软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(7):2936−2952 [doi: 10.13328/j.cnki.jos.007655] [CSTR: 32375.14.jos.007655]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向图分类任务的互补感知证据提取方法*

岳立楠[1,2], 张敏灵[1,2]

[1](东南大学 计算机科学与工程学院, 江苏 南京 211189)

[2](计算机网络和信息集成教育部重点实验室 (东南大学), 江苏 南京 211189)

通信作者: 张敏灵, E-mail: zhangml@seu.edu.cn

摘　要: 图神经网络 (graph neural network, GNN) 在图分类任务中表现出色, 但其“黑箱”性质引发了对预测过程可解释性的广泛关注. 作为一种自解释机制, 证据提取方法近年来受到广泛研究, 其目标是在生成预测结果的同时, 从原始图中提取紧凑、具备判别性的子图结构 (即证据子图) 以作为解释. 然而, 现有方法往往依赖数据中的“捷径”特征进行预测, 导致提取的解释缺乏忠实性, 进而影响模型的可解释性与鲁棒性. 为缓解上述问题, 提出一种互补感知证据提取 (complement-aware rationale extraction, CaR) 方法, 将未被选为证据的子图区域视为互补信息, 并从以下 3 个方面提升反事实建模与解释能力: 首先, 引入对比学习机制以解耦证据表示与互补表示, 增强二者的语义独立性; 其次, 提出回声学习 (echo-learning) 策略, 充分利用 GNN 各层消息传递过程中的中间表示, 捕捉不同深度层次下互补结构的差异性; 最后, 将当前与历史层的互补表示与证据表示结合, 进而构造反事实样本, 提升训练数据的多样性. 在多个真实数据集及一个合成数据集上的实验证明, CaR 能够生成更具忠实性的证据, 验证了其有效性.

关键词: 证据提取; 图神经网络; 捷径学习; 可解释人工智能; 数据挖掘

中图法分类号: TP18

中文引用格式: 岳立楠, 张敏灵. 面向图分类任务的互补感知证据提取方法. 软件学报, 2026, 37(7): 2936–2952. http://www.jos.org.cn/1000-9825/7655.htm

英文引用格式: Yue LN, Zhang ML. Complement-aware Rationale Extraction Method for Graph Classification Tasks. Ruan Jian Xue Bao/Journal of Software, 2026, 37(7): 2936–2952 (in Chinese). http://www.jos.org.cn/1000-9825/7655.htm

Complement-aware Rationale Extraction Method for Graph Classification Tasks

YUE Li-Nan[1,2], ZHANG Min-Ling[1,2]

[1](School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

[2](Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 211189, China)

Abstract: Graph neural networks (GNNs) have achieved remarkable performance on graph classification tasks, but their black-box nature has raised widespread concerns about the explainability of their prediction process. As a self-explaining mechanism, rationale extraction has received increasing attention in recent years. Its goal is to extract concise subgraph structures from the original graph (i.e., rationale subgraphs) as explanations while generating prediction results. However, existing methods often rely on spurious shortcut features in the data, resulting in explanations that lack faithfulness, which in turn compromises both the interpretability and robustness of the model. To address these issues, this study proposes a complement-aware rationale extraction (CaR) method, which treats the subgraph regions not selected as rationales as complement information. The method enhances counterfactual modeling and interpretability from the following three perspectives. First, a contrastive learning mechanism is introduced to disentangle rationale representations from complement representations, enhancing their semantic independence. Second, an echo-learning strategy is proposed to fully leverage the intermediate representations generated during the message-passing process of GNNs, capturing the structural differences in complement parts across different network depths. Finally, the method combines complement and rationale representations from both current and historical layers to

* 基金项目: 国家自然科学基金 (62506072, 62225602)
收稿时间: 2025-09-19; 修改时间: 2026-01-18, 2026-02-02; 采用时间: 2026-02-24; jos 在线出版时间: 2026-04-29
CNKI 网络首发时间: 2026-04-30

construct counterfactual samples, increasing the diversity of the training data. Extensive experiments on multiple real-world benchmark datasets and a synthetic dataset demonstrate the effectiveness of CaR in producing faithful rationales.

Key words: rationale extraction; graph neural network (GNN); shortcut learning; explainable AI; data mining

图神经网络 (graph neural network, GNN) 在多种图分类任务中展现出优异性能[1−4]. 然而, 其预测结果普遍缺乏可解释性, 已成为制约其在实际应用中广泛部署的重要瓶颈. 为提升图神经网络模型的可解释性, 研究者提出了多种方法, 主要包括事后可解释方法[5,6]与自解释方法[7−9]. 其中, 图证据提取 (graph rationale extraction) 方法作为自解释方法的重要分支, 近年来受到广泛关注[10−12]. 该类方法首先从输入图中抽取一张对模型决策具有关键作用的子图 (即证据子图), 然后基于该子图进行任务预测. 该子图通常由若干关键节点或边组成, 可作为模型决策依据的解释性证据, 有助于理解其推理过程.

尽管图证据提取方法在提升模型可解释性方面展现出良好潜力, 但其易受到“捷径学习 (shortcut learning)”问题的干扰[13,14]. 所谓捷径, 是指图中某些结构模式与目标标签之间存在表面显著但非因果的统计关联, 导致模型在预测与解释过程中倾向于依赖这些非因果特征. 已有研究表明, 图证据提取方法在构建证据子图的过程中, 容易捕捉这类捷径关联进行任务预测[10,15]. 因此, 即便模型在预测准确率上表现优异, 其生成的解释结果亦可能缺乏忠实性. 更为严重的是, 在处理分布外 (out-of-distribution, OOD) 样本时, 图结构与标签之间的统计关系可能发生变化, 从而使原先依赖的捷径失效, 显著削弱模型在新场景下的预测与解释能力[16−19].

以图 1 所示的 motif 类型预测任务为例, 模型需根据由 motif 与 base 构成的图结构判别其中的 motif 类型. 在该任务中, House 与 Cycle 分别作为 motif 标签, Tree 与 Wheel 作为 base 子图并不影响任务预测. 在训练集中, House-Wheel 与 Cycle-Tree 的组合频率显著不均, 形成明显的样本不平衡现象. 这一偏差可能导致图证据提取方法在训练过程中过度依赖 motif 与 base 之间的统计共现关系 (即捷径), 而非真正基于 motif 类型本身的判别语义特征进行推理. 因此, 在分布内测试集中, 模型通常能够准确地将 House-Wheel 样本分类为 House, 因其已学习到该组合的频繁共现模式 (捷径). 然而, 在面对分布外样本 (例如 Cycle-Wheel) 时, 模型仍倾向于错误地预测为 House, 主要原因在于训练时学到的解决关系在分布外发生变化, 导致模型错误地将 Wheel 区域提取为证据子图, 从而削弱了解释结果的可信度与有效性.

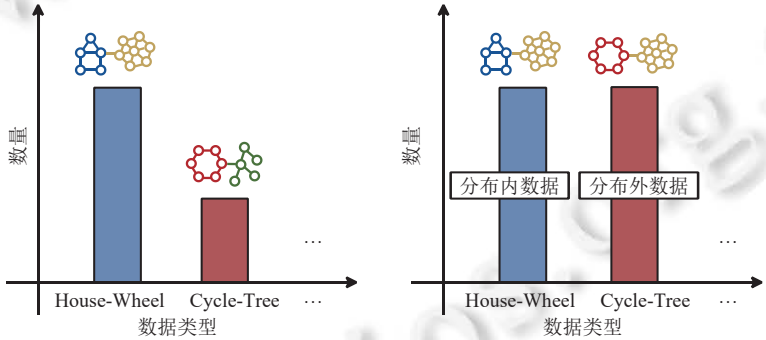


图 1 motif 类型预测任务示意图

为缓解图证据提取方法中的捷径学习问题, 已有研究提出多种改进策略[20,21], 其中基于反事实数据增强的图证据提取方法[11,22,23]通过削弱非因果特征与标签之间的统计相关性, 在抑制捷径学习方面表现出较好的效果. 该类方法通常将输入图划分为证据子图与互补子图 (即非证据部分), 并根据划分粒度的不同, 分为基于节点划分[23]与基于边划分[11]的反事实增强方法. 本文聚焦于前者, 即基于节点划分的反事实数据增强方法. 具体而言, 该类方法通过生成二值掩码向量来预测每个节点是否为证据节点; 随后, 利用图神经网络对图结构进行编码, 获得节点表示. 通过将掩码与节点表示逐元素相乘, 可分别提取证据子图与互补子图的表示. 在此基础上, 引入反事实样本构造机制, 其核心假设是: 在证据部分确定的前提下, 互补部分与标签应当条件独立. 基于该假设, 方法通常在同一批

次内对互补子图施加扰动 (如随机打乱或重组), 并将其与保持不变的证据子图重新组合, 生成标签不变的反事实样本. 该策略通过打破证据与互补之间的统计共现, 显著削弱了模型对非因果特征的依赖, 从而有效缓解捷径学习问题. 例如, 在图 1 所示的示例中, 若将 House 与 Wheel 的组合关系打破, 并用 Tree 替代 Wheel 与 House 组合, 即可阻断模型对 House-Wheel 共现关系的依赖, 促进其关注真正决定 motif 类型的结构性特征.

