

注意力机制在图神经网络模型中的算法研究

陈丹阳, 张长伦*

北京建筑大学理学院, 北京

收稿日期: 2023年10月24日; 录用日期: 2023年11月25日; 发布日期: 2024年1月12日

摘要

图神经网络(Graph Neural Networks, 简称GNN)是一种用于处理图结构数据的深度学习模型。随着图数据的广泛应用, 如社交网络、推荐系统和生物信息学等领域, 研究者们致力于提升GNN的性能和表达能力。注意力机制作为一种强大的工具, 已广泛应用于深度学习领域, 可以帮助模型更加有效地学习和利用关键信息。本文首先分别对图神经网络和注意力机制进行简单的概念阐述, 然后详细介绍了图神经网络中与注意力机制结合的传统图注意力网络, 及近几年提出的一些图注意力网络模型相关改进和变体。最后对未来发展方向以及研究趋势进行总结与展望。

关键词

深度学习, 图神经网络, 注意力机制

Algorithm Research of Attention Mechanism in Graph Neural Network Model

Danyang Chen, Changlun Zhang*

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: Oct. 24th, 2023; accepted: Nov. 25th, 2023; published: Jan. 12th, 2024

Abstract

Graph Neural Networks (GNN) are a deep learning model for processing graph-structured data. With the wide application of graph data, such as social networks, recommendation systems, and bioinformatics, researchers are working to improve the performance and expression of GNN. As a powerful tool, attention mechanisms have been widely used in the field of deep learning to help models learn and utilize critical information more efficiently. In this paper, the concepts of graph neural network and attention mechanism are briefly described, and then the traditional graph at-

*通讯作者。

attention network combined with attention mechanism in graph neural network is introduced in detail. Some improvements and variants of graph attention network models have been proposed in recent years. Finally, the future development direction and research trend are summarized and prospected.

Keywords

Deep Learning, Graph Neural Network, Attention Mechanism

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,随着人工智能和数据科学的快速发展,图数据作为一种重要的数据形式逐渐备受关注。图数据是一种用于表示和分析复杂关系的数据结构,它由节点和连接节点的边组成。这些节点和边可以用于描述各种实际场景,例如社交网络、知识图谱、交通网络、蛋白质相互作用等。图数据被广泛应用于各种领域。在社交网络中,图数据可以帮助我们理解用户之间的关系、发现社区结构和影响传播。在推荐系统中,图数据可以用于建模用户和物品之间的关联,从而实现个性化推荐。在生物信息学中,图数据被用于研究基因之间的相互作用和调控网络。此外,图数据还可应用于网络安全、金融风险分析、交通规划等领域。

作为典型的非欧空间数据,图数据与传统的表格数据或向量数据不同,其特点在于它们能够捕捉到实体之间的复杂关系和相互作用。通过对图数据的分析,我们可以揭示实体之间的关系网络、发现隐藏的模式和结构,以及进行预测和推理。通过深入理解和分析图数据,我们可以获得更深入的洞察和智能决策,推动各个领域的发展和革新。

随着图数据的重要性日益凸显,研究人员开始探索如何利用深度学习技术来处理和分析图数据。在这个背景下,图神经网络应运而生,成为处理图数据的一种强大工具。

2. 图神经网络

2005年,Gori等人[1]舍弃将图转换成向量的传统处理方法,充分利用图的拓扑结构信息,直接对图数据进行处理,用RNN来压缩节点信息和学习图节点标签,首次提出了图神经网络这一概念。

图神经网络作为一种深度图表示学习方法,旨在为各种下游任务(例如,节点分类[2]、链路预测[3]、社区检测[4]和图分类[5]在更低维空间中学习节点/边/图级别的表示。与传统的图嵌入方法(即矩阵分解[6]和随机游走(Random walk-based) [7] [8] [9]不同,GNNs通过不同的神经网络同时学习图的结构和属性特征,由于这种固有的优势,GNNs可以很自然地用于处理图结构[10]。

现有的图神经网络通常分为两大类:频谱域和空间域。谱图神经网络基于谱图理论和图信号处理实现拓扑卷积操作。典型的方法有Spectral CNN [11]、ChebNet [12]、GCN [2]等。谱GNN是基于傅里叶变换的,并且强烈依赖于图拉普拉斯矩阵[2]。一旦确定了图的结构,就很难将此类方法推广到另一个新的图上。而空间图神经网络在消息传递神经网络(MPNN) [13]的作用下,直接从邻域中迭代地聚合和更新节点的信息。考虑到节点/边/图的属性,图网络(Graph Networks) [14]在消息传递机制下对各种图神经网络进行了推广和扩展,支持从简单的构建模块构建复杂的体系结构。

基于空间域的图卷积神经网络不同于从信号处理理论出发的谱域图卷积神经网络, 空间域图卷积神经网络是从图中的节点出发, 设计聚集邻居节点特征的聚合函数, 采用消息传播机制, 思考怎样准确高效地利用中心节点的邻居节点特征来更新表示中心节点特征。CNN 的本质是加权求和, 空间域的图卷积神经网络正是从 CNN 的基本构造过程出发从求和的角度来完成 GNN 聚合邻居节点的目的。

目前我们经常提到的图卷积神经网络(Graph Convolutional Network, GCN)为第三代 GCN [2], 该 GCN 正式将卷积模块用于对图结构数据建模。GCN 对图的拓扑结构和节点特征信息进行学习, GCN 中的卷积模块使用中心节点与其邻居节点的特征作为卷积的输入。每个节点的特征与其邻居节点的特征按照权重进行加权求和, 其中权重表示了节点之间的连接强度。这种加权求和的过程可以看作是一种信息传播, 通过聚合邻居节点的信息来更新节点的表示。通过这样的卷积操作, 每个节点都可以利用其邻居节点的信息进行更新, 从而获得更丰富的表示。Kipf 等人的主要贡献在于对网络的深度加深, 减少邻域宽度, 进一步提升了准确率和学习能力。但是该方法仍然限定在无向图和无权图上。并且要加载整个图才能进行训练, 对于机器存储空间要求较高, 因此超大规模的图可能加载和运时间比较长。整个 GCN 模型可以通过堆叠多个卷积模块来进行深层的图卷积操作, 从而实现更复杂的图结构分析和学习任务, 为将其他深度学习神经网络迁移到图上提供了标准化的框架。

