

基于显式独立性解耦的分布外泛化动态图注意力网络

李昊阳¹, 王鑫^{1*}, 朱文武^{1*}

1. 清华大学计算机科学与技术系, 北京 100084

* 通信作者. E-mail: xin_wang@tsinghua.edu.cn, wwzhu@tsinghua.edu.cn

国家优秀青年科学基金项目 (批准号: 62222209) 和北京信息科学与技术国家研究中心 (批准号: BNR2023TD03006) 资助

摘要 动态图神经网络近年来因其在建模时空动态信息方面的强大表现而备受关注。然而在现实世界中动态图往往存在分布偏移问题, 这种偏移可能来源于时间变化、空间变化, 甚至内在的数据生成机制不一致等。现有动态图神经网络模型在分布偏移下的性能显著下降, 尤其是在复杂场景中, 模型倾向于过度依赖与标签形成虚假关联的变化模式 (即伪相关模式), 难以刻画具有跨分布稳定预测力的时空不变模式, 导致泛化性差。为了解决这一问题, 本文提出了一种基于显式独立性解耦的动态图注意力网络, 所提方法通过设计专门的模块与优化机制, 实现对动态图中时空不变模式的有效建模并显著提升分布偏移下的泛化性能。本文包含三方面的技术创新: 首先, 所提方法设计了解耦动态图注意力网络, 用以建模动态图中的可泛化的时空不变模式, 并通过显式分离不变模式和变化模式而避免模型过度依赖容易受到分布偏移影响的变化模式。其次, 所提方法引入基于因果推断的时空干预机制, 通过生成多种干预分布来模拟分布偏移环境, 帮助消除伪相关性, 提升模型对不变模式的学习能力。最后, 所提方法设计了基于不变性约束和独立性约束的解耦优化模块, 增强模型在跨分布预测任务中的泛化性。本文在多个真实数据集和模拟数据集上进行了广泛的实验。实验结果表明与现有的动态图神经网络模型和分布外泛化方法相比, 本文所提方法在分布偏移场景下显著提高了模型预测泛化性能。此外, 本文还通过消融实验和可视化分析验证了模型中关键模块的有效性和必要性, 并对分布外泛化的动态图神经网络进行了未来研究方向的展望。

关键词 动态图神经网络, 解耦, 分布偏移, 图机器学习, 独立性

1 引言

动态图广泛存在于许多现实世界的应用场景中, 例如金融网络、社交网络以及交通网络等 [1~3]。不同于传统静态图的是, 动态图不仅包含节点和边的时间切片, 还包含与时间相关的结构和特征演化模式。动态图由于其更为复杂的特性, 使得对其建模和分析在实际场景中显得至关重要。近年来,

引用格式: 李昊阳, 王鑫, 朱文武. 基于显式独立性解耦的分布外泛化动态图注意力网络. 中国科学: 信息科学, 在审文章

Haoyang Li, Xin Wang, Wenwu Zhu. Independence-Promoted Disentangled Dynamic Graph Attention Network for Out-of-Distribution Generalization. Sci Sin Inform, for review

图神经网络 (Graph Neural Networks, GNNs) 由于其在图表示上的强大能力而得到了广泛应用, 例如节点分类、链路预测以及图分类等 [4~6]。然而传统的图神经网络主要面向静态图而设计, 无法充分捕捉动态图的时间维度下的动态变化特性。因此研究者们提出了动态图神经网络 (Dynamic Graph Neural Networks) 以同时处理动态图中的结构信息和时间动态特性, 并已在许多任务中取得显著进展 [7,8]。例如在交通网络中车辆流量随着时间、地点的变化而动态变化, 动态图神经网络可以通过建模历史交通数据的时空变化特性, 实现高精度的流量预测; 在金融网络中交易中的支付关系会随市场状态的变化而发生显著变化, 动态图神经网络通过捕捉这些动态关系, 可以帮助识别潜在的异常交易行为; 在社交网络中用户之间的互动关系会因社交趋势、兴趣变化等因素而呈现不同的动态模式, 动态图神经网络能够建模这些复杂变化的社交关系, 从而提升用户行为预测的准确性。

尽管动态图神经网络已经在许多任务中取得成功, 但它们在应对分布外数据 (Out-of-Distribution, OOD) 时仍面临巨大挑战。具体而言, 动态图中广泛存在的时空分布偏移 (spatio-temporal distribution shift) 会极大限制动态图神经网络的泛化能力。分布偏移 (distribution shift) 是指训练数据分布与测试数据分布之间存在差异, 这种现象在动态图中尤为普遍 [9]。例如, 在金融网络中, 不同时期或市场的外部条件会显著影响支付流与交易合法性之间的相关性; 在学术合作者网络中, 新兴领域的快速发展会导致不同研究领域合作模式的显著变化; 在推荐网络中, 用户兴趣的变化会引发行行为模式的变化。上述场景表明, 如果动态图神经网络高度依赖于训练数据中拟合到的预测相关的时空模式, 那么在测试数据中, 当这些模式因分布偏移而发生变化时, 模型的泛化性能将显著下降。

为了应对这一问题, 本文探索通过建模动态图中的不变模式 (invariant patterns) 来提高动态图神经网络在分布偏移下的泛化性能。特别地, 动态图的不变模式是指在不同分布下具有真实稳定预测能力的结构和节点特征模式。这些模式能够帮助动态图模型减少对虚假关联的依赖, 从而增强其泛化能力。然而, 在动态图中识别和建模不变模式是具有挑战的。与静态图相比, 动态图的时空特性使得其不变模式更为复杂; 同时不变模式与变化模式 (variant patterns) 之间往往高度耦合且动态变化。因此要解决动态图中的分布偏移问题, 需要克服以下两个主要难点: 首先如何在动态图中有效地识别解耦 (disentangle) 复杂的时空不变模式 (spatio-temporal invariant patterns) 和时空变化模式 (spatio-temporal variant patterns)? 这些模式既包括随时间变化的图结构模式也包括随时间变化的节点特征模式, 其动态性和复杂性给现有图神经网络带来巨大挑战。其次如何基于识别出的不变模式和变化模式以一种系统且高效的方法应对潜在的时空分布偏移, 从而在测试数据中实现更好的泛化性能? 针对上述挑战, 本文提出了一种新颖的动态图建模方法, 即基于显式独立性解耦的分布外泛化动态图注意力网络 (Independence-Promoted Disentangled Dynamic Graph Attention Network, ID-DyGAT)。所提方法通过设计解耦动态图注意力网络、时空干预模块以及独立性促进解耦模块, 系统性地解决了动态图中的分布偏移问题, 显著提升动态图神经网络的分布外泛化性能。

首先, 本文提出了解耦动态图注意力网络用于捕捉动态图中的不变模式和变化模式。通过这一模块, 图中的每个节点能够通过解耦注意力机制访问其所有的历史邻居, 并对邻居的时空信息进行加权, 有效地提取动态邻域中具有稳定预测能力的不变模式, 同时抑制变化模式对模型的干扰。该模块通过解耦节点的时空邻域特征, 分别计算不变模式和变化模式的注意力权重, 从而在不同时间点捕捉动态图中的对于标签具有稳定泛化预测能力的不变模式。其次, 本文引入了时空干预 (spatio-temporal intervention) 机制, 该机制受因果推断理论的启发, 通过采样和重组动态图中的变化模式, 生成多个干预分布 (intervened distributions), 从而显式地模拟不同的分布偏移环境。在干预 (intervention) 操作中, 变化模式被随机替换而不变模式保持固定。由于不变模式和预测目标保持真实稳定的预测关系, 因此这一干预过程能够有效消除变化模式的虚假关联 (即伪相关) 影响。最后,

本文在训练过程中引入了独立性促进解耦模块, 其中不变性约束通过正则化项对模型的预测结果进行多环境不变性约束, 而独立性约束通过显式鼓励不变模式和变化模式之间的独立性更好地实现不变模式和变化模式之间的解耦, 促使模型更加专注于学习和利用不变模式进行可泛化的预测。为验证所提方法的有效性, 本文在多个真实数据集和模拟数据集上进行了广泛实验。结果表明, 与现有基线方法相比, 所提方法在节点分类和链路预测任务中表现出显著的泛化性能优势。对解耦动态图注意力网络、时空干预模块和独立性促进解耦模块的消融实验验证了各模块对性能提升的关键作用。

本文主要贡献总结如下: 首先本文提出了一种开创性的分布外泛化的动态图神经网络, 首次从不变模式和变化模式独立解耦的视角对动态图时空分布偏移问题进行系统性研究。其次本文设计了解耦动态图注意力网络, 其能够有效捕捉动态图中的不变模式和变化模式; 还设计了基于因果推断的时空干预机制以生成多个干预分布而模拟分布偏移的环境。最后通过包含不变性约束和独立性约束的优化策略, 本文显著提升了动态图神经网络在分布偏移下的泛化能力并在多个真实数据集和模拟数据集上得到验证。下面将具体回顾相关工作, 然后逐步介绍我们所提出的方法并进行实验分析。

本文的后续章节主要包括如下内容: 第 2 节回顾相关工作, 涵盖动态图神经网络、分布外泛化、解耦表征学习的最新研究进展; 第 3 节详细介绍所提出的基于显式独立性解耦的分布外泛化动态图注意力网络, 包括问题定义、模型设计及优化目标; 第 4 节呈现实验设定与结果分析, 并通过消融实验、效率比较与超参敏感性分析等实验进一步验证方法的有效性; 第 5 节展望未来值得探索的方向; 第 6 节总结全文内容。

