图神经网络算法架构及可解释性研究分析

刘 杰[1] 王 敏[2] 唐青梅[1] 张萌月[2]
LIU Jie WANG Min TANG Qingmei ZHANG Mengyue

摘 要 首先，全面概述了图神经网络的基本概念和意义，对基于谱的图神经网络、基于空域的图神经网络和其他图神经网络三个类别进行了系统性介绍，同时总结了训练过程中常用的优化技术。其次，分析了图神经网络领域目前面临的可解释性问题的挑战，从而明确了研究目标。随后，阐述了图形神经网络的可解释性基本概念与基础理论，将其分类为实例级和模型级可解释性技术，并列出了评估图学习方法可解释性的度量指标。最后，在结语部分总结了本文的主要研究脉络，并对该领域的未来研究方向提出了建议。旨在介绍图神经网络的理论基础以及其在可解释性领域的研究。

关键词 深度学习；图神经网络；应用与挑战；可解释性方法；可解释性度量；谱卷积；空域卷积

doi：10.3969/j.issn.1672-9528.2024.01.039

0 引言

最近，由于图结构强大而灵活的建模性质，图神经网络（GNNs）已成为分析结构化数据的一种常用的工具。与传统的神经网络在表格或顺序数据上操作不同，图神经网络可以直接在包含拓扑信息的图结构数据上进行操作，如社会网络、分子结构或引文网络。如图 1 表示社交网络中的连接关系图。这些涉及理解实体之间的依赖关系的数据中，GNNs 利用深度学习方法根据连接关系学习图结构的特征，已成功应用于广泛的任务，例如节点分类、链接预测、推荐系统、自然语言处理和计算机视觉。GNNs 的意义在于，它们能够捕捉图中的局部和全局信息，使它们能够学习到图结构和节点之间的敏感表示。这可以提高许多与图有关的任务的性能，并为复杂系统的结构和功能提供新的解决方案。但是，黑箱模型背后的可解释性问题依然有很大的探索空间。

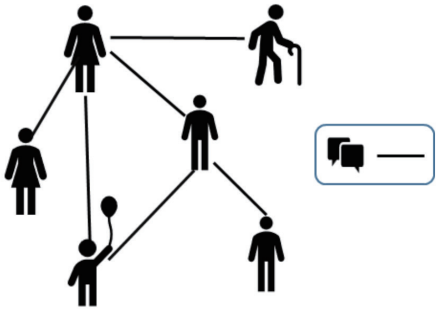

图 1 社交网络图数据

本文旨在全面概述当前 GNNs 的研究理论，会系统介绍 GNNs 的不同架构算法，重点展开对该领域可解释性的挑战的讨论，以探索未来研究的机会，希望在图神经网络领域建立完整的认知框架。

1 相关基础

1.1 图神经网络算法

GNNs将神经网络的黑盒学习方法迁移到图结构数据上。传统的神经网络都是在规则数据的结构上操作的，而 GNNs 可以学习图结构中不规则的拓扑关系。GNNs 的运作方式是根据其邻居的隐藏状态迭代更新每个节点的隐藏状态，使其能够捕捉图中的局部和整体信息。

目前的几种主要 GNNs 架构有：基于频谱的图卷积和基于空域的图卷积以及其他图卷积方法 [1]。基于频谱的卷积网络在频谱域中进行信息的更新，而基于空间的 GNNs 则直接在空间域中聚合相邻之间节点的信息。基于频谱的 GNNs 可以捕捉全局信息，但计算效率不够高，在处理非欧几里得上的数据域时灵活性较差。基于空域的 GNNs 更加灵活，可以处理非欧几里得领域，但可能存在过平滑的问题。

这里介绍几种用于聚合邻近节点信息的算法，包括图卷积网络（GCN）[2]、图注意网络（GAT）[3] 和同构神经网络（GIN）[4]。首先，GCN 是一种使用范围最广泛的基于频谱的卷积学习方法，它计算拉普拉斯矩阵的特征值和向量，并将卷积过滤器应用于相邻节点上完成信息传递。而在基于空域的方法中，GAT 则通过使用注意力机制来权衡相邻节点之间的重要度，按照权重汇集信息。GIN 算法则是一种基于图同构的图学习方法，通过对节点的邻居进行聚合来生成节点的嵌入表示。

