工程科学学报
Chinese Journal of Engineering

利用变分卷积推断局部拓扑结构的图表示方法

侯静怡 唐宇鑫 于欣波 刘志杰

Inferring local topology *via* variational convolution for graph representation

HOU Jingyi, TANG Yuxin, YU Xinbo, LIU Zhijie

引用本文:
侯静怡, 唐宇鑫, 于欣波, 刘志杰. 利用变分卷积推断局部拓扑结构的图表示方法[J]. 工程科学学报, 2023, 45(10): 1750–1758. doi: 10.13374/j.issn2095–9389.2022.07.24.005
HOU Jingyi, TANG Yuxin, YU Xinbo, LIU Zhijie. Inferring local topology *via* variational convolution for graph representation[J]. *Chinese Journal of Engineering*, 2023, 45(10): 1750–1758. doi: 10.13374/j.issn2095–9389.2022.07.24.005

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2022.07.24.005

工程科学学报, 第 45 卷, 第 10 期: 1750−1758, 2023 年 10 月
Chinese Journal of Engineering, Vol. 45, No. 10: 1750−1758, October 2023
https://doi.org/10.13374/j.issn2095-9389.2022.07.24.005; http://cje.ustb.edu.cn

利用变分卷积推断局部拓扑结构的图表示方法

侯静怡[1,2,3)]，唐宇鑫[1,2,3)]，于欣波[1,2,3)]，刘志杰[1,2,3)]✉

1) 北京科技大学智能科学与技术学院, 北京 100083　2) 北京科技大学人工智能研究院, 北京 100083　3) 北京科技大学智能仿生无人系统教育部重点实验室, 北京 100083

✉通信作者，E-mail: liuzhijie2012@gmail.com

摘　要　深度学习技术的长足发展与数据算力的快速提升, 极大地增加了各种结构图神经网络优化和实现的可行性, 使得图结构数据的表示研究工作取得极大进展. 已有的图神经网络方法主要关注图节点之间全局信息的传递, 理论上可证明其强大的信息表示能力. 然而, 面向局部拓扑具有特殊语义的图结构数据表示时, 这些通用方法缺乏灵活的局部结构表示机制, 例如化学反应中组成分子的局部结构—官能团, 其通常能够决定化学分子性质并且参与化学反应过程. 进一步挖掘这些局部结构的信息对基于图表示的各类任务都是非常重要的, 为此提出一个利用变分卷积推断局部拓扑结构的图表示方法, 不仅考虑图节点在全局结构上的关系推理与信息传递, 还基于变分推断自适应地学习图数据的局部拓扑结构, 利用卷积操作对局部结构进行编码, 从而进一步提高图神经网络的表达能力. 本文工作在多个图结构数据集上进行实验, 实验结果表明利用局部结构信息可以有效提升图神经网络在基于图的相关任务上的性能.

关键词　图注意力网络; 局部拓扑结构; 变分推断; 卷积神经网络; 混合结构神经网络

分类号　TP391.4

Inferring local topology *via* variational convolution for graph representation

HOU Jingyi[1,2,3)], *TANG Yuxin*[1,2,3)], *YU Xinbo*[1,2,3)], *LIU Zhijie*[1,2,3)]✉

1) School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China

2) Institute of Artificial Intelligence, University of Science and Technology Beijing, Beijing 100083, China

3) Key Laboratory of Perception and Control of Intelligent Bionic Unmanned Systems (Ministry of Education), University of Science and Technology Beijing, Beijing 100083, China

✉Corresponding author, E-mail: liuzhijie2012@gmail.com

ABSTRACT　The development of deep learning techniques and support of big data computing power have revolutionized graph representation research by facilitating the implementation of the learning of different graph neural network structures. Existing methods, such as graph attention networks, mainly focus on global information propagation in graph neural networks, which have theoretically proven their strong representation capability. However, these general methods lack flexible representation mechanisms when facing graph data with local topology involving specific semantics, such as functional groups in the chemical reaction. Accordingly, it is of great importance to further exploit the local structure representations for graph-based tasks. Several existing methods either use domain expert knowledge or conduct subgraph isomorphism counting to learn local topology representations of graphs. However, there is no guarantee that these methods can easily be generalized to different domains without specific knowledge or complex substructure preprocessing. In this study, we propose a simple and automatic local topology inference method that uses variational convolutions to improve the local representation ability of graph attention networks. The proposed method not only considers the relationship reasoning and message passing on the global graph structure but also adaptively learns the graph's local structure representations with the guidance

