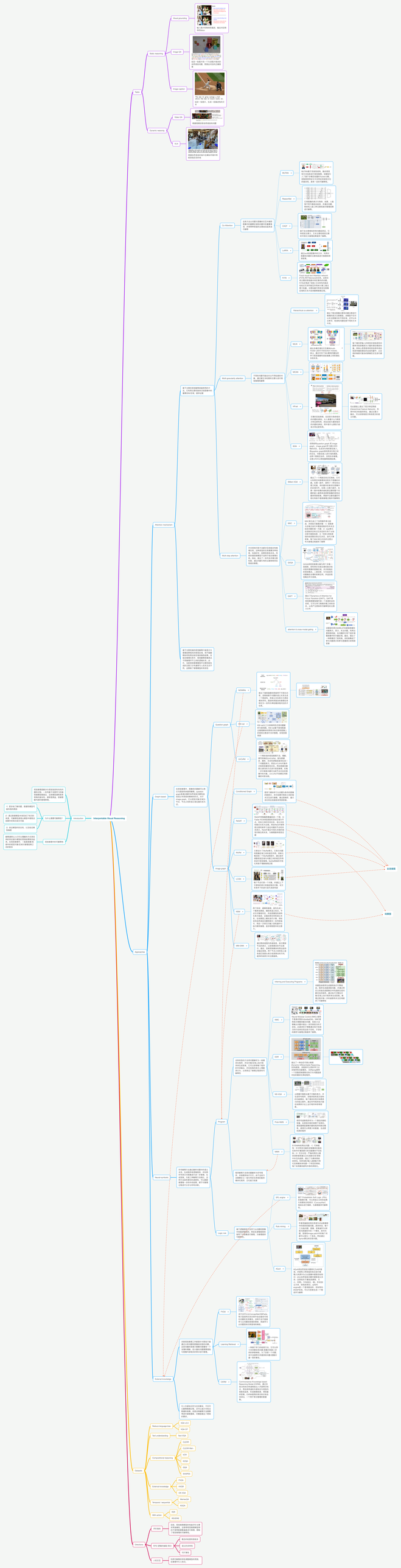
Interpretable Visual Reasoning
Introduction
为什么要研究解释性？
1. 更多地了解问题，根据所得到的知识进行改进
2. 通过理解模型为某些决策下了合理的依据，可解释性是模型中最容易被接受的特性之一
3. 解决模型的安全性，公正性以及隐私问题
最新的研究人工智能的目的不仅是为了提高性能，也是为了理解智能的本质。这包括理解、证明、信任和解释智能决策的能力。
现在常用的可解释性
可视化
Approaches
Basic reasoning
Visual grounding
输入图片和物体的描述，输出对应的BBoxes
Image QA
给定一张图片和一个问题的文本输入，输出一个答案的概率，转换成分类任务解决
Image caption
Video QA
根据视频的内容来回答问题
VLN
根据自然语言指令在场景的环境中导航到指定的目的地
Semantic reasoning
Attention mechanism
基于注意力的视觉推理是在当前任务的条件下，通过学习一个注意力分布来加权处理视觉信息的表示和推理表示，使模型能够聚焦于与当前任务最相关的关键信息，从而提高推理的准确性和可解释性。注意力机制通过为输入的不同部分分配不同的权重，使模型能够动态地关注最重要的信息，这种机制模仿了人类在处理复杂信息时的选择性注意力。
Co-Attention
这类方法以问题引导图片区域的视觉注意力，并利用图片信息来指导问题的理解。
MUTAN
MUTAN基于张量分解的多模态融合，通过将高维交互张量分解为三个低秩张量的乘积，在保持表达能力的同时大幅降低参数量。
Tensorflow
官方实现基于Tensorflow，支持GPU加速，提供了预训练模型和数据集下载脚本，适合复现论文结果。
CAQT
基于注意力机制的问题感知，引导模型关注与问题相关的图像区域，提升视觉问答任务的推理能力。
GitHub
基于Caffe实现的多模态注意力网络，通过交叉注意力机制实现文本和图像特征的对齐与融合。
LaRRA
通过Latent Relation Attention机制，在潜空间中学习实体间的关系表示，提升视觉关系推理的性能。
Pytorch
基于PyTorch实现的Relation-aware Graph Attention Network，通过图注意力网络建模实体间的关系，在VQA任务上取得了较好的性能。
KVTA
Fast and Accurate Multi-modal Tasks with Transformer Adaptation，提出一种轻量级的多模态Transformer适配方法，通过知识蒸馏和参数共享，在保持精度的同时大幅提升推理速度。
Multi-granularity attention
为解决单模态注意力不同粒度的问题，通过在不同层次上建立注意力机制，实现多粒度信息的协同推理。
Hierarchical co-attention
通过三维注意力机制在不同层次上进行多模态信息交互推理，对象级、短语级和词级三个层次的注意力分别建模，并通过层级结构实现跨层信息传递。
MLIN
提出多模态图注意力网络（Multi-modal Latent Interaction Network），通过学习潜在交互矩阵来建模不同模态间的高阶交互关系。
MCAN
提出多模态图注意力网络（Multi-modal Latent Interaction Network），通过学习潜在交互矩阵来建模不同模态间的高阶交互关系。
HPAM
提出一种层次化多模态注意力机制，通过在词级、短语级和句级三个层次上分别计算注意力权重，并通过层级结构实现信息的逐层传递与融合。
Hierarchical Attention Mechanism for Multi-modal Tasks
BAN