GCN 模型利用图的拓扑结构和节点特征信息学习节点嵌入表示且能够捕捉图的全局信息, 但是其要求在固定不变的图上进行学习, 不能直接泛化到未知的节点上, 当图结构发生变化或者有新节点出现时, GCN 模型需重新训练, 从而会带来巨大的计算开销。为此, William L. Hamilton 等人[15]提出了 GraphSAGE 网络(Graph Sample and Aggregate), 该网络主要分两步: 邻居采样和邻居聚合。GraphSAGE 将 GCNs 扩展到归纳的无监督学习任务中, 能够利用邻域节点的特征信息为未知节点高效地生成节点嵌入。其核心思想是如何学习一个函数从目标节点的局部邻域中聚合这些节点的特征信息。GraphSAGE 通过采样节点的邻居子图来进行信息聚合, 在保持计算效率的同时, 捕捉节点的上下文信息。GraphSAGE 网络是一种基于采样的图神经网络架构, 用于处理大规模图数据。该网络的提出为处理大规模图数据提供了一种高效而灵活的建模和分析方法。

大多数 GNN 对邻域节点一视同仁, 在聚合和传播信息时没有考虑不同节点的贡献。然而, 现实世界中的图往往是噪声的, 不相关的节点之间存在连接, 这导致 GNNs 学习到次优的表示, 而不区分邻域中的节点。为解决该问题, Petar Veličković 等人[16]提出了 GAT 网络(Graph Attention Network), 区别于传统的 GCN 网络模型, 不再使用拉普拉斯矩阵更新节点的状态, GAT 网络中每个节点能够选择性地聚合具有不同重要性的邻居信息。GAT 网络的注意力机制为节点间的信息传播和聚合引入了更精细的控制, 使得节点能够根据具体情况有选择性地聚合相邻节点的特征。这种机制在处理复杂关系、大规模图数据和需要考虑节点重要性的任务中具有很大的优势。因此, GAT 网络在图神经网络领域中被广泛应用。到目前为止, 已经有大量的基于注意力机制的图神经网络。GNNs 中的注意力可以应用于不同的节点邻域、结构的不同部分以及节点、边和图的不同表示。

2018 年, Keyulu Xu 等人[17]提出了 GIN 网络(Graph Isomorphism Network), GIN 网络的提出是为了解决图数据中的图同构性质建模问题。传统的图神经网络在处理图同构性较为敏感的任务时可能存在性能限制。而 GIN 网络通过使用多轮图同构聚合来生成节点表示, 不依赖于图的结构信息, 而是将图的结构信息编码为节点特征的变换权重。在 GNN 模型中, GIN 网络是一种重要的存在。它拓展了传统的图神经网络架构, 引入了图同构性建模的思想。GIN 网络的设计思路对后续的研究工作产生了重要的影响, 并催生了许多基于图同构的扩展模型。同时, GIN 网络也为处理需要考虑图同构性质的任务提供了一种有效的工具和方法。因此, GIN 网络在图神经网络领域具有重要的地位和作用。

以上主要对图神经网络发展以来最为经典的图神经网络模型进行简单的梳理和介绍, 基于这些经典

的图神经网络模型, 近几年来研究者们基于这些传统的图神经网络也提出很多变体模型, 同时在众多研究思路的基础上, 图神经网络领域近几年持续涌现了许多创新的架构和模型, 以应对不同的图数据建模和分析需求。这些新的图神经网络架构在模型结构、信息传播方式和聚合策略等方面进行了改进和创新, 推动了图神经网络的发展和应用。

3. 注意力机制

注意力是人类大脑中一项不可或缺的复杂认知功能, 在日常生活中, 人们通过视觉、听觉、触觉等方式接收大量的信息, 但是人们在这些外界的信息轰炸中还能有条不紊地工作, 是因为人脑可以有意或无意地从这些大量输入信息中选择小部分的有用信息来重点处理, 并忽略其他信息, 这种能力叫做注意力。专家学者根据对人类注意力的研究, 提出了注意力机制, 本质上说就是实现信息处理资源的高效分配[18]。

注意力机制是自深度学习快速发展后广泛应用于自然语言处理、统计学习、图像检测、语音识别等领域的核心技术。当人类视野接触到一个场景时, 通常会先注意到场景中的一些重点元素, 例如动态的物体或者突出的颜色, 而对于静态的背景可能会暂时性地忽略[19]。注意力机制具有突出重要信息、忽略不相关信息的能力, 并且能够动态地调整权重, 以适应不同情况下的重要信息选择。因此, 它在可扩展性和鲁棒性方面表现出更高水平[20]。

基本网络框架[21]如图 1 所示。

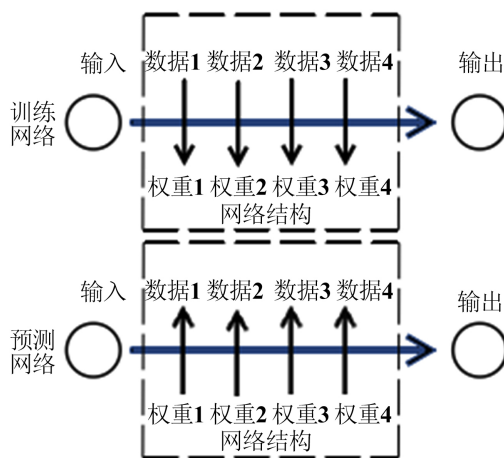


Figure 1. Attention mechanism network architecture
图 1. 注意力机制网络架构图

注意力机制的第一次提出是在视觉图像领域中[22], 它指出注意力的作用就是将之前传统的视觉搜索方法进行优化, 可选择地调整视觉对网络的处理, 减少了需要处理的样本数据并且增加了样本间的特征匹配。文献[23]利用灵长类动物的视觉注意力提供了一个科学解释, 提出的视觉注意模型结构比较简单, 能够对接收到的信息进行特征提取并且快速检测出各种形状和颜色, 但是因为结构的简单, 无法检测到特征间的相关性, 并且没有任何循环机制, 所以在整个视觉识别过程中, 无法重现图像轮廓, 因此人们就想到了将注意力机制与具有循环机制的循环神经网络(RNN)结合, 并且在此基础上进行了一系列研究发现。

Google mind 团队在 2014 年将循环神经网络(RNN)模型与 attention 机制结合[24], 利用注意力机制对特定的区域进行高分辨率处理。该模型受人类注意力的启发, 根据需求将注意力集中在图片中特定的部分, 类似于人类观察衣服图片时, 往往不会看图片中所有因素, 而是先注意到其中的突出因素。所以基于任务需求, 这个模型不是处理全部图像, 而是有针对性地选择相应位置处理, 使整个模型的性能提高。