收稿日期: 2022−07−24

基金项目: 国家自然科学基金资助项目(62106021, U20A20225); 北京科技大学青年教师学科交叉研究培育项目(FRF-IDRY-21-021)

of statistical priors that can be readily accessible. To be more specific, the variational inference is used to adaptively learn the convolutional template size, and the inference is conducted layer-by-layer with the guidance of the statistical priors to make the convolutional template size adaptable to multiple subgraphs with different structures in a self-supervised way. The variational convolution module is easily pluggable and can be concatenated with arbitrary hidden layers of any graph neural network. In contrast, due to the locality of the convolution operations, the relations between graph nodes can be further sparse to alleviate the over-squeezing problem in the global information propagation of the graph neural network. As a result, the proposed method can significantly improve the overall representation ability of the graph attention network using the variational inference of the convolutional operations for local topology representation. Experiments are conducted on three large-scale and publicly available datasets, i.e., the OGBG-MolHIV, USPTO, and Buchwald-Hartwig datasets. Experimental results show that exploiting various kinds of local topological information helps improve the performance of the graph attention network.

KEY WORDS graph attention network; local topology; variational inference; convolutional neural network; hybrid neural network

图神经网络（Graph neural network, GNN）通过信息传递的聚合机制学习图数据节点的低维特征表示，即根据邻居节点聚合后的信息，递归地更新图上的每个节点表示．近年来，随着领域数据的大量收集和算力的大幅提升，基于深度学习的 GNN 广泛应用于各个领域，例如：推荐系统[1]、药物发现[2] 等，甚至计算机视觉[3]、自然语言处理[4] 领域．得益于深度学习开发工具的灵活性，空域 GNN 更容易实现，其结构设计愈发多样化，例如利用循环神经网络[5]、探究多种聚合操作[6]、基于注意力机制[7] 的 GNN，更有文献 [8] 的聚合机制已被证明可超越 Weisfeiler-Lehman（WL）测试，从而大大提高了 GNN 关于图结构数据的表达能力．基于注意力机制[9] 的 GNN，例如图注意力网络（Graph attention network, GAT）[7] 和 Transformer[4]，理论上也能够学得具有通过 WL 测试能力的模型参数，在此基础上，通常还会额外加入位置编码机制来保留节点拓扑结构信息．此类模型灵活性高，可自动学习节点之间的边和边的权重，无需预先定义图的结构．

尽管理论上基于注意力机制的 GNN 学习能力强大，然而在实际应用中，该结构模型并没有普遍应用于所有图结构数据[10]．原因在于这类模型更加注重全局图结构的学习，节点的表示根据邻居节点的阶数度量．而实际应用中，有些图数据的局部子图结构会呈现特定的语义，例如，化学反应预测相关任务中的官能团结构，物理特性预测相关任务中的特定力学结构等，这些语义往往对于下游任务意义重大．通用 GNN 节点间信息传递不考虑对这些局部结构进行特殊处理，导致学得的图表示难以包含此类信息，从而影响模型在下游任务的性能．另外，逐层传递高阶邻居节点信息，会指数级地增加邻节点信息传递量，造成过挤压（Over-squashing）问题[11]．因此，考虑与当前节点最相关邻节点的信息聚合与拓扑结构是非常有必要的．

本文关注提高 GNN 对局部子图拓扑结构的表达能力．近期已有工作设计挖掘图局部结构信息的 GNN，在理论和实践上均证明该机制的有效性．Jin 等[12] 设计联结树变分自编码器（Junction-tree variational autoencoder, JT-VAE），根据化学中的先验知识，将分子图中带有语义的子图合并形成一个无环联结树，该分子图和相应联结树进行编解码操作．Chen 等[13] 通过衡量子图计数能力来判断 GNN 的表达能力，证明了子图结构表示对于 GNN 的必要性．Ying 等[10] 将局部中心、边、节点信息融合到节点之间边权重的计算中来提高 Transformer 网络的表达能力．Bouritsas 等[14] 提出了一种通过子图同构增强局部拓扑表达的 GNN，在理论上证明了该模型的表达能力优于 WL 测试．Yu 等[15] 通过遍历预定义子图结构来学习药物局部分子结构间的相互作用，增强药物相互作用预测模型的可解释性．这些方法充分展示了局部拓扑结构的重要意义，然而在实现上需要额外的领域专家知识，或进行复杂的子图同构计算，在泛化应用和高效计算上存在一定局限性．

