arXiv:2301.02780v2 [cs.LG] 1 Nov 2023

通过非参数子图匹配重新思考图神经网络解释

吴芳 [1 2] **李思远** [1] **金旭瑞** [2] **姜颖慧** [2] **Dragomir Radev** [3] **牛张明** [2] **李泽湘** [1]

摘要

图神经网络（GNN）的成功引发了关于可解释性的问题："输入图的哪一部分对预测起着决定性作用？"尤其是，参数化解释器在现有方法中占据主导地位，因为它们在解读黑箱（即目标图神经网络）方面具有更强的能力。本文基于图通常共享某些常见模体模式这一观察，提出了一种新颖的非参数子图匹配框架，称为 MatchExplainer，用于探索解释性子图。该框架将目标图与其他对应实例相结合，并通过最小化基于节点对应的距离来识别最关键的联合子结构。此外，我们注意到，现有的图采样或节点删除方法通常会遭遇假阳性采样问题。为缓解这一问题，我们设计了一种名为 MatchDrop 的新增强范式。它利用 MatchExplainer 固定图中信息量最大的部分，仅在其余信息量较低的部分进行图增强。在合成数据集和真实世界数据集上的大量实验表明，MatchExplainer 的有效性优于所有最先进的参数化基线方法，且优势显著。实验结果还表明，MatchDrop 是一种可与图神经网络配合使用、以提升性能的通用方案。代码见 https://github.com/smiles724/MatchExplainer。

[1]西湖大学工程学院，中国杭州 [2]MindrankAI，中国杭州 [3]耶鲁大学计算机科学系，美国纽黑文。通讯作者：牛张明 <zhangming@mindrank.ai>，李泽湘 < stan.zq.li@westlake.edu.cn>。

论文集： 40^{th} 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有：作者，2023年。

1. 引言

图神经网络（图神经网络）因其在学习图结构数据表示方面的成功而广受关注，例如社交网络（Fan 等，2019年）、知识图谱（Schlichtkrull 等，2018年）、交通网络（Geng 等，2019年）以及分子图（Gilmer 等，2017年；Wu 等，2023年）。尽管图神经网络效果显著，但其透明性不足，因为人类难以理解其预测结果的推理依据。这不仅使从业者难以了解网络特征，也妨碍他们在将模型部署到现实世界应用之前纠正模型所犯错误中的系统性模式。

大量研究已注意到这一问题，并投入大量努力来解释图神经网络（Yuan 等人，2020b）。研究者致力于回答诸如"输入图中的哪些知识对模型的决策起着最主要的作用？"之类的问题。为此，特征归因与选择（Selvaraju 等人，2017年；Sundararajan 等人，2017年；Ancona 等人，2017年）成为一种普遍采用的范式。这些方法通过类梯度信号（Baldassarre 和 Azizpour，2019年；Pope 等人，2019年；Schnake 等人，2020年）、掩码或注意力分数（Ying 等人，2019年；Luo 等人，2020年），或扰动特征导致的预测变化（Schwab 和 Karlen，2019年；Yuan 等人，2021年），将模型的预测结果归因于输入图，随后选取一个显著子结构作为解释。

除此之外，较新的方法倾向于依赖深度学习网络来参数化解释的生成过程（Vu 和 Thai，2020年；Wang 等人，2021b）。这些基于学习的机制在经验上表现出比上述非参数方法更高的准确率。一些解释器模型针对局部忠实度进行优化（Chen 等人，2018年），例如 GNNExplainer（Ying 等人，2019年）、PGM-Explainer（Vu 和 Thai，2020年）和 SubgraphX（Yuan 等人，2021年）。与此同时，另一些方法致力于提供对模型预测的全局理解，包括 PGExplainer（Luo 等人，2020年）、XGNN（Yuan 等人，2020a）和 ReFine（Wang 等人，2021b）。

尽管取得了富有成效的进展，并且参数化

解释器这一趋势日益流行，但我们观察到，不同的图群之间共享着不同的关键子图模式，这可能是解读图神经网络决策的关键。这些频繁出现的模体蕴含丰富的语义，并体现了整个图实例的特征（Henderson 等，2020年；Zhang 等，2020年；Banjade 等，2021年；Wu 等，2023年）。例如，小分子中的羟基（-OH）通常会带来更高的水溶性，而羧基（-COOH）通常有助于提高稳定性和沸点。此外，官能团在蛋白质结构预测中的关键作用也已得到证实（Senior 等，2020年）。

受这一观察的启发，我们提出以子图匹配的方式挖掘解释性基元，并设计了一种名为 MatchExplainer 的新型非参数算法，其工作流程如图1所示。对于每一对图，我们的 MatchExplainer 都会在高维特征空间中，通过最小化基于节点对应的距离，努力探索最关键的联合子结构。随后，它会将目标图与参考集中的其他对应图逐一配对，以寻找潜在的解释性子图。因此，与传统解释器不同，MatchExplainer 为同一个目标图实例生成的候选解释可能并不唯一。

更进一步，我们利用信息论中的互信息度量来分析 MatchExplainer 的工作原理。具体而言，我们将包含成对图之间所有共享信息的解释定义为充分解释，而将包含共享信息并剔除非共享信息的解释定义为最小充分解释。我们证明，最小充分解释能够在理论保证下近似所需的真实解释。这种紧密关系还为我们提供了一种视角，用于从所有候选解释性子图中筛选出最佳子结构。具体来说，我们通过最大化从原始图中移除解释性子图后预测结果的差异，来优化最终的候选解释。

最后但同样重要的是，我们展示了 MatchExplainer 的一项额外优势应用于增强传统图增强方法。尽管当前的图采样或节点丢弃机制在防止过拟合和过平滑方面表现强劲，但仍受假阳性采样问题困扰。也就是说，信息量最大的子结构中的节点或边会意外地被丢弃或删除，但模型仍需预测原始属性，这可能造成误导。为缓解这一障碍，我们利用 MatchExplainer，提出了一种名为 MatchDrop 的简单技术。具体而言，它首先借助 MatchExplainer 找出解释性子图，并保持该部分不变。然后，仅对其余信息量较少的部分进行图采样或节点丢弃。因此，输入图中揭示标签信息的核心部分不会受到影响，从而有效缓解假阳性采样问题。

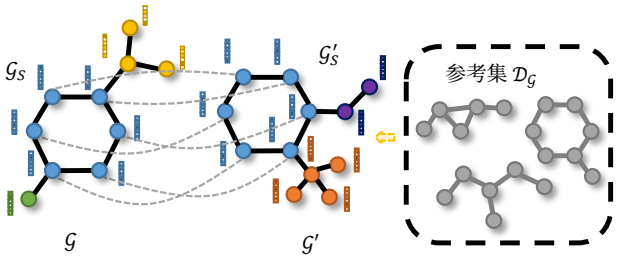


图 1. 我们提出的 MatchExplainer 示意图。解释 $\mathcal{G}_S$ 通过在 $\mathcal{G}$ 与 $\mathcal{G}'$ 之间进行子图匹配获得，其中我们以贪心搜索的方式，最小化高维特征空间中的节点间累计距离。由于将 $\mathcal{G}_S$ 与参考集 $\mathcal{G}$ 中的不同对应图 $\mathcal{G}'$ 进行匹配，可以得到多个 $\mathcal{D}_\mathcal{G}$，因此我们通过最大化公式 10来寻找最优解。

总而言之，据我们所知，我们率先从非参数子图匹配的角度研究了图神经网络的可解释性。在合成数据和真实世界应用上的大量实验表明，我们的 MatchExplainer 能够快速、准确地找到具有最先进性能的解释性子图。此外，我们通过实验证明，MatchExplainer 的实用应用 MatchDrop 可以作为一种有效的方法，促进传统图增强方法的发展。

2. 预备知识与任务描述

本节首先介绍图神经网络解释任务，并简要回顾图匹配和图相似性学习（GSL）的相关背景。

图神经网络的解释。 令 $h_Y: \mathcal{G} \to \mathcal{Y}$ 表示待解释的训练良好的图神经网络，其给出预测 Y 以近似真实标签 Y。不失一般性，我们考虑解释图分类任务。我们的目标是找到一个解释器$h_S: \mathcal{G} \to \mathcal{G}_\mathcal{S}$，从输入图 $\mathcal{G}$ 中发现子图 $\mathcal{G}_S$，使得：

$$\min_{h_S} \mathcal{R}(h_Y \circ h_S(\mathcal{G}), \hat{Y}), \text{s.t.} |h_S(\mathcal{G})| \leq K, \qquad (1)$$