1. 山东大学机电与信息工程学院 山东威海 264209
2. 中国电子科技集团公司第五十四研究所 河北石家庄 050081
[基金项目] 山东自然科学基金面上项目 ZR2020MA064

接下来将结合这三种具体的算法来介绍三类图卷积神经网络。

1.1.1 基于频谱的图神经网络

基于频谱的图神经网络一般利用图结构的拉普拉斯算子来捕捉图的拓扑结构。这个算子是一个实对称矩阵，其对角线条目代表每个节点本身的自我程度，而非对角线条目代表节点之间的连接。通过计算拉普拉斯矩阵的特征值和特征向量，基于频谱的 GNNs 可以捕获关于图的全局信息。

GCN 是一种流行的基于谱域变换的 GNNs。GCN 的运作方式是对每个节点的特征向量进行线性转换，然后使用图拉普拉斯矩阵聚合相邻节点的特征向量。上述过程可以重复进行以构建多层的卷积，使网络在特征维度的变化中不断地捕捉到节点之间不同复杂程度的链接。对于给定一个图 $G=(V, E)$，其中包含节点集合 V 及边集合 E。假设 $\boldsymbol{X}\in R^{n\times d}$ 表示节点的特征矩阵，$\boldsymbol{A}\in R^{n\times n}$ 表示图的邻接矩阵，它可以表示图结构中节点之间的拓扑连接。在第 l+1 层 GCN 中信息传递过程可以表示为：

$$\boldsymbol{H}^{(l+1)}=f\left(\boldsymbol{A},\boldsymbol{X}^{(l)},\boldsymbol{W}^{(l)}\right)=\sigma\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}^{(l)}\boldsymbol{W}^{(l)}\right) \quad (1)$$

式中：每一层的特征嵌入由矩阵 $\boldsymbol{H}^{(l+1)}$ 表示，而在输入层的初始特征为节点的特征矩阵 $\boldsymbol{X}$。$\tilde{\boldsymbol{A}}=\boldsymbol{I}+\boldsymbol{A}$，$\boldsymbol{I}$ 为单位矩阵，对角线 $\hat{D}^{-\frac{1}{2}}$ 为 $\boldsymbol{A}$ 的行标准化版本，$\boldsymbol{W}^{(l)}\in R^{d1\times d2}$ 表示第 l 层中学习的映射权重矩阵，$\sigma(\cdot)$ 为激活函数，将结果映射到某种分布上。

图 2 是使用一层基于谱域的 GCN 卷积方法对图 1 中图模型的操作过程。首先通过与 $\boldsymbol{W}^{(l)}$ 运算实现线性映射对特征维度进行降维，然后与对称归一化后的拉普拉斯矩阵 $\boldsymbol{L}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}$ 进行运算实现信息在谱域上的信号转换。

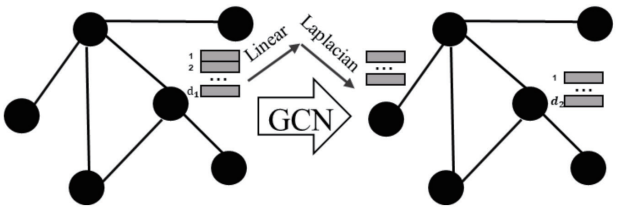

图 2 基于谱域的 GCN 卷积方法

1.1.2 基于空域的图神经网络

基于空域的图神经网络通过聚合相邻节点的信息在空间域进行操作。与基于光谱的图神经网络通过拉普拉斯矩阵一次性对整个图进行操作不同的是，基于空域的图神经网络对每个节点进行单独操作，并根据其邻居的隐藏状态来更新每个节点的隐藏状态。