提出双线性注意力网络（Bilinear Attention Network），通过学习一个双线性注意力矩阵来建模视觉特征和文本特征之间的交互关系，在保持参数效率的同时提升模型表达能力。
Multi-step attention
为多步推理场景设计的注意力机制，通过在多个推理步骤中反复计算注意力权重，逐步聚焦到与当前推理步骤相关的信息上。
Bilinear VQA
提出了一种双线性注意力机制，通过学习一个双线性交互矩阵来建模视觉特征和文本特征之间的交互关系，在保持参数效率的同时提升模型表达能力。
MAC
MAC网络通过三个可微分模块（control、read、write）在内部记忆单元上执行多步推理，每个时间步骤从问题中提取一个指令（如find、relate、compare），read模块从知识库检索信息，write模块将检索结果与当前记忆融合，实现多跳推理。在CLEVR数据集上达到98.9%的准确率，超过人类表现。
DCGA
DCGA采用双重注意力机制，一个注意力模块从问题中提取多个推理步骤，另一个注意力模块在每个推理步骤上从视觉特征中检索相关信息，实现动态的多跳推理。在CLEVR数据集上达到98.8%的准确率，接近人类表现。
SAT
提出了Dynamic of Attention for Focal Transition（DAFT），SAT将注意力机制建模为一个马尔可夫决策过程，从而学习最优的注意力转移策略。
attention & cross-modal gating
结合注意力机制和门控机制的多模态融合方法，通过注意力机制学习不同模态间的交互关系，通过门控机制动态调整各模态信息的融合权重，实现更精细的多模态信息融合。
基于注意力机制的视觉推理方法，通过学习注意力分布来动态地关注输入信息中与当前任务最相关的部分，从而提高推理的准确性和可解释性。这类方法模仿了人类在处理复杂信息时的选择性注意力机制，能够实现多模态信息的有效融合与对齐。
Graph based
在知识推理中，图数据结构能够以非欧几里得的方式表示实体间的复杂关系，这种结构特别适合处理具有丰富关系信息的问题。通过图卷积网络（GCN）、图注意力网络（GAT）等方法，节点之间的信息可以沿着关系边传播，从而实现多跳推理和关系推理。
Question graph
NSVRN
提出了基于图卷积网络的视觉问答方法，通过将问题和场景分别建模为图结构，并通过图卷积网络进行多跳推理。该方法在多个视觉问答数据集上取得了较好的性能。
DP-net
Mn-net(CL)是基于图注意力机制的图推理模型，通过学习节点间的注意力权重来动态调整信息传播的重要性，从而实现多跳推理。
DYCVRD
DYCVRD提出动态推理图（Dynamic Reasoning Graph），通过学习问题条件下的节点重要性权重，实现自适应的多跳推理。在CLEVR数据集上达到98.7%的准确率，接近人类表现。
Conditional Graph
基于条件图卷积网络的视觉推理方法，通过学习条件图卷积核来实现问题条件下的视觉推理。该方法在多个视觉问答数据集上取得了较好的性能。
Image graph
RgNet
Figure 3: The RN-VQA RgNet Graph Reasoning Module. Each Object Proposal is first embedded into a vector. Then, the RgNet Graph Reasoning Module uses a Graph Convolutional Network to propagate information between the object proposals. The Relation Prediction Module predicts the relationships between the object proposals.