其中，$\mathcal{R}$（。）是风险函数，例如交叉熵损失或均方误差（MSE）损失；K 是对 $\mathcal{G}_S$ 大小的约束，以获得紧凑的解释。也就是说，$\mathcal{G}_S$ 最多包含 K 个节点。

图匹配。作为经典的组合优化问题，图匹配通常被认为是 NP难问题（Loiola 等人，

2007年）。这类问题需要耗时、复杂且不切实际的求解器，因而只能得到近似解（Wang 等人，2020年）。给定两个不同的图 $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ 和$\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$，它们分别有 N_1 和 N_2 个节点，其间的匹配通常可表示为二次指派规划（QAP）形式（Wang 等人，2019年）：

$$\min_{\mathbf{T}\in\{0,1\}^{N_1\times N_2}} \text{vec}(\mathbf{T})^T \mathbf{K}\text{vec}(\mathbf{T}),\ s.t., \mathbf{T}\mathbf{1} = \mathbf{1},\ \mathbf{T}^T\mathbf{1} = \mathbf{1}, \tag{2}$$

其中 $\mathbf{T}$ 是一个二元置换矩阵，用于编码节点对应关系，而$\mathbf{1}$ 表示一个所有元素均为 1 的列向量。 $\mathbf{K}$ 是所谓的亲和矩阵（Leordeanu 和 Hebert，2005年），其元素编码了节点之间和边之间的亲和度，在 $\mathcal{G}_1$ 和$\mathcal{G}_2$之间。

图相似性学习。GSL 是一种通用的图表示学习框架，需要对图的结构和语义进行推理（Li 等，2019）。我们需要计算它们之间的相似度分数 $s(\mathcal{G}_1, \mathcal{G}_2)$。这种相似度 $s(.,.)$ 通常通过以下方式定义：针对完整图或子图同构进行精确匹配（Berretti 等，2001年；Shasha 等，2002年），或者采用某种结构相似性度量，例如图编辑距离（Willett 等，1998；Raymond 等，2002年）。在我们的设定中，$s(.,.)$ 完全取决于这两个图是否属于同一类别，或是否具有非常相近的属性。对于类型相同的 $\mathcal{G}_1$ 和 $\mathcal{G}_2$，GSL 致力于最大化其表示之间的互信息，其联合分布 $p(\mathcal{G}_1, \mathcal{G}_2)$ 如下：

$$\max_{f_1, f_2} I(f_1(\mathcal{G}_1), f_2(\mathcal{G}_2), T), \tag{3}$$

其中，f_1 和 f_2 是编码函数。它们可以共享同一参数（即 $f_1 = f_2$），也可以合并为一个架构。 T是表示特定任务所需信息的随机变量，与模型选择无关。

3. MatchExplainer 方法

近期的大多数方法依赖参数化网络来解释图神经网络，而一些早期的图神经网络解释方法基于局部可解释性，并从单图视角出发（Ying 等人，2019年；Baldassarre 和 Azizpour，2019年；Pope 等人，2019年；Schwab 和 Karlen，2019年）。尽管存在这种倾向，我们认为，非参数化的图—图方式同样能够挖掘重要子图，并可能带来更好的可解释性。在本研究中，我们提出 MatchExplainer，通过子图匹配识别共同的关键子结构，以此解释图神经网络（见算法1）。

3.1. MatchExplaine 的理论分析 r

从概率论和信息论的角度来看，公式 1 等价于最大化输入图 $\mathcal{G}$ 与子图 $\mathcal{G}_S$ 在 h_Y背景下的互信息。也就是说，解释器的目标是推导出一个较小的子图 $\mathcal{G}_S$，使其满足：

$$\max_{\mathcal{G}_S\subset\mathcal{G}, |\mathcal{G}_S|\le K} I(\mathcal{G}_S, T_h), \tag{4}$$

其中 $I(。)$ 表示两个随机变量的香农互信息。不同于 T 作为模型无关的解释器， T_h代表图神经网络预测器 h_Y在具体下游任务中学到的知识。值得注意的是，不同于仅仅优化隐藏在 $\mathcal{G}_S$中的信息，另一类研究（Yuan 等人，2021年）旨在降低剩余子图 $\mathcal{G} - \mathcal{G}_S$ 与原始图 $\mathcal{G}$ 之间的互信息，具体如下：

$$\min_{\mathcal{G}_S\subset\mathcal{G}, |\mathcal{G}_S|\le K} I(\mathcal{G} - \mathcal{G}_S, T_h). \tag{5}$$

作为直接优化公式 4的一种近似，MatchExplainer 的核心思想是找到另一个图 $\mathcal{G}'$，使其与 $\mathcal{G}$ (即， $h_Y(\mathcal{G})$ $=h_Y(\mathcal{G}')$具有相同的预测属性，然后提取二者之间最相关的部分作为解释。具体而言，我们旨在寻找最佳匹配图 $\mathcal{G}'$，使输入图 $\mathcal{G}$ 与子图 $\mathcal{G}_S$ 之间的互信息最大化：

$$\max_{\mathcal{G}'\in\mathcal{D}_\mathcal{G}, \mathcal{G}'\neq\mathcal{G}} \left[\max_{\mathcal{G}_S\subset\mathcal{G}, |\mathcal{G}_S|\le K} I(\mathcal{G}_S, \mathcal{G}', T_h)\right], \tag{6}$$

其中 $\mathcal{D}_\mathcal{G}$ 表示由所有可用图组成的参考集，而 $\mathcal{G}_S$ 是通过在 $\mathcal{G}$ 与 $\mathcal{G}'$之间进行子图匹配得到的。与监督学习中的信息瓶颈理论（Tishby & Zaslavsky，2015年；Achille & Soatto，2018年）类似，我们可以在子图匹配的背景下，借助其对应项$\mathcal{G}' \neq \mathcal{G}$，定义 $\mathcal{G}$ 的充分解释和最小充分解释。

定义 3.1（充分解释）。给定 $\mathcal{G}'$，如果且仅如果$I(\mathcal{G}_S^{suf}, \mathcal{G}', T_h) = I(\mathcal{G}, \mathcal{G}', T_h)$，那么 $\mathcal{G}_S^{suf}$ 是 $\mathcal{G}$ 的充分解释。

$\mathcal{G}_S^{suf}$ 对 $\mathcal{G}$ 的充分解释保留了与学习到的信息 T_h相关的、关于 $\mathcal{G}'$ 的全部联合信息。换言之， $\mathcal{G}_S^{suf}$ 包含了 $\mathcal{G}$ 与 $\mathcal{G}'$之间所有共享的信息。对称地， $\mathcal{G}'$的充分解释满足 $I(\mathcal{G}'^{suf}_S, \mathcal{G}', T_h) = I(\mathcal{G}, \mathcal{G}', T_h)$。

定义 3.2（最小充分解释）。 给定 $\mathcal{G}'$，当且仅当 $I(\mathcal{G}_S^{min}, \mathcal{G}, T_h) \le I(\mathcal{G}_S^{suf}, \mathcal{G}, T_h)$ 时， $\mathcal{G}$ 的充分解释 $\mathcal{G}_S^{min}$是最小的。

在所有充分解释中，最小充分解释 $\mathcal{G}_S^{min}$包含关于 $\mathcal{G}$的信息最少

对于学习到的知识 T_h，通常假设 $\mathcal{G}_S^{min}$仅保留 $\mathcal{G}$ 和 $\mathcal{G}'$之间的共享信息，并消除其他非共享信息，即 $I(\mathcal{G}_S^{min}, \mathcal{G}|\mathcal{G}') = 0$。

定理 3.3（解释中的任务相关信息）。（*Wang* 等人，*2022a*）给定 $\mathcal{G}'$，最小充分解释 $\mathcal{G}_S^{min}$包含的信息中，h_Y从输入 $\mathcal{G}$ 中学习到的任务相关信息少于任何其他充分解释 $\mathcal{G}_S^{suf}$。形式化地，我们有：

$$\begin{aligned} I(\mathcal{G}, T_h) &= I(\mathcal{G}_S^{min}, T_h) + I(\mathcal{G}, T_h|\mathcal{G}') \\ &\geq I(\mathcal{G}_S^{suf}, T_h) = I(\mathcal{G}_S^{min}, T_h) + I(\mathcal{G}_S^{suf}, \mathcal{G}, T_h|\mathcal{G}') \\ &\geq I(\mathcal{G}_S^{min}, T_h). \end{aligned} \tag{7}$$