2 相关工作介绍

本文主要包含三方面相关工作, 分别是动态图神经网络、分布外泛化、解耦表征学习。

动态图神经网络。近年来, 动态图神经网络 (Dynamic Graph Neural Networks) 逐渐成为处理复杂动态图数据的重要工具, 其通过建模图结构和时间信息, 在众多领域的预测任务中展现了强大的能力 [4~6, 10, 11]。与静态图不同, 动态图能够表达随时间变化的结构和特征模式, 这种特性使得动态图神经网络在建模现实世界的复杂网络演化方面具有独特的优势。随着数据规模和复杂性的增长, 研究者们提出了多种方法提升动态图神经网络的性能。动态图神经网络的研究方向主要集中在如何有效地捕获动态图中的时空信息。一类经典方法依赖于序列模型和时间编码技术来建模时间维度。例如, GCRN 模型通过图卷积网络捕获图的结构信息, 并结合门控循环单元 (GRU) [12] 沿时间维度建模图的演化 [13]。DySAT 通过结构注意力机制聚合每个时间快照中的邻域信息, 并利用时间自注意力机制动态赋权, 从而适应来自不同时间和邻域的消息 [14]。另一类方法则引入时间编码以表示时间函数, 并结合空间模块如图神经网络或记忆模块处理结构信息。例如, TGAT 提出了一种时间编码方法, 从而使得节点表征能够内在表示为时间的函数 [15]。EGCN 通过演化图卷积网络的参数而非节点表征, 避免了对全时间跨度节点知识的依赖, 适用于节点集合频繁变化的情况 [16]。研究者们还提出了动态图的多尺度建模方法。例如, 通过结合局部和全局信息, 模型能够在多种时间尺度下捕获动态图中的动态特征。这些方法包括基于卷积的时间窗口分割以及递归的时间依赖性建模, 从而适应不同时间跨度的预测任务 [17]。这些方法已广泛应用于动态异常检测、事件预测、动态推荐等实际场景。例如, DGEL 提出一种动态图演化学习框架, 通过内在交互潜力、时间衰减邻域增强和共生局部结构学习, 从不同视角高效更新节点表示, 用于动态环境下的推荐任务 [18]。DynShare 进一步扩展到社交平台中, 通过动态分享推荐模型实现基于时间戳的好友推荐, 有效提高社交电商平台的用户黏性 [19]。然而, 现有方法通常假设训练和测试数据分布相同, 而在现实应用中, 这一假设难以成立。从金融网络中的市场波动到社交网络中的社区动态, 分布偏移广泛存在, 导

致模型在测试阶段性能显著下降。这种问题凸显了对动态图分布外泛化问题研究的重要性。

分布外泛化。分布外泛化 (Out-of-Distribution Generalization, OOD) 是机器学习领域近年来备受关注的研究方向, 其核心问题是如何在训练和测试分布不同的情况下保证模型的泛化能力 [20]。现有研究主要集中在以下两类技术路径上, 即风险最小化和不变性学习。具体来说, IRM 通过在多个训练域中最小化经验风险并学习不变预测器, 使得分类器能够在所有环境中泛化 [21]。GroupDRO 则通过对误差较大的训练域赋予更高权重, 最小化最差组的风险 [22]。针对图数据, 现有研究探索了结构分布和特征分布的变化对图模型的影响。一些方法基于因果推断假设提出结构因果模型 (SCM), 从而学习具有外推能力的图表示 [23]。例如, EERM 通过引入对抗性训练的上下文探索器生成多个虚拟环境以最大化风险方差, 从而提高单一观察环境下的模型外推能力 [24]。DIR 通过捕捉在结构分布变化下不变的因果模式, 并过滤掉不稳定的伪相关模式来实现分布外泛化 [25]。当前的研究主要针对静态图的结构分布变化而对动态图中时空分布变化的研究尚属空白。

解耦表征学习。解耦表征学习旨在表征数据中的多种潜在解释因素, 这些因素通常以向量的形式独立表示 [26]。经典工作如 β -VAE 和其改进版本 [27, 28] 通过在变分自编码器中引入正则化约束实现潜在因子的解耦, 使得不同潜在因子分别对应输入实例的不同语义属性, 从而支持可控生成与因子干预。InfoGAN [29] 则通过最大化潜在因子与生成样本间的互信息而学习解耦潜在因子。扩散模型同样被用于解耦表征学习, 例如 DisDiff [30] 和 DisenBooth [31] 等方法在图像与视频生成中通过解耦表征实现对生成过程的控制。在自然语言处理领域, 研究者同样尝试将语义与句法、内容与风格等因素解耦, 以提升模型的可解释性和泛化性 [32, 33]。近年来, 解耦表征学习在图数据上的研究逐渐兴起。例如, DisenGCN 通过邻域路由机制识别导致节点与邻居链接的潜在因素, 从而实现节点表征的解耦 [34]。此外, GRACES 设计了一种自监督解耦图编码器, 用于表征多样图结构中的不变因素, 从而促进后续的图神经架构搜索 [35]。在动态图背景下, DisenCTR 通过提取用户兴趣的多样性并建模其历史行为的时间效应, 显著提升了动态推荐任务的性能 [36]。一些研究还提出了结合生成模型和解耦表征的方法, 通过分解动态图中的节点、边、时间和空间等因素, 实现动态图生成的可解释性 [37]。然而现有方法大多不能直接被应用于动态图分布外泛化问题, 因此本文提出了一种基于显式独立性解耦的动态图注意力网络, 通过区分时空不变模式和时空变化模式, 并利用不变模式进行预测而提高分布外泛化能力。

3 基于显式独立性解耦的分布外泛化动态图注意力网络

本节中我们提出基于显式独立性解耦的动态图注意力网络以应对动态图中的时空分布偏移问题 (图 1)。首先我们设计了解耦动态图注意力网络用于提取不变与变化的时空模式。然后我们提出了时空干预机制用于生成多个干预数据分布, 并结合不变性损失函数最小化各干预分布之间的预测方差。最后, 我们设计了显式独立性解耦正则项用于提升所学习不变模式的质量, 并将不变性正则项和独立性正则项与预测损失函数联合优化, 以引导模型依赖不变模式进行分布外稳定可泛化的预测。

3.1 问题定义

在介绍所提方法前, 本节将首先介绍动态图及相关预测任务的定义, 并形式化地描述动态图中时空分布偏移下的预测问题。同时, 本文所使用的符号在表 1 中进行了汇总。设图 $\mathcal{G}$ 包含节点集合 $\mathcal{V}$ 和边集合 $\mathcal{E}$, 则动态图可以定义为: $\mathcal{G} = (\{\mathcal{G}^t\}_{t=1}^T)$, 其中 T 表示时间步数量, $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$ 表示时间步 t 的图快照, $\mathcal{V} = \bigcup_{t=1}^T \mathcal{V}^t$, $\mathcal{E} = \bigcup_{t=1}^T \mathcal{E}^t$ 。我们用 $\mathbf{G}^t$ 表示 $\mathcal{G}^t$ 对应的随机变量。在动态图中,

预测任务可概括为利用过去的图结构进行未来的预测, 即 $p(\mathbf{Y}^t | \mathbf{G}^1, \mathbf{G}^2, \dots, \mathbf{G}^t) = p(\mathbf{Y}^t | \mathbf{G}^{1:t})$. 标签 $\mathbf{Y}^t$ 表示节点的属性或节点之间在 $t+1$ 时刻的链接关系. 本文主要关注节点级别的任务, 这是动态图相关文献中较为常见的设定 [7, 8]. 我们遵循 [24, 38] 的处理方式, 将图轨迹的整体分布分解为以节点为中心的自我图轨迹 (ego-graph trajectory) 分布: $p(\mathbf{Y}^t | \mathbf{G}^{1:t}) = \prod_v p(\mathbf{y}^t | \mathbf{G}_v^{1:t})$. 从节点 v 在时间 t 引出的自我图由其 L 阶邻居 ($\mathcal{N}_v^t$) 之间的邻接关系以及这些邻居节点的特征构成. 模型的优化目标是通过经验风险最小化学习一个最优的预测函数: $\min_{\theta} \mathbb{E}_{(\mathbf{y}^t, \mathcal{G}_v^{1:t}) \sim p_{tr}(\mathbf{y}^t, \mathbf{G}_v^{1:t})} \mathcal{L}(f_{\theta}(\mathcal{G}_v^{1:t}), \mathbf{y}^t)$, 其中 f_{θ} 表示可学习的动态图神经网络模型. 我们用 $\mathbf{G}_v^{1:t}$ 和 $\mathbf{y}^t$ 表示自我图轨迹及其标签的随机变量, $\mathcal{G}_v^{1:t}$ 和 y^t 表示其对应的具体实例.

当存在分布偏移问题时, 基于训练分布学习到的最优预测函数无法很好地泛化到测试分布. 在动态图相关研究中, 已有研究致力于建模网络演化中的稳定规律 [39, 40]. 受到上述研究启发, 我们假设条件分布保持一致, 即 $p_{tr}(\mathbf{Y}^t | \mathbf{G}^{1:t}) = p_{te}(\mathbf{Y}^t | \mathbf{G}^{1:t})$, 且仅考虑协变量偏移问题, 即 $p_{tr}(\mathbf{G}^{1:t}) \neq p_{te}(\mathbf{G}^{1:t})$. 除了时间变化数据中普遍存在的时间分布偏移 [41, 42], 以及非欧几里得数据中的结构分布偏移 [25], 动态图中还存在更为复杂的时空分布偏移, 例如不同时间段或不同社区中的自我图轨迹分布可能存在显著差异.

3.2 应对时空分布偏移的目标函数

动态图研究中已有工作尝试总结网络演化过程中的关键模式, 以反映真实世界网络随时间变化的演化规律 [43, 44]. 例如, 三元闭包规律指出在社交网络中拥有共同邻居的两个节点 (即特定结构模式) 往往倾向于在未来建立连接 [45, 46]. 除了结构信息外, 节点属性同样构成时空模式的重要组成部分, 例如社交互动也可能受到性别、年龄等因素影响 [47]. 与手动总结模式不同, 我们希望通过动态图神经网络自动学习时空模式, 从而在动态图中挖掘出结构与特征混合的复杂时空模式. 因此, 我们将用于节点级预测的时空模式定义为自我图轨迹的一个子集: $P^t(v) = m_v^t(\mathcal{G}_v^{1:t})$, 其中 $m_v^t(\cdot)$ 表示从节点 v 的自我图轨迹中选择结构与特征的函数. 在工作 [46] 中, 该模式可以解释为具有相似邻域结构的开放三元组, 模型根据三元闭包原理倾向于预测是否连接该三元组中尚未连接的节点对, 即 $\hat{g}_{u,v}^t = f_{\theta}(P^t(u), P^t(v))$. 动态图神经网络的目标是利用具有预测性的时空模式来提升预测性能. 然而源于时空分布偏移的影响, 某些模式的预测能力在不同时间段或社区之间可能存在差异. 受因果理论 [48, 49] 启发, 我们提出如下假设:

假设 1 对给定任务存在一个预测器 $f(\cdot)$ 使得对来自任意分布的样本 $(\mathcal{G}_v^{1:t}, y^t)$, 存在不变模式 $P_I^t(v)$ 和变化模式 $P_V^t(v)$ 满足 $y^t = f(P_I^t(v)) + \epsilon$ 且 $P_I^t(v) = \mathcal{G}_v^{1:t} \setminus P_V^t(v)$, 即 $\mathbf{P}_I^t(v) \perp \mathbf{P}_V^t(v)$ 成立.

在假设 1 中, $P_I^t(v) = \mathcal{G}_v^{1:t} \setminus P_V^t(v)$ 表示动态图由不变模式与变化模式共同构成. 该假设指出不变模式 $\mathbf{P}_I^t(v)$ 对于预测标签 y^t 是充分的, 且可在不同时间段与环境间通用而无需重新调整预测函数, 而变化模式 $\mathbf{P}_V^t(v)$ 与不变模式独立从而对于标签 y^t 不具有真实的预测能力.