GAT 是一种典型的基于空域的 GNNs 方法。GAT 通过使用注意力机制对相邻节点的贡献进行加权，使 GAT 能够捕获节点之间更复杂的关系。具体来说，首先，它根据学习到的权重矩阵和非线性激活函数，为每一对连接关系计算出所连接的两个点之间的注意力系数。然后，在更新每个节点的隐藏状态时，注意系数被用作加权系数作为相邻节点的贡献度。GAT 模型的节点嵌入可以表示为：

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\sum_{i=1}^{n}\sum_{j\in N_i}a_{ij}^{(l)}\boldsymbol{W}^{(l)}h_i^{(l)}\right) \quad (2)$$

式中：l 表示第 l 层神经网络，N_i 表示与节点 i 相邻的节点集合，$h_i^{(l)}$ 表示节点 i 在第 l 层的嵌入表示。

在式（2）中，第 l 层中的节点 i 和节点之间的权重值 $a_{ij}^{(l)}$，可以表示为：

$$a_{ij}^{(l)}=\frac{\exp\left(\text{LeakyReLU}\left(f\left(\left[W^{(l)}h_i^{(l)}\,\|\,W^{(l)}h_j^{(l)}\right]\right)\right)\right)}{\sum_{k\in N_i}\exp\left(\text{LeakyReLU}\left(f\left(\left[W^{(l)}h_i^{(l)}\,\|\,W^{(l)}h_k^{(l)}\right]\right)\right)\right)} \quad (3)$$

式中：$\|$ 表示矩阵拼接操作，f 表示单层的简单前馈神经网络，LeakyReLU 表示具有负斜率的 *ReLU* 函数，用于防止梯度消失。注意力权重的计算将节点 i 和节点 j 的特征向量进行拼接作为分子，比值表达了节点 j 在所有与节点 i 邻接的节点的重要程度。同时，计算是基于 *Softmax* 的，即每个节点的权重值是在所有邻居节点的权重值上进行归一化的。图 3 是 GAT 方法中信息的传递过程，A 点与直接相邻的 B、C、D 点按照它们之间的权重值来不同程度地传递信息。

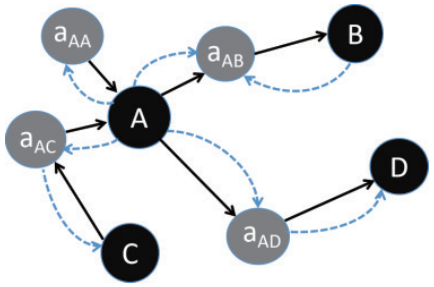

图 3 基于空域的 GAT 卷积方法

1.1.3 其他图神经网络

除了 GCN 和 GAT 之外，最近还有一些文献中提出了其他的图神经网络模型。一个比较经典的模型是 GAC（graph attention convolution，GAC）[5]，它结合了 GAT 的注意力机制和 GCN 的卷积过滤器。GAC 通过计算相邻节点的注意力系数，然后对产生的特征向量应用卷积滤波器来生成最后的聚合信息。

另一个较受欢迎的模型是图形同构网络（GIN），GIN 算法则是一种基于图同构的图神经网络，通过对节点的邻居进行聚合来生成节点的嵌入表示。它通过使用一个加和函数以及一个具有共享参数的多层感知器（MLP）来聚合相邻节点的特征向量。GIN 可以被看作是 GCN 和 GraphSAGE[6] 的泛化，已被证明在某些任务上优于这些模型。其中 GraphSAGE 是一种基于空域的方法，它通过对相邻节点进行采样并计算每个节点的固定长度的嵌入来聚合信息。

1.2 图神经网络训练

在处理不规则的图数据时，图神经网络的强大表示学习能力让它们成了重要的深度学习工具。然而，由于图结构本身的复杂性，图神经网络的训练过程面临着挑战。为了确保其有效性和鲁棒性，一些特殊技巧被提出，例如采样方法、参数共享和半监督学习等。这些技巧可以帮助更好地训练图神经网络。