定理 3.3 表明，$\mathcal{G}$ 与 T_h 之间的互信息可分为两部分。一部分是 $\mathcal{G}_S^{min}$，即与学习到的知识 T_h相关的 $\mathcal{G}$ 与 $\mathcal{G}'$ 之间的交互。另一部分则由 $\mathcal{G}$ 与 $\mathcal{G}'$ 相对于学习到的信息 T_h的互不相交结构决定。我们的子图匹配致力于最大化 $I(\mathcal{G}_S^{min}, T_h)$，它是 $I(\mathcal{G}, T_h)$的下界。值得注意的是，$I(\mathcal{G}, T_h|\mathcal{G}')$与$I(\mathcal{G}_S^{min}, T_h)$并非完全独立，而是 $I(\mathcal{G}_S^{min}, T_h)$相对于$I(\mathcal{G}, T_h)$的偏移。因此，如果增大 $I(\mathcal{G}_S^{min}, T_h)$，就会同时最小化 $I(\mathcal{G}, T_h|\mathcal{G}')$。由此，$I(\mathcal{G}_S^{min}, T_h)$不仅可用于提高 $I(\mathcal{G}, T_h)$的下界，还可用于近似 $I(\mathcal{G}, T_h)$，后者正是我们最终的解释对象。这为我们的 MatchExplainer 通过图匹配方法挖掘最具解释力的子结构提供了坚实的理论基础。

3.2. 非参数子图探索

前言。值得注意的是，我们通过子图匹配挖掘解释的方法，与图匹配或 GSL 都存在显著差异。一方面，图匹配算法（Zanfir 和 Sminchisescu，2018年；Sarlin 等人，2020年；Wang 等人，2020年；2021a）通常会在一个完整图 $\mathcal{G}_1$ 与另一个完整图 $\mathcal{G}_2$之间建立节点对应关系。然而，我们旨在构建 $\mathcal{G}_1$ 的子图与 $\mathcal{G}_2$ 的子图之间的部分节点对应关系。另一方面，GSL 关注由 f_1 和 f_2编码的图表示，以及真实标签信息 T，而非T_h 图神经网络预测器学习到的信息 h_Y。

此外，现有的大多数图匹配架构（Zanfir 与 Sminchisescu，2018年；Li 等人，2019年；Wang 等人，2020年；Pa-pakis 等人，2020年；Liu 等人，2021a）都基于深度学习。它们利用网络预测节点之间或图之间的关系，但存在若干缺陷。例如，训练网络需要大量计算资源。更重要的是，其有效性并不可靠，而且如果网络设计不够精巧，在某些情况下可能会失效。为克服这些局限，我们采用非参数子图匹配范式。该范式完全无需训练，能够快速探索任意一对输入实例共享的、信息量最大的联合子结构。

子图匹配框架。我们将目标图神经网络 h_Y 拆分为两个连续部分：$h_Y = \phi_G \circ \phi_X$，其中 ϕ_G 是用于计算图级表示并预测属性的聚合器，而 ϕ_X 是用于更新节点特征和边特征的特征函数。给定一个图 $\mathcal{G}$，其节点特征为 $\mathbf{h}_i \in \mathbb{R}^{\psi_v}$，$\forall i \in \mathcal{V}$，边特征为 $\mathbf{e}_{ij} \in \mathbb{R}^{\psi_e}$，$\forall (i, j) \in \mathcal{E}$，更新后的输出计算为 $\{\mathbf{h}'_i\}_{i\in\mathcal{V}}$，$\{\mathbf{e}'_{ij}\}_{(i,j)\in\mathcal{E}} = \phi_X\left(\{\mathbf{h}_i\}_{i\in\mathcal{V}}, \{\mathbf{e}_{ij}\}_{(i,j)\in\mathcal{E}}\right)$，随后将其传入 ϕ_G。

如前文所述，我们的主要目标是找到一个包含 K 个节点的子图$\mathcal{G}_S$，以最大化 $I(\mathcal{G}_S, \mathcal{G}', T_h)$。基于如下假设：最优对应项 $\mathcal{G}'$ 应当与 $\mathcal{G}$共享相同的解释性子结构。我们的目标等价于优化 $I(\mathcal{G}_S, \mathcal{G}'_S, T_h)$，其中包含 $\mathcal{G}_S \subset \mathcal{G}$ 和 $\mathcal{G}'_S \subset \mathcal{G}'$。为此，我们使用基于节点对应关系的距离 d_G 来替代对 $I(\mathcal{G}_S, \mathcal{G}'_S, T_h)$ 的度量，即 $\mathcal{G}_S$ 与 $\mathcal{G}'_S$ 之间共享的学习信息。接着，给定一对 $\mathcal{G}$ 和 $\mathcal{G}'$，d_G 定义如下，并将其最小化：

$$\min_{\mathcal{G}_S\subset\mathcal{G}, \mathcal{G}'_S\subset\mathcal{G}'} d_G(\mathcal{G}_S, \mathcal{G}'_S) = \min_{\mathcal{G}_S\subset\mathcal{G}, \mathcal{G}'_S\subset\mathcal{G}'} \left(\min_{\mathbf{T}\in\Pi(\mathcal{G}_S, \mathcal{G}'_S)} \langle \mathbf{T}, \mathbf{D}^{\phi_X} \rangle \right), \tag{8}$$

其中，$\mathbf{D}^{\phi_X}$ 是 $\mathcal{G}_S$ 与 $\mathcal{G}'_S$ 的节点特征之间所有成对距离构成的矩阵。其元素计算为$\mathbf{D}_{ij}^{\phi_X} = d_X(\mathbf{h}_i, \mathbf{h}'_j)$ $\forall i \in \mathcal{V}$，$j \in \mathcal{V}'$，其中 d_X 是标准的向量空间相似度度量，例如欧几里得距离和汉明距离。内部优化针对 $\Pi(\cdot, \cdot)$ 进行，该集合包含所有具有指定边缘分布的矩阵，定义如下：

$$\Pi(\mathcal{G}_S, \mathcal{G}'_S) = \left\{ \mathbf{T} \in \{0, 1\}^{K\times K} \,|\, \mathbf{T}\mathbf{1} = \mathbf{1}, \mathbf{T}^T\mathbf{1} = \mathbf{1} \right\}. \tag{9}$$

由于图匹配具有 NP 难性（Loiola 等人，2007年），我们采用贪心策略来优化 $d_G(\mathcal{G}_S, \mathcal{G}'_S)$并得到子图 $\mathcal{G}_S$。值得注意的是，贪心算法不能保证得到全局最优解（Bang-Jensen 等人，2004年），但能在合理的时间内以 $O(K)$的复杂度得到局部最优解。

之后，我们将 $\mathcal{G}_S$ 输入 h_Y并检查其正确性。如果 $h_Y(\mathcal{G}_S) = h_Y(\mathcal{G})$，则将 $\mathcal{G}_S$ 视为候选解释。否则，将放弃 $\mathcal{G}_S$，因为它无法恢复 h_Y预测 $\mathcal{G}$所需的信息。

算法1 MatchExplainer 的工作流程

输入： 目标图神经网络 h_Y、图 $\mathcal{G}$、参考集 $\mathcal{D}_\mathcal{G}$
初始化一个空的候选列表 $\mathcal{D}$ S.
对于 $\mathcal{G}' \in \mathcal{D}_\mathcal{G}$，**执行**
$\mathcal{G}_S \leftarrow$ 最小值$_{\mathcal{G}_S \subset \mathcal{G},\ \mathcal{G}'_S \subset \mathcal{G}'}\ d_G(\mathcal{G}_S,\ \mathcal{G}'_S)$ 见公式8
如果 $h_Y(\mathcal{G}_S) = h_Y(\mathcal{G})$ **则**
将 $\mathcal{G}_S$ 添加到 $\mathcal{D}_S$
结束条件判断
结束循环
$\mathcal{G}_S^+ \leftarrow$ 最大值$_{\mathcal{G}' \in \mathcal{D}_S, \mathcal{G}' \neq \mathcal{G}}\ \Delta_\mathcal{G}(\mathcal{G}', h_Y)$ 见公式 10
返回： $\mathcal{G}_S^+$