为了获得更强的泛化能力, 模型应依赖于不变模式而非变化模式, 因为前者在不同分布下始终具有良好的预测能力, 而后者在分布偏移下的预测性能可能会显著变化. 因此优化目标可以表达为:

$$\min_{\theta_1, \theta_2} \mathbb{E}_{(\mathbf{y}^t, \mathcal{G}_v^{1:t}) \sim p_{tr}(\mathbf{y}^t, \mathbf{G}_v^{1:t})} \mathcal{L}(f_{\theta_1}(\tilde{P}_I^t(v)), y^t), \quad s.t. \quad \phi_{\theta_2}(\mathcal{G}_v^{1:t}) = \tilde{P}_I^t(v), \quad \tilde{\mathbf{P}}_I^t(v) \perp \tilde{\mathbf{P}}_V^t(v), \quad (1)$$

其中 $f_{\theta_1}(\cdot)$ 表示基于不变模式的预测函数, $\phi_{\theta_2}(\cdot)$ 表示用于提取不变模式的函数. 然而实现该目标具有两方面挑战: 首先不变模式与变化模式并没有显式标签来区分, 模型需要通过优化过程自动区分. 其次为了使模型真正依赖于不变模式进行预测, 需设计特定机制来实现不变且充分的预测. 为此我们从两个层面引入正则化项, 分别受到因果推断理论和解耦表征学习理论的启发以引导以优化

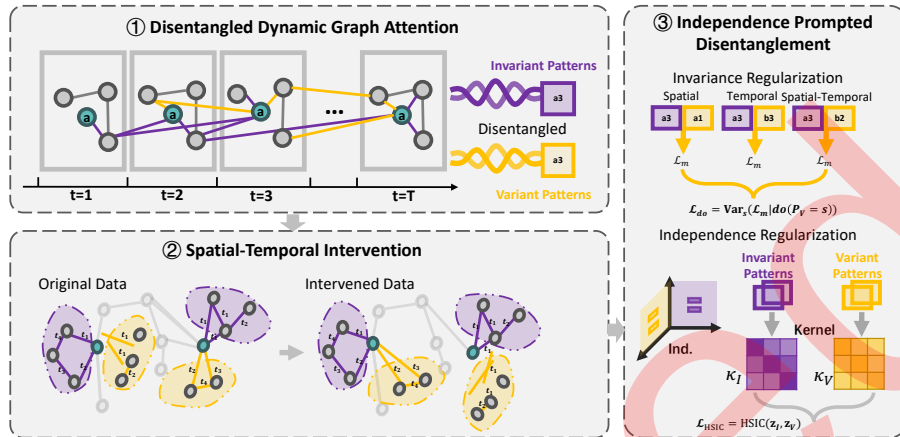


图 1 模型框架。所提模型主要包含三个模块：解耦动态图注意力网络模块、时空干预模块和独立性促进解耦模块
Figure 1 Model Framework. The proposed model mainly consists of three modules: disentangled dynamic graph attention network, spatio-temporal intervention module, independence-promoted disentanglement module.

过程帮助模型识别不变模式进行分布外泛化的预测。根据因果推断理论 [48, 49], 公式 (1) 可实现为:

$$\begin{aligned} \min_{\theta_1, \theta_2} \mathbb{E}_{(y^t, \mathcal{G}_v^{1:t}) \sim p_{tr}} \mathcal{L}(f_{\theta_1}(\phi_{\theta_2}(\mathcal{G}_v^{1:t})), y^t) \\ + \beta \text{Var}_{s \in \mathcal{S}} (\mathbb{E}_{(y^t, \mathcal{G}_v^{1:t}) \sim p_{tr}} \mathcal{L}(f_{\theta_1}(\phi_{\theta_2}(\mathcal{G}_v^{1:t})), y^t) | \text{do}(\mathbf{P}_V^t = s)) + \lambda \mathcal{L}_{ind}(\mathbf{P}_I^t(v), \mathbf{P}_V^t(v)), \end{aligned} \quad (2)$$

其中, ‘do’ 表示通过 do 演算 (do-calculus) 对原始分布进行干预 [49, 50], $\mathcal{S}$ 为干预集合。根据公式 (1), 在已知不变模式 $\mathbf{P}_I^t$ 的条件下, 变化模式 $\mathbf{P}_V^t$ 不影响标签 y^t , 那么对 $\mathbf{P}_V^t$ 进行干预 (如置换、扰动) 不应影响最终预测。由于该损失在样本级别对分布进行干预, 并追求每个样本中的不变模式稳定性, 因此我们将公式 (2) 中的方差项 ($\text{Var}(\cdot)$) 称为不变性约束。为了促进不变模式和变化模式的解耦, 还需要引入额外的独立性约束 $\mathcal{L}_{ind}(\cdot, \cdot)$ (具体实现见第 3.5 节) 以实现 $\hat{\mathbf{P}}_I^t(v) \perp \hat{\mathbf{P}}_V^t(v)$, 从而让模型可以只关注于不变模式进行预测而实现分布外泛化。

3.3 解耦动态图注意力网络

为了同时考虑时间空间信息, 我们定义节点 u 在时刻 t 的动态邻域为 $\mathcal{N}^t(u) = \{v : (u, v) \in \mathcal{E}^t\}$, 即所有在该时间与节点 u 发生交互的节点。对于时刻 t_1 的节点 u , 其历史动态邻域 $\mathcal{N}^t(u), t \leq t_1$ 描述了 u^t 的结构演化轨迹, 使得模型可基于当前时刻聚合不同历史结构信息。例如, u^{t_2} 与 u^{t_3} 在处理相同邻域 $\mathcal{N}^{t_1}(u)$ 时可能聚合不同的消息, 因为用户的兴趣随时间演化, 即便是来自相同邻域的消息也可能被不同程度地聚合。因此模型需具备对邻域演化模式的感知能力。为简化符号记法, 定义中的动态邻域仅考虑一阶邻居, 但该定义可通过引入 n 跳路径上的节点而自然扩展为 n 阶邻居。我们遵循经典消息传递网络的设计, 通过堆叠多层以引入高阶邻域信息。

解耦时空图注意力层。为捕捉每个节点的时空模式, 我们提出一种时空图注意力机制, 使得每个节点能够并行关注其动态邻域中的节点。对于时刻 t 的节点 u 及其历史邻居 $v \in \mathcal{N}^{t'}(u), \forall t' \leq t$, 我们计算查询-键-值向量:

$$\mathbf{q}_u^t = \mathbf{W}_q(\mathbf{h}_u^t \parallel \text{TE}(t)), \quad \mathbf{k}_v^{t'} = \mathbf{W}_k(\mathbf{h}_v^{t'} \parallel \text{TE}(t')), \quad \mathbf{v}_v^{t'} = \mathbf{W}_v(\mathbf{h}_v^{t'} \parallel \text{TE}(t')), \quad (3)$$

其中 $\mathbf{h}_u^t$ 表示节点 u 在时刻 t 的表征, $\mathbf{q}, \mathbf{k}, \mathbf{v}$ 分别表示查询、键和值向量, 省略偏置项以简化表示且本文中所有向量均以行向量表示。TE(t) 表示时间编码函数, 用于获得时刻 t 的时间表征, 从而

隐式考虑链接发生的时间信息 [15,51]。我们计算动态邻域节点之间的注意力分数并构建结构掩码:

$$\mathbf{m}_I = \text{Softmax}\left(\frac{\mathbf{q} \cdot \mathbf{k}^T}{\sqrt{d}}\right), \quad \mathbf{m}_V = \text{Softmax}\left(-\frac{\mathbf{q} \cdot \mathbf{k}^T}{\sqrt{d}}\right), \quad (4)$$

其中 d 是特征维度, $\mathbf{m}_I$ 与 $\mathbf{m}_V$ 分别是不变模式与变化模式的结构掩码。在不变模式中得分高的邻居在变化模式中得分则相对较低, 体现两种模式之间的负相关性。为捕捉不变特征模式我们引入可学习的特征掩码 $\mathbf{m}_f = \text{Softmax}(\mathbf{w}_f)$ 用于从邻域消息中选择稳定特征, 则动态邻域中的消息聚合为:

$$\mathbf{z}_I^t(u) = \text{Agg}_I(\mathbf{m}_I, \mathbf{v} \odot \mathbf{m}_f), \quad \mathbf{z}_V^t(u) = \text{Agg}_V(\mathbf{m}_V, \mathbf{v}), \quad (5)$$

其中 $\text{Agg}(\cdot)$ 表示从动态邻域中聚合消息的函数。为了进一步解耦两种模式, 我们分别设计了不同的聚合函数 $\text{Agg}_I(\cdot)$ 与 $\text{Agg}_V(\cdot)$ 。两种模式的聚合结果相加作为表征传入下一层 $\mathbf{h}_u^t \leftarrow \mathbf{z}_I^t(u) + \mathbf{z}_V^t(u)$ 。

解耦动态图注意力网络的整体模型架构由多个上述时空图注意力层堆叠组成。类似于经典图神经网络的消息传递机制, 该设计保证了每个节点可以间接访问高阶动态邻域。在第 l 层中, $\mathbf{z}_I^t(u)$ 与 $\mathbf{z}_V^t(u)$ 可视为节点 u 在 l 阶邻域中对应的不变与变化模式表征。在实际应用中, 注意力机制可扩展为多头注意力机制 [52], 以提升训练稳定性并建模多维度的图演化信息 [14]。

3.4 时空干预机制

实现公式 (2) 中对变化模式分布的干预最直接的方法是显式生成并替换变化模式, 但该方案在实际中难以实现, 原因如下: 首先, 对每个节点的邻域结构与特征逐一进行干预, 其计算复杂度极高; 其次, 生成时间变化的结构与特征本身就是一个难解的问题。为克服上述问题, 我们提出利用第 3.3 节中得到的模式表征向量 $\mathbf{z}^t$ 来近似表示 $\mathbf{P}^t$ 。由于 $\mathbf{z}_I^t(u)$ 与 $\mathbf{z}_V^t(u)$ 分别代表了节点 u 在时刻 t 的不变与变化模式表征, 我们通过替换变化模式的表征向量来近似干预过程, 而不需显式修改原始结构与特征。具体地, 我们收集所有节点在所有时间的变化模式表征, 从中采样一个用于替换其他节点的变化模式。例如, 我们可使用节点 v 在时间 t_2 的变化模式替换节点 u 在时间 t_1 的变化模式:

$$\mathbf{z}_I^{t_1}(u), \mathbf{z}_V^{t_1}(u) \leftarrow \mathbf{z}_I^{t_1}(u), \mathbf{z}_V^{t_2}(v). \quad (6)$$

由于不变模式保持固定, 标签也不应发生变化。得益于可解耦的时空图注意力机制, 我们在不同邻域与时间点获得了丰富的变化模式, 可作为天然的干预样本, 有效避免了复杂的多环境模拟问题。通过多次执行公式 (6), 我们可以获得多个干预后数据分布, 供后续训练优化使用。

3.5 独立性促进解耦

不变性约束。基于前述多个干预数据分布 (由不同变化模式构成), 我们接下来优化模型以聚焦于不变模式进行预测。此处我们引入不变性损失以具体实现公式 (2) 中的不变性约束项。令 $\mathbf{z}_I$ 和 $\mathbf{z}_V$ 分别表示不变与变化模式表征, 任务损失定义为仅使用不变模式进行预测:

$$\mathcal{L} = \ell(f(\mathbf{z}_I), \mathbf{y}), \quad (7)$$

其中 $f(\cdot)$ 为预测函数。该损失项鼓励模型依赖不变模式进行预测。接下来我们计算混合损失项:

$$\mathcal{L}_m = \ell(g(\mathbf{z}_V, \mathbf{z}_I), \mathbf{y}), \quad (8)$$

表 1 关键符号总结

Table 1 Summary of the key notations.