同年, Bahdanau 等人[25]提出了一种基于注意力机制的神经网络模型, 这篇论文将注意力机制用于机器翻译任务, 该模型能够根据输入句子的不同部分分配不同的注意力权重, 以便在翻译过程中有选择性地关注输入句子的相关部分。通过引入注意力机制, 该模型能够更好地处理长句子的翻译任务, 提高了机器翻译的性能和表达能力。这项工作为后续的注意力机制研究奠定了基础。2017年, Vaswani 等人[26]提出了一种名为 Transformer 的神经网络架构, 介绍了以自注意力机制为基本单元的 Transformer 模型, 使得注意力机制得到真正的成功运用, 将注意力机制推向了诸多应用方向的热潮。注意力机制的引入使得 Transformer 在自然语言处理和其他序列相关任务中取得了显著的性能提升, 并成为了许多领域中的重要模型架构。

2021年, Kitaev [27]针对 Transformer 模型在处理长序列文本时存在的注意力机制计算复杂度高和内存效率低的问题进行改进, 提出了一个新的模型架构 Reformer, 将 Transformer 中的点积注意力替换为局部位置敏感哈希注意力。Transformer 中的多头注意力是并行计算并叠加, 它计算两个数据点之间的 attention score 需要将多个自注意力连起来因此导致计算量很大, 所占内存较多。Reformer 中选了用局部敏感哈希注意力, 代替多头注意力。局部敏感哈希通过计算一个哈希函数, 将相似的向量匹配在一起, 而不是搜索所有可能的向量对。从而更有效地使用可用内存。

在自然语言处理等方面, Transformer 和 Reformer 算法已经在效率和使用内存上做出了巨大的贡献。这些算法也同样给其他专家研究学者提供了新的思路, 以注意力机制为基本框架的神经网络算法也将会被不断地改进更新, 从而探寻更优的算法。

在神经网络中, 注意力机制能够关注相关的信息而忽略不相关的信息, 直接建立输入与输出之间的依赖关系, 而不再通过循环, 并行化程度增强, 运行速度有了很大提高[28] [29]。其次, 注意力机制克服了传统神经网络中的一些限制。传统神经网络在输入长度增加时性能下降, 而注意力机制能够有效地建模可变长度的序列数据, 从而提高系统的性能。此外, 传统神经网络在处理不合理的输入顺序时计算效率低下, 而注意力机制能够减少这种效率问题。另外, 传统神经网络在特征提取和强化方面存在不足, 而注意力机制可以进一步增强系统对特征的提取和强化能力。因此, 注意力机制不仅减少了层次深度, 还有效提高了系统的精度, 同时捕获了远程依赖信息[30] [31]。由此可见, 注意力机制在深度学习及神经网络中得以迅速发展。

自图神经网络研究兴起, 便有不少学者致力于将注意力机制与图神经网络进行融合, 在图神经网络中, 注意力机制可以应用于节点级别和图级别的任务, 以帮助模型更好地处理图结构数据。在节点级别的任务中, 注意力机制可以用来对节点之间的关系进行建模和加权。在图级别的任务中, 注意力机制可以用来对图中不同部分的重要性进行建模。综上所述, 注意力机制通过加权节点之间的关系或调整节点特征表示来帮助模型更好地处理图结构数据, 从而提升图神经网络的能力。

4. 图注意力网络模型

4.1. 经典图注意力网络

2018年 Veličković 首次将注意力机制(Attention)与图神经网络模型相结合[16], 提出了图注意力网络模型(Graph Attention Network, GAT)。GAT 的核心工作原理是通过注意力机制来计算节点间的关系, GAT 会计算一个节点与其邻居节点之间的注意力权重, 然后根据这个权重来更新节点的状态。通过计算权重而更新信息的方式使得 GAT 能更好地捕捉到图中的结构信息。在计算权重分值和捕捉信息的方面, GAT 采用了类似于 Transformer 的掩蔽自注意力机制, 由堆叠在一起的图注意力层构成, 每个图注意力层获取节点嵌入作为输入, 输出转换后的嵌入, 节点嵌入会关注到它所连接的其他节点的嵌入。图注意力网络如下图 2 所示。

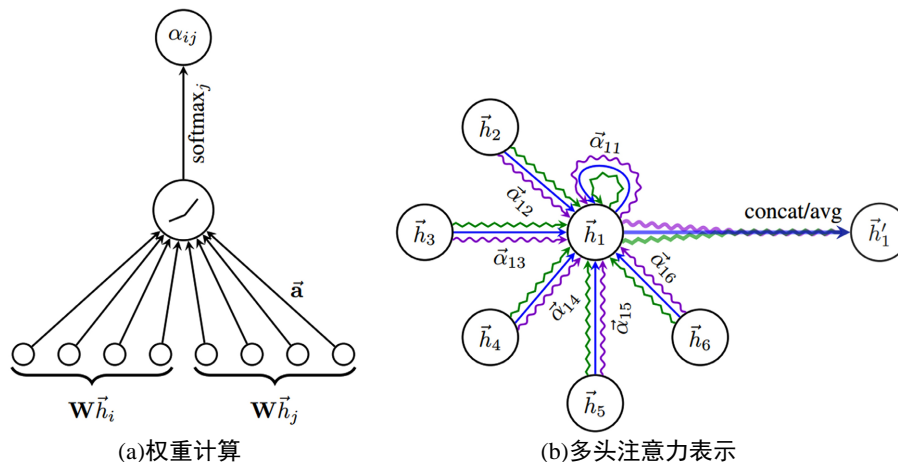


Figure 2. GAT network structure
图 2. GAT 网络结构

上图 2(a)表示节点 i 和节点 j 间权重的计算, 图 2(b)表示某一节点在其邻域内采用了多头的注意力机制来更新自身表示。

节点 j 相对于节点 i 的注意力因子求解为:

$$a_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^T \left[W h_i \parallel W h_j \right] \right)\right)}{\sum_{k \in N_i} \exp\left(\text{LeakyReLU}\left(a^T \left[W h_i \parallel W h_k \right] \right)\right)}$$