图神经网络解释的非唯一性。不同于以往基于学习的图神经网络解释方法（Vu 和 Thai，2020年；Wang 等人，2021b；2022b）生成唯一的子图 $\mathcal{G}_S$ 用于 $\mathcal{G}$，我们对 $\mathcal{G}_S$ 的选择会随着对应对象 $\mathcal{G}' \in \mathcal{D}_\mathcal{G}$的选择而变化。因此，一旦给定一组对应对象，MatchEx-plainer 就能为单个图提供多对一的解释 $\mathcal{G}$ 。这带来了一种新的认识：图神经网络预测结果的决定因素并非唯一，图神经网络可以基于多个大小相同但彼此不同的解释性子图获得正确的预测结果。

图神经网络解释的优化。由于我们的 Match-Explainer 能够发现多种可能的解释性子图，如何筛选出信息量最大的子图便成为一个关键问题。如定理 3.3 所示，$I(\mathcal{G}_S^{min}, T_h)$ 是 $I(\mathcal{G}, T_h)$ 的下界，而二者的差异 $I(\mathcal{G}, T_h|\mathcal{G}')$ 完全取决于匹配对应图 $\mathcal{G}'$ 的选择。理想情况下，$\mathcal{G}'$ 应与 $\mathcal{G}$具有完全相同的解释性子结构，即 $\mathcal{G}_S = \mathcal{G}'_S$。同时，在 $\mathcal{G}'$ 条件下的 $\mathcal{G}$ 与学得的知识 T_h相互独立，即 $I(\mathcal{G}, T_h|\mathcal{G}') = 0$。因此，选择对应图有两项截然不同的原则。

第一项原则是寻找 $\mathcal{G}'$，使其解释性子图尽可能接近 $\mathcal{G}$。第二项原则是确保在 $\mathcal{G}'$ 条件下的 $\mathcal{G}$ 所保留的与学得信息 T_h相关的信息尽可能少。然而，如果缺乏足够的领域知识，无法判断哪种子结构对图属性起主要作用，我们就无法手动选择满足 $\mathcal{G}_S \approx \mathcal{G}'_S$的对应图$\mathcal{G}'$ 。

作为补救措施，我们考虑优化公式 5中描述的相反目标。也就是说，我们希望最小化 $\mathcal{G} - \mathcal{G}_S$与 T_h之间的交集，即，$I(\mathcal{G} - \mathcal{G}_S,\ T_h)$。为实现这一目标，我们从$\mathcal{G}$ 中移除提取出的子图 $\mathcal{G}_S$，并希望干扰图神经网络对剩余部分 $\mathcal{G} - \mathcal{G}_S$ 的预测。从数学上讲，最优的 $\mathcal{G}'$ 会最大化整个图的预测与移除 $\mathcal{G}_S$后的图的预测之间的差异。换言之

w我们希望检索到最佳解释 $\mathcal{G}$ ^+_S 经由：

$$\max_{\mathcal{G}' \in \mathcal{D}_\mathcal{S}, \mathcal{G}' \neq \mathcal{G}} \Delta_\mathcal{G}(\mathcal{G}', h_Y) = \max_{\mathcal{G}' \in \mathcal{D}_\mathcal{S}, \mathcal{G}' \neq \mathcal{G}} \left[h_Y^{c^*}(\mathcal{G}) - h_Y^{c^*}(\mathcal{G} - \mathcal{G}_S) \right], \tag{10}$$

其中，c^* 是 $\mathcal{G}$ 的真实类别，$\mathcal{G}_S$ 是通过与 $\mathcal{G}'$进行子图匹配得到的子结构。 $\mathcal{D}_\mathcal{S}$ 是候选子图集。

总而言之，给定任意图 $\mathcal{G}$ 和参考图集$\mathcal{D}_\mathcal{G}$，我们首先通过将 $\mathcal{G}$与 $\mathcal{D}_\mathcal{G}$中可用的对应图进行匹配，获取所有可能的子图。完成成对子图匹配后，我们计算它们对应的 $\Delta_\mathcal{G}(., h_Y)$，并选出能产生最大 $\Delta_\mathcal{G}(., h_Y)$的图作为最优对应图。值得注意的是， $\mathcal{D}_\mathcal{G}$中的图并非都符合对应图的条件，而且有若干直观条件需要 $\mathcal{G}'$ 满足。首先， $\mathcal{G}$ 和 $\mathcal{G}'$ 应属于 h_Y预测的同一类别。此外， $\mathcal{G}'$至少需要包含 K 个节点。否则， G_S 会小于给定的约束大小。

有效性与效率。 时间复杂度始终是评估解释器实用性的重要议题。对于我们的 MatchExplainer，参考集的大小（即$|\mathcal{D}_\mathcal{G}|$）在决定时间开销方面起着重要作用，因为总时间开销为 $O(K|\mathcal{D}_\mathcal{G}|)$。然而，对应图数量有限也可能妨碍其探索更优的解释性子图。因此，选择合适的 $\mathcal{D}_\mathcal{G}$大小，以平衡 MatchExplainer 的有效性与效率，并非易事。

4. MatchDrop方法论

防止假阳性采样。深度图学习面临着独特的挑战，例如特征数据不完整、结构数据稀疏性以及过平滑。为了解决这些问题，图领域中提出了越来越多的数据增强技术（Hamilton等人，2017年；Rong 等人，2019年），并取得了令人鼓舞的成果。图采样和节点丢弃（Feng等人，2020年；Xu 等人，2021年）是两种常用机制。然而，以往的大多数方法都完全采用随机方式，导致假阳性采样，并在训练过程中注入虚假信息。例如，*1,3*-二硝基苯（C_6氢$_4$$N_2$氧$_4$）是一种致突变分子，其解释依据是 NO_2 基团（Debnath等人，1991年）。如果 NO_2 基团的任何边或节点被意外丢弃或破坏，致突变性就不复存在。因此，在节点或边采样后，若仍将原始标签赋予该分子图，就会误导图神经网络。

为了解决这一缺陷，请注意，我们的 MatchExplainer 提供了一种便捷的方法来发现给定图中最关键的部分

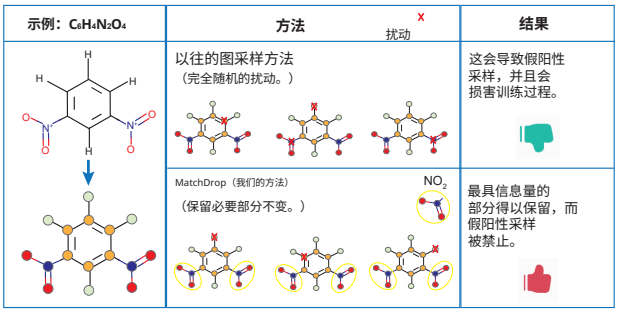


图 2。比较使用和不使用 MatchDrop 时的图增强。

。我们可以保留这一关键部分不变，仅删除其余部分中的节点或边，这是很自然的做法。基于这一思路，我们提出了一种简单但有效的方法，称为 MatchDrop。该方法保留由我们的 MatchExplainer 找到的图中信息量最大的部分，并修改信息量较低的部分（见图2）。

我们的 MatchDrop 流程如下所述。首先，我们训练一个 GNN h_Y 若干训练轮次，直到其收敛至可接受的准确率，从而确保后续子图选择的有效性。然后，对于训练集 $\mathcal{D}_{\text{train}}$中的每个图 $\mathcal{G}$，我们随机选择另一个与其类别相同的图 $\mathcal{G}' \in \mathcal{D}_{\text{train}}$ 作为对应图。随后，我们通过 MatchExplainer 探索该图的子图 $\mathcal{G}_S$，保留比例为 ρ（即 $|\mathcal{G}_S| = \rho|\mathcal{G}|$），并将其用作模型输入来训练 h_Y。

值得注意的是，与旋转和翻转等典型的图像增强技巧类似（Shorten 和 Khoshgoftaar，2019年），MatchDrop 是一种用于图神经网络训练的新型数据增强技术。然而，MatchDrop 并非随机增强 $\mathcal{G}$，而是保留信息量最大的部分，只改变不太重要的子结构。这显著降低了假阳性采样的可能性。此外，与其他用于检查子图的可学习机制不同，我们的 MatchDrop 完全不依赖参数，因此可以在训练过程的任何阶段部署。

训练目标。 图神经网络的训练由交叉熵（CE）损失监督。假设总共有 M 个类别，则损失形式如下：

$$\mathcal{L}_S = -\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{\mathcal{G} \in \mathcal{D}_{\text{train}}} \sum_{c=1}^{M} Y_{\mathcal{G}} \log \left(h_Y^c \left(h_S(\mathcal{G}, \rho) \right) \right), \quad (11)$$

