵的维度，同时保持了捕捉特征间复杂交互的能力。
- 多模态压缩双线性融合：在视觉问答任务中，研究者们应用双线性池化方法进行多模态特征融合，即多模态压缩双线性融合（Multimodal Compact Bilinear Pooling, MCBP）。MCBP通过细粒度融合不同模态的特征，提高了模型对复杂跨模态问题的推理能力。

主要论文如下：

- **Bilinear CNN Models for Fine-Grained Visual Recognition (ICCV 2015)**: 提出了双线性池化方法，通过增强不同模态间的交互，提高了图像识别的精度。在视觉问答任务中，双线性池化可以增强图像特征和文本特征的交互，使模型更好地理解复杂问题。
- **Compact Bilinear Pooling (CVPR 2016)**: 提出了压缩双线性池化方法，解决了双线性池化维度过高的问题。通过引入随机投影和张量压缩技术，大幅降低了特征维度，同时保持了特征融合的效果。这显著提高了计算效率和存储效率，适用于大规模数据和实时应用场景。
- **Multimodal Compact Bilinear Pooling (EMNLP 2016)**: 将压缩双线性池化方法应用于多模态任务，实现了视觉问答中的图像和文本特征融合。在回答“图片中有多少个人？”时，模型能够有效结合图像中的人物特征和问题中的语言描述，提供准确的答案。
- **Multimodal Factorized Bilinear Pooling with Co-Attention Learning (ICCV 2017)**: 引入了联合注意力机制，进一步提升多模态特征融合的效果。联合注意力机制使模型在特征融合过程中动态调整不同模态的权重，从而提高推理的准确性和鲁棒性。

- **MUTAN: Multimodal Tucker Fusion for Visual Question Answering (ICCV 2017)**: 提出了MUTAN模型，使用Tucker分解进行多模态融合，增强了模型的表现力。通过Tucker分解对高维特征进行压缩和融合，有效提升了模型在视觉问答任务中的性能。
- **BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection (AAAI 2019)**: 提出了BLOCK模型，进行双线性超对角融合。该方法通过超对角张量分解技术，进一步提升了特征融合的效率和效果，尤其在视觉关系检测任务中表现突出。

记忆网络

当你学习一门新语言时，你会通过不断的练习和记忆来掌握新单词和语法规则。在遇到新的句子时，你会调用之前学过的知识来理解和翻译这些句子。传统的深度学习模型如RNN、LSTM和GRU使用隐层状态作为记忆，但其记忆能力有限，无法精确记录复杂信息。为了增强模型的记忆和推理能力，Jason Weston等人提出了记忆网络（Memory Network），通过增加记忆模块来实现对复杂信息的长期记忆和推理，通过动态更新记忆状态，使模型能够更好地处理复杂的推理任务。记忆网络包括记忆模块和推理模块：

- **记忆模块**：负责存储输入的记忆信息。
- **推理模块**：根据输入问题从记忆中选择相关信息进行推理。

基于多模态知识的记忆网络需要从视觉、语义和事实等多方面收集信息，并基于这些知识进行推理。例如，基于目标检测的信息、图像描述生成的事件信息和知识库检索的事实信息进行推理。

主要论文包括：

- **Memory Networks (ICLR 2015)**: 提出了记忆网络，解决了神经网络缺乏长期记忆能力的问题。记忆网络通过引入记忆模块，可以在需要时调取相关信息进行推理，从而显著提高模型的推理能力。
- **End-To-End Memory Networks (NIPS 2015)**: 提出了端到端记忆网络，实现了各模块的联合优化，并通过多层堆叠实现多步推理。端到端记忆网络通过将输入文本编码成向量并存储在记忆中，然后在推理过程中动态检索相关记忆，生成最终的答案。
- **Cross-modal Knowledge Reasoning for Knowledge-Based Visual Question Answering (PR 2020)**: 引入了多模态知识推理，将视觉、语义、事实等多方面信息进行整合，实现复杂的跨模态知识推理。例如，基于图像和文本的联合表示，模型能够更好地回答涉及多个模态的信息查询。

注意力机制

在一个吵闹的房间里与朋友交谈时，你会自动忽略周围的噪音，专注于朋友的声音。这类似于注意力机制，它帮助你过滤掉不相关的信息，只关注重要的部分。注意力机制通过对输入信息进行加权选择，强调相关信息，抑制无关信息，从而提高推理的准确性和效率。在跨模态推理中的应用包括自注意力、协同注意力等。

- **基于自注意力机制的推理**：将不同模态的信息统一处理，输入到Transformer结构中实现推理。例如，LXMERT模型通过跨模态编码器进行信息交互，充分挖掘模态间的信息，消除数据异构性带来的挑战。

  协同注意力机制：为了实现更好的视觉语言推理，协同注意力机制（Co-Attention）通过计算图像和文本特征之间的相似性，使两者联系起来。协同注意力机制有并行协同注意力和交替协同注意力两种形式，分别产生图像和文本的注意力分布。