为简单高效地增强基于注意力机制的 GNN 的表达能力，本文提出一个启发式的神经网络模块，称其为自适应卷积模块．无需领域专家知识和子图同构计算，自适应卷积模块利用卷积模板对不同局部模式响应不同的机制，来近似学习不同局部拓扑结构包含的特定语义信息．为了使卷积模板的尺寸可适应多种不同结构的子图，采用变分推断自适应地学习卷积模板尺寸，并且仅利用基于统计的先验分布即可自监督地完成推断．自适应卷积模块具有可插拔性质，能够嵌入 GNN 的各层．另外，由于卷积操作的局部性质，可进一步稀疏化全局信息传递中的关系矩阵，减轻过挤压

问题. 与基于卷积的 GNN 中卷积层对所有节点均采用固定大小的卷积模板不同, 本方法不仅能够自适应学习卷积模板尺寸, 而且能够根据节点的位置不同有针对性进行卷积计算, 便于局部拓扑结构的提取. 本文工作在 OGBG-MolHIV[16]、USPTO[17] 和 Buchwald-Hartwig[18] 数据集上分别进行了验证.

本文的主要贡献包括 3 个方面:

(1) 关注利用局部结构增强基于注意力机制的 GNN 的表达能力, 据此提出一个全新的自适应变分卷积推断的模块;

(2) 提出的模块简单有效, 只利用基于统计和启发式的先验即可实现, 无需额外的专家知识和复杂的预处理过程;

(3) 在公共数据集上的实验结果展示了所提出的模块在学习图局部结构信息方面的强大能力.

1　相关工作

1.1　图网络中引入卷积方法

本文方法采用了图网络(即 Transformer 模型)与卷积操作相结合的方式, 使模型同时具备全局建模和局部信息捕捉的能力. 这种全局–局部的建模方式在自然语言处理[19–20]、语音识别[21]、图像分类[22–23] 等任务上已证明具有良好的性能.

Wu 等[19] 将自注意力图模块中的多头注意力替换为卷积层, 用于多个自然语言处理任务. Wu 等[20] 将自注意力图模型和卷积层并行计算, 用于处理自然语言信息. 以上两种模型均为轻量级模型, 减轻了自然语言处理任务对 Transformer 等大规模模型的依赖, 同时建模全局和局部信息, 有效提升了模型的性能. Gulati 等[21] 交替进行自注意力图信息传播与卷积操作, 弥补了全局特征表示丢失局部细节和局部卷积难以学得全局表示的缺点, 减少参数量的同时提高了语音识别的效果. Wang 等[22] 将自注意力图网络学得的注意力映射作为图像通道输入卷积进行处理, 该方法在图像分类、机器翻译等多种任务上均取得了良好效果. Wu 等[23] 使用卷积映射计算注意力操作中的各个参数, 用于在大规模图像分类. Si 等[24] 将卷积、最大池化和自注意力机制并行的方式, 从而同时捕捉低频全局特征和高频局部特征, 该方法在图像分类、分割、检测任务上取得了出色的性能.

上述方法主要关注自然界信号特征提取, 该类信号表示较为稠密, 而很多图结构数据表示十分紧凑, 每一个节点信息都十分关键. 与上述方法不同, 本文主要关注同一语义层级多种尺度局部信息精确选择与提取, 不局限于自然界的信号, 应用范围更加广泛.

1.2　多尺度卷积方法

提取多尺度信息的方法被广泛应用于计算机视觉相关任务, 该方法可以获取尺度不变的鲁棒特征. 在神经网络的一层中使用多种尺度卷积, 最有代表性的工作为 Inception 网络[25–26], 这类网络使用多尺度卷积作为一个模块, 并行提取不同尺度的信息, 将多个模块串联构成整体网络, 其中多尺度卷积输出特征直接拼接. Zhou 等[27] 采用多尺度时序进行视频中的时序关系检测, 将不同时间尺度的视频帧序列卷积后加权相加得到特征表示. Kim[28] 使用多种尺度卷积模板对文本进行信息抽取, 将不同模板得到的信息池化为一个标量后, 直接级联得到定长特征用于句子级别文本分类.