其中， h_Y^c(。)表示 $\mathcal{G}_S$ 属于类别 c 的预测概率， Y_G 是真实值。 h_S 使用 MatchExplainer，通过将 $\mathcal{G}$ 与训练集$\mathcal{D}_{\text{train}}$ 中随机选取的对应图 $\mathcal{G}'$ 按预设比例 ρ进行匹配，来挖掘子图 $\mathcal{G}_S$ 。

5. 实验分析

5.1. 数据集和实验设置

遵循Wang 等人 （2021b），我们使用四个标准数据集和多种目标图神经网络。

• **分子图分类**：MUTAG（Debnath等人，1991年；Kazius等人，2005年）是用于图分类问题的分子数据集。每个图表示一个分子，其中节点代表原子，边代表化学键。标签由分子对细菌的致突变效应决定。经过充分训练的图同构网络（GIN）（Xu 等人，2018年）的测试准确率约为 82%。

• **基元图分类**。： Wang 等人(2021b) 创建了一个包含 3000 个图的合成数据集 BA-3Motif。他们以 Barabasi-Albert（BA）图为基础，并为每个基础图附加三种基元之一：房屋、环或网格。我们训练了一个 ASAP 模型 (Ranjan 等人，2020年)，其测试准确率达到 99.75%。

• **手写图分类**： Knyazev 等人 (2019年) 将 MNIST 图像转换为 7 万个超像素图，每个图最多包含 75 个节点。节点是超像素，边表示它们之间的空间距离。数字标签共有十种类型。我们采用了基于样条的图神经网络 (Fey 等人，2018年)，在测试集上取得了约 98% 的准确率。

• **场景图分类**：Wang 等人 (2021b) 从 Visual Genome (Krishna 等人，2017年) 中选取了 4443 对图像和场景图，以构建 VG-5 数据集 (Pope 等人，2019年)。每个图都标注为五个类别之一：体育场、街道、农场、冲浪和森林。物体区域表示为节点，边则表示物体节点之间的关系。我们训练了一个 AAPNP (Klicpera 等人，2018年)，其测试准确率达到 61.9%。

我们将 MatchExplainer 与若干最先进且广受欢迎的解释基线方法进行比较，具体如下：

• **SA** (Baldassarre &Azizpour，2019年）直接将模型预测相对于输入图邻接矩阵的梯度作为边的重要性。

• **Grad-CAM** (Selvaraju 等人，2017年；Pope 等人，2019年）利用任何目标概念（例如图中的基元）流入最后一个卷积层时的梯度，生成粗略的定位图，突出显示图中对预测该概念至关重要的区域。

[1]这些结果是通过复现得到的

	MUTAG	VG-5	MNIST	BA-3Motif	
	ACC-AUC	ACC-AUC	ACC-AUC	ACC-AUC	召回率@5
SA	0.769	0.769	0.559	0.518	0.243
Grad-CAM	0.786 ± 0.011	0.909 ± 0.005	0.581 ± 0.009	0.533 ± 0.003	0.212 ± 0.002
GNNExplainer	0.895 ± 0.010	0.895 ± 0.003	0.535 ± 0.013	0.528 ± 0.005	0.157 ± 0.002
PG-Explainer	0.631 ± 0.008	0.790 ± 0.004	0.504 ± 0.010	$\underline{0.586} \pm 0.004$	0.293 ± 0.001
PGM-Explainer	0.714 ± 0.007	0.792 ± 0.001	0.615 ± 0.003	0.575 ± 0.002	0.250 ± 0.000
ReFine	$\underline{0.955} \pm 0.005$	$\underline{0.914} \pm 0.001$	$\underline{0.636} \pm 0.003$	0.576 ± 0.013^{1}	$\underline{0.297} \pm 0.000^{1}$
MatchExplainer	**0.997**	**0.993**	**0.938**	**0.634**	**0.305**
相对提升	4.5%	8.6%	48。9%	8.1%	2.6%

表*1.* 我们提出的 MatchExplainer 与其他基线解释器的比较。

• **GNNExplainer** (Ying 等人，2019年)通过优化边和节点特征的软掩码，最大化原始预测结果与新预测结果之间的互信息。

• **PGExplainer** (Luo 等人，2020年) 使用参数化模型来判定一条边是否重要。该模型使用包含所有边的多个待解释实例进行训练。

• **PGM-Explainer**(Vu 和 Thai，2020年) 收集随机节点扰动所引起的预测变化，然后根据这些扰动—预测观测结果学习贝叶斯网络，以捕捉节点与预测之间的依赖关系。

• **ReFine** (Wang 等人，2021b) 运用预训练和微调的思路开发了一种多粒度 GNN 解释器。它能够全局理解模型的工作方式，并深入洞察特定实例。

由于真实解释通常未知，因此很难定量评估解释的优劣。在此，我们遵循 Wang 等人(2021b) 的做法，并采用**预测准确率（ACC@ρ）**和**Recall@N** 作为评估指标。具体而言，ACC@ρ 通过将解释性子图输入目标模型，并检验模型恢复目标预测的效果，来衡量这些子图的忠实度。ACC-AUC 是不同选择比例下 ACC 曲线下面积的报告值 $\rho \in \{0.1, 0.2, ..., 1.0\}$。Recall@$N$ 的计算方式为 $\mathbb{E}_{\mathcal{G}}\left[|\mathcal{G}_s \cap \mathcal{G}_S^*| \,/\, |\mathcal{G}_S^*|\right]$，其中 $\mathcal{G}_S^*$是真实解释性子图。值得注意的是，Recall@N 仅适用于 BA3-Motif，因为该数据集是合成的，且其中的基元是预先设定的。

5.2. MatchExplainer 能找到更好的解释吗?

量化结果。为考察 MatchExplainer 的有效性，我们在四个数据集上进行了广泛实验，比较结果如表 1所示。对于 MUTAG、VG-5 和 BA3-Motif，我们将全部训练集和验证集数据用作参考集。对于 MNIST，我们随机选取 10% 的可用样本作为参考集，以加快匹配速度。结果表明，MatchExplainer 在所有情况下均优于每一种基线方法。尤其是，先前的解释器无法很好地解释 MNIST 上的图神经网络，其 ACC-AUC 均低于 65%，而 MatchExplainer 的 ACC-AUC 最高可达 93.8%。如果将 MNIST 中全部训练集和验证集数据用作参考集，其 ACC-AUC 还可提升至 97.2%。这一现象表明，当数据集具有明显的解释性子图模式时，子图匹配在解释图神经网络方面具有优势。此外，与最强基线方法相比，MatchExplainer 在 VG-5 和 BA3-Motif 上分别取得了 8.6% 和 8.1% 的显著相对提升。

此外，值得注意的是，除了 BA-3Motif 任务外，MatchExplainer 在每项任务中都实现了近 100% 的 ACC-AUC。对于 BA-3Motif，我们发现其预测准确率会随着选择比例的不同而变化 $[0.31, 0.31, 0.31, 0.34, 0.49, 0.71, 0.97, 1.0, 1.0, 1.0]$。这与该任务中的大多数模体占据整个图中较大比例这一事实相符。当选择比例大于 0.7 时，MatchExplainer 就能找出正确的解释子图。

可视化。此外，我们在附录C中展示了 MatchExplainer 对 MUTAG 的解释，以便进行定性评估。我们还在附录3中将 MatchExplainer 的效率与其他参数化方法进行了比较。结果表明，MatchExplainer 推理速度快，具有竞争力，且无需额外训练，因此适用于大规模部署。

5.3. MatchDrop 能提升图神经网络的性能吗?

实现。我们考虑两种骨干网络：图卷积网络（Kipf & Welling，2016年）和图同构网络（Xu 等人，2018年），深度均为 6。与 Rong 等人 （2019年）类似，我们针对每种架构采用随机超参数搜索，以实现更稳健的比较。其中，*DropNode* 表示随机采样子图，也可以视为节点丢弃的一种特定形式。假阳性丢弃（*FP-Drop*）是我们 MatchDrop 的反向操作，其中子图采样或节点丢弃仅在解释性子图中进行，而其余部分保持不变。