符号	说明
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Graph, node set and edge set
$\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$	Graph slice at time t
$\mathcal{G}^{1:t}, Y^t$	Graph trajectory, label
$\mathcal{G}_v^{1:t}, y^t$	Ego-graph trajectory, node label
$f(\cdot), g(\cdot)$	Prediction functions
$\text{do}(\cdot)$	do-calculus
ℓ	Loss function
$\phi(\cdot)$	Function finding invariant patterns
d	Dimensionality
$\mathcal{N}^t(u)$	Neighborhood of node u at time t
$\mathbf{m}_I, \mathbf{m}_V, \mathbf{m}_f$	Structure and feature mask
$\mathbf{z}_I^t(u), \mathbf{z}_V^t(u)$	Pattern summarizations
$\text{Agg}_I(\cdot), \text{Agg}_V(\cdot)$	Aggregation functions

算法 1 ID-DyGAT 的算法流程

输入: 训练轮数 L , 干预样本数量 S

主迭代:

```

1: for  $l = 1, \dots, L$  do
2:   计算得到节点在每个时间步的变化与不变模式的表征  $\mathbf{z}_V^t, \mathbf{z}_I^t$  (公式 (5))
3:   计算任务损失  $\mathcal{L}$  和混合损失  $\mathcal{L}_m$  (公式 (7) 和 (8))
4:   从  $\mathbf{z}_V^t$  中采样  $S$  个变化模式, 构造干预集合  $\mathcal{S}$ 
5:   for 每个干预样本  $s \in \mathcal{S}$  do
6:     替换节点的变化模式表征为  $s$  (公式 (6))
7:     重新计算混合损失  $\mathcal{L}_m$ 
8:   end for
9:   计算不变性约束正则化项  $\mathcal{L}_{do}$  (公式 (9))
10:  计算独立性约束正则化项  $\mathcal{L}_{ind}$  (公式 (13))
11:  根据整体训练目标更新模型参数 (公式 (14))
12: end for

```

输出: 参数优化后的模型

其中另一个预测器 $g(\cdot)$ 使用不变模式与变化模式共同进行预测。混合损失衡量了在暴露于变化模式时模型的预测能力。最终, 不变性约束正则化项为:

$$\mathcal{L}_{do} = \text{Var}_{s_i \in \mathcal{S}}(\mathcal{L}_m | \text{do}(\mathbf{P}_V^t = s_i)), \quad (9)$$

其中 ‘do’ 表示第 3.4 节中的干预机制。该损失项衡量模型在多个干预分布下的预测稳定性。

独立性约束。 增强不变模式 $\mathbf{P}_I^t$ 和变化模式 $\mathbf{P}_V^t$ 的独立性可以显式地鼓励动态图神经网络解耦不变模式和变化模式, 并仅通过不变模式进行预测从而实现分布外泛化。由于直方图方法在高维空间中的计算不可行性 [53], 我们采用希尔伯特-施密特 (Hilbert-Schmidt) 独立性准则 (HSIC) [54] 来促进不变模式 $\mathbf{P}_I^t$ 和变化模式 $\mathbf{P}_V^t$ 的图表征之间的独立性。设 κ 是定义在 $\mathbf{P}^t$ 空间上的可测正定核函数, 对应的再生核希尔伯特空间 (RKHS) 为 $\mathcal{H}$, 映射函数 $\phi(\cdot)$ 将 $\mathbf{P}^t$ 映射到 $\mathcal{H}$ 。

对于不变模式和变化模式, 随机变量 $\mathbf{P}_I^t$ 与 $\mathbf{P}_V^t$ 来自联合分布 $p(\mathbf{P}_I^t, \mathbf{P}_V^t)$, 其在对应再生核希尔伯特空间上的互协方差算子定义如下:

$$\mathcal{C}_{\mathbf{P}_I^t, \mathbf{P}_V^t} = \mathbb{E}_{p(\mathbf{P}_I^t, \mathbf{P}_V^t)} [(\phi_{k_I}(\mathbf{P}_I^t) - \mu_{\mathbf{P}_I^t})^\top \cdot (\phi_{k_V}(\mathbf{P}_V^t) - \mu_{\mathbf{P}_V^t})], \quad (10)$$

其中 $\mu_{\mathbf{P}_I^t} = \mathbb{E}_{p(\mathbf{P}_I^t)} [\phi_{k_I}(\mathbf{P}_I^t)]$, $\mu_{\mathbf{P}_V^t} = \mathbb{E}_{p(\mathbf{P}_V^t)} [\phi_{k_V}(\mathbf{P}_V^t)]$ 。HSIC 即该互协方差算子的希尔伯特-施密特范数定义为:

$$\text{HSIC}(\mathbf{P}_I^t, \mathbf{P}_V^t) := \|\mathcal{C}_{\mathbf{P}_I^t, \mathbf{P}_V^t}\|_{\text{HS}}^2, \quad (11)$$

其中 $\|\mathbf{M}\|_{\text{HS}}^2 = \sum_{i,j} \mathbf{M}_{i,j}^2$ 。我们根据如下命题判断变量间的独立性:

命题 1 若 $\mathbb{E}[\kappa_{\mathbf{P}_I^t}(\mathbf{P}_I^t, \mathbf{P}_I^t)] < \infty$, $\mathbb{E}[\kappa_{\mathbf{P}_V^t}(\mathbf{P}_V^t, \mathbf{P}_V^t)] < \infty$, 且 $\kappa_{\mathbf{P}_I^t} \kappa_{\mathbf{P}_V^t}$ 是特征核 (characteristic kernel), 则有:

$$\text{HSIC}(\mathbf{P}_I^t, \mathbf{P}_V^t) = 0 \Leftrightarrow \mathbf{P}_I^t \perp\!\!\!\perp \mathbf{P}_V^t. \quad (12)$$

在实际实现中, 我们采用工作 [55] 提出的 HSIC 无偏估计量:

$$\text{HSIC}(\mathbf{P}_I^t, \mathbf{P}_V^t) = \frac{1}{m(m-3)} \left[\text{tr}(\tilde{U}\tilde{V}^T) + \frac{1^T \tilde{U} \mathbf{1} \mathbf{1}^T \tilde{V}^T \mathbf{1}}{(m-1)(m-2)} - \frac{2}{m-2} \mathbf{1}^T \tilde{U} \tilde{V}^T \mathbf{1} \right], \quad (13)$$

其中 $\tilde{U}$ 与 $\tilde{V}$ 是 $\kappa_{\mathbf{P}_I^t}$ 与 $\kappa_{\mathbf{P}_V^t}$ 对应的格拉姆 (Gram) 矩阵, 其对角元素被置为零。我们使用径向基函数 (RBF) 核。特别地, 采用该准则来衡量图表征间的独立性具有以下三大优势 [54, 56]: 首先动

态图模式通过 RKHS 映射进行比较, 可捕提高阶联合矩及非线性依赖; 其次 HSIC 无需显式估计变量间的联合分布, 计算非常高效; 最后我们使用的 HSIC 估计量是无偏的 [55], 优于其他经验估计方法 [54]。最终公式 (2) 中的独立性约束实现为 $\mathcal{L}_{ind} = \text{HSIC}(\mathbf{P}_I^t, \mathbf{P}_V^t)$ 。

3.6 整体训练目标

所提方法的训练目标为:

$$\min_{\theta} \mathcal{L} + \beta \mathcal{L}_{do} + \lambda \mathcal{L}_{ind}, \quad (14)$$

其中任务损失 $\mathcal{L}$ 用于引导模型学习不变模式, 不变性约束正则项 $\mathcal{L}_{do}$ 和独立性约束正则项 $\mathcal{L}_{ind}$ 则用于帮助模型解耦不变与变化模式, 超参数 β 和 λ 控制正则项的强度。值得注意的是, 在推理阶段模型仅使用不变模式进行预测, 从而实现分布外泛化的预测性能。整体训练流程如算法 1 所示。

复杂度分析。所提方法的计算复杂度分析如下: 设图中节点总数为 $|V|$, 边总数为 $|E|$, 隐藏表示维度为 d 。时空聚合模块的时间复杂度为 $O(|E|d + |V|d^2)$ 。可解耦模块仅引入一个常数乘因子 2, 对聚合过程的时间复杂度无影响。记 $|E_p|$ 为待预测的边数量, $|S|$ 为干预样本集合大小。时空干预模块在训练时的时间复杂度为 $O(|E_p||S|d)$, 独立性促进解耦模块的时间复杂度为 $O(d^2)$ 。这两个模块仅在训练阶段使用, 因此不会给推理过程带来额外开销。因此, 所提方法总体时间复杂度为 $O(|E|d + |V|d^2 + |E_p||S|d + d^2)$ 。其中 $|S|$ 为超参数且通常被设置为较小的常数。综上, 所提方法在节点数与边数上的时间复杂度是线性的, 与现有动态图神经网络为同一量级。