然而, 该类方法的有效性高度依赖于互补部分与标签之间的独立性假设, 而该假设在实际图神经网络建模过程中往往难以严格成立. 一方面, 互补表示并非仅由原始互补节点构成, 而是由掩码向量与节点表示共同决定; 另一方面, 节点表示是通过多层图神经网络的消息传递机制获得的. 在这一过程中, 互补节点不可避免地接收到来自证据节点的信息, 且随着传播层数的增加, 这种信息渗透效应将持续累积, 从而使互补表示逐渐混入与证据子图相关的语义信息. 由此, 互补表示与标签之间可能产生潜在相关性, 导致构造的反事实样本偏离理想的反事实分布, 进而削弱反事实增强策略的干扰效果与整体鲁棒性.

为缓解上述问题, 本文提出一种面向图分类任务的互补感知证据提取 (complement-aware rationale extraction, CaR) 方法. 该方法的核心目标在于增强互补表示与标签之间的独立性, 从而提升反事实样本构造的质量. 具体而言, 本文引入对比学习机制 [24], 以促使证据表示与互补表示在特征空间中实现解耦, 从而提升反事实样本的构造质量. 此外, 考虑到图神经网络中消息传递层数的增加将加剧证据信息对互补节点的干扰, 本文进一步提出一种回声学习 (echo-learning) 策略. 即在固定层数的图网络结构下 (例如 5 层), 较早层 (如第 3 层) 的互补表示受到的证据干扰相对较少, 因而相较当前层 (如第 5 层) 更能保持与标签的独立性. 基于该观察, 不同于现有方法仅依赖最终层互补表示的限制, 本文设计出一种融合历史层互补表示的反事实生成机制, 以进一步增强模型抵御捷径干扰的能力.

本文的主要贡献如下.

(1) 从图神经网络多层消息传递的角度, 分析现有基于反事实增强的图证据提取方法中互补表示与标签之间可能产生相关性的原因, 指出随着传播层数增加, 互补节点表征中逐渐混入与证据子图相关的语义信息, 从而削弱“互补子图与标签条件独立”假设的成立程度, 进而影响对捷径学习的抑制效果.

(2) 提出一种互补感知的图证据提取方法 CaR, 通过引入对比学习约束证据表示与互补表示在特征空间中的解耦, 增强互补表示与标签之间的独立性, 提高反事实样本构造的有效性与鲁棒性. 同时提出一种回声学习策略, 利用不同层次互补表示中证据信息渗透程度的差异, 融合历史层互补特征进行反事实生成, 进一步削弱捷径共现关系对模型预测与解释的干扰.

(3) 在多个真实图分类基准数据集 [1,25]及合成数据集 [10]上的实验证明, 所提出的 CaR 方法在生成忠实解释方面具有显著优势, 有效缓解了由捷径引发的解释不可靠问题.

1 相关工作

图神经网络近年来在图分类、节点分类、链路预测等任务中取得了显著进展 [26−31], 其强大的结构建模能力为复杂图数据的表示与处理提供了全新路径. 然而, 图神经网络的“黑箱性”使其预测过程难以理解, 严重制约了模型在安全敏感与高可信度要求场景中的部署与应用. 为应对这一问题, 研究者提出了多种图可解释性技术, 旨在提升模型的透明度与决策过程的可理解性 [32−36]. 现有图可解释方法主要分为两类: 事后可解释方法与自解释方法. 事后可解释方法 [37]通常在模型训练完成后, 通过干预或分析已训练模型的内部机制 (如注意力权重提取、梯度反向传播、输入扰动等), 识别模型在预测中关注的结构区域. 这类方法适配性强, 可广泛应用于多种图神经网络架构, 适用于通用的模型解释需求. 然而, 其解释过程通常计算复杂, 尤其在处理大规模图数据时, 面临效率低下与鲁棒性差的问题. 此外, 由于解释生成过程与模型预测解耦, 往往难以保证解释结果的忠实性与一致性.

因此, 近年来越来越多研究关注自解释方法. 自解释方法在模型结构中引入显式的解释模块或结构性约束, 使模型在进行预测的同时生成解释, 通常具备更强的内在一致性与可追溯性. 例如, 李平等人 [38]提出一种层次化自解释图表示学习方法, 能够在图分类过程中输出基于图结构的层次化解释, 增强了模型推理逻辑的透明性与可信度. 此类方法的优势在于可通过结构设计精确约束解释生成机制, 从而实现更具语义性的结构区域定位与分析.

在自解释方法中, 图证据提取方法因其解释结果直观、结构清晰, 受到广泛关注. 该类方法通常通过学习结构

掩码, 识别模型预测所依赖的关键子图结构, 即“证据子图”, 并将其作为模型决策的依据呈现给用户. 图证据提取方法具有良好的可读性与可视化能力, 有助于用户理解、验证并调整模型的预测逻辑, 已逐渐成为图可解释性研究的核心方向之一. 然而, 已有研究指出, 当前图证据提取方法易受到捷径学习的影响[15]. 在训练过程中, 图神经网络模型可能并未捕捉到与标签具有因果关系的语义结构, 而是依赖数据中某些与标签高度共现但缺乏语义解释的伪特征或统计捷径进行预测. 尽管这些捷径在训练集上可能有效, 但在分布外样本或真实场景中常失效, 导致模型生成错误或误导性的解释结果, 严重削弱其可信性.

为缓解捷径问题, 近年来已有多项研究[20,39−42]尝试通过结构设计增强解释的忠实性与鲁棒性. 例如, DARE[40]提出自引导训练框架, 基于互信息策略实现输入解耦, 引导模型捕捉与标签高度相关且信息充分的证据信号; GSAT[20]引入信息瓶颈理论, 构建“最小化输入与解释之间冗余信息、最大化解释与标签之间关键信息”的优化目标, 并通过可学习的低随机性注意机制提升解释稳定性; DIVE[41]则关注解释的多样性, 引入子图多样性正则项, 鼓励模型识别多个具有互补性的判别子图, 降低对单一捷径路径的依赖. 与此同时, 另一类研究路径尝试从环境扰动的角度引入反事实数据增强 (counterfactual data augmentation), 以提升模型的泛化能力与解释结果的鲁棒性. 该类方法通常将输入图刻画为“证据+环境”双结构形式, 通过在训练过程中引入多样化的环境样本, 强化模型对与标签稳定相关的核心证据信息的识别能力. 例如, DIR[10]提出“不变证据发现”原则, 将互补子图视为环境干扰源, 并与固定的证据子图重新组合生成反事实样本, 引导模型学习在不同环境下保持标签不变的结构模式. Faith-DARE[43]在DARE 基础上, 利用解耦后的互补表示构造反事实样本, 以削弱模型对捷径路径的依赖, 从而进一步提升反事实干预的有效性. DisC[22]与 RGDA[23]同样基于互补结构构造反事实图, 以削弱模型对捷径路径的依赖. 不同于 Faith-DARE 等在图级层面进行反事实构造, FIG[44]在节点级粒度上生成反事实样本, 实现了更精细的环境扰动建模, 从而提高了反事实样本对捷径抑制与解释鲁棒性的贡献. 此外, 还有方法通过引入显式或隐式的环境表示辅助反事实构造. 例如, GIL[33]与 C2R[12]在小批量样本中聚类互补表示, 进而学习局部环境表示, 并与证据子图组合构造反事实样本; HSE[45]与 EQuAD[46]则进一步引入环境标签生成机制, 通过环境推断为每个样本显式赋予环境类别, 提升模型的泛化能力.

尽管上述方法在增强模型解释的忠实性与稳定性方面取得了积极进展, 但仍面临关键挑战: 部分基于互补表示的反事实构造方法中, 互补子图本身可能蕴含较强的标签相关信息, 导致模型在缺失证据子图的情况下仍保持较高预测精度. 这一现象违背了“互补结构不携带标签信息”的基本假设, 削弱了反事实样本干扰模型捷径行为的能力, 进而影响增强策略的有效性与泛化能力.

2 预备知识

2.1 问题定义

令数据集 $\mathcal{D}$ 中每个样本为图-标签对 $(G,Y)\in\mathcal{D}$, 其中图 $G=(\mathcal{V},\mathcal{E})$ 由节点集合 $\mathcal{V}$ 和边集合 $\mathcal{E}$ 组成, 节点数和边数分别为 $|\mathcal{V}|$ 和 $|\mathcal{E}|$. 本文旨在通过证据提取器 $f_s(G)$ 结合节点特征表示 $H_G\in\mathbb{R}^{|\mathcal{V}|\times d}$ 学习节点掩码 $M\in\mathbb{R}^{|\mathcal{V}|}$, 以获得任务相关的证据子图表示. 具体而言, 通过对节点表示与掩码进行逐元素乘积得到证据子图表示 $M\odot H_G$, 并基于该表示训练任务预测器 $f_p(\cdot)$ 完成任务预测. 模型的训练目标可形式化为最小化如下期望损失:

$$f_s^*,f_p^*=\arg\min\mathbb{E}_{(G,Y)\sim\mathcal{D}}\left[\ell\left(f_p\left(f_s(G)\right),Y\right)\right] \tag{1}$$

其中, $\ell(\cdot)$ 表示交叉熵损失函数.