基于采样的方法被广泛应用于处理大规模图数据，因为这些方法可以从原始图中随机采样一些子图来减少计算复杂度，同时保持原始图的结构特征。例如，FastGCN[7] 使用随机游走采样的方法来构建子图，从而加速图卷积层的计算。

参数共享的策略可以被用来处理具有不同大小和形状的图。GCN 和 GraphSAGE 等方法使用参数共享的策略对图中的节点进行聚合，从而降低模型的计算复杂度。在进行不同图结构的信息预测时，这些方法都使用了公共的权重矩阵来聚合不同的节点特征。

在半监督学习中，利用已知标签的节点来训练模型，并利用未标记节点的信息来提高模型的预测能力。半监督学习的目标是利用带标签的节点来训练一个分类器，然后使用该分类器来预测未标记节点的标签，可以通过将已标记节点的标签作为监督信号来实现。与监督学习不同，半监督学习在训练时不需要所有节点都带有标签，因此可以在数据集较少或标记成本较高的情况下实现良好的性能。例如，DropEdge[8] 使用通过在图结构中的节点特征向量上添加噪声来提高模型的鲁棒性和泛化性能。

1.3 应用与挑战

这些模型的性能与任务和数据集密切相关，一些模型在某些任务上的表现比其他模型更优。图谱神经网络已被证明是一种通用而强大的方法，可在各种现实世界的应用中分析和处理数据。本文将讨论一些效果显著的 GNNs 应用案例，并探讨研究面临的挑战。

在推荐系统中，GNNs 被用来根据用户过去的行为向其推荐产品、服务或内容。例如，在娱乐平台上，GNNs 已被用来向用户推荐电影、歌曲或书籍；在电子商务网站上，它可向客户推荐可能感兴趣的产品；在交通预测和管理中，GNNs 可分析交通流的复杂模式，检测瓶颈和拥堵点，并优化车辆路线；在医疗保健中，GNNs 的应用目标是诊断疾病、预测健康结果和个性化治疗，分析病人数据，识别模式和相关性，并评估干预措施的有效性。

从推荐系统到医疗保健，这些研究证明了 GNNs 在各种现实世界应用中的有效性。相比于传统的机器学习方法，在这些应用中 GNNs 具有独特的优势，例如能够对图中实体之间的复杂关系和依赖关系进行建模。尽管最近 GNNs 的研究取得了很多进展，但仍有一些差距和局限性需要在未来的工作中加以解决，其中一个重要的挑战是 GNNs 模型的可解释性问题。由于图结构的复杂性和非线性，GNNs 学习到的信息可能变得难以理解。例如，在招聘决策中，公司已经开始使用 GNNs 来分析简历和工作申请，以帮助他们筛选候选人并确定最适合的职位。然而，这种方法带来一些负面影响。由于模型可能从原本带有偏见的数据中学习，因此在决策时可能会选择符合特定条件的候选人而延续了这些偏见，不能选择最合适的候选人。此外，在招聘决策中使用 GNNs 可能导致缺乏问责制和透明度。因此，GNNs 能否真正地应用于实际取决于对可解释性问题的研究。

2 图神经网络的可解释性

尽管深度学习模型的准确率在不断地提高，但是这种黑箱模型的可解释性逐渐下降。在医学领域中，可解释性问题关乎用户切身的生命安全，而深度学习因其黑箱机制受到实际应用的排斥。若不关注模型内部形成的小概率对象，则会对医疗的诊断造成极大隐患。由于深度学习模型缺乏对预测结果背后底层机制的推理，影响了使用者理解和信任，其合理性一般的可解释性方法是使用注意力机制衡量图中对预测而言重要的部分，或是使用可视化技术来帮助直观地理解图的学习过程。但是这些方法对理论的探究都不够深入。