其中, α_{ij} 表示节点 j 相对于节点 i 的注意力因子, W 是一个用于降维的仿射变换, a^T 表示权重向量参数, \parallel 表示向量的拼接操作, LeakyReLU 为修正线性单元。利用学习到的注意力因子就可以实现对中心节点 i 的更新:

$$h'_i = \sigma \left(\sum_{j \in N_i} a_{ij} W^k h_j \right)$$

在 GAT 的实际运算中, 作者还应用了多头注意力机制, 注意力分数的计算是通过一个名为“注意力头”的结构完成, 每个注意力头便会计算出一组注意力分数, 并且在最后的结果中, 所有的注意力头的结果会被平均或者拼接起来, 以得到最终的节点嵌入。这样做的好处是, 每个注意力头可以关注到不同的特征或者模式, 从而使得 GAT 能够捕捉到更多的信息。这里, 若同时计算 K 个不同的注意力机制, 则可得到:

$$h'_i = \parallel_{k=1}^K \sigma \left(\sum_{j \in N_i} a_{ij}^k W^k h_j \right)$$

传统的图神经网络方法在节点特征聚合时通常采用简单的平均值或固定权重来提取邻居信息。然而, 这种方法无法有效地捕捉节点之间的复杂关系。注意力机制通过计算节点之间的相似度, 并根据相似度分配权重, 可以实现对节点之间关系的精细建模。相似度较高的节点在特征聚合时获得更高的权重, 从而突出了与当前节点最相关的邻居节点的信息。这样能够更准确地表达节点之间的关系, 提取出更有意义的特征。

在图中, 节点之间的相互作用与信息传递是非常重要的。然而, 传统的图神经网络方法通常只考虑节点的局部邻域而忽略了较远邻居之间的关系。注意力机制可以基于节点的相对重要性, 自适应地控制信息的传递程度。通过注意力权重的计算, 模型可以学习到每个节点之间更精细的关系和上下文信息,

并实现全局信息的有效传递。这样, 每个节点可以通过聚合邻居节点的特征来更新自己的表示, 从而实现更全面和准确的特征表达。

注意力机制具有较好的鲁棒性和适应性, 可以处理图中的不均衡性和噪声。在现实世界的图数据中, 节点之间的连接通常存在不均衡性, 即一些节点比其他节点具有更多的关联关系。注意力机制通过相似度计算和权重分配, 能够更好地适应不同节点之间的连接情况, 从而提高对不均衡数据的处理能力。同时, 注意力机制可以通过自适应地调整权重, 减少噪声节点的影响, 提升模型对噪声的鲁棒性。

4.2. 注意力机制在 GNN 中的其他研究

自 GAT 模型提出以来, 研究者们同样也试图将注意力机制融入至图神经网络中的各个模块, 以下部分对近几年来注意力机制在图神经网络各个方向的融合做了梳理。

文献[32]在图自编码器(Graph Auto-Encoder, GAE)基础上引入自注意力机制, 提出了一种无监督表示学习的图注意力自编码器(Graph Attention Auto-encoder, GATE)。GATE 在编码器和解码器层同时引入自注意力机制, 对节点属性和图结构信息进行重新构建。在编码器中, 将原始的高维节点特征映射到一个低维表示空间, 并利用注意力机制对节点之间的关系进行建模, 将节点属性作为初始表达。每一层都使用注意力机制, 根据邻域中节点之间的相关性生成新的节点表示。在解码器中, 将低维表示重建为原始的高维节点特征, 反向编码以重建节点属性。解码器的每一层逆向操作与编码器正向操作对应。此外, 对节点表示进行正则化以重构图结构。由于 GATE 框架不需要先验知识图结构, 因此适用于归纳学习。在基准数据集的节点分类任务中, GATE 表现出色, 对于直推式和归纳式任务都具有强大的竞争力。

2018 年, 提出了门控注意力网络(Gated Attention Networks, GaAN) [33], GaAN 不同于传统 GAT 采用求平均或者拼接的方式确定不同注意力因子的权重, 均衡的消耗所有的注意力, 它使用一个卷积子网络来控制每个注意力头的重要性, GaAN 认为, 多头注意力机制虽然能够捕获不同的特征信息, 但并不是每头注意力机制的贡献都是相同的, 某一头注意力可能会捕获到无用的信息, 输入一个捕捉无用表示的注意力的输出会误导模型的最终预测。因此 GaAN 为每个注意力头分别赋予了不同的权重, 以此来聚合多头注意力中捕获的邻居节点的信息, 完成对中心节点的更新。GaAN 模型计算一个介于 0 (低重要性) 和 1 (高重要性) 之间的附加软门, 为每个头部分配不同的重要性。结合多头注意聚合器, 我们得到了门控注意聚合器的公式:

$$y_i = FC_{\theta_0} \left(\tilde{x}_i \oplus \bigoplus_{k=1}^K \left(g_i^{(k)} \sum_{j \in N_i} \omega_{i,j}^{(k)} FC_{\theta_v}^h(z_j) \right) \right)$$

$$g_i = (g_i^{(1)}, g_i^{(2)}, \dots, g_i^{(K)}) = \Psi_g(\tilde{x}_i, z_{N_i})$$

其中, $FC_{\theta_0}(\cdot)$ 表示线性变换后不应用激活函数, \oplus 是连接操作, K 是注意力机制的数量, $\omega_{i,j}^{(k)}$ 是节点 i 和节点 j 间的第 k 头的门控值, 应用卷积网络 Ψ_g 可根据自己的实际需要来进行设计, 采用了平均池化和最大池化进行构造:

$$g_i = FC_{\theta_g}^\delta \left(\tilde{x}_i \oplus \max_{j \in N_i} \left(\{ FC_{\theta_m}(z_j) \} \right) \oplus \frac{1}{|N_i|} \sum_{j \in N_i} z_j \right)$$

其中, θ_m 表示将邻居节点的特征向量映射到 d_m 维, θ_g 表示拼接好的特征向量映射到第 k 个门。最后, 作者构造门控递归单元, 成功应用到交通速度预测问题上。

2019 年, Busbridge 等人[34]提出一种比较特殊的关系型图注意力网络(Relational Graph Attention Network, RGAT), 以关系型图卷积网络(Relational Graph Convolutional Network, RGCN) [35]为基础, 将注

注意力机制扩展到关系型图域中。在关系型图注意力网络中, 每个节点都与其他节点之间的关系形成了一个关系矩阵。关系矩阵被用作注意力机制的输入, 以计算节点之间的关系权重。不同关系的权重是通过学习得到的, 模型可以自适应地为不同关系分配不同的注意力权重。这种关系型注意力机制使得模型能够更好地捕捉节点之间的关系, 从而提高图数据的表示能力和预测性能。

关系型图注意力网络通常包含多个图注意力层, 每个图注意力层由编码器和解码器组成。编码器负责学习节点的低维表示, 并利用关系型注意力机制将节点的特征和关系进行融合。解码器则负责将低维表示重建为原始的高维节点特征, 以实现节点重构和预测任务。通过多个图注意力层的叠加, 模型可以逐步提取更抽象和高层次的特征表示。