2.2 基础图证据提取方法

本节介绍基础的图证据提取框架, 其核心模块包括证据提取器和任务预测器. 目标是在保持模型准确性的同时, 从输入图中提取出对预测决策最关键的子结构 (证据子图). 具体而言, 证据提取器首先采用图神经网络编码器 $GNN_m(\cdot)$ 将图中各节点编码为 d 维特征向量, 随后通过线性变换权重矩阵 $W_m\in\mathbb{R}^{2\times d}$ 计算每个节点作为证据的选择概率, 具体表示为:

$$\widetilde{M}=Softmax(W_m(GNN_m(G))) \tag{2}$$

为实现可微的二值掩码采样过程, 本文采用 Gumbel-Softmax[47]技术, 在概率分布 $\widetilde{M}$ 上加入 Gumbel 噪声并通过温度参数 τ 控制采样的平滑度, 进而采样掩码变量 M. 之后, 利用另一图神经网络编码器 $GNN_G(\cdot)$ 对原始图 G 进行编码, 获得节点特征矩阵 H_G. 证据子图的节点表示通过掩码与节点特征逐元素相乘得到, 即 $M\odot H_G$, 对应的互补 (非证据) 节点表示为 $(1-M)\odot H_G$.

任务预测器模块由读出函数 ($READOUT(\cdot)$) 和分类器组成. 读出函数对证据子图及其互补节点表示分别进行图级聚合, 生成对应的图表示向量 h_r 和 h_e:

$$h_r = READOUT(M\odot H_G) \tag{3}$$

$$h_e = READOUT((1-M)\odot H_G) \tag{4}$$

随后, 分类器 $\phi(\cdot)$ 基于证据子图表示 h_r 进行标签预测: $\hat{Y}_r=\phi(h_r)$, 对应的监督损失定义为:

$$\mathcal{L}_r = \mathbb{E}_{(G,Y)\sim\mathcal{D}}\left[\ell\left(\hat{Y}_r, Y\right)\right] \tag{5}$$

为鼓励生成的证据子图尽可能简洁, 引入稀疏性正则项, 约束掩码 M 的平均值接近预设稀疏率 α:

$$\mathcal{L}_{\text{sp}} = \left|\frac{1}{|M|}\sum_{i=1}^{|M|} m_i - \alpha\right| \tag{6}$$

综上, 基础图证据提取模型的整体优化目标为:

$$\mathcal{L}_{\text{rat}} = \mathcal{L}_r + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} \tag{7}$$

其中, λ_{sp} 为稀疏性约束的权重系数.

在推断阶段, 模型仅依据证据子图表示进行预测.

2.3 基础图证据提取方法

为提升模型的鲁棒性和泛化能力, 在基础图证据提取模型的框架上, 本文引入一种基于反事实数据增强的方法 CDA-GR (counterfactual data augmentation based graph rationalization). 该方法基于以下假设: 在给定证据子图的条件下, 互补子图与标签相互独立. 基于此, 通过跨样本组合证据表示与互补表示, 构造反事实样本, 提升模型对于捷径学习现象的抵抗能力. 具体而言, 考虑一个包含 B 个样本的训练批次 $\{(G^i,Y^i)\}_{i=1}^{B}$ 及其对应的证据表示和互补表示 $\{(h_r^i,h_e^i)\}_{i=1}^{B}$, 该方法从上述批次中随机选取另一个样本 $j\neq i$ 的互补表示 h_e^j, 并将其与对应证据表示 h_r^i 相加, 形成反事实样本表示:

$$h_{\widetilde{G}}^{(i,j)} = h_r^i + h_e^j \tag{8}$$

构造出的反事实样本对应的标签保持不变, 即 $\widetilde{Y}^{(i,j)}=Y^i$. 将该反事实表示输入分类器生成预测:

$$\hat{Y}^{(i,j)} = \phi\left(h_{\widetilde{G}}^{(i,j)}\right) \tag{9}$$

并计算反事实样本的监督损失:

$$\mathcal{L}_e = \mathbb{E}_{(G^i,Y^i)\sim\mathcal{D}}\left[\ell\left(\hat{Y}^{(i,j)}, \widetilde{Y}^{(i,j)}\right)\right] \tag{10}$$

最终, 模型训练通过结合基础图证据提取损失和反事实增强损失进行联合优化:

$$\mathcal{L}_{\text{com}} = \mathcal{L}_{\text{rat}} + \lambda_e\mathcal{L}_e \tag{11}$$

其中, λ_e 控制反事实增强损失的权重. 推断时, 无需构造反事实样本, 仍仅利用证据子图表示进行预测.

3 互补感知的图证据提取方法

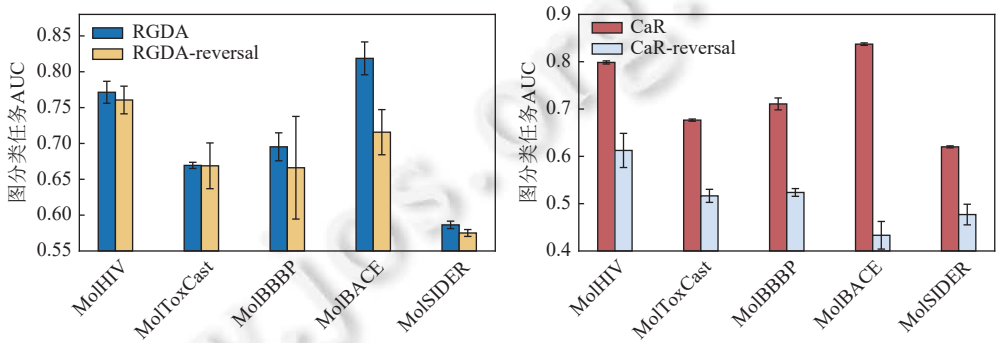
3.1 重新审视 CDA-GR 方法

如第 2.3 节所述, CDA-GR 方法的有效性建立在一个关键假设之上: 在给定证据子图的条件下, 互补子图与标签之间是条件独立的. 基于该假设, CDA-GR 通过将不同样本的互补表示与证据表示进行组合, 构造多样化的反事实样本, 从而打破原始数据中的虚假关联, 缓解模型对“捷径”特征的依赖, 提升泛化性能.

然而, 在实际应用中, 这一独立性假设往往难以严格满足, 进而可能影响反事实增强策略的效果. 具体而言, 回

顾公式 (4) 可知, 互补表示源自节点表示 H_G, 而 H_G 通常通过图神经网络在多层信息传播和聚合的过程中获得. 在这一过程中, 互补节点不可避免地接收到来自证据节点的信息, 使其表示中逐渐混入与证据子图相关的语义特征. 这种由消息传递机制引入的信息渗透, 从结构上破坏了“互补子图与标签条件独立”的建模假设, 从而可能限制 CDA-GR 方法在实际图分类任务中的增强效果.

为验证上述分析, 本文设计了一系列实证实验, 系统评估 CDA-GR 方法的实际表现. 具体而言, 本文选取了 5 个 OGB 图分类数据集[1], 并报告了 CDA-GR 的代表性实现方法 RGDA[23]的预测结果, 结果如图 2(a) 所示. 为了进一步检验互补表示中是否携带标签相关的信息, 本文引入对照实验方法 RGDA-reversal, 将原本用于预测的证据表示 h_r 替换为互补表示 h_e, 用于分类任务. 实验结果表明, 在大多数数据集上, RGDA-reversal 的性能与 RGDA 相当, 甚至在部分数据集上表现略优. 这一观察进一步印证了互补表示中确实包含与标签相关的证据信息, 从而从实证层面印证了“互补子图与标签条件独立”的假设在实际应用中难以严格满足. 这一发现揭示了当前 CDA-GR 方法在建模假设层面的潜在局限性.



(a) RGDA与RGDA-reversal的性能表现　　(b) CaR与CaR-reversal的性能表现

图 2　RGDA、RGDA-reversal、CaR 与 CaR-reversal 在 OGB 数据集上的性能表现

3.2　CaR 方法的模型架构

基于第 3.1 节对 CDA-GR 方法的重新审视, 如图 3 所示, 本文提出一种新的基于证据提取的图神经网络方法 CaR, 旨在更好地满足“互补子图与标签条件独立”的关键假设. CaR 方法主要包括两个核心模块: (1) 基于对比学习的解耦机制, 实现证据表示与互补表示的有效分离; (2) 回声学习 (echo-learning) 策略, 从图神经网络的早期传播层中提取互补相关信息, 用以提升反事实表示的稳定性.

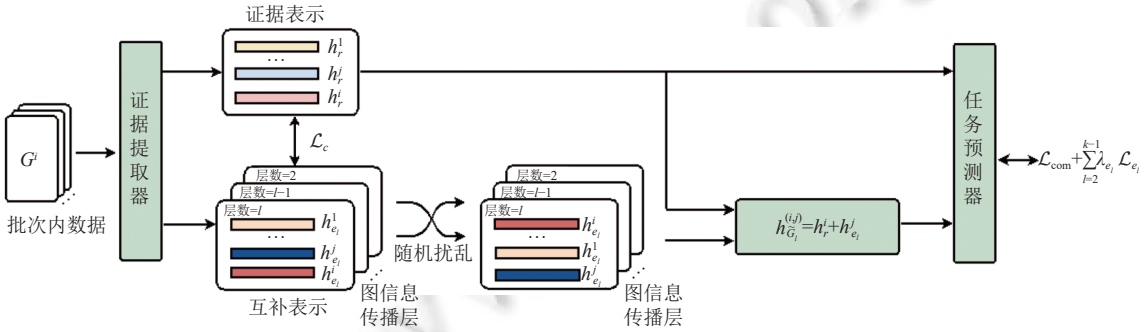


图 3　CaR 方法的模型架构

1) 基于对比学习的证据-互补解耦机制

为了实现证据表示 h_r 与互补表示 h_e 之间的显式解耦, 本文引入对比学习约束机制, 旨在在保持证据语义不变的前提下, 使证据表示与任意互补表示在嵌入空间中相互远离, 从而压缩二者的共享信息成分, 避免互补表示中混

入可预测标签的证据信息. 具体而言, 本文设计如下对比损失函数 $\mathcal{L}_c$:

$$\mathcal{L}_c = -\log \frac{\exp\left(\frac{h_r^{\top}\widetilde{h}_r}{\tau}\right)}{\exp\left(\frac{h_r^{\top}\widetilde{h}_r}{\tau}\right)+\sum_{h_e\in\mathcal{T}}\exp\left(\frac{h_r^{\top}h_e}{\tau}\right)} \tag{12}$$

其中, $\widetilde{h}_r$ 为 h_r 的正样本, $\mathcal{T}$ 表示当前 mini-batch 中所有样本的互补表示 $\left\{\left\{h_{e_l}^i\right\}_{i=1}^{B}\right\}_{l=2}^{k-1}$, k 表示图神经网络的总传播层数, τ 为温度系数, 用于调节对比学习中相似度分布的敏感度.