4 实验设计

本节中, 我们通过定量实验和定性实验证明了所提方法的有效性。实验数据集如表 2 所示。

4.1 基线方法

我们选取了若干具有代表性的图神经网络模型以及分布外泛化方法作为对比基线。第一类方法为静态图神经网络。**GAE** [57] 是一种经典的静态图神经网络模型, 通过堆叠多层图卷积操作来捕捉图结构与节点属性之间的关系, 从而无监督地学习图表征; **VGAE** [57] 在 GAE 的基础上将变分自编码器的思想引入图结构建模, 通过在隐空间中引入随机变量来建模节点表征的不确定性, 以获取更具鲁棒性和泛化能力的图表征。第二类方法为动态图神经网络。**GCRN** [13] 首先利用图卷积从静态图快照中提取节点表征, 然后通过门控循环单元 [12] 来捕捉图随时间演化的动态过程。这种“空间建模 + 时间序列建模”的框架能够同时学习图的关联结构和时间依赖关系。**EGCN** [16] 则采用另一种思路, 它不是直接学习节点的时序表征, 而是通过循环神经网络动态更新图卷积的参数, 使得模型能够在不同时间快照下自适应演化。这一机制特别适合节点集合频繁变化的动态图场景。**DySAT** [14] 在建模过程中引入了双重注意力机制, 在每个时间快照中利用结构注意力自适应地聚合不同邻居的信息, 同时在时间维度上使用自注意力机制捕捉跨时间的依赖关系, 使得在聚合过程中可以为来自不同邻居的信息自适应地分配权重, 从而既能关注局部邻域, 又能灵活建模长期和短期的动态特征。第三类方法为分布外泛化方法。**IRM** [21] 的目标是学习一个在所有训练环境中都能保持一致的预测器, 使其在不同分布下都能最小化经验风险。这种方法强调提取跨环境稳定的因果关系, 从而提升模型的分布外泛化能力。**GroupDRO** [22] 在优化过程中引入了鲁棒性思想, 在最小化经验风险的同时, 对误差较大的训练环境分配更高的权重, 从而实现对最坏情况风险的优化。**VREx** [58] 则通过在目标函数中约束不同训练环境之间的风险差异, 来降低模型对分布偏移的敏感

表 2 数据集统计摘要。“演化特征”表示节点特征是否随时间变化;“未见节点”表示测试集中是否包含在过去从未出现或部分未出现的节点。

Table 2 Summarization of dataset statistics. Evolving features denote whether the node features vary through time. Unseen nodes denote whether the test nodes are partially or fully unseen in the past.

Dataset	COLLAB	Yelp	Synthetic	OGBN-Arxiv	Aminer	Reddit
# Timestamps	16	24	16	20	17	10
# Nodes	23,035	13,095	23,035	168,195	43,141	8,291
# Links	151,790	65,375	151,790	3,127,274	851,527	264,050
Granularity	Year	Month	Year	Year	Year	Day
Feature dimension	32	32	64	128	128	20
Evolving features	No	No	Yes	No	No	No
Unseen nodes	Partial	Partial	Partial	All	All	All
Task level	Link	Link	Link	Node	Node	Node

性。这种策略避免了模型在某些训练环境上过拟合的情况,从而提升整体的泛化表现。**DIDA** [9] 针对图结构数据进行了专门设计,通过捕捉在不同环境下保持一致的动态图表征,提升模型的预测稳定性。上述具有代表性的分布外泛化方法旨在提升模型在面对分布偏移时的泛化能力,但这类方法大多依赖显式的环境标签以计算损失函数。因为所有方法均无法获得真实的环境信息,为了公平对比,我们将样本随机划分为多个训练域。考虑到这些方法是通用的分布外泛化策略,我们采用在训练集上表现最优的动态图神经网络作为其主干网络 (backbone network)。

4.2 真实数据集上的链路预测实验

我们使用了两个真实世界的动态图数据集,分别为 **COLLAB** 和 **Yelp**, 并采用具有挑战性的归纳式未来链路预测任务,要求模型利用历史图结构预测下一时间步的链路。每个数据集都可以根据其领域信息划分为多个子动态图。为简明起见,我们使用 ‘w/ DS’ 表示含有分布偏移的测试数据, ‘w/o DS’ 表示不含分布偏移的测试数据。为了评估模型在时空分布偏移下的表现,我们从数据集中选择一个领域作为 ‘w/ DS’, 其余数据按时间顺序划分为训练集、验证集和测试集作为 ‘w/o DS’。需要注意的是, ‘w/o DS’ 是合并后的动态图,不包含领域信息;而 ‘w/ DS’ 在训练过程中是不可见的,这在现实应用中更为贴近真实情况且更具挑战性。以下是对两个真实世界数据集的简要介绍:**COLLAB** [59] 是一个学术合作数据集,包含 1990 年至 2006 年间发表的论文。图中的节点和边分别表示作者和合著关系。根据合作论文的研究领域,每条边附带一个领域标签,包括“数据挖掘”、“数据库”、“医学信息学”、“理论”以及“可视化”。时间粒度为年,总计包含 16 个时间片。我们选取“数据挖掘”作为 ‘w/ DS’, 其余作为 ‘w/o DS’。我们使用 word2vec [60] 从论文摘要中提取 32 维特征,并取平均获得作者特征。我们使用 10、1、5 个时间片分别作为训练集、验证集和测试集。该数据集共包含 23,035 个节点和 151,790 条边。**Yelp** [14] 是一个商业评论数据集,包含顾客对商家的评论行为。图中的节点和边分别表示顾客/商家与评论行为。我们选取了 2019 年 1 月至 2020 年 12 月之间 5 个商业类别的交互数据,包括“比萨”、“新派美式食品”、“咖啡与茶”、“寿司店”和“快餐”。时间粒度为月,总计包含 24 个时间片。我们将“比萨”类别作为 ‘w/ DS’, 其余作为 ‘w/o DS’。我们使用 word2vec [60] 从评论文本中提取 32 维特征,并取平均值获得用户和商家的特征表示。我们筛选交互次数大于 10 次的用户与商家,并使用 15、1、8 个时间片分别作为训练集、验证集和测试集。该数据集共包含 13,095 个节点和 65,375 条边。

实验结果。根据表 3 中真实世界链路预测数据集的实验结果,可以发现基线方法在分布偏移下表现显著下降。尽管动态图神经网络基线方法在无分布偏移的测试集 (‘w/o DS’) 上表现良好,但其在存在分布偏移的测试集 (‘w/ DS’) 上的性能大幅下降。特别地, **DySAT** 在 ‘w/o DS’ 中表现较

表 3 不同方法在真实世界链路预测数据集上的结果 (AUC%)。最优结果以加粗标出。‘w/o DS’ 表示无分布偏移的测试数据, ‘w/ DS’ 表示有分布偏移的测试数据。

Table 3 Results (AUC%) of different methods on real-world link prediction datasets. The best results are in bold. ‘w/o DS’ and ‘w/ DS’ denote test data with and without distribution shift.

Model \ Dataset	COLLAB		Yelp	
Test Data	w/o DS	w/ DS	w/o DS	w/ DS
GAE	77.15±0.50	74.04±0.75	70.67±1.11	64.45±5.02
VGAE	86.47±0.04	74.95±1.25	76.54±0.50	65.33±1.43
GCRN	82.78±0.54	69.72±0.45	68.59±1.05	54.68±7.59
EGCN	86.62±0.95	76.15±0.91	78.21±0.03	53.82±2.06
DySAT	88.77±0.23	76.59±0.20	78.87±0.57	66.09±1.42
IRM	87.96±0.90	75.42±0.87	66.49±10.78	56.02±16.08
VREx	88.31±0.32	76.24±0.77	79.04±0.16	66.41±1.87
GroupDRO	88.76±0.12	76.33±0.29	79.38±0.42	66.97±0.61
DIDA	91.97±0.05	81.87±0.40	78.22±0.40	75.92±0.90
ID-DyGAT	92.19±0.31	82.43±0.51	79.52±0.46	76.95±1.65

优, 而在 ‘w/ DS’ 中的性能下降幅度达到了近 12%。在 Yelp 数据集上, GCRN 与 EGCN 的表现存在分布偏移时甚至不如静态图神经网络 GAE 和 VGAE。这一现象说明, 现有动态图神经网络过拟合于对标签有虚假关联的变化模式, 因此在面对分布偏移时难以泛化。此外, 分布外泛化基线方法并非专门针对动态图中的时空分布偏移设计, 因此在 Yelp 上仅带来有限性能提升。特别地, 这些方法依赖于真实的环境标签以实现分布外泛化, 而在真实动态图中环境标签是不可获得的。其较差的表现进一步表明, 在缺乏环境标签的情形下, 这些方法难以取得理想效果, 也验证了缺乏环境标签是动态图处理分布偏移时的一大挑战。我们的方法在处理分布偏移方面优于现有基线, 尤其是在分布偏移较强的场景中表现更为突出。所提方法在所有数据集的 ‘w/ DS’ 条件下均显著优于所有对比方法。需要指出的是, Yelp 数据集中由于 COVID-19 的爆发导致中期消费行为发生剧烈变化, 表现出更强的时间分布偏移; 而在此情形下, 所提方法在 ‘w/ DS’ 上相比基线方法 GroupDRO 仍有 9% 的提升。相比之下, Yelp 中各类别 (均为餐饮类) 之间的领域信息较为接近, 而 COLLAB 数据集中不同领域间差异更大, 体现出更强的空间分布偏移; 而在此条件下, 所提方法在 ‘w/ DS’ 上相比经典基线方法 DySAT 提升了 5%。另外尽管 DIDA 在分布偏移下取得的较为理想的泛化性能, 但由于其缺乏显式的解耦机制, 可能导致不变模式和变化模式解耦不彻底, 从而影响预测效果。最终所提方法取得了一致最优的分布外泛化性能。

4.3 真实数据集上的节点分类实验

我们使用了三个真实世界的动态图数据集进行节点分类实验, 包括 OGBN-Arxiv [61]、Aminer [62] 和 Reddit [63]。前两个数据集均为引文网络, 其中节点表示论文, 时间戳为 t 的从 u 到 v 的边表示论文 u 在年份 t 发表, 并引用了论文 v 。动态图上的节点分类任务具有挑战性, 因为测试节点来自未来, 例如新论文在未来被发表, 模型需利用时空信息进行分类预测。我们遵循 [24] 的设置, 采用归纳学习 (inductive learning) 策略, 即测试节点在训练阶段完全不可见, 这在真实动态图场景中更具实用性和挑战性。以下是三个真实数据集的简要介绍。OGBN-Arxiv [61] 是一个计算机科学领域的 arXiv 论文引文网络, 由 MAG [64] 索引。每篇论文通过对标题和摘要中的词向量取平均获得一个 128 维的特征向量, 词向量由 skip-gram 模型 [65] 在 MAG 语料库上训练得到。任务是预测该论文所属的 40 个主题领域, 如 cs.AI、cs.LG 和 cs.OS。我们在 2001 年至 2011 年间发表的论文上训练, 2012 年至 2014 年间的论文用于验证, 2015 年之后的论文用于测试。由于科学论文发表数

表 4 不同方法在真实世界节点分类数据集上的准确率 (ACC%)。最优结果以加粗标出。
Table 4 Accuracy (ACC%) on real-world node classification datasets. Best results are in bold.