目前，有很多对于图深度学习模型的可解释性问题的工作，这些方法从模型中不同角度的信息来分析理解和解释。这些方法通常从下面几个问题考虑实现对图模型进行解释，包括：对于给定的图结构数据，如何识别其中的哪些边、节点以及特征对模型表现来说是更重要的？以及哪种图模式可以最优地精准预测某个类别？在文献 [9] 中，图模型的可解释性技术被分为实例级方法与模型级方法两个大类。其中实例级解释方法中从输入的具体样例角度解释学习到数据中的重要性出发，而模型级方法则从方法本身的角度讨论高层次的运行机制分解及其普遍性的规律理解。接下来的内容将按照图 4 中的 GNNs 可解释性技术分类框架展开讲解。

图 4 图神经网络可解释性分类框架

2.1 实例级方法

实例级方法的目的是找到输入图样例中对预测结果影响显著的特征，基于输入样例来解释每个图数据。它从输入的个体数据中探究给定图的重要特征来解释 GNNs 的预测结果。根据模型对输入图的重要特征的衡量与识别方式，实例级可解释性方法可以分为下面四种类型。

2.1.1 基于梯度 / 特征的可解释性方法

在图像和文本任务中，常用的模型解释方法是使用梯度或者是特征的变化来间接地计算不同输入图数据的特征重要度。这种方法的核心思想是认为隐层中特征图的值的梯度变化包含了对应输入特征的重要性信息。这类方法主要在梯度变化反馈给网络的过程中，将所有隐藏特征图结合的策略上存在区别。

SA[10] 认为当绝对梯度变化值越大时，对应位置表示的特征就应当越重要，它简单地通过计算梯度的平方值来作为不同输入节点 / 边 / 节点特征的重要性得分。这种梯度的计算只能对预测产生局部的解释，而忽略了全局的重要信息，而且，在实际应用过程中 SA 还存在一些其他的局限。首先，SA 只能通过观察输出结果对输入图数据的灵敏度来间接地反应某些特征的重要性，这样不能直接准确地给出这些特征的重要程度。其次，该方法存在饱和问题，即在解释模型的饱和区位置上，重要的输入图特征发生细微的变化时，梯度很难反映出对应特征对模型的显著影响。因此，在使用 SA 时需要注意模型在训练空间的位置，以便更准确地评估不同输入特征的重要性。

2.1.2 基于扰动的可解释性方法

通过生成大量的噪声数据可以测试模型是否学习到输入图数据中的重要特征。观察每种类型的输入扰动下，模型的解释预测值的差异情况。若输入图数据中的重要特征未被扰动，好的解释模型中的预测结果应与未扰乱前的预测结果近似。在现有的计算机视觉领域中，一般学习生成掩码来选择重要的输入像素以解释深度图像模型。但是，由于图像的像素数据之间是一种规则的连接结构，而图数据的结构是包含拓扑信息的不规则结构，这种方法不能直接应用于图结构数据。图数据是由一组节点和一组边的组合来表示，结构信息是图结构数据中至关重要的信息，这些连接关系决定了图数据的根本功能。

对于图模型，扰动可以通过添加 / 删除节点和边来实现。一些工作使用节点 / 边删除操作来生成输入扰动，以评估每个节点 / 边的重要性。这些方法通过逐步删除节点 / 边并重新计算模型预测值来获得特征重要性分数，它们都非常直观且易于实现。但是，在大规模图上计算时，这些方法计算成本非常高，因为需要对每个节点 / 边进行单独计算。另外，对于那些大规模图，这种基于扰动的方法存在着无法有效处理大量扰动的缺陷。

2.1.3 基于分解的可解释性方法

基于分解的方法用 GNNs 学习到的参数信息来衡量输入图中节点 / 边 / 节点特征的重要程度，以解释 GNNs 对于输入图结构的预测结果。由于训练网络的最后一个隐藏层的神经元通常包含最丰富的信息，该方法通常将分解到最后一个隐藏层的神经元的概率作为预测分数，它的值就作为输入的分解揭示了输入特征的重要性。这种方法可以帮助理解 GNNs 模型是如何对输入图结构进行分类和预测的，同时根据分解值的大小帮助识别出对于模型预测结果最为关键的数据。这种方法旨在观察模型参数中包含输入图数据的分解来探究输入图的特征与输出的解释预测之间的关系。