受公式 (8) 启发, 考虑到互补表示在理想情况下不应影响模型预测结果, 因此本文将正样本 $\widetilde{h}_r$ 构造为证据表示 h_r 与随机采样的互补表示 $\hat{h}_e \in \mathcal{T}$ 的加和, 即:

$$\widetilde{h}_r = h_r + \hat{h}_e \tag{13}$$

构造后的样本由于保留了证据信息, 且不受互补信息的影响, 因此可作为证据表示 h_r 的正样本. 在负样本方面, mini-batch 中所有互补表示 h_e 均被视为与 h_r 语义无关的对立样本. 通过最小化上述对比损失 $\mathcal{L}_c$, 模型能够有效压缩证据与互补表示之间的共性信息, 并增强它们在嵌入空间中的区分性, 从而为后续反事实增强提供更加可靠的表示基础.

2) 回声学习策略

现有基于反事实组合的方法通常仅利用最终一层图表示来构造互补表示. 然而, 在图神经网络深层消息传递过程中, 节点表示会不断聚合邻域信息, 使得本应与标签条件独立的互补分量中混入大量证据信息, 从而破坏“证据-互补可加且互不干扰”的建模假设. 为缓解该问题, 本文提出回声学习 (echo-learning) 策略, 从图神经网络的中间传播层中回溯并提取互补表示, 用于构造反事实样本. 具体而言, 该策略基于第 3.1 节的实证分析, 并引入如下关键假设.

假设 1: 给定图神经网络消息传递的层数 l 与 k ($l < k$), 本文假设第 l 层生成的互补表示相较于第 k 层包含更少的证据信息.

在此假设基础上, 定义回声互补表示 (echo complement representation) 如下.

定义 1. 回声互补表示. 设图神经网络的总传播层数为 k, 则第 l 层 ($l < k$) 的互补表示 h_{e_l} 被定义为回声互补表示, 计算公式为:

$$h_{e_l} = READOUT\left((1-M)\odot H_{G_l}\right) \tag{14}$$

其中, H_{G_l} 表示图在第 l 层的节点表示, M 为证据掩码, $\odot$ 表示逐元素乘法操作.

基于上述假设与定义, 本文认为, 相较于最终层互补表示, 早期传播层得到的回声互补表示更能保持与标签之间的独立性. 因此, CaR 利用不同层次的回声互补表示参与反事实样本构造, 以在增强表达多样性的同时削弱由过度消息传播引入的虚假关联.

具体而言, 模型首先从第 l 层中采样一个 mini-batch 的回声互补表示 $\left\{h_{e_l}^i\right\}_{i=1}^{B}$, 之后对于每个 mini-batch $\left\{(h_r^i, h_{e_l}^i, Y^i\right\}_{i=1}^{B}$, 在每一层 l 上执行交叉组合以构造反事实样本并进行任务结果预测:

$$h_{\widetilde{G}_l}^{(i,j)} = h_r^i + h_{e_l}^j \tag{15}$$

$$\hat{Y}_l^{(i,j)} = \phi\left(h_{\widetilde{G}_l}^{(i,j)}\right) \tag{16}$$

$$\mathcal{L}_{e_l} = \mathbb{E}_{(G^i,Y^i)\sim\mathcal{D}}\left[\ell\left(\hat{Y}_l^{(i,j)}, \widetilde{Y}^{(i,j)}\right)\right] \tag{17}$$

其中, $\widetilde{Y}^{(i,j)}$ 表示构造样本的伪标签, $\phi(\cdot)$ 表示分类器. 该过程在多个中间层 $l = \{2,3,\ldots,k-1\}$ 上执行, 使模型同时受到来自不同语义层级互补表示的反事实约束, 从而抑制深层表示中由过度消息传播引入的虚假相关.

最终, 综合原始分类损失、证据-互补对比解耦损失以及多层回声反事实损失, CaR 的最终训练目标为:

$$\mathcal{L}_{\text{car}} = \mathcal{L}_{\text{com}} + \lambda_c\mathcal{L}_c + \sum_{l=2}^{k-1}\lambda_{e_l}\mathcal{L}_{e_l} \tag{18}$$

其中, $\mathcal{L}_{\text{com}}$ 为原始分类损失, λ_c 与 λ_{e_l} 为损失权重系数, 用于平衡对比学习与反事实增强的训练目标.

4 实验与评估

为系统评估所提出的 CaR 方法的有效性与适用性, 本文设计了一系列实验, 围绕以下关键研究问题 (research question, RQ) 展开.

- RQ1: CaR 方法在提升任务预测性能及证据提取质量方面表现如何?
- RQ2: CaR 方法中不同模块的设计选择及其超参数设置对模型性能有何影响?
- RQ3: 与现有的 CDA-GR 方法相比, CaR 所学习的互补表示是否在语义上与标签具有更强的条件独立性?
- RQ4: 所提出的回声学习策略是否能够有效增强 CDA-GR 框架下的反事实生成能力, 从而进一步提升模型性能?
- RQ5: 在分布外 (OOD) 场景中, CaR 是否能够准确识别出对预测决策起关键作用的证据信息?

4.1 数据集与基准模型

1) 数据集

为全面评估 CaR 方法的有效性, 本文在合成数据集和真实世界数据集上开展了系统实验, 相关数据集的统计信息如表 1 所示.

表 1 合成数据集与真实世界数据集的统计信息

数据集	训练/验证/测试	类别数	平均节点数	平均边数
Spurious-Motif (b=0.5)	3 000/3 000/6 000	3	29.6	42.0
Spurious-Motif (b=0.7)	3 000/3 000/6 000	3	30.8	45.9
Spurious-Motif (b=0.9)	3 000/3 000/6 000	3	29.4	42.5
MolHIV	32 901/4 113/4 113	2	25.5	27.5
MolToxCast	6 860/858/858	617	18.8	19.3
MolBBBP	1 631/204/204	2	34.1	26.0
MolBACE	1 210/151/152	2	34.1	36.9
MolSIDER	1 141/143/143	27	33.6	35.4

合成数据集: 在本文中, 采用 Spurious-Motif 数据集[5,10]作为合成实验平台, 其任务为 motif 类型预测. 该数据集中, 每个图由两个结构性子图组成: motif 子图 R 与 base 子图 C. 其中, motif 子图是预测目标的主要依据, 可视为模型推理所依赖的“证据”, 其结构类型包括 Cycle、House 和 Crane (分别对应 $R=\{0, 1, 2\}$); base 子图结构则与 motif 类型无关, 视为互补部分, 包括 Tree、Ladder 和 Wheel (即 $C=\{0, 1, 2\}$).

为验证 CaR 在缓解“捷径”问题方面的能力, 本文手动构建了多个具有不同伪相关程度的数据分布. 具体而言, motif 子图按均匀分布采样, 而 base 子图的生成遵循以下分布:

$$P_{\text{data}} = b \cdot \mathbb{I}(C = R) + \frac{1-b}{2} \cdot \mathbb{I}(C \neq R) \tag{19}$$

其中, b 控制 base 与 motif 子图之间的伪相关强度, 当 $b=1$ 时, base 子图与对应的 motif 类型完全共现, 即每种 motif 类型总是对应唯一一种 base 类型, 形成强伪相关关系; 而当 $b=0$ 时, base 子图与 motif 子图的结构独立, 伪相关性为 0. 本文构建了 3 类数据集, 分别对应 $b=\{0.5, 0.7, 0.9\}$, 以分析 CaR 在不同伪相关程度下的鲁棒性. 为评估模型在真实泛化场景下的表现, 本文构建了一个去偏测试集, 即设置 $b=1/3$, 使得 base 与 motif 子图之间不存在统计相关性, 从而模拟分布外场景, 确保测试结果的公平性和无偏性.

OGB 数据集: 为进一步验证 CaR 在真实图分类任务中的适用性, 本文在 Open Graph Benchmark (OGB)[1] 提供的 OGB-Mol 系列数据集上开展实验. 该基准广泛用于图级分子属性预测任务, 本文选用了其中的 5 个子集 MolHIV、MolToxCast、MolBBBP、MolBACE 和 MolSIDER, 涵盖毒性、药理活性、血脑屏障穿透性等多个分子性质. 为保证实验公平性, 本文采用 OGB 默认的 scaffold 分割策略, 该策略在划分训练集与测试集时构造了分布

差异, 符合 OOD 设定, 有助于评估模型的泛化能力. 鉴于本文聚焦于图分类任务, 故未在工业级图数据 (如包含亿级节点的图) 上开展实验, 因为此类数据集主要用于节点分类而非图分类任务.