Model \ Dataset	OGBN-Arxiv				Aminer			Reddit		
	Split	2015-2016	2017-2018	2019-2020	2015	2016	2017	8	9	10
GCRN		46.77±2.03	45.89±3.41	46.61±3.29	47.96±1.12	51.33±0.62	42.93±0.71	34.38±1.35	31.08±1.58	33.98±2.50
EGCN		48.70±2.12	47.31±3.45	46.93±5.17	44.14±1.12	46.28±1.84	37.71±1.84	29.24±0.59	31.45±0.72	30.21±0.68
DySAT		48.83±1.07	47.24±1.24	46.87±1.37	48.41±0.81	49.76±0.96	42.39±0.62	36.10±0.14	33.82±0.30	33.62±1.52
IRM		49.57±1.02	48.28±1.51	46.76±3.52	48.44±0.13	50.18±0.73	42.40±0.27	30.24±1.26	32.21±0.78	29.76±0.83
VREx		48.21±2.44	46.09±4.13	46.60±5.02	48.70±0.73	49.24±0.27	42.59±0.37	30.72±1.94	31.41±0.77	30.33±0.88
GroupDRO		49.51±2.32	47.44±4.06	47.10±4.39	48.73±0.61	49.74±0.26	42.80±0.36	31.65±0.35	30.68±1.33	31.55±0.32
DIDA		51.46±1.25	49.98±2.04	50.91±2.88	50.34±0.81	51.43±0.27	44.69±0.06	36.14±0.32	34.74±0.14	34.67±0.46
ID-DyGAT		51.56±1.05	50.49±1.14	51.93±1.96	51.19±0.53	52.41±0.72	45.14±0.18	42.13±0.24	40.52±1.32	40.28±0.53

量每 12 年翻一番, 该数据集中自然存在显著的时空分布偏移。该数据集总共包含 168,195 个节点和 3,127,274 条边。**Aminer** [62] 是从 DBLP、ACM、MAG 及其他来源中提取的引文网络。我们使用 word2vec [60] 从论文摘要中提取 128 维特征, 并取平均以获得每篇论文的特征表示。我们筛选了排名前 20 的会议作为分类目标, 任务为预测论文的发表会议。与 OGBN-Arxiv 数据集类似, 我们在 2001 年至 2011 年间的论文上训练, 2012 年至 2014 年用于验证, 2015 年之后用于测试。由于测试节点在训练中完全不可见, 模型需挖掘不变的时空模式, 以实现分布偏移下的稳定预测。该数据集共包含 43,141 个节点和 851,527 条边。**Reddit** [63] 是一个从知名社交媒体平台 Reddit 上提取的大型数据集, 包含基于文本的帖子、评论、用户信息以及社区详情。图的构建方式是将每个节点视为一个帖子, 当两个帖子共享关键词时, 在它们之间建立一条边。节点特征通过 word2vec 方法生成 [60]。该数据集共包含 10 个时间步。我们在时间步 1 至 6 的数据上进行训练, 在时间步 7 上进行验证, 并在时间步 8 至 10 上进行测试。整个数据集共包含 8,291 个节点和 264,050 条边。

实验结果。根据表 4 中真实节点分类数据集的实验结果, 我们发现大多数基线方法随着时间推移性能明显下降。在 OGBN-Arxiv 数据集上, EGCN 的准确率从 2015 年的 48.70% 逐渐下降至 2020 年的 46.93%。这一现象可能源于动态图中时空分布的逐渐偏移, 例如学术论文的发表数量显著增长, 研究主题与引文模式也可能与过去发生了变化。此外, 尽管一般的分布外泛化方法在某些场景下优于动态图神经网络基线, 但由于这些方法并非专门为处理动态图中的复杂时空分布偏移而设计, 其改进仍远未令人满意。所提方法显著缓解了随着时间推移的性能下降问题, 在分布外的测试集上取得了一致且最优的预测性能。这验证了我们的方法能够有效捕捉数据中的不变与变化的时空模式, 并利用不变模式在分布偏移下实现更稳定的预测。我们的方法在多数情况下具有更小的标准差, 这可能得益于不变性约束和独立性约束正则项降低虚假相关性的影响。

4.4 模拟数据集上的实验

为了进一步评估模型在时空分布偏移下的泛化能力, 我们参考 [24], 在合并所有领域后的 COLLAB 数据集上引入人为设计的分布偏移。设原始特征和结构分别为 $\mathbf{X}_1^t \in \mathbb{R}^{N \times d}$ 和 $\mathbf{A}^t \in \{0, 1\}^{N \times N}$ 。对于每一个时间点 t , 我们从 $\mathbf{A}^{t+1}$ 中均匀采样 $p(t)|\mathcal{E}^{t+1}|$ 条正样本边和 $(1-p(t))|\mathcal{E}^{t+1}|$ 条负样本边。接着, 我们将其构造为结构保持性的变化特征 $\mathbf{X}_2^t \in \mathbb{R}^{N \times d}$ 。最终的输入节点特征为 $\mathbf{X}^t = [\mathbf{X}_1^t, \mathbf{X}_2^t]$, 用于训练与推理。采样概率定义为 $p(t) = \text{clip}(\bar{p} + \sigma \cos(t), 0, 1)$ (clip 为截取函数), 用以表示变化强度, 其中更高的 $p(t)$ 意味着 $\mathbf{X}_2^t$ 与未来链路 $\mathbf{A}^{t+1}$ 的相关性更强。我们在测试阶段设置 $\bar{p}_{test} = 0.1, \sigma_{test} = 0, \sigma_{train} = 0.05$, 并在训练阶段将 $\bar{p}_{train}$ 从 0.4 到 0.8 进行变化以做评估。由于 $\mathbf{X}_2^t$ 与标签 $\mathbf{A}^{t+1}$ 的相关性随时间与邻域而变化, 因此其中的模式为分布偏移下的变化模式。由于静态图

表 5 不同方法在合成数据集上的曲线下面积 (AUC%)。最优结果以加粗标出。更大的 $\bar{p}$ 表示更高程度分布偏移。
Table 5 Area under the curve (AUC%) of different methods on the synthetic dataset. The best results are in bold. Larger $\bar{p}$ denotes higher distribution shift level.

Model \ $\bar{p}$	0.4		0.6		0.8	
Split	Train	Test	Train	Test	Train	Test
GCRN	69.60 $\pm$ 1.14	72.57 $\pm$ 0.72	74.71 $\pm$ 0.17	72.29 $\pm$ 0.47	75.69 $\pm$ 0.07	67.26 $\pm$ 0.22
EGCN	78.82 $\pm$ 1.40	69.00 $\pm$ 0.53	79.47 $\pm$ 1.68	62.70 $\pm$ 1.14	81.07 $\pm$ 4.10	60.13 $\pm$ 0.89
DySAT	84.71 $\pm$ 0.80	70.24 $\pm$ 1.26	89.77 $\pm$ 0.32	64.01 $\pm$ 0.19	94.02 $\pm$ 1.29	62.19 $\pm$ 0.39
IRM	85.20 $\pm$ 0.07	69.40 $\pm$ 0.09	89.48 $\pm$ 0.22	63.97 $\pm$ 0.37	95.02 $\pm$ 0.09	62.66 $\pm$ 0.33
VREx	84.77 $\pm$ 0.84	70.44 $\pm$ 1.08	89.81 $\pm$ 0.21	63.99 $\pm$ 0.21	94.06 $\pm$ 1.30	62.21 $\pm$ 0.40
GroupDRO	84.78 $\pm$ 0.85	70.30 $\pm$ 1.23	89.90 $\pm$ 0.11	64.05 $\pm$ 0.21	94.08 $\pm$ 1.33	62.13 $\pm$ 0.35
DIDA	87.92 $\pm$ 0.92	85.20 $\pm$ 0.84	91.22 $\pm$ 0.59	82.89 $\pm$ 0.23	92.72 $\pm$ 2.16	72.59 $\pm$ 3.31
ID-DyGAT	88.61$\pm$0.53	85.31$\pm$0.15	92.34$\pm$0.98	83.21$\pm$0.38	95.06$\pm$0.21	74.89$\pm$1.42

神经网络无法处理时变特征, 我们省略其在该实验中的结果。对于变化特征 $\mathbf{X}_2^t$ 的构造过程, 我们在 COLLAB 数据集中使用与 $\mathbf{X}_1^t$ 相同的初始特征和结构 $\mathbf{A}^t$, 并引入与监督信号具有可控相关性的变化特征 $\mathbf{X}_2^t$ 。 $\mathbf{X}_2^t$ 是通过最小化重建损失 $\ell(\mathbf{X}_2 \mathbf{X}_2^T, \tilde{\mathbf{A}}^{t+1})$ 来学习得到的, 其中 $\tilde{\mathbf{A}}^{t+1}$ 表示采样后的边集, ℓ 表示交叉熵损失函数。我们使用 Adam 优化器进行训练, 学习率为 1e-1, 权重衰减为 1e-5, 早停容忍度为 50。通过这种方式, 我们实证发现仅使用 $\mathbf{X}_2^t$ 来预测 $\tilde{\mathbf{A}}^{t+1}$, 其内积预测器可达到超过 99% 的 AUC, 从而表明所生成的特征与采样边具有很强的相关性。通过控制 p , 我们可以实现训练阶段与测试阶段之间特征 $\mathbf{X}^t$ 与标签 $\mathbf{A}^{t+1}$ 的相关性差异, 从而构造分布偏移场景。

实验结果。根据表 5 中在合成数据集上的实验结果, 我们发现所提方法在处理分布偏移方面优于所有基线方法。尽管各基线方法在训练阶段表现较好, 但其在测试阶段的性能大幅下降, 说明现有的动态图神经网络方法难以有效应对分布偏移。**ID-DyGAT** 在不同的偏移强度下均显著优于动态图神经网络基线, 性能提升幅度明显。对于一般的分布外泛化方法, 尽管在部分情况下能降低方差, 但其整体性能提升有限。相比之下, **ID-DyGAT** 是专门为动态图设计的, 能够有效利用不变模式, 在不同分布偏移强度下持续缓解由变化模式带来的负面影响。随着偏移程度的增加, 几乎所有基线方法在训练阶段的性能提升而在测试阶段显著下降。这一现象表明, 随着变化特征与标签之间的关系增强, 现有动态图神经网络方法在训练中愈加依赖于变化模式, 从而导致测试阶段性能退化。相比之下, **ID-DyGAT** 在训练表现的提升和测试表现的下降幅度明显优于基线方法, 表明其能有效挖掘不变模式, 做出可泛化的预测。相比于基线方法 DIDA, 所提方法也能取得一致更优的泛化性能, 这主要得益于显式地对不变模式和变化模型进行独立性解耦。