与这类方法不同, 本方法关注多尺度卷积信息的选择, 为降低多尺度卷积信息的冗余问题, 采用变分推断的方法使得模型根据数据本身自适应地学习化学反应式官能团层级的语义信息.

2　模型结构

局部卷积操作需要根据图节点位置进行计算, 即对相邻位置的图节点进行同一组卷积计算. 在实现时可以先对图节点位置进行排序, 根据情况采用 1 维排序或高维排序, 然后进行卷积计算. 例如 3D 点云结构可以直接根据位置排列, 采用 3D 卷积进行计算. 为了表述方便, 本文以化学分子相关预测任务为例详细介绍本方法的实现过程. 化学分子式根据 SMILES 格式进行表示与线性排序, 对其进行 1D 卷积计算即可.

本文采用基于 Transformer 编码网络作为主干模型, 在该网络的隐藏层中插入自适应卷积模块, 从而在局部传递化学分子间信息、推理分子两两关系的基础上, 编码化学式局部结构, 充分挖掘化学反应式序列中局部拓扑和全局结构关系表示, 应用于化学反应相关预测任务上. 分以下三部分介绍本方法: 基于 Transformer 的分词特征提取、自适应卷积模块、回归预测任务. 具体而言, 首先将化学式转化为 SMILES 格式表示并分词, 使用 Transformer 模型提取分词后的节点特征和全局特征; 然后将节点特征输入自适应卷积模块对化学反应式的局部结构进行编码表示; 最后将其与全局特征融合用于回归预测.

2.1　分词特征提取

给定化学反应式的 SMILES 格式作为输入, 对

该表达式进行分词得到各原子层级节点，并使用Transformer模型的编码器提取特征，即将该模型作为本方法的骨干（Backbone）模型，经自注意力网络层的信息传递得到各原子节点的特征表示，定义为 $\boldsymbol{X}^{\mathrm{b}}=\left[\boldsymbol{x}_{\mathrm{cls}}^{\mathrm{b}},\boldsymbol{x}_{1}^{\mathrm{b}},\boldsymbol{x}_{2}^{\mathrm{b}},\cdots,\boldsymbol{x}_{N_{\mathrm{b}}}^{\mathrm{b}}\right]\in\mathbf{R}^{d_{\mathrm{b}}\times(N_{\mathrm{b}}+1)}$. 其中，$d_{\mathrm{b}}$表示特征维度，$N_{\mathrm{b}}$为化学反应式中分词节点的个数，$\boldsymbol{x}_{\mathrm{cls}}^{\mathrm{b}}$为全局分类节点特征，$\boldsymbol{x}_{i}^{\mathrm{b}}|_{i=1}^{N_{\mathrm{b}}}$表示各原子节点分词特征. Transformer模型对各原子进行关系推理，其目的在于从反应式全局角度获取分类节点特征$\boldsymbol{x}_{\mathrm{cls}}^{\mathrm{b}}$，以表征化学反应式中各原子之间的相互作用. 该过程如图1所示.

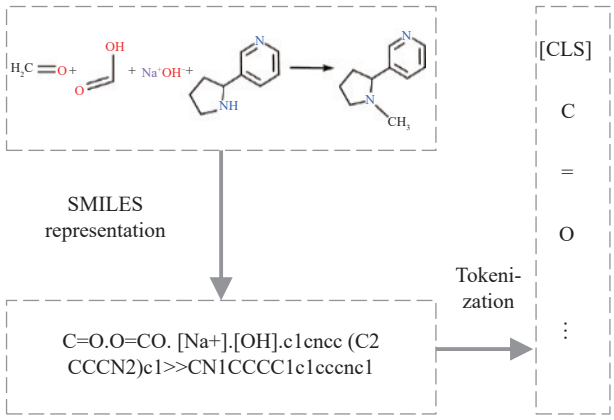

图1 化学反应分子式的SMILES格式转化与分词示意图

Fig.1 Example of SMILES format transformation and tokenization of chemical reaction