但是由于在图结构数据中具有不规则的拓扑信息，因此不能直接确认输入与分解结果之间确切的关系。因此需要预先提供一些先验策略来辅助采样的过程，比如，前文提到的 GraphSAGE 算法中基于邻接关系的下采样策略。

2.1.4 基于代理的可解释性方法

基于代理的方法从输入的图数据中取一个数据集子集作为样本，然后用可解释的模型来拟合这个采样数据集合，如机器学习中的决策树算法，而这个可解释模型将被作为原模型的近似代理解释。其基本思想是用简单且可解释的已存在模型作为代理来近似解释所提出的复杂的深层模型，这样可以作为对输入实例图数据中邻近空间的近似预测。

这类方法建立在原本方法与代理方法的输入实例之间的相邻空间关系不复杂的前提下，这样才能简单地用解释性强的代理模型捕捉这些关系并近似表达要解释的模型。但是图形数据包含拓扑结构的离散信息，所以在原始黑箱方法与代理方法之间存在着鸿沟。因此，如何将代理方法应用于图结构数据领域是一个可以扩展的灵活问题。同时，如何定义输入图结构空间中的邻域概念，以及明确哪种类型的代理模型是合适用于解释的，这些问题仍然没有得到充足的探索。

2.2 模型级方法

模型级方法的目的是基于解释深层图的模型来给出高层次的理解，而不是像实例级解释方法那样给出针对单个输入示例的解释。这类方法探究一般性的普遍规律，比如研究某种输入图模式和 GNNs 的某种行为之间的必然或间接联系。这些方法通常基于模型的结构和参数来进一步分析，例如网络层次结构、卷积核的形状和大小等。通过分析上述数据，可以获得模型对输入数据的整体处理方式和特征提取性能的

了解，从而提供模型级的解释。目前的研究中，模型级可解释 GNNs 方法还只有 XGNNs[11]，因此模型级的 GNNs 解释方法还有很大的探索空间。

XGNNs 提出了一种通过设计生成器来生成图的方式以解释 GNNs 的方法。与直接对输入图进行优化的方法不同，XGNNs 在训练图生成器时，设置对生成图的目标图预测最优化作为训练目标，因此生成器中得到的新图可以看作对设定目标的解释。相应地，在判别器中包含了预先期望的图模式。XGNNs 将图生成问题当作强化学习来设计训练过程和策略，生成器会对当前图结构中的数据改变其连接关系，并用这些新生成的图送入训练好图神经网络模型中预测，以进一步强化其功能。在训练过程中，通过梯度反向传播获得反馈优化模型。从本质上来说，XGNNs 是一种基于生成的框架方法，所以它除了可以看作模型级的解释方法外，还可以作为一种通用的图生成算法。该解释方法提供了对训练的 GNNs 的全局理解，并在图分类模型上证明了其有效性，但其在节点分类任务上的适用性还需要进一步探索。

3 图神经网络可解释性技术评估方法

在介绍图模型解释技术的评估指标之前，先介绍一些相关的数据集。经过训练的 GNNs 模型可以捕捉图形结构进行预测。因此通过观察和对比每种数据集的构建规则与对应设置的预测结果，可以对模型结果进行规律性的分析从而解释 GNNs 模型。目前在 GNNs 可解释领域，通常使用以下几种数据集：情感图数据、分子图数据以及合成数据。其中，合成数据是为了按照特定的目标研究某个问题而被人为地建立，所有节点的特征向量中的值全为 1。接下来，将介绍精度度量、稳定性度量、忠实度度量以及稀疏度度量这四种 GNNs 的可解释性的度量方法。

首先，从模型输出的预测表现上评估模型整体的性能，精度度量是模型对数据集判断的预测效果的指标。在数据集中包含真实标签的情况下，可以将其与模型预测值进行对比来证明模型的有效性。这类指标通常有 Accuracy，ROC-AUC score 和 F_1 score 等。度量值越高则表示模型的预测解释与事实越相接近。但是在现实世界中，由于缺乏真实值标签，这种度量方式无法得到广泛的应用。