4.5 进一步实验

消融实验。我们进行了消融实验以验证所提出的解耦图注意力网络、时空干预机制和显式独立性促进解耦模块的有效性。首先我们移除了第 3.5 节中提出的独立性约束。从图 2 中可以看出, 在缺少不变模式和变化模式显式独立性约束的情况下, 模型 (w/o HSIC) 的性能相比所提方法 (如 Ours 所示, 下同) 出现显著下降, 验证了我们提出的显式独立性促进解耦模块对不变模式识别建模的有效性。然后我们移除了时空干预机制。在缺乏时空干预机制时, 模型 (w/o Intervention) 泛化效果不佳, 特别是对于三个节点分类数据集, 性能下降显著。这验证了我们提出的时空干预机制有助于模型聚焦于不变模式以进行更稳定的预测。我们进一步移除了提出的解耦注意力机制 (见 w/o Disen.)。解耦注意力机制是模型设计中的关键模块, 如果移除该设计几乎所有数据集上的动态图神经网络的泛化性能都出现大幅下降。在没有解耦模块的情况下, 模型无法有效获取不变模式与变化模式, 从

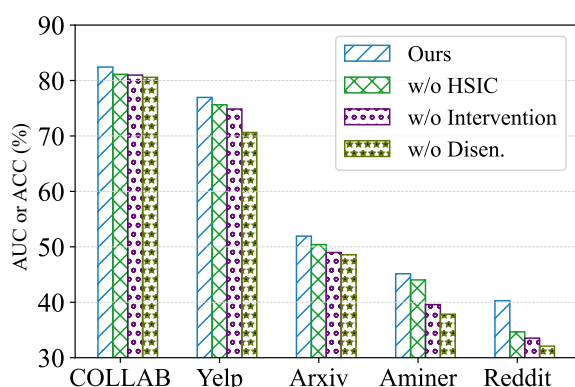


图 2 消融实验分析

Figure 2 Ablation studies on the HSIC regularizer, intervention mechanism and disentangled attention

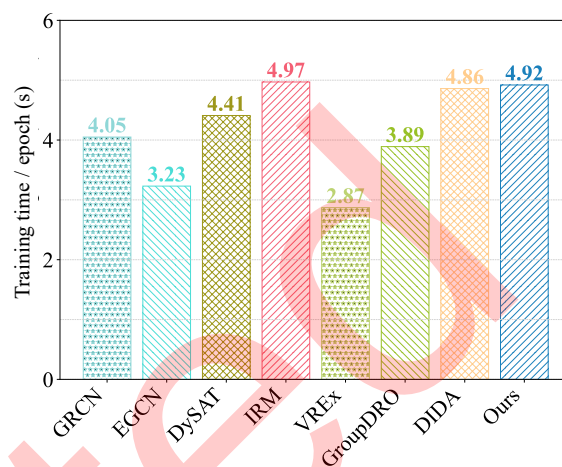


图 3 所提方法和基线方法每个迭代轮次训练时间比较

Figure 3 The comparisons of empirical time cost

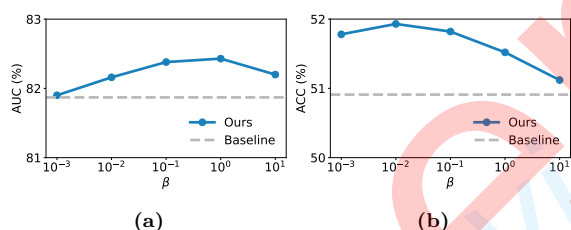
图 4 超参数 β 的敏感性分析。(a) COLLAB 的结果; (b) Arxiv 的结果。

Figure 4 Sensitivity analysis of hyperparameter β . (a) Results of COLLAB; (b) Results of Arxiv.

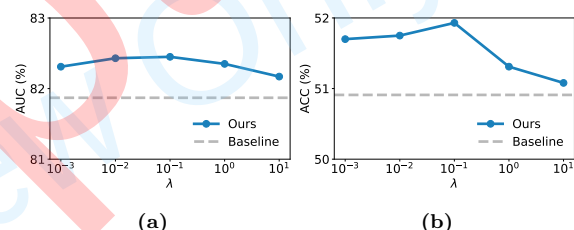
图 5 超参数 λ 的敏感性分析。(a) COLLAB 的结果; (b) Arxiv 的结果。

Figure 5 Sensitivity analysis of hyperparameter λ . (a) Results of COLLAB; (b) Results of Arxiv.

而影响后续的时空干预过程和独立性促进解耦过程, 无法实现理想的分布外泛化预测效果。

模型效率比较。为了进一步验证所提方法在实际中的效率, 我们在 Arxiv 数据集上对所提方法及其他具有竞争力的基线方法在训练时间上的表现进行了比较, 结果如图 3 所示, 我们报告了每个迭代轮次的平均训练时间。可以从图 3 中观察到所提方法不仅在分布外泛化能力上优于现有方法, 同时在训练时间成本上也十分高效, 进一步体现了所提方法的优越性。

超参数敏感性分析。选择合适的正则化项超参数对模型的预测能力至关重要。为了探究我们方法的超参数敏感性, 我们在优化目标 (14) 中固定其他默认超参数, 并在较大范围内调整不变性正则项系数 β 和独立性正则项系数 λ 的取值以比较其对应的模型性能。结果如图 4 和图 5 所示 (蓝色实线为所提方法结果, 灰色虚线为最优基线结果)。可以看出, 当超参数 β 或 λ 取值较小时, 模型难以有效解耦准确且具有泛化能力的不变模式; 当 β 或 λ 过大时, 模型可能缺乏对于不变模式的表征学习能力。所提方法在较宽泛的超参数设置范围内都能获得最优表现, 显示出较强的超参数设定鲁棒性。

节点表征可视化分析。为了进一步验证我们方法的有效性, 我们可视化了所提方法及部分基线方法的表征学习结果。如图 6 所示 (以 Arxiv 数据集为例), 我们以节点的类别标签为颜色区分, 发现所提方法输出的节点表征在类别上更加分离, 不同类别标签之间呈现明显聚类特征, 说明其捕捉到的不变模式相比基线方法更加准确。此外, 我们还计算了轮廓系数 (Silhouette Coefficient) [66], 这是聚类任务中常用的评估指标, 用于进一步衡量模型效果。在节点分类任务中, 所提方法学习出的表征的轮廓系数明显高于其他基线方法 (如 VREx、GroupDRO 和 EGCN), 表明所提方法通过

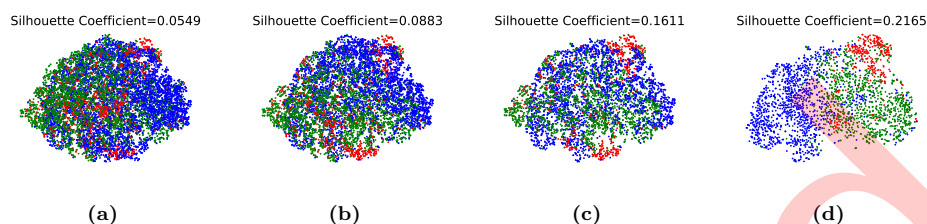


图 6 所提方法和基线方法节点表征在 Arxiv 数据集上的可视化分析。(a) VREx 的结果; (b) GroupDRO 的结果; (c) EGCN 的结果; (d) ID-DyGAT 的结果。

Figure 6 Visualization of node embeddings produced by ID-DyGAT and baselines on Arxiv. (a) Results of VREx; (b) Results of GroupDRO; (c) Results of EGCN; (d) Results of ID-DyGAT.

显式独立性解耦来学习不变模式对于实现分布外泛化预测具有重要意义。

5 未来研究方向

本文的研究方向上仍存在若干值得继续探索的方向。首先, 现有动态图神经网络分布外泛化理论仍不完备, 需要在可学习的动态图分布外泛化问题上推导出更多基础理论, 并进一步发展具有理论保证的动态图分布外泛化方法。还应在方法设计中更详细地建模分布偏移 (如协变量偏移、概念偏移、标签偏移等), 研究建立在特定分布偏移的具体假设上的分布外泛化的动态图神经网络理论。其次, 大多数分布外泛化算法依赖于数据集提供的多个训练环境。然而为现实世界中的动态图数据收集准确的环境标签是非常昂贵的, 限制了现有算法的直接应用。因此开发单环境图分布外泛化方法或在训练过程中准确推断环境划分标签都是值得研究的话题。另外, 大语言模型近年来在自然语言处理和多模态学习中展现出强大的学习能力和跨任务泛化能力, 为动态图研究提供了新的方向。动态图大模型有望成为推动图机器学习发展的全新范式。未来可以考虑构建多领域的大规模动态图数据集, 学习具有普适性的动态图表征。由于动态图大模型在多样环境中训练, 其学习到的表征具备更强的泛化性, 有潜力缓解由时空分布偏移带来的性能退化问题, 实现更稳定的跨环境预测。最后, 一些经典的动态图神经网络已经在各种现实应用中被广泛部署并取得很大成功。因为分布外的动态图数据广泛存在于我们的日常生活中, 所以在具有分布偏移的实际应用中部署分布外泛化的动态图神经网络是一个很有前景的方向。很多领域包括社交网络、交通预测、以及风险敏感的金融、医疗等, 都亟需部署具有分布外泛化能力的动态图神经网络, 以获得更有效、可信和令人满意的效果。而如何引入适当的领域知识是将分布外泛化的动态图神经网络应用于这些领域的一个主要挑战。一个可能的思路是将领域知识作为额外的先验来对待, 从而指导动态图神经网络模型和优化策略的设计。

6 总结

本文提出了一种基于显式独立性解耦的分布外泛化动态图注意力网络, 用于应对动态图中的时空分布偏移下的泛化问题。我们首先设计了一种解耦的动态图注意力网络, 以捕捉动态图中分布偏移下对于标签不变和变化的时空模式。其次我们构建了时空干预机制以生成多个干预分布来模拟不同分布的训练环境。进一步地我们引入不变性正则项, 引导模型在分布偏移下聚焦于不变模式; 同时引入独立性正则项, 以引导不变和变化时空模式的解耦, 忽略变化模式带来的虚假关联。在多个真实世界数据集和模拟数据集上的大量实验结果证明, 所提方法在应对时空分布偏移方面显著优于现有基线方法, 对于未来动态图神经网络在实际场景中的部署具有重要潜在价值。