2) 基准模型

为验证 CaR 方法的性能优势, 本文选取若干具有代表性的图分类方法作为对比基准模型, 这些方法均以证据提取为核心, 以提升模型在分布外场景下的泛化能力. 具体如下.

● DIR[10]提出通过对训练分布进行干预以构造多个干预环境的策略, 从中挖掘任务不变证据, 以提升模型的泛化能力.

● DisC[22]提出一种通用的解耦框架, 用于区分因果子结构与偏差子结构. 该方法通过合成反事实样本强化训练过程, 从而降低因果因素与伪相关因素之间的耦合.

● CAL[11]基于因果注意力机制, 旨在识别对预测结果具有因果贡献的图结构特征, 同时削弱与标签呈伪相关关系的信息, 从而缓解“捷径”问题.

● GSAT[20]引入随机屏蔽机制, 在信息瓶颈理论[48,49]指导下, 选择性地保留与标签高度相关的图结构, 从而过滤掉无关信息.

● DARE[40]是一种自引导的证据提取方法, 基于解耦机制从输入中提取富含信息的子结构. 本文将其由自然语言理解任务迁移至图级分类任务.

● Faith-DARE[43]是 DARE 的扩展版本, 旨在通过缓解捷径依赖来提取更加可靠的证据子图, 其核心思想是将 DARE 识别出的非证据特征视为与预测结果去相关的“环境”因素, 并基于其解耦的互补表示构造反事实样本.

● InterRAT[21]基于因果干预提出一种证据提取方法, 采用后门调整策略[50,51]移除数据中的伪相关性, 挖掘真实的因果子结构. 本文同样将其从自然语言理解任务扩展至图级分类任务.

● RGDA[23]提出一套通用于节点级与图级任务的反事实数据增强框架. 在图级分类场景中, RGDA 利用因果子结构与“捷径”结构合成反事实样本, 提升模型对因果因素的敏感性.

● C2R[12]构建了一个协同分类与证据提取框架, 通过融合多环境信息生成忠实的证据, 提升模型的泛化与可解释能力.

总体来看, DIR、DisC、CAL 与 GSAT 属于基于边划分 (edge-based partitioning) 的方法, 侧重通过边级操作识别关键子图; 而 DARE、InterRAT、RGDA 与 C2R 属于基于节点划分 (node-based partitioning) 的方法, 更强调通过节点级别的分离实现证据与非证据结构的解耦.

4.2 实验设置

在所有实验中, 本文在 CaR 及各对比方法中统一采用图同构网络 (graph isomorphism network, GIN)[52]作为主干图神经网络结构, 并使用均值池化 (mean pooling) 操作作为读出函数. 默认超参数设置如下: 对比损失的权重参数 $\lambda_c = 0.1$, 反事实监督项的权重 $\lambda_e = 1$, 稀疏性控制项的权重 $\lambda_{\text{sp}} = 1$, 各层回声互补表示项的权重 $\lambda_{e_l} = 0.5$. 隐藏层维度 d 在 Spurious-Motif 数据集上设置为 32, 在 OGB 数据集上设置为 128.

在训练过程中, 本文在 Spurious-Motif 和 OGB 上均使用 Adam 优化器[53], 初始学习率设置为 0.01. 针对不同数据集的特性设置稀疏性参数 α, 其中 MolHIV 设置为 0.1, MolSIDER、MolBACE 和 MolToxCast 设置为 0.5, MolBBBP 设置为 0.2, 其余数据集统一设置为 0.4. 在图神经网络的层数设置方面, 对于 OGB 数据集采用 4 层或 5 层消息传递模块, 对于 Spurious-Motif 数据集则统一设为 4 层.

在模型评估阶段, 本文在 Spurious-Motif 数据集上采用准确率 (accuracy, ACC) 作为主要性能指标, 在 OGB 数据集上使用曲线下面积 (area under curve, AUC) 指标衡量预测效果. 此外, 由于 Spurious-Motif 数据集已经提供人工标注的真实证据信息, 本文采用 Precision@5 指标评估模型提取的前 5 条证据与真实证据之间的一致性, 从而全面评估模型的证据提取能力. 为确保实验结果的可靠性与稳定性, 本文在单张 A100 GPU 上对所有模型进行训练, 并在 5 组不同的随机种子下重复实验. 最终报告的是验证集上取得最优性能的模型在测试集上的平均值与标准差.

4.3 整体性能表现 (RQ1)

1) 任务预测性能

为系统评估 CaR 在图分类任务中的预测能力, 本文在多个数据集上开展了广泛实验, 结果如表 2 所示. 其中, 加粗表示最优结果. 实验结果表明, CaR 在所有数据集上均取得领先性能. 例如, 在 MolBBBP 数据集上, CaR 相较于当前最优方法 C2R 提升了 1.09%; 在 Spurious-Motif 数据集 ($b = 0.9$) 上, CaR 更是实现了 6.33% 的显著性能增益. 上述结果充分验证了本文提出的基于对比学习的解耦方法以及引入的回声学习策略在提升模型泛化能力方面的有效性. 进一步地, 与现有方法相比, Faith-DARE 同样通过解耦机制削弱了互补信息与证据信息之间的相关性, 但其整体性能仍不及本文提出的 CaR 方法. 其主要原因在于, Faith-DARE 并未显式建模图神经网络多层消息传递过程中证据信息向互补表示逐层渗透的问题. 相比之下, CaR 通过引入回声学习策略, 融合来自不同传播层次的互补表示, 有效减弱了由深层消息传递引入的干扰, 从而在抑制捷径学习的同时进一步提升了预测性能与解释忠实性. 此外, 从不同类型对比方法的整体表现来看, 基于节点划分的方法 (如 RGDA 和 C2R) 与基于边划分的方法 (如 DisC 和 CAL) 在预测性能上大致相当, 说明两类方法在增强模型可解释性方面均具有应用潜力. 然而, CaR 在所有评估任务中始终保持领先, 进一步验证了增强互补表示与标签之间独立性的核心设计理念在构建鲁棒模型方面的关键作用.

表 2 CaR 及基准模型在合成数据集与真实世界数据集上的性能表现

模型	Spurious-Motif (ACC)			OGB-Mol (AUC)				
	b=0.5	b=0.7	b=0.9	MolHIV	MolToxCast	MolBBBP	MolBACE	MolSIDER
DIR	0.3950±0.0471	0.3872±0.0531	0.3768±0.0447	0.6303±0.0607	0.5451±0.0092	0.6460±0.0139	0.7391±0.0282	0.4989±0.0115
DisC	0.4585±0.0660	0.4885±0.1154	0.3859±0.0400	0.7731±0.0101	0.6662±0.0089	0.6963±0.0206	0.8293±0.0171	0.5846±0.0169
CAL	0.4734±0.0681	0.5541±0.0323	0.4474±0.0128	0.7339±0.0077	0.6476±0.0066	0.6582±0.0397	0.7848±0.0107	0.5965±0.0116
GSAT	0.4517±0.0422	0.5567±0.0458	0.4732±0.0367	0.7524±0.0166	0.6174±0.0069	0.6722±0.0197	0.7922±0.0239	0.6041±0.0096
DARE	0.4843±0.1080	0.4002±0.0404	0.4331±0.0631	0.7836±0.0015	0.6677±0.0058	0.6820±0.0246	0.8239±0.0192	0.5921±0.0260
Faith-DARE	0.5382±0.1101	0.5755±0.0919	0.5257±0.1204	0.7915±0.0021	0.6691±0.0039	0.6974±0.0047	0.8293±0.0088	0.61633±0.0010
InterRAT	0.4628±0.0234	0.5182±0.0214	0.4983±0.0523	0.7446±0.0131	0.6531±0.0045	0.6753±0.0034	0.7958±0.0201	0.5821±0.0031
RGDA	0.4251±0.0458	0.5331±0.1509	0.4568±0.0779	0.7714±0.0153	0.6694±0.0043	0.6953±0.0229	0.8187±0.0195	0.5864±0.0052
C2R	0.5203±0.1437	0.5913±0.0413	0.5601±0.0979	0.7919±0.0006	0.6709±0.0052	0.6999±0.0122	0.8143±0.0096	0.6131±0.0117
CaR	**0.6793±0.0316**	**0.6433±0.0707**	**0.6234±0.0936**	**0.7985±0.0033**	**0.6766±0.0023**	**0.7108±0.0126**	**0.8372±0.0026**	**0.6201±0.0021**
w/o echo	0.4879±0.0258	0.4143±0.0621	0.4347±0.0797	0.7876±0.0028	0.6686±0.0073	0.6901±0.0147	0.8246±0.0205	0.5993±0.0138
w/o dis	0.5923±0.0997	0.6125±0.0644	0.5957±0.1274	0.7932±0.0035	0.6711±0.0048	0.7046±0.0158	0.8304±0.0048	0.6147±0.0039

在评估任务预测性能的同时, 本文进一步分析了引入对比学习与回声学习机制后所带来的计算开销. 具体而言, 本文在 MolHIV 数据集上对比了不同方法的训练时间与 GPU 显存占用情况, 并以 RGDA 作为基准方法, 将其计算成本归一化为 1.00×, 报告其他方法相对于该基准的倍数开销 (例如 1.20×表示训练时间或显存消耗约为 RGDA 的 1.2 倍). 如表 3 所示, 尽管本文方法引入了对比学习约束及多层互补表示的回声融合机制, 但其训练时间与显存开销仅较基线方法略有增加, 整体仍处于同一数量级范围内. 这表明在获得显著性能提升与解释忠实性增强的同时, 本文方法仅引入了较为有限的额外计算成本, 具有较好的实际可部署性.