其次，好的解释在应对输入数据的微小扰动时，数据发生变化前后的输出结果会保持稳定性（stability）的解释结果，而不会造成显著的影响。通过比较加入扰动前后的模型预测结果变化的程度，可以衡量模型对噪声的鲁棒性。在现实世界中的数据往往存在噪声、不完整性、错误等问题，如果模型对这些变化非常敏感，那么模型的准确性和可靠性就会受到质疑。而一个鲁棒性较强的模型可以更好地应对这些问题，提高模型的可靠性和实用性。

再次，当这种数据扰动设置为输入图的重要特征时，好的解释方法应当做到对其准确地识别。在 Fidelity 度量中[12-13]，首先，设置其为去除重要的输入图数据的点 / 边 / 节点特征和保留不重要的特征的前提，然后来观察解释预测的变化，最后根据预测结果的变化量衡量解释是否有识别重要的输入特征的能力。Fidelity 表示为模型对原始的输入数据的预测与设置为去掉输入图结构的重要特征之后的新预测之间的对比：

$$\boldsymbol{F_{acc}} = \frac{1}{N}\sum_{i=1}^{N}\left(f\left(\hat{y}_i = y_i\right) - f\left(\hat{y}_i^{1-m_i} = y_i\right)\right) \quad (4)$$

式中：N 表示输入图的数量，i 表示图的序号，$f(\cdot)$ 表示计算差值损失的方式，y_i 表示图的真实标签值，$\hat{y}_i$ 表示模型的解释预测值，1-m_i 表示丢掉输入图的重要特征，$\hat{y}_i^{1-m_i}$ 表示将按上述方式构建的图数据放入训练好的网络模型中得到的预测结果。式（4）的运算差值度量了预测准确率在改变结构信息前后的模型表现的变化。若其计算数值越高，则要解释的模型对重要的特征是忠诚的，证明模型确实是通过学习这些重要特征去训练网络，这样模型可解释是好的[14]。

从相反的角度来考虑，Infidelity 度量则是设置为去除输入图数据中不重要信息：

$$\boldsymbol{I_{acc}} = \frac{1}{N}\sum_{i=1}^{N}\left(f\left(\hat{y}_i = y_i\right) - f\left(\hat{y}_i^{m_i} = y_i\right)\right) \quad (5)$$

与前面 Fidelity 计算类似，m_i 表示要保留重要的输入特征。其计算数值越低，则输入数据中去掉不重要特征之后，对解释预测的影响越小，则这样解释模型是好的。

最后，从分析模型本身性能的角度，综合 Fidelity 度量中捕捉重要特征和 Infidelity 度量中忽略不相关特征的思路，优秀的解释性方法可以直接衡量模型捕捉输入图结构数据中重要特征的能力，即是满足稀疏性（Sparsity）的[15]。稀疏度指标衡量了解释模型所选择的重要要素的比例，这可以反映模型的解释结果的简洁性和学习特征的能力。对于给定图数据 G_i 和对应的贡献映射分数 m_i，稀疏度被定义为与图中预设的重要特征的数量与图中所有输入的特征数量之比：

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|m_i|}{|M_i|}\right) \quad (6)$$

式中：$|m_i|$ 是输入的 m_i 中重要特征的数量，$|M_i|$ 是在 G_i 中所有输入特征的数量。Sparsity 的值越大，则需要的重要输入特征 $|m_i|$ 数量越少，也就意味着它对重要的输入特征有更强的捕捉能力。

针对不同的需求和场景，选择适合的度量来评估解释性方法是十分重要的。全面综合考虑不同的度量方法的特点，既全面又客观地评估所提出的 GNNs 可解释性方法的性能，为实际应用提供更有针对性的指导 [16]。