由于Transformer模型中的编码器网络结构在本质上属于自注意力图网络，其主要作用是传递节点之间长时和全局的相互关系的信息，因此缺乏对局部邻域结构信息的表达. 而在化学反应式中，除了原子之间的相互作用外，若干个相邻原子所形成的语义信息对化学式的表达也有较大影响，在化学分子式中，该语义是显式的，称为官能团. 有机化学反应主要发生在官能团上，因此，官能团对有机化合物的性质起决定作用. 需要注意的是，本方法旨在面向所有图结构数据的局部拓扑信息表示，设计模型算法来自适应地学习局部表示，无需借助官能团本身的语义信息等领域知识，即只需统计各官能团的结构和原子组成作为先验知识. 为表述方便，后文中称化学分子式局部结构为官能团.

2.2 自适应卷积模块

采用基于1D卷积操作对原子特征序列进行编码，以得到化学反应式中各官能团的特征表示，即局部结构特征. 卷积核的尺寸对应于组成官能团的原子数量，自适应局部结构信息编码模块能够在计算卷积过程中自动推断适合的卷积核尺寸. 图2所示为以2个尺度的卷积核为例的自适应卷积模块示意图.

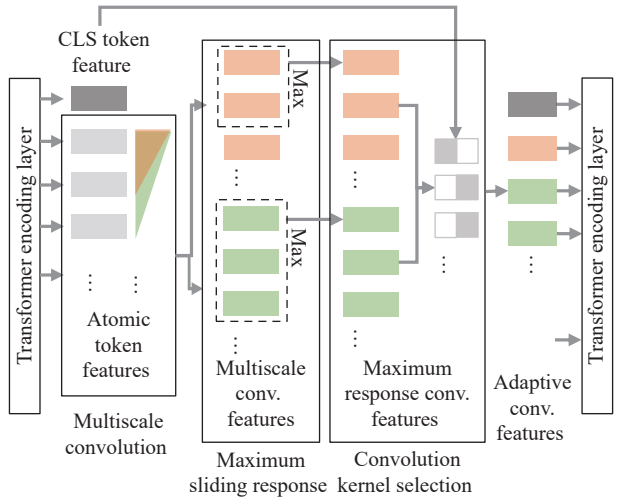

图2 自适应卷积模块示意图

Fig.2 Illustration of the self-adaptive convolutional module

具体地，给定不同尺寸的1D卷积核模板集合，其中各卷积核模板的尺寸定义为$\{k_i|i=1,\cdots,n\}$. 首先，对除分类节点以外的全部原子节点序列在全部的n个卷积模板上进行卷积操作以进行局部信息特征表示，探索官能团的多种组合. 由于同一局部结构中各个节点的局部结构表示应是相同的，传统的卷积操作无法实现. 因此，增加一个长度为k_i的滑动窗口，取窗口中卷积模板的最大响应值，来近似获取一定局部范围内与卷积模板最匹配的局部结构，称该操作为滑动最大响应卷积. 不失一般性，第j个原子节点在第i个卷积模板上的滑动最大响应卷积操作为

$$\boldsymbol{x}_{j}^{i}=\max\left(\mathrm{Conv}_{k_i}\left(\boldsymbol{x}_{j-k_i+1}^{\mathrm{b}},\cdots,\boldsymbol{x}_{j}^{\mathrm{b}};\theta_{k_i}\right),\cdots,\mathrm{Conv}_{k_i}\left(\boldsymbol{x}_{j},\cdots,\boldsymbol{x}_{j+k_i-1}^{\mathrm{b}};\theta_{k_i}\right)\right)\tag{1}$$

其中，θ_{k_i}为可学习的卷积核参数，$\max(\cdot)$表示对卷积输出每一通道取最大值，等价于步长为1的最大池化操作.