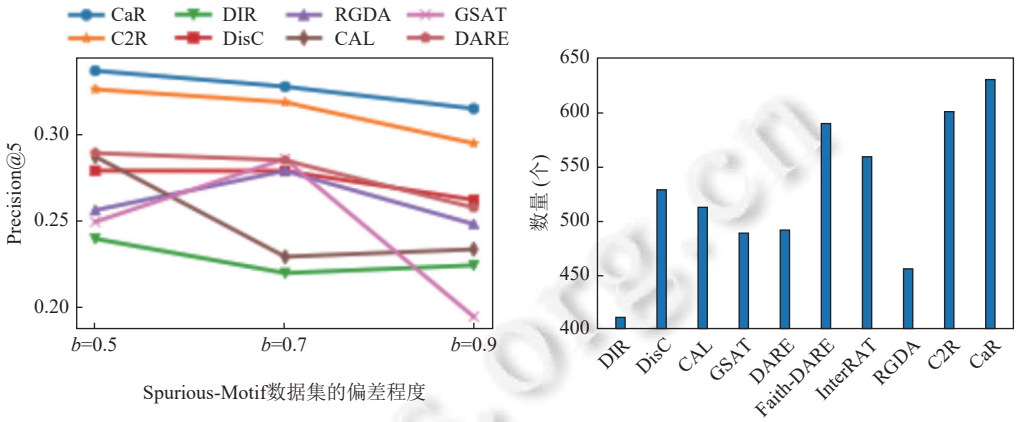
表 3 不同方法在 MolHIV 数据集上的计算开销对比

模型	训练时长	GPU显存占用
RGDA	1.00×	1.00×
DARE	1.05×	1.08×
Faith-DARE	1.54×	1.42×
CaR	1.17×	1.20×

2) 证据提取性能

为验证 CaR 是否能够准确识别对预测具有关键影响的因果证据, 而非依赖伪相关信息 (即“捷径”), 本文在包

含真实证据信息标注的 Spurious-Motif 数据集上进行了深入评估. 图 4(a) 展示了使用 Precision@5 作为指标的实验结果, 用于衡量模型所提取前 5 个证据与真实证据之间的一致性, 从而定量评估证据定位精度. 实验表明, 在不同伪相关强度 (即不同 b 值) 设定下, CaR 始终显著优于所有对比方法, 表现出更高的证据准确性. 这一结果进一步说明, CaR 能有效过滤伪相关结构, 聚焦于与预测真正相关的关键子图, 从而在提高模型可解释性的同时增强其决策可信度.



(a) 提取证据与真实证据之间的Precision@5　　(b) 提取证据的忠实性评估

图 4　提取证据的准确性与忠实性评估

此外, 为进一步评估模型所提取证据的忠实性, 本文基于 Spurious-Motif 数据集, 首先构建一个包含 3 类 motif (Cycle、Crane、House) 的平衡数据集, 每类各 1 000 个样本 (b = 1/3). 随后额外引入 1 000 个带偏样本, 其中所有样本的真实类别均为 Cycle, 但其图结构中嵌入了 Tree 子结构 (即 Cycle-Tree), 从而在人为构造的“Tree 结构-Cycle 标签”之间形成伪相关性. 在该数据集下进行训练时, 依赖捷径的模型往往会错误地将 Tree 子结构作为主要判别依据. 为定量评估不同方法提取证据的忠实性, 本文从第 4.1 节所述测试集中选取全部 Cycle-Tree 样本 (共 679 个), 并在这些样本上运行各模型以提取其预测的证据子图. 借助数据集中提供的真实 motif 标注, 本文进一步判断预测证据与伪相关区域 (Tree) 及真实判别区域 (Cycle) 之间的重叠关系. 具体而言, 当节点的被选中概率超过 0.55 时, 将其视为证据节点; 若超过 60% 的证据节点落在 Tree 子结构而非 Cycle motif 中, 则认为该解释主要依赖捷径, 忠实性不足. 基于上述判定准则, 本文统计并对比了各方法中“非捷径驱动”的 Cycle-Tree 样本数量. 实验结果如图 4(b) 所示, RGDA 及其他基线方法 (如 DIR 和 DARE) 在带偏样本上普遍倾向于将 Tree 基结构作为主要证据, 表现出明显的捷径依赖. 相比之下, 本文提出的 CaR 方法能够更加稳定地识别真正决定标签的 Cycle motif, 其对伪相关性的依赖显著降低. 上述定量结果从忠实性角度进一步验证了本文方法在抑制捷径学习、提升证据子图可靠性方面的优势.

4.4　消融实验与超参数敏感性分析 (RQ2)

1) 消融实验

为进一步验证 CaR 各组成模块的有效性, 本文设计了消融实验, 重点分析以下两个关键模块的贡献.

(i) 首先, 移除回声学习策略 (即公式 (17)), 该变体记作 CaR w/o echo. 实验结果如表 2 所示. 尽管移除了该策略, CaR w/o echo 在多个任务中仍优于现有多种基于解耦的证据提取方法, 如 CAL、DisC 与 DARE, 验证了 CaR 框架中对比学习机制在建模互补表示与证据表示解耦方面的有效性.

(ii) 在此基础上, 进一步移除证据-互补解耦模块 (即公式 (12)), 该变体记作 CaR w/o dis. 结果表明, 该变体在多数任务中仍优于基线方法和 CaR w/o echo, 这一现象表明, 从图的早期信息传播层中构造互补表示, 在保持关键信息完整性的同时, 有助于缓解后续证据提取中的信息冗余问题, 从而提升模型整体性能.

上述消融实验结果表明, CaR 的两个核心设计模块在提升模型预测能力和证据提取准确性方面均发挥关键作用. 尽管每个模块在独立作用时仍具一定效果, 但完整集成后能实现更优的协同增强, 显著提升模型的泛化能力和解释性能.

2) 超参数敏感性分析

本节重点分析图神经网络中两个关键超参数对 CaR 方法性能的影响: 图信息传播层数 l 以及批次大小. 具体而言, 本文首先将传播层数分别设置为 3、4、5 和 6, 并在 MolBACE 与 MolSIDER 两个数据集上对 CaR 与 RGDA 的性能进行对比, 实验结果如图 5(a) 和 (b) 所示. 可以观察到, 随着传播层数的增加, 两种方法的性能均呈现先上升后下降的趋势. 该现象主要源于深层图神经网络中普遍存在的过度平滑 (over-smoothing) 问题[54], 即节点表示在多次消息传递后逐渐趋于一致, 从而削弱模型的判别能力. 尽管如此, 在所有层数设置下, CaR 始终显著优于 RGDA, 表明本文方法在不同网络深度条件下均具有更强的建模能力与泛化性能. 进一步地, 在较高层数 (如 $l=5$) 下, CaR 的性能相比于较浅层数 (如 $l=3$) 显著提升. 本文认为, 该现象主要得益于回声学习策略: 在更深的消息传递过程中, 早期层与中间层能够提供更加多样化且语义层次不同的互补表示, 用于构造更丰富的反事实样本, 从而有效提升模型对结构与语义多样性的刻画能力. 该结果表明, 所提出的方法通过增强反事实样本的多样性, 有助于打破原始数据中潜在的虚假相关, 缓解模型对“捷径”特征的依赖, 从而提升预测结果的泛化性与因果忠实性.

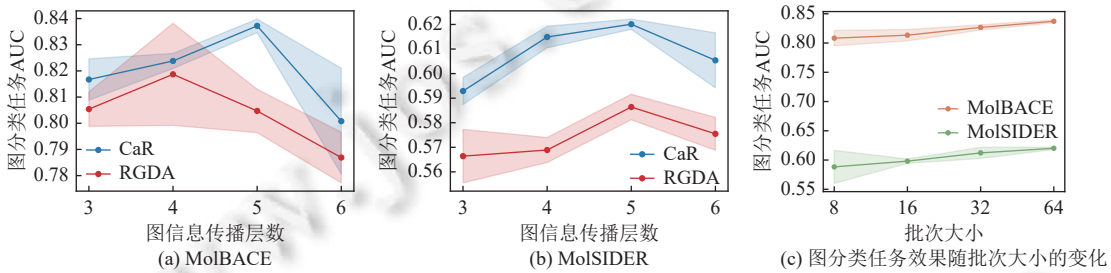


图 5 图信息传播层数和图批次大小的超参数敏感性分析

随后, 本文进一步考察了不同批次大小对模型性能的影响, 并在 MolBACE 和 MolSIDER 数据集上分别将批次大小设置为 8、16、32 和 64. 实验结果如图 5(c) 所示. 随着批次大小的增大, 模型性能整体呈现稳定上升趋势. 这是由于在更大的批次大小中可采样到更多互补表示, 用于构造数量更为丰富、分布更为多样的反事实样本, 从而增强反事实一致性约束的有效性. 该结果进一步验证了 CaR 方法通过提升反事实样本多样性来削弱虚假关联、抑制捷径学习的机制, 并从超参数敏感性角度印证了所提出方法在提升泛化性能与推理忠实性方面的有效性.