4 总结

本文的探究重点是图神经网络的学习理论和其在可解释性领域的发展。首先，介绍了图神经网络的基本概念和应用领域，对目前主流的图神经网络模型进行了分析和比较。然后，重点讨论了图神经网络的可解释性问题，阐述了几种图神经网络可解释性技术，包括实例级的可解释性和模型级的可解释性技术。最后引入了一些实用的 GNNs 可解释性的重要性评估技术。虽然本文对 GNNs 的可解释性问题进行了一些讨论，但是仍存在一些待解决的问题需要进一步研究。例如，在图神经网络的可解释性和精度效果之间的冲突上，可以进一步探索新的技术方法使它们在某种程度上达到平衡最优解。在 GNNs 的应用方面，可以将图神经网络与其他领域的技术相结合，探索新的应用场景和方法。

参考文献：

[1] WU Z, PAN S, CHEN F, et al. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2020, 32(1): 4-24.

[2] KIPF T N, WELLING M. Semi-supervised classification with grap convolutional networks[EB/OL]. (2016-09-09)[2023-05-21]. https://arxiv.org/abs/1609.02907.

[3] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. (2017-10-30)[2023-05-26].https://arxiv.org/abs/1710.10903.

[4] XU K, HU W, LESKOVEC J, et al. How powerful are graph neural networks?[EB/OL]. (2018-10-01)[2023-04-22].https://arxiv.org/abs/1810.00826.

[5] WANG L, HUANG Y C, HOU Y L , et al. Graph attention convolution for point cloud semantic segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE,2019:10288-10297.

[6]HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[J]. Advances in neural information processing systems, 2017, 30:1025-1035.

[7] CHEN J, MA T, XIAO C. Fastgcn: fast learning with graph convolutional networks via importance sampling[EB/OL]. (2018-11-30)[2023-05-22].https://arxiv.org/abs/1801.10247.

[8] RONG Y, HUANG W, XU T, et al. Dropedge: Towards deep graph convolutional networks on node classification[EB/OL]. (2019-07-25)[2023-05-22].https://arxiv.org/abs/1907.10903.

[9] YUAN H, YU H, GUI S, et al. Explainability in graph neural networks: A taxonomic survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2023,45(5):5782-5799.

[10] BALDASSARRE F,AZIZPOUR H.Explainability techniques for graph convolutional networks[EB/OL]. (2019-05-31)[2023-05-16].https://arxiv.org/abs/1905.13686.

[11] YUAN H,TANG J,HU X,et al.XGNNs: Towards model-level explanations of graph neural networks[EB/OL]. (2020-06-03)[2023-05-05].https://arxiv.org/abs/2006.02587.

[12] HOOKER S,ERHAN D,KINDERMANS P J,et al.A benchmark for interpretability methods in deep neural networks[EB/OL].(2018-06-28)[2023-05-07].https://arxiv.org/abs/1806.10758.

[13] POPE P E,KOLOURI S,ROSTAMI M,et al.Explainability methods for graph convolutional neural networks[EB/OL]. (2019-05-31)[2023-05-23].https://arxiv.org/abs/1905.13686.

[14] 窦慧 , 张凌茗 , 韩峰 , 等 . 卷积神经网络的可解释性研究综述 [J]. 软件学报 ,2024,35(1):159-184.

[15] 徐冰 , 岑科廷 , 黄俊杰 , 等 . 图卷积神经网络综述 [J]. 计算机学报 , 2020, 43(5):755-780.

[16] LIU N, FENG Q, HU X. Interpretability in graph neural networks[EB/OL].(2022-01-03)[2023-04-26].https://link.springer.com/chapter/10.1007/978-981-16-6054-2_7.

【作者简介】

刘杰（1979—），男，山东高密人，高级实验师，研究方向：大数据处理等。

王敏（1989—），女，山东枣庄人，高级工程师，研究方向：图像处理。

唐青梅（2000—），女，湖南邵阳人，研究方向：图神经网络和 3D 检索识别等。

张萌月（1996—），女，通信作者，河北邯郸人，工程师，研究方向：遥感图像处理、小样本学习等。

（收稿日期：2023-08-13）