关系型注意力网络逐渐发展成两种变体: 内部关系图注意力机制(Within-Relation Graph Attention, WIRGAT)和交叉关系图注意力机制(Across-Relation Graph Attention, ARGAT)。WIRGAT 和 ARGAT 使用了附加注意力机制和乘性注意力机制。WIRGAT 用来聚集节点的邻居并使用注意力系数作用于节点聚合矩阵。ARGAT 用来阐述不同节点表示上实现单个概率分布和关系的重要性, 并将图局部先验知识进行编码。由于在实际表现中, RGAT 在归纳式图分类任务上产生了边际效应, 因此 Busbridge 等人使用完全累积分布函数(Cumulative Distribution Function, CDF)搜索合适的超参数, 并做了统计假设检验。

针对图领域中的图池化模块, 一个 GNN 生成节点特征后, 我们就可以使用它们进行最终的任务。但是直接使用所有这些特征可能在计算上很有挑战性, 因此需要一个下采样策略。其中图池化操作便是一个下采样策略。池化操作旨在通过对节点进行下采样来减少参数的大小, 以生成更小的表示, 从而避免过拟合、置换不变性和计算复杂性问题。

2015 年 Li 等人[36]在门控图序列神经网络(Gated Graph Sequence Neural Networks, GGS-NNs)这一模型中, 对于该模型生成节点特征后的输出, 作者引用了软注意力机制。对于图分类的任务, 定义了一个图级别的向量表示:

$$h_{\zeta} = \tanh \left(\sum_{v \in \gamma} \sigma \left(i \left(h_v^{(T)}, x_v \right) \right) \otimes \tanh \left(j \left(h_v^{(T)}, x_v \right) \right) \right)$$

其中, $\sigma \left(i \left(h_v^{(T)}, x_v \right) \right)$ 是一个软注意力机制, 决定了那些节点和当前的图级别任务相关。 i 和 j 是将 $h_v^{(T)}$ 和 x_v 拼接起来作为输入并输出实值向量的神经网络。该注意力机制在图池化过程中, 使得节点之间的关系能够动态地选择和聚合特征, 减少了图的规模和复杂度, 生成更简洁、更具表达力的图级特征表示。

2019 年, Lee 等人[37]发表论文首次将注意力机制加入图池化中, 具体的图池化层如下图 3 所示

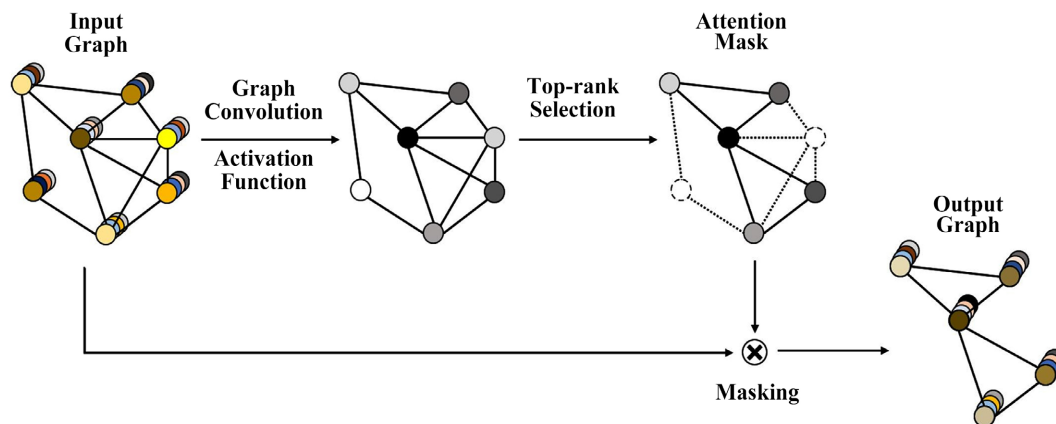


Figure 3. Diagram pooling layer flow
图 3. 图池化层流程图

该图池化层使得我们可以更多地关注重要的特征, 而更少地关注不重要的特征。特别是, 自我注意, 通常被称为内部注意, 允许输入特征成为注意本身的标准[26]。作者使用图卷积来获得自注意分数。例如, 如果使用 Kipf & Welling [2] 的图卷积公式, 则自注意得分计算如下:

$$Z = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X \Theta_{att} \right)$$

其中, σ 是激活函数, $\tilde{A} \in R^{N \times N}$ 是带自连接的邻接矩阵, $\tilde{D} \in R^{N \times N}$ 是 \tilde{A} 的度矩阵, $X \in R^{N \times F}$ 是 N 个节点, F 是输入特征维度, $\Theta_{att} \in R^{F \times 1}$ 是唯一参数。通过初始化图卷积来获得自注意力分数, 池化的结果是基于图特征和图结构的。同时这篇论文中关于图注意力池化, 提出了多种方式计算注意力分数, 可以不仅仅使用相邻节点, 还可以使用多跳相连结点等。

针对层次注意力机制, Do 等人[38]提出一种用于多标签学习的图注意力模型(Graph Attention Model for Multi-Label Learning, GAML)。在多标签分类问题中, 分类的准确性和可解释性始终是一个难点, 这与标签和子图的未发掘关系之间存在某种关联。因此 GAML 将所有标签视作一个个的节点, 称为标签节点。普通图节点称作数据节点, 然后将标签节点与数据节点一起作为神经网络的输入端。之后 GAML 运用消息传递算法[39]构建了 MPGNN(Message Passing Graph Neural Network)。在 MPGNN 中, 更新每个节点的局部结构是通过在标签节点和数据节点之间传递子结构消息来实现的。同时, MPGNN 采用了分层注意力机制, 其目的是确定不同子结构消息的重要性。这种分层注意力机制利用中间注意力因子来保存计算结果, 并将其应用于标签节点, 以便能够抽取与之最相关的子结构来更新节点状态。随后, 更新后的状态被用来预测标签节点的最相关类别, 在状态更新步骤中, 节点状态更新公式为:

$$x'_t = g^u(x_t^{t-1}, m_t^i)$$

其中, x'_t 是第 t 轮节点的状态向量, m_t^i 是节点的消息向量。 $g^u(\cdot)$ 表示网络框架, 这里 Do 等人使用了高速网络(Highway Network)获取数据之间的长距离依赖特征, 更方便地表达标签和输入子图在不同尺度上的关系, 使得 GAML 具有较为优秀的分类能力。更重要的是 GAML 可以提供较为直观的可视化结果, 这有助于了解不同标签子结构关系以及解释模型的内在原理。