数据集	骨干网络	原始	FPDrop	DropNode	PGDrop	MatchDrop
MUTAG	GCN	0.828 ± 0.004	0.803 ± 0.017	0.832 ± 0.008	0.825 ± 0.02	**0.844±0.006**
	GIN	0.832 ± 0.003	0.806 ± 0.020	0.835 ± 0.009	0.828 ± 0.01	**0.845±0.007**
VG-5	GCN	0.619 ± 0.003	0.587 ± 0.014	0.623 ± 0.007	0.604 ± 0.002	**0.638±0.008**
	GIN	0.621 ± 0.004	0.593 ± 0.018	0.622 ± 0.006	0.600 ± 0.004	**0.630±0.003**
MNIST	GCN	0.982 ± 0.001	0.955 ± 0.008	0.982 ± 0.002	0.975 ± 0.003	**0.986±0.002**
	GIN	0.988 ± 0.001	0.959 ± 0.005	0.989 ± 0.001	0.979 ± 0.002	**0.990±0.001**

表 *2.* 不同骨干网络使用和不使用 MatchDrop 时的测试准确率（%）对比。

我们加入 FPDrop 作为基线方法，以帮助揭示 MatchDrop 有效的原因。*PGDrop* 与 MatchDrop 类似，但使用固定的 PGExplainer（Luo 等人，2020年）来探索信息丰富的子结构。选择比例 ρ 对于 FPDrop、PGDrop 和 MatchDrop 均设为 0.95。

总体结果。 表2 展示了除 BA-3Motif 外所有数据集上的性能，因为该数据集的测试准确率已接近 100%。可以观察到，MatchDrop 在所有情况下都能持续提高测试准确率。值得注意的是，FPdrop 对图神经网络的性能产生了负面影响。这表明，假阳性采样会损害传统的图增强方法，而我们提出的 MatchDrop 能有效克服这一问题。另一方面，PGDrop 也会导致准确率下降。一个可能的原因是，PGExplainr 这类参数化解释器是在图神经网络能够正确预测的样本上训练的，因此无法在图神经网络预测错误的未见图上探索解释性子图。

6. 相关工作

6.1. GNN 可解释性

可解释性和特征选择对于揭示复杂深度学习模型的奥秘日益重要，人们对解释图神经网络的兴趣也不断增长（Ying 等人，2019年；Wu 等人，2022）。尽管取得了丰硕进展，与图像和自然语言领域相比，这一领域的研究仍然不足。总体而言，研究主要沿着两条主流路线展开。如今广泛采用的是参数化解释方法。这类方法运行参数化模型，以挖掘信息丰富的子结构或生成显著性图。例如，GNNExplainer（Ying 等人，2019年）为每个实例学习软掩码，并将其应用于邻接矩阵。PGExplainer（Luo 等人，2020年）利用概率图生成模型对多个样本进行整体解释。XGNN（Yuan 等人，2020a）利用图生成器输出按类别划分的图模式，以解释各类别的图神经网络。PGM-Explainer（Vu 和 Thai，2020年）针对图扰动与预测变化的配对关系提出了一种贝叶斯网络。另一条路线则是非参数化解释方法，这类方法不涉及任何额外的可训练模型。它们采用一些启发式方法，例如通过反向传播获得类似梯度的分数，并将其作为特定实例的特征贡献（Baldassarre 和 Azizpour，2019年；Pope 等人，2019年；Schnake 等人，2020）。如前所述，后者通常不太受青睐，因为其性能远逊于前述参数化方法。相比之下，我们的 MatchExplainer 竟然取得了最先进的成果。

6.2. 图增强

近年来，数据增强在图表示学习中日益受到关注，用于应对数据噪声和数据稀缺等问题（Zhao等人，2022）。相关工作根据增强模态（Ding等人，2022）大致可分为特征级（Zhang等人，2017；Liu等人，2021b；Taguchi等人，2021）、结构级（You等人，2020；Zhao等人，2021a）和标签级（Verma等人，2019）三类。其中，许多研究致力于增强图结构。与添加或删除边（Xu 等人，2022）相比，针对节点集的增强操作更加复杂。一种典型应用是通过插入超节点来促进整个图的传播（Gilmer等人，2017），而 Zhao 等人（2021b）则对节点进行插值，以丰富少数类。相反，一些方法通过丢弃节点来实现图或子图采样，以达到不同目的，例如扩展图神经网络（Hamilton等人，2017）、实现对比学习（Qiu等人，2020），以及抑制过拟合和过平滑（Rong 等人，2019）。然而，这些图采样或节点丢弃方法中，很少有方法能够从输入图中找到既能最大程度保留原始属性、又经过增强的图实例。

7. 结论

本文提出了一种前景广阔的子图匹配技术 MatchExplainer，用于图神经网络解释。不同于使用缺乏可解释性的参数化网络这一流行趋势，我们设计了一种非参数算法，在一对图之间搜索信息量最大的联合子图，并提供理论保证。此外，我们将 MatchExplainer 与经典的

图增强方法相结合，并展示了其在缓解假阳性采样难题方面的强大能力。实验有力地证明，MatchExplainer 的效果显著优于参数化方法。我们希望这项工作能够推动非参数方法在解释深度学习模型方面的发展。

参考文献

Achille, A. 和 Soatto, S. 深度表征中不变性与解耦的涌现。机器学习研究期刊，19(1):1947–1980，2018年。

¨Ancona, M., Ceolini, E., Oztireli, C., 和 Gross, M. 为了更好地理解深度神经网络的基于梯度的归因方法。*arXiv* 预印本 *arXiv:1711.06104*，2017年。

Baldassarre, F. 和 Azizpour, H. 图卷积网络的可解释性技术。*arXiv* 预印本 *arXiv:1905.13686*，2019年。

Bang-Jensen, J., Gutin, G., 和 Yeo, A. 贪心算法何时失效。离散优化，1(2):121–127，2004年。

Banjade, H. R., Hauri, S., Zhang, S., Ricci, F., Gong, W., Hautier, G., Vucetic, S. 和 Yan, Q. 面向无机晶体系统的结构基元中心学习框架。《科学进展》，7(17):eabf1754，2021年。

Berretti, S., Del Bimbo, A. 和 Vicario, E. 基于内容检索中图模型的高效匹配与索引。*IEEE* 模式分析与机器智能汇刊，23(10):1089–1105，2001年。

Chen, J., Song, L., Wainwright, M. 和 Jordan, M. 学习如何解释：模型解释的信息论视角。载于国际机器学习会议，第883–892页。PMLR，2018年。

Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J. 和 Hansch, C. 致突变性芳香族和杂芳香族硝基化合物的构效关系：与分子轨道能量和疏水性的相关性。《药物化学杂志》，34(2):786–797，1991年。

Ding, K., Xu, Z., Tong, H., and Liu, H. 深度图学习的数据增强：综述。*arXiv*预印本 *arXiv:2202.08235*，2022。

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, E., Tang, J., and Yin, D. 用于社交推荐的图神经网络。载于万维网大会，第417–426页，2019年。

Feng、W.、Zhang、J.、Dong、Y.、Han、Y.、Luan、H.、Xu、Q.、Yang、Q.、Kharlamov、E. 和 Tang、J. 图随机神经网络用于图上的半监督学习。神经信息处理系统进展，33：22092–22103，2020年。Fey、M.、Lenssen、J. E.、Weichert、F. 和 M¨uller、H. Splinecnn：利用连续B样条核进行快速几何深度学习。载于 *IEEE*计算机视觉与模式识别会议论文集，第869–877页，2018年。

Geng, X., Li, Y., Wang, L., Zhang, L., Yang, Q., Ye, J., and Liu, Y. 用于网约车需求预测的时空多图卷积网络。载于*AAAI*人工智能会议论文集，第33卷，第3656–3663页，2019年。

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. 用于量子化学的神经消息传递。载于国际机器学习会议，第1263–1272页。PMLR，2017年。

Hamilton, W., Ying, Z., 和 Leskovec, J. 在大型图上进行归纳式表示学习。神经信息处理系统进展，30，2017年。

Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C., 和 Li, L. Rolx：大型图中的结构角色提取与挖掘。载于第*18*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第 1231–1239 页，2012年。

Kazius, J., McGuire, R., 和 Bursi, R. 用于致突变性预测的毒效团推导与验证。《药物化学杂志》，48(1):312–320，2005年。

Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。

Kipf, T. N. 和 Welling, M. 使用图卷积网络进行半监督分类。*arXiv* 预印本 *arXiv:1609.02907*，2016年。

Klicpera, J.、Bojchevski, A. 和 G¨unnemann, S. 先预测，再传播：图神经网络与个性化PageRank相结合。*arXiv* 预印本 *arXiv:1810.05997*，2018年。