主要论文包括：

- **Show, Attend and Tell: Neural Image Caption Generation with Visual Attention (ICML 2015)**: 提出了基于视觉注意力的图像描述生成方法，通过注意力机制选择相关的图像区域生成描述。该方法显著提升了图像描述生成的准确性和流畅性。

- **Stacked Attention Networks for Image Question Answering (CVPR 2016)**: 在上述工作的基础上，提出了堆叠注意力网络，实现了多步推理，进一步提升了图像问答任务的性能。堆叠注意力网络通过逐步聚焦相关图像区域和文本片段，提供了更精确的答案。
- **Hierarchical Question-Image Co-Attention for Visual Question Answering (CVPR 2016)**: 引入了协同注意力机制，进行多层次的双向跨模态交互，提高了视觉问答的准确性。该方法通过同时对图像和文本进行注意力分配，实现了更高效的多模态信息融合。
- **MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering (EMNLP 2020)**: 结合了BERT编码，提出了多模态融合Transformer，实现了图像和文本的联合表示与推理。该模型通过结合Transformer和BERT的优点，显著提升了视觉问答任务的表现。
- **LXMERT: Learning Cross-Modality Encoder Representations from Transformers (EMNLP 2019)**: 将视觉信息直接输入到Transformer中进行跨模态推理，充分挖掘模态之间的信息交互。LXMERT模型通过多模态编码器实现了图像和文本信息的深度融合，提升了跨模态推理的性能和鲁棒性。

视觉问答

跨模态推理能够解决多种复杂的实际问题，特别是在视觉问答（Visual Question Answering, VQA）领域中。VQA任务要求模型根据输入图像和相关内容回答自然语言问题，是面向用户的人机交互式视觉系统的重要途径。

VQA技术演进

早期的VQA发展主要集中于视觉特征提取能力的不断提升，缺乏对于这个问题本质困难的探索。

论文列表：

1. **VQA: Visual Question Answering, ICCV 2015**：这篇论文提出了视觉问答的基础模型，通过联合图像和问题的特征进行答案生成。该方法主要依赖于深度学习的图像和文本特征提取。模型输入图像和自然语言问题，经过特征提取后进行融合，通过分类器生成答案。这一方法开创了视觉问答领域的研究，为后续工作奠定了基础。
2. **Visual Dialog, CVPR 2017**：该研究将视觉问答扩展到多轮对话形式，需要模型不仅理解当前问题，还需要结合对话历史上下文进行推理。通过引入对话历史的上下文信息，增强了模型在多轮对话中的连贯性和上下文理解能力。这种方法使得视觉对话系统能够进行更复杂的多轮交互，提高了实际应用的可行性。
3. **Where to Look: Focus Regions for Visual Question Answering, CVPR 2016**：引入了注意力机制，通过对问题进行注意力加权，找出与问题相关的图像区域，增强了视觉和语言之间的交互。通过这种方法，模型能够更精确地定位图像中的关键区域，显著提高了答案的准确性和相关性。这种机制被广泛应用于后续的VQA研究中，成为提升模型性能的重要手段。
4. **Stacked Attention Networks for Image Question Answering, CVPR 2016**：这篇论文提出了多步注意力机制，通过堆叠多个注意力层来逐步推理出答案。该方法解决了一些复杂问题需要多步推理的挑战，模型通过逐层关注图像的不同区域，逐步聚焦到最相关的部分，从而更准确地回答问题。这一创新为处理复杂场景提供了新的思路。
5. **Hierarchical Question-Image Co-Attention for Visual Question Answering, CVPR 2016**：引入了联合注意力机制，实现了图像和问题之间的多层次双向交互。该方法通过层次化注意力机制来增强多模态之间的信息传递，使得模型能够更全面地理解图像和问题之间的复杂关系，进一步提高了回答的准确性和多样性。

6. **MuRel: Multimodal Relational Reasoning for Visual Question Answering, CVPR 2019**：提出了基于视觉关系和场景图的关系推理模型，通过视觉关系建模和多模态融合进行推理。该模型通过构建视觉关系图，捕捉图像中物体之间的关系，提高了模型对复杂场景中物体关系的理解和推理能力。这种方法在处理涉及多个物体关系的问题时表现尤为出色。
7. **LXMERT: Learning Cross-Modality Encoder Representations from Transformers, EMNLP 2019**：基于预训练的方法，在大规模视觉数据集和文本数据集上进行预训练，然后迁移到下游任务上。该模型利用跨模态编码器实现视觉和语言之间的高效表示和推理，通过预训练学习到的知识，显著提升了模型在各种VQA任务中的性能。