4.5 CaR 中互补信息的性能表现 (RQ3)

前文图 2(a) 展示了 CDA-GR 方法在满足“在给定证据条件下, 互补信息与标签独立”这一假设时所面临的挑战. 即便依赖互补信息进行任务预测, 模型仍能取得较好的性能, 说明其互补表示中仍包含与标签高度相关的信息. 该现象反映出, 现有方法在互补信息建模过程中存在证据与互补信息混淆的问题, 未能实现二者之间的充分解耦, 从而违背了原始的建模假设. 针对这一问题, 为进一步验证本文提出的 CaR 方法是否能够缓解上述问题, 本文引入 CaR-reversal 变体, 即仅使用 CaR 学习得到的互补表示进行预测, 而完全移除证据信息. 图 2(b) 所示结果清晰表明: 该变体在所有评估数据集上的性能均显著低于完整的 CaR 模型, 且相较于 CDA-GR 的互补预测结果下降更为明显. 这一现象表明, CaR 方法通过证据-互补解耦机制与回声学习策略, 有效地抑制了互补信息中与标签相关的部分, 从而更好地满足了“互补子图与标签条件独立”的核心假设.

4.6 回声学习策略的结构泛化能力 (RQ4)

消融实验已经验证了本文提出的回声学习策略在 CaR 方法中的有效性. 基于这一结果, 本文进一步提出一个更具探索性的研究问题: 该策略是否具备良好的结构泛化能力, 即能否被成功应用于其他基于节点划分的图级证

据提取方法, 从而提升其性能? 为此, 本文选取了 4 种代表性的基于节点划分的证据提取方法 DARE、InterRAT、RGDA 和 C2R, 并在其原有模型结构中引入回声学习策略, 并在 OGB 系列真实分子数据集上进行系统实验. 实验结果如表 4 所示, 引入回声学习策略后, 4 种方法在几乎所有数据集与评估指标上均取得了不同程度的性能提升. 这一现象表明, 本文提出的回声学习策略不仅在 CaR 框架内有效, 其机制本质上具有较强的通用性和适应性, 能够帮助不同方法更好地解耦互补信息与证据表示, 进而增强模型对真实因果结构的捕捉能力. 由此可见, 该策略体现出良好的通用性与结构泛化能力, 具备在更广泛的图分类与证据提取任务中推广应用的潜力.

表 4 回声学习策略的结构泛化能力分析

方法	MolHIV	MolToxCast	MolBBBP	MolBACE	MolSIDER
DARE	0.783 6	0.667 7	0.682 0	0.823 9	0.592 1
DARE+echo	0.790 1 (+0.65%)	0.672 1 (+0.44%)	0.693 7 (+1.17%)	0.829 9 (+0.60%)	0.610 3 (+1.82%)
InterRAT	0.744 6	0.653 1	0.675 3	0.795 8	0.582 1
InterRAT+echo	0.768 3 (+2.37%)	0.668 2 (+1.51%)	0.688 7 (+1.34%)	0.808 4 (+1.26%)	0.606 9 (+2.48%)
RGDA	0.771 4	0.669 4	0.695 3	0.818 7	0.586 4
RGDA+echo	0.795 6 (+2.42%)	0.673 7 (+0.43%)	0.707 3 (+1.20%)	0.835 6 (+1.69%)	0.617 3 (+3.09%)
C2R	0.791 9	0.670 9	0.699 9	0.814 3	0.613 1
C2R+echo	0.802 4 (+1.05%)	0.679 2 (+0.83%)	0.715 8 (+1.59%)	0.832 8 (+1.85%)	0.619 7 (+0.66%)

4.7 可视化验证 (RQ5)

本节展示了在 Spurious-Motif 数据集 (伪相关强度参数 $b = 0.9$) 上训练得到的 CaR 模型的可视化结果. 图 6 直观呈现了 CaR 提取出的若干关键证据子图. 每个示例图包含一个 motif 类型子图 (Cycle、House、Crane 这 3 种结构之一) 和对应的 base 子图 (Tree、Wheel、Ladder 这 3 种结构之一). 其中, 模型认为与预测相关的重要节点以藏青色高亮表示, 具体判定标准为节点的证据概率 $\widetilde{m}_i$ 超过阈值 0.55. 同时, 节点之间的连接边用红色线条直观展示.

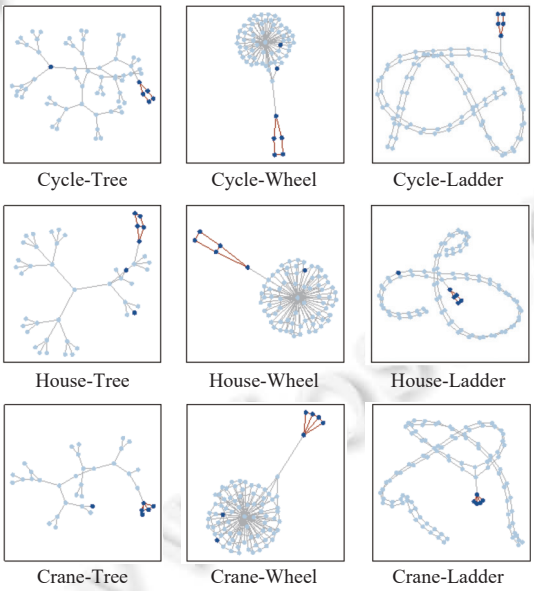


图 6 CaR 提取的证据子图的可视化展示

从这些可视化结果中可以明显观察到, CaR 能够精准地识别出对任务预测至关重要的证据结构, 成功区分真实的因果子图 (motif) 与互补信息子图 (base), 这表明模型并未受到强伪相关结构的干扰, 而是更偏向于基于因果有效性的子图进行判别, 体现出较强的因果判别能力与特征解耦能力. 上述实验结果验证了本文提出的证据-互补

解耦机制和回声学习策略的有效性.

5 总 结

本文提出了一种面向图分类任务的互补感知证据提取方法 CaR, 旨在通过引入反事实数据增强策略, 有效提升图神经网络模型的可解释性与泛化能力. 该方法基于一个核心假设: 在给定证据的条件下, 互补部分与标签应保持相互独立. 该假设的引入有助于构造更具判别性的反事实样本, 从而缓解模型对伪相关特征的依赖问题. 为实现上述目标, 本文首先设计了基于对比学习的证据-互补解耦机制, 从表示空间层面促使互补表示远离证据表示, 增强了互补部分与标签之间的独立性, 有效提升了模型对因果结构的辨识能力. 进一步地, 针对现有方法普遍存在仅依赖最终消息传递层获取互补表示, 导致互补信息可能包含证据信息的问题, 本文提出回声学习 (echo-learning) 策略. 该策略通过在多个早期图神经网络层中采样互补表示, 显著提升了互补信息的多样性, 从而增强了反事实样本的表达能力, 提升了模型在复杂图结构下的适应能力与泛化能力. 在多个真实图分类数据集与构造的合成数据集上进行的系统实验证明, 本文所提 CaR 方法在任务预测性能与证据提取准确性等方面均显著优于现有代表性方法, 说明了提出方法的有效性.

尽管本文提出的回声学习与互补感知反事实增强策略在缓解捷径学习和提升解释忠实性方面取得了良好效果, 但仍存在一定局限性. 首先, 回声学习中历史层的选择及融合权重主要依赖经验设定, 不同数据集和网络深度下的最优层数可能存在差异, 尚缺乏自适应的层选择机制, 影响方法的通用性与可复现性. 其次, 虽然通过打乱或重组互补表示能够有效破坏证据与环境之间的捷径共现关系, 从而削弱模型对非因果相关性的依赖, 但所构造的反事实样本在部分情况下可能偏离真实数据分布, 可能对训练过程的稳定性以及模型在真实应用场景中的泛化性能产生一定影响, 有待在后续工作中结合分布约束或生成式建模进一步加以缓解.

References

[1] Hu WH, Fey M, Zitnik M, Dong YX, Ren HY, Liu BW, Catasta M, Leskovec J. Open graph benchmark: Datasets for machine learning on graphs. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1855.

[2] Guo ZC, Zhang CX, Yu WH, Herr J, Wiest O, Jiang M, Chawla NV. Few-shot graph learning for molecular property prediction. In: Proc. of the 2021 Web Conf. Ljubljana: ACM, 2021. 2559–2567. [doi: 10.1145/3442381.3450112]

[3] Yu B, Yin HT, Zhu ZX. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 3634–3640.

[4] Wu SW, Sun F, Zhang WT, Xie X, Cui B. Graph neural networks in recommender systems: A survey. ACM Computing Surveys, 2023, 55(5): 97. [doi: 10.1145/3535101]

[5] Ying R, Bourgeois D, You JX, Zitnik M, Leskovec J. GNNExplainer: Generating explanations for graph neural networks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 829.

[6] Fang JF, Liu W, Gao Y, Liu ZM, Zhang A, Wang X, He XN. Evaluating post-hoc explanations for graph neural networks via robustness analysis. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2024. 3168.

[7] Lee S, Wang XT, Han S, Yi XY, Xie X, Cha M. Self-explaining deep models with logic rule reasoning. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 232.

[8] Gui SR, Liu M, Li XN, Luo YZ, Ji SW. Joint learning of label and environment causal independence for graph out-of-distribution generalization. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2024. 174.

[9] Sui YD, Tang CZ, Chu ZX, Fang JF, Gao Y, Cui Q, Li LF, Zhou J, Wang X. Invariant graph learning for causal effect estimation. In: Proc. of the 2024 ACM Web Conf. Singapore: ACM, 2024. 2552–2562. [doi: 10.1145/3589334.3645549]

[10] Wu YX, Wang X, Zhang A, He XN, Chua TS. Discovering invariant rationales for graph neural networks. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[11] Sui YD, Wang X, Wu JC, Lin M, He XN, Chua TS. Causal attention for interpretable and generalizable graph classification. In: Proc. of the 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2022. 1696–1705. [doi: 10.1145/3534678.3539366]

[12] Yue LN, Liu Q, Liu Y, Gao WB, Yao FZ, Li WF. Cooperative classification and rationalization for graph generalization. In: Proc. of the

2024 ACM Web Conf. Singapore: ACM, 2024. 344–352. [doi: 10.1145/3589334.3645332]

[13] Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2020, 2(11): 665–673. [doi: 10.1038/s42256-020-00257-z]

[14] Friedrich F, Stammer W, Schramowski P, Kersting K. A typology for exploring the mitigation of shortcut behaviour. Nature Machine Intelligence, 2023, 5(3): 319–330. [doi: 10.1038/s42256-023-00612-w]

[15] Chang SY, Zhang Y, Yu M, Jaakkola TS. Invariant rationalization. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 135.