Knyazev, B.、Taylor, G. W. 和 Amer, M. 理解图神经网络中的注意力与泛化。神经信息处理系统进展，第32卷，2019年。

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., 等。Visual Genome：利用众包密集图像标注连接语言与视觉。国际计算机视觉杂志，123(1):32–73，2017年。

Leordeanu, M. 和 Hebert, M. 一种利用成对约束解决对应问题的谱技术。2005年。

Li, Y., Gu, C., Dullien, T., Vinyals, O. 和 Kohli, P. 用于学习图结构对象相似度的图匹配网络。载于国际机器学习会议，第 3835–3845 页。PMLR，2019年。

Liu, L., Hughes, M. C., Hassoun, S. 和 Liu, L. 随机迭代图匹配。载于 国际机器学习会议，第 6815–6825 页。PMLR，2021a。

Liu, S.,Dong, H., Li, L., Xu, T., Rong, Y., Zhao, P., Huang, J. 和 Wu, D. 面向图神经网络的局部增强。*arXiv* 预印本 *arXiv:2109.03856*，2021b。

Loiola, E. M., de Abreu, N. M. M., Boaventura-Netto, P.O., Hahn, P. 和 Querido, T. 二次指派问题综述。《欧洲运筹学杂志》，176(2):657–690，2007年。

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H. 和 Zhang, X. 图神经网络的参数化解释器。神经信息处理系统进展，33:19620–19631，2020年。

Papakis, I., Sarkar, A. 和 Karpatne, A. GCNNMatch：基于Sinkhorn归一化的多目标跟踪图卷积神经网络。*arXiv*预印本 *arXiv:2010.00067*，2020年。

Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E. 和 Hoffmann, H. 图卷积神经网络的可解释性方法。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10772–10781页，2019年。

Qiu, J., Chen, Q., Dong, Y., Zhang, J., Yang, H., Ding, M., Wang, K. 和 Tang, J. GCC：用于图神经网络预训练的图对比编码。载于第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第1150–1160页，2020年。

Ranjan, E., Sanyal, S. 和 Talukdar, P. ASAP：用于学习层次化图表示的自适应结构感知池化。载于*AAAI*人工智能会议论文集，第34卷，第5470–5477页，2020年。

Raymond, J. W., Gardiner, E. J. 和 Willett, P. Rascal：利用最大公共边子图计算图相似性。《计算机杂志》，45(6)：631–644，2002年。

Rong, Y., Huang, W., Xu, T., and Huang, J. Dropedge：面向节点分类的深层图卷积网络。*arXiv* 预印本 *arXiv:1907.10903*，2019年。

Sarlin, P.-E., DeTone, D., Malisiewicz, T., and Rabinovich, A. Superglue：利用图神经网络学习特征匹配。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 4938–4947 页，2020年。Schlichtkrull, M., Kipf, T. N., Bloem, P., Berg, R. v. d., Titov, I., and Welling, M. 利用图卷积网络建模关系数据。载于欧洲语义网会议，第 593–607 页。施普林格，2018年。Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Sch¨utt, K. T., M¨uller, K.-R., and Montavon, G. 通过相关游走解释图神经网络的高阶解释。*arXiv* 预印本 *arXiv:2006.03589*，2020年。

Schwab, P. 和 Karlen, W. Cxplain：不确定性条件下用于模型解释的因果解释。神经信息处理系统进展，32，2019年。

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. 和 Batra, D. Grad-CAM：通过基于梯度的定位，从深度网络中获得视觉解释。载于*IEEE*国际计算机视觉会议论文集，第618–626页，2017年。

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., ˇZ´ıdek, A., Nelson, A. W., Bridg- land, A., 等. 利用深度学习得出的势能改进蛋白质结构预测。《自然》，577(7792)：706–710，2020年。

Shasha, D., Wang, J. T. 和 Giugno, R. 树与图搜索的算法学及应用。载于第二十一届 *ACM SIGMOD-SIGACT-SIGART*数据库系统原理研讨会论文集，第39–52页，2002年。

Shorten, C. 和 Khoshgoftaar, T. M. 深度学习图像数据增强综述。《大数据杂志》，6(1)：1–48，2019年。

Sundararajan, M., Taly, A. 和 Yan, Q. 深度网络的公理化归因。载于国际机器学习会议，第3319–3328页。PMLR，2017年。

Taguchi, H.、Liu, X. 和 Murata, T. 面向包含缺失特征的图的图卷积网络。*FutureGeneration Computer Systems*，117:155–168，2021年。

Tishby, N. 和 Zaslavsky, N. 深度学习与信息瓶颈原理。载于*2015 IEEE*信息论研讨会 *(ITW)*，第 1–5 页。IEEE，2015 年。

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D. 和 Bengio, Y. 流形混合：通过插值隐藏状态获得更好的表示。载于 国际机器学习会议，第 6438–6447 页。PMLR，2019年。

Vu, M. 和 Thai, M. T. PGM-Explainer：用于图神经网络的概率图模型解释。神经信息处理系统进展，33：12225–12235，2020年。

Wang, H., Guo, X., Deng, Z.-H. 和 Lu, Y. 重新思考对比学习中的最小充分表示。*arXiv* 预印本 *arXiv:2203.07004*，2022a。

Wang, R., Yan, J. 和 Yang, X. 学习用于深度图匹配的组合嵌入网络。载于*IEEE/CVF*国际计算机视觉会议论文集，第 3056–3065 页，2019年。

Wang, R., Yan, J. 和 Yang, X. 鲁棒深度图匹配的组合学习：一种基于嵌入的方法。*IEEE*《模式分析与机器智能汇刊》，2020年。

Wang, R., Yan, J. 和 Yang, X. 神经图匹配网络：求解Lawler二次指派问题，并扩展至超图和多图匹配。*IEEE*《模式分析与机器智能汇刊》，2021a。

Wang, X., Wu, Y., Zhang, A., He, X., 和 Chua, T.-S. 面向图神经网络的多粒度可解释性。神经信息处理系统进展，34，2021b。

Wang, X., Wu, Y., Zhang, A., Feng, F., He, X., 和 Chua, T.- S. 面向图神经网络的强化因果解释器。*IEEE* 模式分析与机器智能汇刊，2022b。

Willett, P., Barnard, J. M., 和 Downs, G. M. 化学相似性搜索。《化学信息与计算机科学杂志》，38(6):983–996，1998。

Wu, F., Li, S., Wu, L., Radev, D. R., 和 Li, S. Z. 通过多阶交互发现并解释图神经网络的表示瓶颈。*arXiv* 预印本 *arXiv:2205.07266*，2022。

Wu, F., Radev, D. 和 Li, S. Z. Molformer：面向三维异质分子图的基于基序的Transformer。*Rn*，1:1，2023年。

Xu, D., Cheng, W., Luo, D., Chen, H. 和 Zhang, X. In-fogcl：信息感知图对比学习。神经信息处理系统进展，34，2021年。

Xu, K.、Hu, W.、Leskovec, J. 和 Jegelka, S. 图神经网络的表达能力有多强？*arXiv*预印本 *arXiv:1810.00826*，2018年。

Xu, Z.、Du, B. 和 Tong, H. 图清理及其在节点分类中的应用。载于*ACM Web Conference 2022* 论文集，第 1136–1147 页，2022。Ying, Z.、Bourgeois, D.、You, J.、Zitnik, M. 和 Leskovec, J. GNNExplainer：为图神经网络生成解释。神经信息处理系统进展，32，2019。You, Y.、Chen, T.、Sui, Y.、Chen, T.、Wang, Z. 和 Shen, Y. 通过增强进行图对比学习。神经信息处理系统进展，33：5812–5823，2020。

Yuan, H., Tang, J., Hu, X. 和 Ji, S. Xgnn：迈向图神经网络的模型级解释。在第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第430–438页，2020a。

Yuan, H., Yu, H., Gui, S. 和 Ji, S. 图神经网络的可解释性：分类综述。*arXiv*预印本 *arXiv:2012.15445*，2020b。

Yuan, H., Yu, H., Wang, J., Li, K., andJi, S. 通过子图探索理解图神经网络的可解释性。载于国际机器学习会议，第 12241–12252 页。PMLR，2021年。

Zanfir, A. 和 Sminchisescu, C. 图匹配的深度学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 2684–2693 页，2018年。