多变场景的稳定性

在视觉问答任务中，当问题涉及不同的场景或变化的环境（如光照变化、背景复杂度等）时，模型需要在不同条件下仍能提供准确的答案。为此，可以采用记忆网络的多步推理和动态更新机制，或引入循环一致性方法，使模型在复杂和多变的场景中保持稳定性。

- **End-To-End Memory Networks, NIPS 2015**：这项研究提出了端到端记忆网络，通过将记忆网络与循环神经网络（RNN）相结合，实现了在复杂场景中的多步推理和动态记忆更新。该方法增强了模型在处理动态和复杂场景时的稳定性，确保模型在面对多变环境时仍能提供准确的回答。例如，在回答涉及多变场景的问题时，模型可以调用不同步的记忆状态，结合多次推理结果，提供稳定且准确的答案。
- **Cycle-Consistency for Robust VQA, CVPR 2019**：这篇论文引入循环一致性（cycle-consistency）的方法，使模型在面对复杂和多变的场景时，能够保持一致的推理能力。该方法通过在训练过程中加入循环一致性约束，使模型在处理不同光照条件下的图像时，能够保持对图像内容的稳定理解，确保回答的一致性和准确性。

知识引入与常识推理

在回答"这张图片中显示的节日是什么？"等问题时，模型不仅需要理解图像和文本信息，还需要引入外部知识（如节日相关知识）进行推理。通过外部知识和常识推理能力，模型可以利用知识图谱等外部资源进行更深层次的推理。

- **OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge, CVPR 2019**：该研究提出了OK-VQA数据集，这是一个需要外部知识的视觉问答基准。数据集强调了外部知识在推理中的重要性，例如在回答涉及特定节日的问题时，模型可以利用外部知识库中的相关信息，提供准确答案。OK-VQA挑战了传统VQA方法，使其不仅依赖于图像和文本的内部信息，还需要调用外部知识库进行推理。
- **From Recognition to Cognition: Visual Commonsense Reasoning, CVPR 2019**：提出了视觉常识推理（Visual Commonsense Reasoning, VCR）任务。该任务推动了VQA从简单的图像识别任务向更复杂的认知推理任务发展。例如，模型在处理需要常识推理的问题时，可以调用预训练的常识知识库，提供更加准确和合理的回答。VCR数据集包含了丰富的上下文和常识推理任务，旨在评估模型在复杂情境下的推理能力。
- **FVQA: Fact-based Visual Question Answering, TPAMI 2018**：这篇论文提出了基于知识库检索的视觉问答方法，通过问题检索知识库获得补充信息完成问题回答。模型通过联合表征、问题类型预测及知识编码进行答案预测，从而实现基于事实的视觉问答。
- **Mucko: Multi-layer Cross-modal Knowledge Reasoning for Fact-based Visual Question Answering, IJCAI 2020**：提出了多模态图谱构建和跨模态异构图推理的方法。通过构建视觉图、语义图和事实图，实现多层次跨模态知识推理，增强了模型对复杂问题的回答能力。

- **Unicoder-vl: A Universal Encoder for Vision and Language by Cross-modal Pre-training, AAAI 2020**：基于预训练隐式知识的方法，通过大规模数据上的预训练学习不同模态之间的语义对应关系，并利用这种隐式关系实现跨模态推理。该方法在视觉和语言任务中表现出了显著的性能提升。

语义对齐与多模态融合

为了准确回答跨模态问题，模型需要对视觉和语言信息进行有效的语义对齐和多模态融合。基于注意力机制和联合表示的技术能够显著提升模型的对齐和融合能力。

- **Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, CVPR 2017**：该研究强调了在视觉问答任务中提升图像理解的重要性。通过强化图像特征的表示，改进了VQA的性能。引入联合注意力机制，使模型能够更好地对齐图像和文本信息，从而提供更准确的回答。
- **Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering, CVPR 2018**：提出了克服VQA任务中先验偏见的方法。通过精确的视觉和语言对齐，模型能够更好地处理图像和问题之间的复杂关系，提高回答的准确性。该研究引入了新的训练方法，使模型能够在没有强先验假设的情况下进行更可靠的推理。

对话理解与多轮推理

VQA任务中常涉及多轮对话理解和推理，这要求模型能够处理连续的问题和回答。基于对话的视觉问答模型可以实现更自然和连贯的多轮推理。

- **Visual Dialog (CVPR 2017)**：提出了视觉对话（Visual Dialog）任务。这一任务要求模型在多轮对话中理解和回答问题，提升了模型的对话理解和推理能力。Visual Dialog数据集包含了大量的多轮对话，模型需要理解对话上下文，并在连续的对话轮次中提供准确的回答。
- **CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog (NAACL 2019)**：提供了一个多轮推理的数据集，通过连续的对话轮次，评估模型在多轮推理中的表现。CLEVR-Dialog数据集设计用于评估模型的多轮推理能力，通过复杂的对话场景，测试模型在多轮对话中的表现和推理能力。