[16] Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. arXiv:1907.02893, 2019.

[17] Feng WZ, Zhang J, Dong YX, Han Y, Luan HB, Xu Q, Yang Q, Kharlamov E, Tang J. Graph random neural networks for semi-supervised learning on graphs. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1853.

[18] Li HY, Wang X, Zhang ZW, Zhu WW. Out-of-distribution generalization on graphs: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2025, 47(11): 10490–10512. [doi: 10.1109/TPAMI.2025.3593897]

[19] Fan SH, Wang X, Shi C, Cui P, Wang B. Generalizing graph neural networks on out-of-distribution graphs. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2024, 46(1): 322–337. [doi: 10.1109/TPAMI.2023.3321097]

[20] Miao S, Liu M, Li P. Interpretable and generalizable graph learning via stochastic attention mechanism. In: Proc. of the 39th Int'l Conf. on Machine Learning. Baltimore: PMLR, 2022. 15524–15543.

[21] Yue LN, Liu Q, Wang L, An YQ, Du YC, Huang ZY. Interventional rationalization. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 11404–11418. [doi: 10.18653/v1/2023.emnlp-main.700]

[22] Fan SH, Wang X, Mo YH, Shi C, Tang J. Debiasing graph neural networks via learning disentangled causal substructure. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1808.

[23] Liu G, Inae E, Luo TF, Jiang M. Rationalizing graph neural networks with data augmentation. ACM Trans. on Knowledge Discovery from Data, 2024, 18(4): 86. [doi: 10.1145/3638781]

[24] van den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

[25] Knyazev B, Taylor GW, Amer MR. Understanding attention and generalization in graph neural networks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 378.

[26] Zhao G, Wang QG, Yao F, Zhang YF, Yu G. Survey on large-scale graph neural network systems. Ruan Jian Xue Bao/Journal of Software, 2022, 33(1): 150–170 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6311.htm [doi: 10.13328/j.cnki.jos.006311]

[27] Sun YH, Li GY, Du JJ, Ning B, Chen H. A subgraph matching algorithm based on subgraph index for knowledge graph. Frontiers of Computer Science, 2022, 16(3): 163606. [doi: 10.1007/s11704-020-0360-y]

[28] Fang P, Wang F, Shi Z, Feng D, Yi QX, Xu XH, Zhang YX. An efficient memory data organization strategy for application-characteristic graph processing. Frontiers of Computer Science, 2022, 16(1): 161607. [doi: 10.1007/s11704-020-0255-y]

[29] Liu XK, Zhang K, Liu Y, Chen EH, Huang ZY, Yue LN, Yan JX. RHGN: Relation-gated heterogeneous graph network for entity alignment in knowledge graphs. In: Findings of the Association for Computational Linguistics: ACL 2023. Toronto: ACL, 2023. 8683–8696. [doi: 10.18653/v1/2023.findings-acl.553]

[30] Gao WB, Wang H, Liu Q, Wang F, Lin X, Yue LN, Zhang Z, Lv R, Wang SJ. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In: Proc. of the 46th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Taipei: ACM, 2023. 983–992. [doi: 10.1145/3539618.3591774]

[31] Zhou CY, Zeng C, He P, Zhang Y. GKCI: An improved GNN-based key class identification method. Ruan Jian Xue Bao/Journal of Software, 2023, 34(6): 2509–2525 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6846.htm [doi: 10.13328/j.cnki.jos.006846]

[32] Chen YQ, Zhang YG, Bian YT, Yang H, Ma KL, Xie BH, Liu TL, Han B, Cheng J. Learning causally invariant representations for out-of-distribution generalization on graphs. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. 2022. 22131–22148.

[33] Li HY, Zhang ZW, Wang X, Zhu WW. Learning invariant graph representations for out-of-distribution generalization. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 859.

[34] Chen JL, Ying R. TempME: Towards the explainability of temporal graph neural networks via motif discovery. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 1263.

[35] Chen YQ, Bian YT, Zhou KW, Xie BH, Han B, Cheng J. Does invariant graph learning via environment augmentation learn invariance? In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 3130.

[36] Chen YQ, Bian YT, Han B, Cheng J. How interpretable are interpretable graph neural networks? In: Proc. of the 41st Int'l Conf. on Machine Learning. PMLR, 2024. 6413–6456.

[37] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 1135–1144. [doi: 10.1145/2939672.2939778]

[38] Li P, Song SH, Zhang Y, Cao HW, Ye XC, Tang ZM. HSEGRL: A hierarchical self-explainable graph representation learning model. Journal of Computer Research and Development, 2024, 61(8): 1993–2007 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.202440142]

[39] Wang YY, Zhang M, Yang JR, Xu SK, Chen YX. Research on fairness in deep learning models. Ruan Jian Xue Bao/Journal of Software, 2023, 34(9): 4037–4055 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6872.htm [doi: 10.13328/j.cnki.jos.006872]

[40] Yue LN, Liu Q, Du YC, An YQ, Wang L, Chen EH. DARE: Disentanglement-augmented rationale extraction. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1929.

[41] Sun X, Wang L, Liu Q, Wu S, Wang ZL, Wang L. DIVE: Subgraph disagreement for graph out-of-distribution generalization. In: Proc. of the 30th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Barcelona: ACM, 2024. 2794–2805. [doi: 10.1145/3637528.3671878]

[42] Yang YH, Wu L, Liao YX, He ZZ, Shao PY, Hong RC, Wang M. Invariance matters: Empowering social recommendation via graph invariant learning. In: Proc. of the 48th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Padua: ACM, 2025. 2038–2047. [doi: 10.1145/3726302.3730013]

[43] Yue LN, Liu Q, Du YC, Wang L, An YQ, Chen EH. Towards reliable and faithful explanations: A disentanglement-augmented approach for selective rationalization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2025, 47(11): 9906–9921. [doi: 10.1109/TPAMI.2025.3592313]

[44] Xu Z, Pan MH, Chen YZ, Chen HY, Yan YC, Das M, Tong HH. Fine-grained graph rationalization. In: Proc. of the 34th ACM Int'l Conf. on Information and Knowledge Management. Seoul: ACM, 2025. 3708–3719. [doi: 10.1145/3746252.3761307]

[45] Piao YH, Lee S, Lu YJX, Kim S. Improving out-of-distribution generalization in graphs via hierarchical semantic environments. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 27631–27640. [doi: 10.1109/CVPR52733.2024.02609]

[46] Yao TJ, Chen YQ, Chen ZH, Hu K, Shen ZQ, Zhang K. Empowering graph invariance learning with deep spurious infomax. In: Proc. of the 41st Int'l Conf. on Machine Learning. PMLR, 2024. 56860–56884.

[47] Jang E, Gu SX, Poole B. Categorical reparametrization with Gumble-Softmax. arXiv:1611.01144, 2016.

[48] Tishby N, Pereira FC, Bialek W. The information bottleneck method. arXiv:physics/0004057, 2000.

[49] Alemi AA, Fischer I, Dillon JV, Murphy K. Deep variational information bottleneck. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[50] Pearl J, Glymour M, Jewell NP. Causal Inference in Statistics: A Primer. Chichester: John Wiley & Sons, 2016.

[51] Wang T, Huang JQ, Zhang HW, Sun QR. Visual commonsense R-CNN. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10760–10770. [doi: 10.1109/CVPR42600.2020.01077]

[52] Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[53] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[54] Li QM, Han ZC, Wu XM. Deeper insights into graph convolutional networks for semi-supervised learning. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 433.

附中文参考文献

[26] 赵港, 王千阁, 姚烽, 张岩峰, 于戈. 大规模图神经网络系统综述. 软件学报, 2022, 33(1): 150–170. http://www.jos.org.cn/1000-9825/6311.htm [doi: 10.13328/j.cnki.jos.006311]

[31] 周纯英, 曾诚, 何鹏, 张龑. GKCI: 改进的基于图神经网络的关键类识别方法. 软件学报, 2023, 34(6): 2509–2525. http://www.jos.org.cn/1000-9825/6846.htm [doi: 10.13328/j.cnki.jos.006846]

[38] 李平, 宋舒寒, 张园, 曹华伟, 叶笑春, 唐志敏. HSEGRL: 一种分层可自解释的图表示学习模型. 计算机研究与发展, 2024, 61(8): 1993–2007. [doi: 10.7544/issn1000-1239.202440142]

[39] 王昱颖, 张敏, 杨晶然, 徐晟恺, 陈仪香. 深度学习模型中的公平性研究. 软件学报, 2023, 34(9): 4037–4055. http://www.jos.org.cn/1000-9825/6872.htm [doi: 10.13328/j.cnki.jos.006872]

作者简介

岳立楠, 博士, 副教授, CCF 专业会员, 主要研究领域为图表示学习, 机器学习, 数据挖掘.

张敏灵, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为机器学习, 数据挖掘.