Zhang, H., Cisse, M., Dauphin, Y. N., 和 Lopez-Paz, D. mixup：超越经验风险最小化。*arXiv* 预印本 *arXiv:1710.09412*，2017年。

Zhang, S., Hu, Z., Subramonian, A., 和 Sun, Y. 基于基元的图表示对比学习。*arXiv*预印本 *arXiv:2012.12533*，2020年。

Zhao, T., Liu, Y., Neves, L., Woodford, O., Jiang, M., 和 Shah, N. 图神经网络的数据增强。载于*AAAI* 人工智能会议论文集，第 35 卷，第 11015–11023 页，2021a。

Zhao, T., Zhang, X., 和Wang, S. Graphsmote：基于图神经网络的图不平衡节点分类。载于第 *14* 届 *ACM*网络搜索与数据挖掘国际会议论文集，第 833–841 页，2021b。

Zhao, T., Liu, G., G¨unnemann, S. 和 Jiang, M. 图机器学习的图数据增强：综述。*arXiv* 预印本 *arXiv:2202.08871*，2022。

A. 子图匹配贪心算法

这里给出了贪心方式进行子图匹配的伪代码。给定两个图 $\mathcal{G}$ 和 $\mathcal{G}'$，我们首先将它们输入图神经网络 h_Y，并获得各自对应的节点特征，分别记为 $\{\mathbf{h}_i\}_{i\in\mathcal{V}}$ 和 $\{\mathbf{h}'_i\}_{i\in\mathcal{V}'}$。我们的目标是找到子图 $\mathcal{G}_S$ 和 $\mathcal{G}'_S$，使得 $d_G(\mathcal{G}_S,\ \mathcal{G}'_S)$ 最小。

算法 2探索共享子图的贪心算法

输入： 节点特征 $\{\mathbf{h}_i\}_{i\in\mathcal{V}}$ 和 $\{\mathbf{h}'_i\}_{i\in\mathcal{V}'}$，子图大小 K
初始化一个空列表 g，用于存储所选节点。
计算距离矩阵 $\mathbf{D}^{\phi x}$，其中 $\mathbf{D}^{\phi x}_{ij} = d_X(\mathbf{h}'_i,\ \mathbf{h}'_j)\forall i \in \mathcal{V},\ \forall j \in \mathcal{V}'$。
对 $t = 1, ..., K$ **do**
$(i,\ j) \leftarrow$ 最小值$_{i\in\mathcal{V},\ j\in\mathcal{V}'}\ \mathbf{D}^{\phi x}_{ij}$,
将 i 添加到 g,
从 $\mathcal{V}$中移除 i ,
从 $\mathcal{V}'$中移除 j ,
结束循环
返回： g

B. 实验细节与补充结果

解释图神经网络。所有实验均在单块 A100 PCIe GPU（40GB）上进行。对于包含 GNNExplainer、PGExplainer、PGM-Explainer 和 ReFine 的参数化方法，我们采用 Wang 等人（2021b）中报告的性能。关于在 BA-3Motif 上重新实现 ReFine，我们使用原始代码和相同的超参数，并采用 Adam 优化器（Kingma 和 Ba，2014），将预训练和微调的学习率分别设为 1e-3 和 1e-4。

图增强。所有实验也均在单块 A100 PCIe GPU（40GB）上实现。我们采用三种不同的图神经网络变体（图卷积网络、图注意力网络和图同构网络）来拟合这些数据集，并验证各种图增强方法的有效性。我们使用 Adam 优化器训练模型。对于 MUTAG，批量大小为 128，学习率为 1e-3。对于 BA3-Motif，批量大小为 128，学习率为 1e-3。对于 VG-5，批量大小为 256，学习率为 0.5 * 1e-3。所有数据集的训练轮次均固定为 100。

效率研究。 我们计算了不同方法在每个数据集上为单个图生成解释所需的平均推理时间，并统计了训练和推理的总耗时，结果汇总于表 3。具体而言，我们对 GNNExplainer 和 PG-Explainer 训练 10 个训练轮次，并在评估前对 ReFine 进行 50 个训练轮次的预训练。可以看出，尽管先前的方法推理速度较快，但训练阶段耗时较长。作为替代方案，我们的 MatchExplainer 完全无需训练。比较总耗时可知，MatchExplainer 在 MUTAG、VG-5 和 MNIST 上的计算开销最低。然而，由于 BA-3Motif 中的大多数基元尺寸较大，MatchExplainer 必须遍历大型参考集以获取合适的对应图，这不可避免地导致耗时大幅增加。

C. 图分类模型的解释

本节展示图 3 中解释的可视化结果。

表 3. 不同方法的效率研究（单位：秒）。

方法	阶段	MUTAG	VG-5	MNIST	BA-3Motif
GNNExplainer	训练	186.0	1127.2	1135.4	66.1
	推理（每个图）	1.290	0.565	0.732	0.517
	训练 +推理（总计）	703.4	1644.6	1782.1	271.6
PG-Explainer	训练	186.3	286.3	1154.1	112.4
	推理（每个图）	0.056	0.094	0.025	0.020
	训练 +推理（总计）	208.6	309.5	1162.1	**120.4**
ReFine	训练	1191.6	1933.3	5025.8	763.0
	推理（每个图）	0.068	0.107	0.026	0.027
	训练 +推理（总计）	1218.9	1959.7	5051.2	773.8
MatchExplainer	训练	–	–	–	–
	推理（每个图）	0.485	0.732	0.682	7.687
	训练 +推理（总计）	**194.6**	**180.3**	**667.8**	224.7

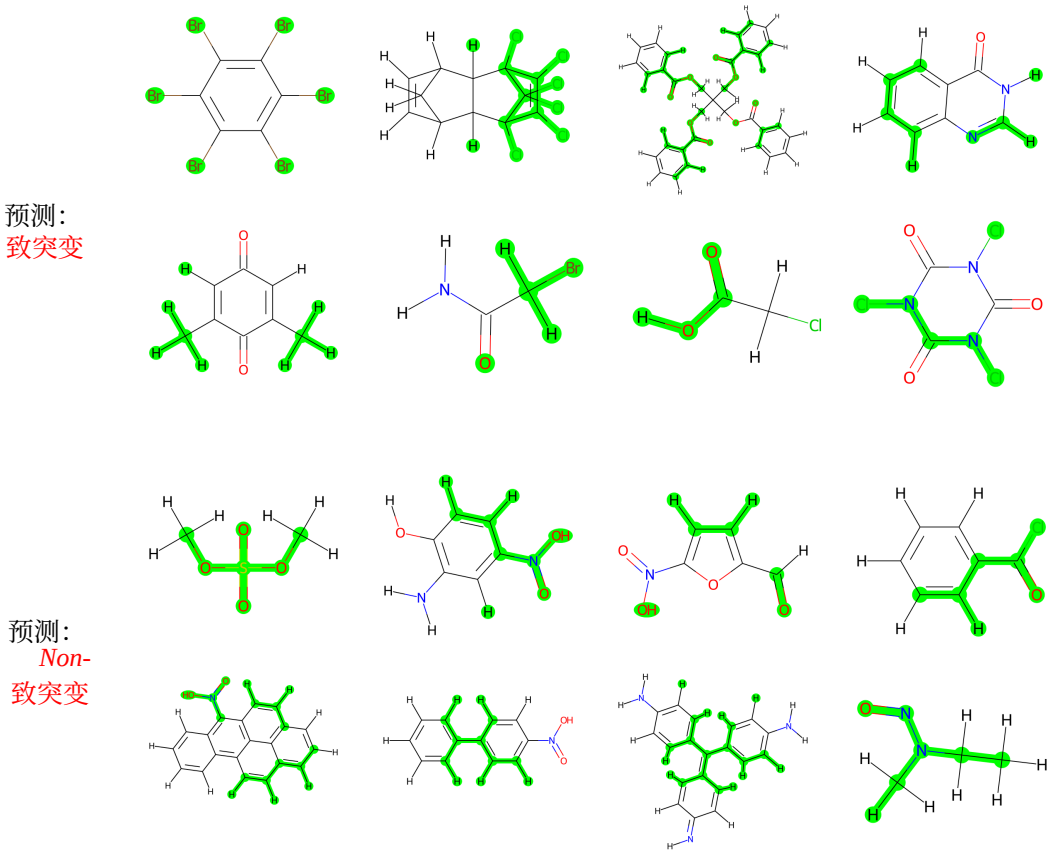


图 3。 Mutagenicity 中的解释性子图，其中 50% 的节点被突出显示。