在对图注意力网络模型进一步研究中, 噪声干扰会使得模式分析局限在图上一小块范围。针对此问题, 文献[40]提出图注意力模型(Graph Attention Model, GAM)。GAM 利用注意力机制引导的随机游走采样节点, 以强制游走区域限定在与任务相关的区域, 并忽略无用的数据节点, 从而捕获图结构中关键的区域连通信息。通过动态选择有效的节点序列, GAM 能够自适应地选择一些节点信息, 并只处理图的部分内容。这种方法在不需要对整个图进行全局分析的情况下, 大幅度减少了计算复杂度和存储开销。

Samy Bengio 等人于 2018 年提出了一个基于注意力机制的节点嵌入方法[41], 该方法结合了随机游走的思想。该方法通过使用注意力机制, 为每个游走步骤中的邻居节点分配权重, 然后将权重与节点特征进行结合, 以产生更丰富的节点表示。它还提出了一种名为“Watch Your Step”的新游走策略, 以在游走过程中更好地探索图的结构。这篇论文在图嵌入领域引起了广泛的关注, 并为后续的相关研究提供了启示。它将图的随机游走和注意力机制相结合, 改善了节点嵌入的质量和表达能力, 为图结构的学习和应用提供了新的思路。

表 1 从模型、基准网络、注意力机制、图种类、学习模式、激活函数以及任务对图注意力网络模型进行了汇总。

Table 1. Summary of graph attention network model**表 1.** 图注意力网络模型总结

模型	基准网络	注意力机制	图种类	学习模式	激活函数	任务
GAT	GCN	自注意力 多头注意力	异质无向	有监督	Sigmoid LeakyReLU	节点分类
GATE	GCN	自注意力	异质无向	无监督	Sigmoid Softmax	节点分类
GaAN	GCN	多头注意力	异质无向	-	Sigmoid Softmax	节点分类
RGAT	GAT	附加注意力 乘性注意力	异质无向	半监督	Sigmoid LeakyReLU	节点分类 图分类
GGs-NN	GCN	自注意力 门控递归单元 全局状态表示	有向	-	Sigmoid Softmax	节点级别 图级别
GAM	RNN	注意力引导 随机游走	无向有向	有监督	Softmax	图分类
GAML	MPGNN	分层注意力	异质有向	有监督	Sigmoid ReLU	多标签分类 图预测
Attention based graph pooling	GNNs	自注意力	-	-	Sigmoid LeakyReLU	图分类
Attention based node embedding	节点嵌入	注意力引导 随机游走	-	无监督	Softmax	链接预测

4.3. 图注意力网络的最新研究进展

近两年, 自图神经网络方向的经典模型提出后, 研究学者们进而不断发现各模型所存在的问题及缺陷, 例如, GCN 及 GAT 模型只有在层数为 2 时效果达到最优, 模型不能学到高阶的特征; 在图卷积网络的学习过程中, 如何对邻域节点进行区分, 以学习到更好的表示; 图注意力网络中模型学到的特征该如何解释等等。在众多问题及研究思路的基础上, 图神经网络领域近几年持续涌现了许多创新的架构和模型, 以应对不同的图数据建模和分析需求。下面简单介绍近两年为解决图神经网络中出现的某些问题而提出的模型。

2022 年 Brody 等人[42]发现 GAT 模型中注意力计算存在问题: GAT 模型只计算一种受限制的“静态”注意力形式, 对于任何查询节点, 注意力函数关于邻居(键)分数是单调的。即注意力系数的排序对于图中的所有节点都是相同的, 注意力得分排名不受查询节点的影响, 这一事实严重损害了 GAT 的表现力。比如, 在一个控制问题中, 作者表明了静态注意力甚至会阻碍 GAT 拟合训练数据。为了摆脱这些限制, 作者便提出了一种基于动态图神经网络变体。作者针对 GAT 模型发现的这个问题进行改进, 通过修改 GAT 模型中的注意力权重计算公式, 最后得到更优的模型。修改后的节点特征更新函数如下所示:

$$\text{GAT (Velickovic et al., 2018)} \quad e(h_i, h_j) = \text{LeakyReLU}(\alpha^T \cdot [Wh_i \parallel Wh_j])$$

$$\text{GATv2 (Brody et al., 2018)} \quad e(h_i, h_j) = \alpha^T \text{LeakyReLU}(W \cdot [h_i \parallel h_j])$$

2023 年, Javaloy 等人[43]分析和 GCN 和 GAT 两种模型的优势与劣势, 希望能够将两类模型相结合,

充分利用与发挥两种模型的优势, 将两个模型进行结合, 提出了图卷积注意力网络(Graph Convolutional Attention Networks, CAT)基于此, 作者又提出可学习的图卷积注意力网络(Learning Graph Convolutional Attention Networks, L-CAT): 一个 GNN 架构, 通过添加两个标量参数, 自动在 GCN、GAT 和 CAT 之间进行插值。我们的结果表明, L-CAT 能够有效地结合网络上不同的 GNN 层, 在广泛的数据集上优于竞争的方法, 从而产生一个更稳健的模型, 减少了交叉验证的需要。此模型特征更新函数如下:

$$\Psi(h_i, h_j) = \lambda_1 \cdot \alpha(W\tilde{h}_i, W\tilde{h}_j),$$

$$\tilde{h}_i = \frac{h_i + \lambda_2 \sum_{t \in N_i} h_t}{1 + \lambda_2 |N_i|}$$

$\lambda_1, \lambda_2 \in [0, 1]$ 为两个标量参数, 当 $\lambda_1 = 0$ 时, 该传播函数即为 GCN 模型的传播函数; 当 $\lambda_1 = 1, \lambda_2 = 0$ 时, 该传播函数即为 GAT 模型的传播函数。通过引入插值, 有效结合了 GNN 的两种不同模型。

图注意力网络模型同图卷积网络相同, 在模型层数为 2 时性能达到最优, 图注意力网络正如现有的大部分图神经网络模型相同处于两难境地, 直观上越深的模型表现越好, 但不加任何技巧模块的深层 GNN 模型反而表现越来越糟[44]。

代表性的基于注意力机制的 GNNs 的注意力功能能否在更深的层上保持表达能力? 已有研究提出了导致深度 GNN 性能下降的可能原因, 包括节点特征的过度平滑、过度压缩[45]和过度相关[46]。然而, 专门针对深度图注意力的理论局限性的讨论却很少。