其次，令$\boldsymbol{a}_j\in\{0,1\}^{n\times1}$为独热（One-hot）编码，表示卷积核选择器，则第j个原子节点卷积后的特征计算为

$$\boldsymbol{x}_{j}^{\mathrm{c}}=\left[\boldsymbol{x}_{j}^{1},\cdots,\boldsymbol{x}_{j}^{n}\right]\boldsymbol{a}_j\tag{2}$$

利用卷积后特征$\boldsymbol{x}_{j}^{i}$计算当前节点下取该卷积核的概率

$$\boldsymbol{\pi}_j[i]=\mathrm{softmax}_{k_i}\left(\phi\left(\boldsymbol{x}_{j}^{i},\varphi\left(\boldsymbol{x}_{\mathrm{cls}}^{\mathrm{b}};\theta_{\varphi}\right);\theta_{\phi}\right)\right)\tag{3}$$

其中，$\phi(\cdot)$和$\varphi(\cdot)$表示两个空间映射，$\varphi(\cdot)$用于将分

类节点特征$\boldsymbol{x}_{\text{cls}}^{\text{b}}$映射到局部结构特征$\boldsymbol{x}_j^i$所在空间, $\phi(\cdot)$将接收到的$\boldsymbol{x}_{\text{cls}}^{\text{b}}$和$\boldsymbol{x}_j^i$进行级联并映射到一维空间以便于计算概率, 两个空间变换均通过前馈神经网络层来实现, θ_ϕ和θ_φ分别表示两个全连接层的可学习参数. 根据卷积核的概率$\boldsymbol{\pi}_j[i]$计算卷积核选择器$\boldsymbol{a}_j$对应元素的取值, 从而选择当前节点最合适的卷积核, 并进行卷积计算.

将原子节点分词特征$\boldsymbol{x}_j^{\text{b}}$和自适应卷积操作计算得到的特征$\boldsymbol{x}_j^{\text{c}}$融合, 得到自适应卷积模块特征$\boldsymbol{x}_j$, 融合方法为特征级联后进行非线性映射. 得到的自适应卷积模块特征可以继续输入 Transformer 的自注意力网络层中进行全图节点之间的信息传递.

2.3　回归预测

交替叠加自注意力层和自适应编码层后, 将得到的图节点特征池化为全图表示, 进而用于下游分类或回归任务. 使用自适应平均池化操作, 将图节点表示为定长特征$\boldsymbol{v}$, 使用该操作可以尽可能保留局部结构特征. 给定表示化学式全局信息的分类节点特征$\boldsymbol{x}_{\text{cls}}$与局部信息的特征$\boldsymbol{v}$, 对二者加权后级联得到的化学反应式的最终特征表示:

$$\boldsymbol{f}=\left[\beta\boldsymbol{x}_{\text{cls}}{}^{\text{T}};(1-\beta)\boldsymbol{v}^{\text{T}}\right]^{\text{T}} \tag{4}$$

其中β是超参数, 用于调节两种特征的权重. 得到的特征$\boldsymbol{f}$通过非线性映射层即可得到化学反应的预测值$\hat{y}$.

3　模型优化

考虑到卷积核选择器的学习中, 直接使用不同卷积核的概率$\boldsymbol{\pi}_j[i]$对卷积核选择器$\boldsymbol{a}_j$对应的元素赋值易导致模型学习不稳定, 使得模型优化失效. 本文使用变分推断方法, 将卷积核选择器作为随机变量来估计其后验分布, 通过最大化变分下界来近似估计其后验分布的参数. 记变分下界为

$$V_{\text{LB}}(\theta)=E_{q_\theta(\boldsymbol{a}_j|\boldsymbol{x}_j)}\left[\ln p_\psi\left(y|\boldsymbol{x}_{1:N_{\text{b}}},\boldsymbol{a}_{1:N_{\text{b}}}\right)\right]-\sum_{j=1}^{N_{\text{b}}}\text{KL}\left[q_\theta\left(\boldsymbol{a}_j|\boldsymbol{x}_j\right)\|p\left(\boldsymbol{a}_j\right)\right] \tag{5}$$

其中, $\ln p_\psi\left(y|\boldsymbol{x}_{1:N_{\text{b}}},\boldsymbol{a}_{1:N_{\text{b}}}\right)$为关于预测$y$的近似后验概率的对数似然函数, $\text{KL}\left[q_\theta\left(\boldsymbol{a}_j|\boldsymbol{x}_j\right)\|p\left(\boldsymbol{a}_j\right)\right]$是关于$\boldsymbol{a}_j$的近似后验与先验的 Kullback-Leibler 散度.