参考文献

- 1 Nascimento D C, Pimentel B A, Souza R M, et al. Dynamic graph in a symbolic data framework: An account of the causal relation using covid-19 reports and some reflections on the financial world. *Chaos, Solitons & Fractals*, 2021, 153: 111440
- 2 Zhang S, Suzumura T, Zhang L. Dyngraphtrans: Dynamic graph embedding via modified universal transformer networks for financial transaction data. 2021: 184–191
- 3 Peng H, Du B, Liu M, et al. Dynamic graph convolutional network for long-term traffic flow prediction with reinforcement learning. *Information Sciences*, 2021, 578: 401–416
- 4 Cai D, Qian S, Fang Q, et al. User cold-start recommendation via inductive heterogeneous graph neural network. *TOIS*, 2023, 41: 1–27
- 5 Ma T, Huang L, Lu Q, et al. Kr-gcn: Knowledge-aware reasoning with graph convolution network for explainable recommendation. *TOIS*, 2023, 41: 1–27
- 6 Yang T, Hu L, Shi C, et al. Hgat: Heterogeneous graph attention networks for semi-supervised short text classification. *TOIS*, 2021, 39: 1–29
- 7 Skarding J, Gabrys B, Musial K. Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access*, 2021: 79143–79168
- 8 Zhu Y, Lyu F, Hu C, et al. Learnable encoder-decoder architecture for dynamic graph: A survey. *arXiv preprint arXiv:2203.10480*, 2022
- 9 Zhang Z, Wang X, Zhang Z, et al. Dynamic graph neural networks under spatio-temporal distribution shift. In: *NeurIPS*, 2022
- 10 Chen X, Xiong K, Zhang Y, et al. Neural feature-aware recommendation with signed hypergraph convolutional network. *TOIS*, 2020, 39: 1–22
- 11 Zhang G, Li Z, Huang J, et al. efraudcom: An e-commerce fraud detection system via competitive graph neural networks. *TOIS*, 2022, 40: 1–29
- 12 Cho K, van Merriënboer B, Çaglar Gülçehre, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: *EMNLP*, 2014
- 13 Seo Y, Defferrard M, Vandergheynst P, et al. Structured sequence modeling with graph convolutional recurrent networks. 2018: 362–373
- 14 Sankar A, Wu Y, Gou L, et al. Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In: *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020. 519–527
- 15 Xu D, Ruan C, Körpeoglu E, et al. Inductive representation learning on temporal graphs. 2020
- 16 Pareja A, Domeniconi G, Chen J, et al. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In: *AAAI*, 2020, volume 34. 5363–5370
- 17 You J, Wang Y, Pal A, et al. Hierarchical temporal convolutional networks for dynamic recommender systems. In: *The world wide web conference*, 2019. 2236–2246
- 18 Tang H, Wu S, Xu G, et al. Dynamic graph evolution learning for recommendation. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. 1589–1598
- 19 Zhao Z, Zhu X, Xu T, et al. Time-interval aware share recommendation via bi-directional continuous time dynamic graphs. In: *SIGIR*, 2023. 822–831
- 20 Shen Z, Liu J, He Y, et al. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021
- 21 Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization. *arXiv preprint*, 2019
- 22 Sagawa S, Koh P W, Hashimoto T B, et al. Distributionally robust neural networks. In: *ICLR*
- 23 Li H, Wang X, Zhang Z, et al. Out-of-distribution generalization on graphs: A survey. *arXiv preprint*, 2022
- 24 Wu Q, Zhang H, Yan J, et al. Handling distribution shifts on graphs: An invariance perspective. *ICLR*, 2022
- 25 Wu Y X, Wang X, Zhang A, et al. Discovering invariant rationales for graph neural networks. *ICLR*, 2022
- 26 Wang X, Chen H, Tang S, et al. Disentangled representation learning. *IEEE TPAMI*, 2024
- 27 Higgins I, Matthey L, Pal A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework. In: *International conference on learning representations*, 2017
- 28 Kumar A, Sattigeri P, Balakrishnan A. Variational inference of disentangled latent concepts from unlabeled observations. *International conference on learning representations*, 2018
- 29 Chen X, Duan Y, Houthoofd R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NeurIPS*, 2016, 29
- 30 Yang T, Wang Y, Lv Y, et al. Disdiff: Unsupervised disentanglement of diffusion probabilistic models. *NeurIPS*, 2023
- 31 Chen H, Zhang Y, Wu S, et al. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In: *The Eleventh International Conference on Learning Representations*, 2023
- 32 Bao Y, Zhou H, Huang S, et al. Generating sentences from disentangled syntactic and semantic spaces. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019

- 33 Cheng P, Min M R, Shen D, et al. Improving disentangled text representation learning with information-theoretic guidance. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020
- 34 Ma J, Cui P, Kuang K, et al. Disentangled graph convolutional networks. 2019: 4212–4221
- 35 Qin Y, Wang X, Zhang Z, et al. Graph neural architecture search under distribution shifts. In: *International Conference on Machine Learning*, 2022. 18083–18095
- 36 Wang Y, Qin Y, Sun F, et al. Disenctr: Dynamic graph-based disentangled representation for click-through rate prediction. In: *SIGIR*, 2022. 2314–2318
- 37 Zhang W, Zhang L, Pfoser D, et al. Disentangled dynamic graph deep generation. 2021: 738–746
- 38 Huang K, Zitnik M. Graph meta learning via local subgraphs. *NeurIPS*, 2020, 33: 5862–5874
- 39 Wang Y, Chang Y, Liu Y, et al. Inductive representation learning in temporal networks via causal anonymous walks. In: *9th International Conference on Learning Representations*, 2021
- 40 Huang H, Fang Z, Wang X, et al. Motif-preserving temporal network embedding. In: *IJCAI*, 2020. 1237–1243
- 41 Gagnon-Audet J C, Ahuja K, Darvishi-Bayazi M J, et al. Woods: Benchmarks for out-of-distribution generalization in time series tasks. *arXiv preprint arXiv:2203.09978*, 2022
- 42 Venkateswaran P, Muthusamy V, Isahagian V, et al. Environment agnostic invariant risk minimization for classification of sequential datasets. In: *KDD*, 2021. 1615–1624
- 43 Kovanen L, Karsai M, Kaski K, et al. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011, 2011: P11005
- 44 Paranjape A, Benson A R, Leskovec J. Motifs in temporal networks. In: *Proceedings of the tenth ACM international conference on web search and data mining*, 2017. 601–610
- 45 Huang H, Tang J, Liu L, et al. Triadic closure pattern analysis and prediction in social networks. *IEEE TKDE*, 2015, 27: 3374–3389
- 46 Zhou L, Yang Y, Ren X, et al. Dynamic network embedding by modeling triadic closure process. In: *AAAI*, 2018, volume 32
- 47 Kovanen L, Kaski K, Kertész J, et al. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences*, 2013, 110: 18070–18075
- 48 Pearl J, et al. *Models, reasoning and inference*. Cambridge, UK: CambridgeUniversityPress, 2000, 19: 2
- 49 Glymour M, Pearl J, Jewell N P. *Causal inference in statistics: A primer*. 2016
- 50 Tian J, Kang C, Pearl J. A characterization of interventional distributions in semi-markovian causal models. 2006
- 51 Rossi E, Chamberlain B, Frasca F, et al. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020
- 52 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *NeurIPS*, 2017
- 53 Bahng H, Chun S, Yun S, et al. Learning de-biased representations with biased representations. 2020: 528–539
- 54 Gretton A, Bousquet O, Smola A, et al. Measuring statistical dependence with hilbert-schmidt norms. In: *ALT*, 2005. 63–77
- 55 Song L, Smola A, Gretton A, et al. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 2012, 13: 1393–1434
- 56 Liu Y, Wang X, Wu S, et al. Independence promoted graph disentangled networks. In: *AAAI*, 2020, volume 34. 4916–4923
- 57 Kipf T N, Welling M. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016
- 58 Krueger D, Caballero E, Jacobsen J H, et al. Out-of-distribution generalization via risk extrapolation (rex). In: *ICML*, 2021. 5815–5826
- 59 Tang J, Wu S, Sun J, et al. Cross-domain collaboration recommendation. In: *KDD'2012*, 2012
- 60 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. In: Y Bengio, Y LeCun, eds., *ICLR*, 2013
- 61 Hu W, Fey M, Zitnik M, et al. Open graph benchmark: Datasets for machine learning on graphs. *NeurIPS*, 2020, 33: 22118–22133
- 62 Sinha A, Shen Z, Song Y, et al. An overview of microsoft academic service (mas) and applications. 2015: 243–246
- 63 Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *NeurIPS*, 2017, 30
- 64 Wang K, Shen Z, Huang C, et al. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 2020, 1: 396–413
- 65 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. *NeurIPS*, 2013, 26
- 66 Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 1987, 20: 53–65

Independence-Promoted Disentangled Dynamic Graph Attention Network for Out-of-Distribution Generalization

Haoyang Li¹, Xin Wang^{1*} & Wenwu Zhu^{1*}

1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

* Corresponding author. E-mail: xin_wang@tsinghua.edu.cn, wwzhu@tsinghua.edu.cn

Abstract Dynamic graph neural networks (DyGNNs) have attracted increasing attention in recent years due to their strong ability to model graph structural and temporal dynamics. However, the existing DyGNNs fail to generalize under distribution shifts, which naturally exist in dynamic graphs, since the patterns exploited by DyGNNs may be variant with respect to labels under distribution shifts. These distribution shifts lead to substantial performance degradation of existing DyGNNs when applied to out-of-distribution (OOD) settings. Specifically, the existing DyGNNs tend to overfit to spurious correlations between variant patterns and labels, making it difficult to capture invariant patterns that are stable under distribution shifts. As a result, current DyGNNs struggle with OOD generalization. To address these challenges, in this paper, we propose an independence-promoted disentangled dynamic graph attention network which can effectively capture invariant patterns in dynamic graphs and significantly improve OOD generalization performance under distribution shifts. First, we introduce a disentangled dynamic graph attention network that explicitly separates invariant patterns from variant ones. Second, we introduce a causality-inspired spatio-temporal intervention mechanism that generates diverse intervened distributions and minimizes the variance of predictions among these distributions to eliminate spurious correlations. Third, we design an independence-promoted disentanglement optimization framework based on invariance and independence regularizers to enhance the model's ability to capture invariant patterns for OOD generalization. We conduct extensive experiments on several real-world and synthetic datasets. The experimental results show that our method consistently outperforms state-of-the-art DyGNNs and OOD generalization baselines under various distribution shift scenarios. In addition, ablation studies and visualization analyses validate the effectiveness of each key module. We also highlight promising directions for future research on OOD generalization in DyGNNs.

Keywords Dynamic Graph Neural Network, Disentanglement, Distribution Shift, Graph Machine Learning, Independence