Zhao 等人[47]探索了图变换的模型深度和注意力函数之间的关系, 探讨更深层的模型是否对 Graph Transformer 有益, 并发现当前的 Graph Transformer 在提高性能方面存在瓶颈。进一步的分析揭示, 这是由于全局注意力的消失能力限制了 Graph Transformer 集中关注关键子结构并获得表达性特征的能力, 限制了深度 Graph Transformer 的发展。因此, 本文提出了一种名为深度图(DeepGraph)的新型 Graph Transformer 模型, 该模型在编码表示中明确地使用子结构标记, 并在相关节点上应用局部注意力, 以获得基于子结构的注意力编码。提出的模型增强了全局注意力集中关注子结构的能力, 促进了表示的表达能力, 解决了随着 Graph Transformer 加深而自注意力限制的问题。尽管这篇论文探索了模型深度和注意力函数之间的关系, 但是理解深度图注意力的理论基础尚不清楚。

2023 年, Lee 等人[48]在理论层面上研究了深度图注意力, 聚焦于深度和注意力将两个研究方向连接起来, 解决两个未被充分探索的问题: (a)深度图注意力有哪些独特的挑战, 以及(b)如何设计可证明更具表现力的深度图注意力。通过理论研究发现, 深度图注意力存在对过度平滑特征的脆弱性和平滑累积注意力两个问题。通过理论和实证分析, 发现各种基于注意力的 GNN 都存在这些问题。为避免上述问题的发生, 提出了基于注意力的深度传播图神经网络(Attentive Deep Propagation-GNN, AERO-GNN 模型), 即将边缘注意力与跳注意力相结合, 通过对边缘注意力函数进行改进, 使得模型在面对平滑特征具有抵抗能力; 其次引入跳注意力, 防止模型出现平滑累积。

5. 研究展望

注意力机制和图神经网络是近年来受到广泛关注的研究领域, 在解决图结构数据分析和建模问题方面展现出了巨大的潜力。未来, 随着深度学习和图神经网络的发展, 注意力机制将继续在图数据分析中发挥重要作用, 并且在以下几个方面有着研究展望。

首先, 提升与图注意力网络相关模型的表达能力, 目前的注意力机制主要关注节点之间的关系, 但对于复杂的图结构, 节点和边缘之间的关系也非常重要。未来的研究可以探索如何将注意力机制应用于边缘信息的建模, 以更好地捕捉节点和边缘之间的复杂关系。另外, 可以进一步研究基于注意力机制的

高阶关系建模, 例如节点对之间的二阶关系、三阶关系等, 以提升模型的表达能力。

其次, 将注意力机制与图神经网络相结合, 我们常常无法解释该模型所学习到的内容, 即缺乏一定的可解释性[49]。为此需要强化图神经网络的可解释性和可视化能力。未来的研究可以探索如何利用注意力机制来解释模型的预测结果和特征选择过程, 并提供对图数据分析的可解释性。此外, 可以进一步研究如何通过可视化注意力权重, 帮助研究人员和从业者理解图数据的特征重要性和结构特征。

最后, 当前的注意力机制主要关注节点级别的注意力权重, 但对于图数据中的多尺度结构和多层次的关系, 仅使用节点级别的注意力机制可能不足以捕捉到所有的信息。未来的研究可以考虑设计基于多尺度和多层次的注意力机制, 以更好地处理复杂图结构中的不同级别的关系和特征。

综合而言, 图神经网络与注意力机制作为当前研究方向, 具有广泛的研究前景和实际应用价值。通过提升模型表达能力, 强化模型的可解释性和可视化能力, 构建多尺度结构和多层关系以及解决实际应用, 我们可以预见图注意力相关网络将在未来取得更多突破和进展, 为人工智能领域的发展做出更大贡献。

参考文献

- [1] Gori, M., Monfardini, G. and Scarselli, F. (2005) A New Model for Learning in Graph Domains. *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks*, Montreal, QC, 31 July-4 August 2005, 729-734. <https://doi.org/10.1109/IJCNN.2005.1555942>
- [2] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks. Arxiv:1609.02907
- [3] Lv, Q., Ding, M., Liu, Q., et al. (2021) Are We Really Making Much Progress? Revisiting, Benchmarking and Refining Heterogeneous Graph Neural Networks. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Singapore, 14-18 August 2021, 1150-1160. <https://doi.org/10.1145/3447548.3467350>
- [4] Su, X., Xue, S. and Liu, F. (2022) A Comprehensive Survey on Community Detection with Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21. <https://doi.org/10.1109/TNNLS.2021.3137396>
- [5] Ying, C., Cai, T., Luo, S., et al. (2021) Do Transformers Really Perform Badly for Graph Representation? *Advances in Neural Information Processing Systems*, **34**, 28877-28888.
- [6] Roweis, S.T. and Saul, L.K. (2000) Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, **290**, 2323-2326. <https://doi.org/10.1126/science.290.5500.2323>
- [7] Perozzi, B., Al-Rfou, R. and Skiena, S. (2014) Deepwalk: Online Learning of Social Representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 24-27 August 2014, 701-710. <https://doi.org/10.1145/2623330.2623732>
- [8] Tang, J., Qu, M., Wang, M., et al. (2015) LINE: Large-Scale Information Network Embedding. *Proceedings of the 24th International Conference on World Wide Web*, Florence, 18-22 May 2015, 1067-1077. <https://doi.org/10.1145/2736277.2741093>
- [9] Grover, A. and Leskovec, J. (2016) node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 855-864. <https://doi.org/10.1145/2939672.2939754>
- [10] Yang, R., Shi, J., Yang, Y., et al. (2021) Effective and Scalable Clustering on Massive Attributed Graphs. *Proceedings of the Web Conference*, Ljubljana, 19-23 April 2021, 3675-3687. <https://doi.org/10.1145/3442381.3449875>
- [11] Bruna, J., Zaremba, W., Szlam, A., et al. (2013) Spectral Networks and Locally Connected Networks on Graphs. Arxiv:1312.6203 <https://doi.org/10.48550/arXiv.1312.6203>
- [12] Defferrard, M., Bresson, X. and Vandergheynst, P. (2016) Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Advances in Neural Information Processing Systems*, **29**, 3884-3852.
- [13] Gilmer, J., Schoenholz, S.S., Riley, P.F., et al. (2017) Neural Message Passing for Quantum Chemistry. *International Conference on Machine Learning*, **70**, 1263-1272.
- [14] Battaglia, P.W., Hamrick, J.B., Bapst, V., et al. (2018) Relational Inductive Biases, Deep Learning, and Graph Networks. Arxiv:1806.01261
- [15] Hamilton, W., Ying, Z. and Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems*, **30**, 1024-1034.