对于计算y的近似后验概率密度函数, 假设

$$p_\psi\left(y|\boldsymbol{x}_{1:N_{\text{b}}},\boldsymbol{a}_{1:N_{\text{b}}}\right)=\mathcal{N}\left(y;\hat{y},\sigma^2\right) \tag{6}$$

即真实值y服从以预测值$\hat{y}$为均值、标准差为固定值σ的正态分布, 则对数似然函数计算为

$$\ln p_\psi\left(y|\boldsymbol{x}_{1:N_{\text{b}}},\boldsymbol{a}_{1:N_{\text{b}}}\right)=-\ln\sigma-\frac{\ln(2\pi)}{2}-\frac{(\hat{y}-y)^2}{2\sigma^2} \tag{7}$$

由于公式 (7) 的第一、二项为常量, 于是在实际计算中, 公式 (5) 中的第一项可直接使用带有系数γ的负均方误差(Mean squared error)的期望代替.

对于公式 (5) 的第二项—KL 散度, 本文采用观测统计的官能团长度信息计算先验分布. 该观测结果包含噪声和不确定性, 设定随机先验服从多项分布, 即

$$p\left(\boldsymbol{a}_j\right)=\text{Multinomial}\left(\boldsymbol{\pi}_j^{\text{prior}}\right) \tag{8}$$

其中, $\boldsymbol{\pi}_j^{\text{prior}}$是多项分布的参数. 类似地, 定义近似后验分布为

$$q_\theta\left(\boldsymbol{a}_j|\boldsymbol{x}_j\right)=\text{Multinomial}\left(\boldsymbol{\pi}_j\right) \tag{9}$$

为了降低估计参数θ的波动性, 采用重参数技巧[29−30]. 由于该方法无法用于离散分布, 本文将多项分布连续化[31], 即令$p\left(\boldsymbol{a}_j\right)=\text{Concrete}(\boldsymbol{a}_j;\boldsymbol{\pi}_j^{\text{prior}},\tau^{\text{prior}})$, 以及$q_\theta(\boldsymbol{a}_j|\boldsymbol{x}_j)=\text{Concrete}(\boldsymbol{a}_j;\boldsymbol{\pi}_j,\tau)$, 其中, $\tau^{\text{prior}},\tau>0$分别为两连续分布的尺度因子参数, 连续分布$\text{Concrete}(\boldsymbol{a};\boldsymbol{\pi},\tau)$定义为$\boldsymbol{a}[i]=\text{softmax}_i(\boldsymbol{g}[i])$, $\boldsymbol{g}[i]$服从 Gumbel 分布, 可利用$\boldsymbol{g}[i]=(\ln(\boldsymbol{\pi}[i])-\ln(-\ln(U)))/\tau$的方式进行采样, 其中$U\sim(0,1)$. 因此, 公式 (5) 中的第二项计算为

$$\sum_{j=1}^{N_{\text{b}}}\text{KL}\left[q_\theta\left(\boldsymbol{a}_j|\boldsymbol{x}_j\right)\|p\left(\boldsymbol{a}_j\right)\right]=\sum_{j=1}^{N_{\text{b}}}E_{q_\theta(\boldsymbol{a}_j|\boldsymbol{x}_j)}\left[\ln\frac{q_\theta\left(\boldsymbol{a}_j|\boldsymbol{x}_j\right)}{p\left(\boldsymbol{a}_j\right)}\right] \tag{10}$$

利用蒙特卡洛采样近似计算期望$E_{q_\theta(\boldsymbol{a}_j|\boldsymbol{x}_j)}$, 从而完成对该 KL 散度的计算:

$$\sum_{j=1}^{N_{\text{b}}}\text{KL}\left[q_\theta\left(\boldsymbol{a}_j|\boldsymbol{x}_j\right)\|p\left(\boldsymbol{a}_j\right)\right]\approx\frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{N_{\text{b}}}\left((n-1)\ln\frac{\tau}{\tau^{\text{prior}}}+\sum_{i=1}^{n}\ln\frac{\boldsymbol{\pi}_j[i]\left(\boldsymbol{a}_j^{(l)}[i]\right)^{\tau^{\text{prior}}-\tau}\sum\limits_{k=1}^{n}\boldsymbol{\pi}_j^{\text{prior}}[k]\left(\boldsymbol{a}_j^{(l)}[k]\right)^{-\tau^{\text{prior}}-1}}{\boldsymbol{\pi}_j^{\text{prior}}[i]\sum\limits_{k=1}^{n}\boldsymbol{\pi}_j[k]\left(\boldsymbol{a}_j^{(l)}[k]\right)^{-\tau-1}}\right) \tag{11}$$