RgNet（关系图网络）是一种用于视觉推理的图神经网络模型，通过学习实体间的关系表示来进行多跳推理。该模型在多个视觉问答数据集上取得了较好的性能。
MuRel
注意力模块（Attention）：计算出一个注意图，用来从图像特征中选择与问题相关的区域特征。
LCGN
图卷积网络（GCN）：对第一步生成的这个带有注意权重的图，进行信息传播与更新。
MAC
MAC（记忆、注意和组合）网络是一种用于视觉推理的神经网络，通过在一个循环框架内交替进行记忆读取、注意力计算和信息组合来实现多步推理。该模型在CLEVR数据集上达到98.9%的准确率，超过人类表现。
MAC-GNN
通过图神经网络学习实体间的关系表示，并利用这些关系表示进行多跳推理。该方法在多个视觉问答数据集上取得了较好的性能。
Neural-symbolic
符号推理方法通过解析问题中的语义结构，将自然语言问题转换为可执行的逻辑表达式，再在知识库上执行推理。这类方法结合了神经网络的语义理解能力和符号推理的可解释性和精确性，能够处理复杂的多步推理任务。
Program
这种方法将问题分解为一系列可执行的子任务，并为每个子任务训练一个专门的神经网络模块。这种方法结合了神经网络的语义理解能力和符号推理的可解释性和精确性，能够处理复杂的多步推理任务。
Learning and Executing Programs
Neural Module Control (NMC) 解决了传统神经模块网络（NMN）需要专家提供布局的问题，为每个问题生成一个特定的计算图，实现端到端的视觉推理。NMC在CLEVR数据集上达到98.9%的准确率，接近人类表现。
NMC
Neural Module Control (NMC) 解决了传统神经模块网络（NMN）需要专家提供布局的问题，为每个问题生成一个特定的计算图，实现端到端的视觉推理。NMC在CLEVR数据集上达到98.9%的准确率，接近人类表现。
DDR
提出了一种结合符号推理与神经网络的方法，通过学习可微分的逻辑规则来进行多步推理，在保持可解释性的同时提升推理性能。
NB-FOL
NB-FOL（Neural Belief Fuzzy Logic）是一种结合神经网络与模糊逻辑的推理方法，通过学习模糊规则来进行多步推理，在保持可解释性的同时提升推理性能。
Prob-NMN
Prob-NMN（Probabilistic Neural Module Network）是一种结合神经网络与概率图模型的推理方法，通过学习概率图模型来进行多步推理，在保持可解释性的同时提升推理性能。
SMNN
NeSyMoNN（NeSy Modular Neural Network）是一种结合神经网络与符号推理的方法，通过学习模块化的神经网络来进行多步推理，在保持可解释性的同时提升推理性能。NMN和NB-FOL分别从不同角度探索了神经网络与符号推理的结合，为可解释性研究提供了多种思路。
Logic rule
基于逻辑规则进行多步推理的方法，通过学习一组可微分的逻辑规则来进行多跳推理，在保持可解释性的同时提升推理性能。
SPL engine
基于Probabilistic Soft Logic（PSL）的推理引擎，可以将自然语言推理问题建模为概率软逻辑（Probabilistic Soft Logic）问题，并通过概率软逻辑推理引擎求解。
Rule mining
Rule mining（规则挖掘）是从知识图谱中自动发现逻辑规则的技术，通过学习实体间的关系模式，发现诸如"如果A是B的父亲，且B是C的父亲，则A是C的祖父"这样的规则。这些规则可以用于知识图谱的补全和推理。
NSAR
NSAR（Neural Symbolic Action Recognition）是一种结合神经网络与符号推理的方法，通过学习一组可微分的逻辑规则来进行多步推理，在保持可解释性的同时提升推理性能。
Knowledge
在视觉推理任务中，外部知识库通常以图结构的形式存储实体间的关系信息，这种结构能够支持多跳推理和关系推理。知识图谱（Knowledge Graph）是一种以节点表示实体、边表示关系的有向图结构，能够高效地存储和检索复杂的关系信息。
FVQA
基于图卷积网络（GCN）的视觉问答方法，通过将问题和场景分别建模为图结构，并通过图卷积网络进行多跳推理。该方法在多个视觉问答数据集上取得了较好的性能。
Learning Retrieval
一种基于学习的检索方法，它不以单一步骤从知识库中检索信息，而是通过多轮检索和推理，逐步收集与问题相关的信息，最终回答问题。
CMRN
Commonsense Knowledge Aware Multi Relation Network（CMRN）是一种结合外部知识图谱的多关系图卷积网络，通过学习实体间的关系权重来进行多跳推理。该模型在多个视觉问答数据集上取得了较好的性能。
OLA-KB
OLA-KB（Object Level Attention Knowledge Base）是一种结合外部知识图谱的视觉问答方法，通过学习实体间的关系权重来进行多跳推理。该模型在多个视觉问答数据集上取得了较好的性能。
Datasets
Natural language task
VQA v2.0
VQA-CP
Text understanding
Text-VQA
OCR-VQA
Compositional reasoning
CLEVR
CLEVR-Man
VCR
JHOCR
SQA
SceneNet
External knowledge
FVQA
VKQR
OK-VQA
Temporal / sequential
MarioQA
MSQA
RPS action
SQR
REVERIE
Directions
四个国家
美国、英国、德国和加拿大的研究小组在所有领域，这些研究报告解释智能决策的能力，以及解释智能决策的公平性、隐私和安全性的重要性。
可视化理解模型（黑盒）
模型内部表示探究
因果关系研究
可扩展性
人类因素
为解决解释性问题，需要跨学科的合作，包括人工智能、认知科学、人机交互和伦理学等领域的专家。