- [16] Veličković, P., Cucurull, G., Casanova, A., *et al.* (2017) Graph Attention Networks. Arxiv:1710.10903
- [17] Xu, K., Hu, W., Leskovec, J., *et al.* (2018) How Powerful Are Graph Neural Networks? Arxiv:1810.00826
- [18] 申翔翔, 侯新文, 尹传环. 深度强化学习中状态注意力机制的研究[J]. 智能系统学报, 2020, 15(2): 317-322.
- [19] 高芬, 苏依拉, 牛向华, 等. 基于 Transformer 的蒙汉神经机器翻译研究[J]. 计算机应用与软件, 2020, 37(2): 141-146+225.
- [20] Chaudhari, S., Mithal, V., Polatkan, G., *et al.* (2021) An Attentive Survey of Attention Models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **12**, 1-32. <https://doi.org/10.1145/3465055>
- [21] 任欢, 王旭光. 注意力机制综述[J]. 计算机应用, 2021, 41(S1): 1-6.
- [22] Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., *et al.* (1995) Modeling Visual Attention via Selective Tuning. *Artificial Intelligence*, **78**, 507-545. [https://doi.org/10.1016/0004-3702\(95\)00025-9](https://doi.org/10.1016/0004-3702(95)00025-9)
- [23] Itti, L., Koch, C. and Niebur, E. (1998) A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1254-1259. <https://doi.org/10.1109/34.730558>
- [24] Mnih, V., Heess, N. and Graves, A. (2014) Recurrent Models of Visual Attention. *Advances in Neural Information Processing Systems*, **27**, 1106-1114.
- [25] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. Arxiv:1409.0473
- [26] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 3104-3112.
- [27] Kitaev, N., Kaiser, L. and Levskaya, A. (2020) Reformer: The Efficient Transformer. Arxiv:2001.04451
- [28] 王明申, 牛斌, 马利. 一种基于词级权重的 Transformer 模型改进方法[J]. 小型微型计算机系统, 2019, 40(4): 744-748.
- [29] Zhan, Q., Cai, X., Chen, C., *et al.* (2017) Commented Content Classification with Deep Neural Network Based on Attention Mechanism. *Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference*, Chongqing, 25-26 March 2017, 2016-2019. <https://doi.org/10.1109/IAEAC.2017.8054369>
- [30] Hao, S., Lee, D.-H. and Zhao, D. (2019) Sequence to Sequence Learning with Attention Mechanism for Short-Term Passenger Flow Prediction in Large-Scale Metro System. *Transportation Research Part C: Emerging Technologies*, **107**, 287-300. <https://doi.org/10.1016/j.trc.2019.08.005>
- [31] 周才东, 曾碧卿, 王盛玉, 等. 结合注意力与卷积神经网络的中文摘要研究[J]. 计算机工程与应用, 2019, 55(8): 132-137.
- [32] Salehi, A. and Davulcu, H. (2020) Graph Attention Auto-Encoders. 2020 *IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, Baltimore, MD, 9-11 November 2020, 989-996. <https://doi.org/10.1109/ICTAI50040.2020.00154>
- [33] Zhang, J., Shi, X., Xie, J., *et al.* (2018) GaAN: Gated Attention Networks for Learning on Large and Spatiotemporal Graphs. Arxiv:1803.07294
- [34] Busbridge, D., Sherburn, D., Cavallo, P., *et al.* (2019) Relational Graph Attention Networks. Arxiv:1904.05811
- [35] Schlichtkrull, M., Kipf, T.N., Bloem, P., *et al.* (2018) Modeling Relational Data with Graph Convolutional Networks. *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, 3-7 June 2018*, 593-607. https://doi.org/10.1007/978-3-319-93417-4_38
- [36] Li, Y., Tarlow, D., Brockschmidt, M., *et al.* (2015) Gated Graph Sequence Neural Networks. Arxiv:1511.05493
- [37] Lee, J., Lee, I. and Kang, J. (2019) Self-Attention Graph Pooling. *International Conference on Machine Learning*, **97**, 3734-3743.
- [38] Do, K., Tran, T., Nguyen, T., *et al.* (2019) Attentional Multilabel Learning over Graphs: A Message Passing Approach. *Machine Learning*, **108**, 1757-1781. <https://doi.org/10.1007/s10994-019-05782-6>
- [39] Pham, T., Tran, T., Phung, D., *et al.* (2017) Column Networks for Collective Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**, 2485-2491. <https://doi.org/10.1609/aaai.v31i1.10851>
- [40] Lee, J.B., Rossi, R. and Kong, X. (2018) Graph Classification Using Structural Attention. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 August 2018, 1666-1674. <https://doi.org/10.1145/3219819.3219980>
- [41] Abu-El-Haija, S., Perozzi, B., Al-Rfou, R., *et al.* (2018) Watch Your Step: Learning Node Embeddings via Graph Attention. *Advances in Neural Information Processing Systems*, **31**, 9180-9190.
- [42] Brody, S., Alon, U. and Yahav, E. (2021) How Attentive Are Graph Attention Networks? Arxiv:2105.14491

- [43] Javaloy, A., Sanchez-Martin, P., *et al.* (2022) Learnable Graph Convolutional Attention Networks. Arxiv:2211.11853
- [44] Zhang, W., Sheng, Z., Yin, Z., *et al.* (2022) Model Degradation Hinders Deep Graph Neural Networks. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington DC, 14-18 August 2022, 2493-2503. <https://doi.org/10.1145/3534678.3539374>
- [45] Alon, U. and Yahav, E. (2021) On the Bottleneck of Graph Neural Networks and Its Practical Implications. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2006.05205>
- [46] Jin, W., Liu, X., Ma, Y., Aggarwal, C. and Tang, J. (2022) Feature Overcorrelation in Deep Graph Neural Networks: A New Perspective. *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Washington DC, 14-18 August 2022, 709-719. <https://doi.org/10.1145/3534678.3539445>
- [47] Zhao, H., Ma, S., Zhang, D., Deng, Z.-H. and Wei, F. (2023) Are More Layers Beneficial to Graph Transformers? *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2303.00579>
- [48] Lee, S.Y., Bu, F., Yoo, J., *et al.* (2023) Towards Deep Attention in Graph Neural Networks: Problems and Remedies. Arxiv:2306.02376
- [49] Kim, D. and Oh, A. (2022) How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision. Arxiv:2204.04879