本文做出关于先验设定的假设: 不同层级的特征表示中, 图数据固有的局部结构是不会改变的, 所以不同层级的自适应卷积模块的卷积核选择器是一致的. 据此, 将较高层级的卷积核选择器的后验作为较低层级选择器的先验, 这是因为接近预测层的特征表示抽象程度高, 表达的语义更加完整、更有意义. 而最高层级的先验, 本文根据在已有的公开数据集上分子式官能团长度的统计

结果，设置近似指数分布的先验. 图 3 所示为 USPTO 数据集上的统计结果示例.

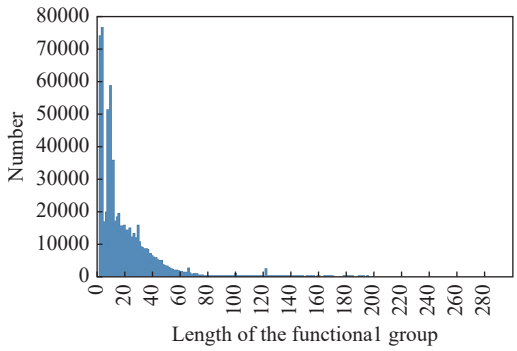

图 3 USPTO 数据集克级分子式官能团长度分布图

Fig.3 Scale distribution of functional groups of the molecules in the USPTO gram set

4 实验

为了验证本文方法的有效性，在 OGBG-MolHIV[16]、USPTO[17] 和 Buchwald-Hartwig[18] 数据集上进行实验，在第一个数据集上使用 ROC-AUC 评价指标对方法性能进行评估，对于后两个数据集采用 R^2 评价指标评估.

4.1 数据集

OGBG-MolHIV 数据集用于预测分子特性是否抑制 HIV 病毒任务，该数据集包含有 41127 种化合物，并将这些化合物标注为 2 类：非活动性（确诊非活动性）和活动性（确诊中度活动性和确诊活动性），该数据集的划分准则与文献 [16] 相同，即训练集、验证集和测试集数据的比例为 8 ： 1 ： 1.

USPTO 数据集用于化学反应收率预测任务，其中化学反应收率是由多位科学家用不同的设备测量的，该数据集中的数据存在一定程度的缺失，如反应产物、反应条件等，以至于整体数据噪声大、结构稀疏. 另外，该数据集包含了克级（Gram）和亚克级（Sub-gram）的化学反应，每一级别所包含的反应种类多，且数量多，整个数据集的反应式总数量多于 50 万个，因此极具挑战性. 根据文献 [32]，进行实验验证时，分别在克级、亚克级反应上预测收率，并使用不同的训练–测试数据划分方式进行验证：一种为随机划分，另一种为时间划分，其中时间划分是以 2012 年为界，在此之前发布的数据作为训练数据，在此之后发布的数据作为测试数据. 另外，由于该数据集中数据噪声过大，对数据的收率进行了平滑操作并报告实验结果，即

$$\bar{y}_i = \frac{1}{5}\left(\sum_{y_i^{\mathrm{NN}}\in \mathrm{NN}(y_i)} y_i^{\mathrm{NN}} + 2y_i\right) \tag{12}$$

其中 y_i 和 $\bar{y}_i$ 分别表示平滑前后的第 i 个化学反应式的收率值，$\mathrm{NN}(y_i)$ 表示 y_i 的 3-近邻数据收率的集合.

为了验证本文方法在面向一类特定的反应的收率预测的有效性，在 Buchwald-Hartwig 数据集上也进行了实验，该数据集只包含 Buchwald-Hartwig 胺化反应. 与已有方法[18] 相同，本文分别在该数据集的 4 个测试集上进行测试，并计算平均 R^2 值，以评价和比较方法性能.

4.2 实验设置

使用 RDKit[33] 工具对化学反应式进行 SMILES 分词操作[34]，在 USPTO 和 Buchwald-Hartwig 数据集上直接微调反应指纹（Reaction fingerprint, rxnfp）模型[35]，在 OGBG-MolHIV 数据集上采用无预训练的相同结构模型. 将 1D 卷积层输出通道数设置为 512，将自适应池化层输出长度 N_p 设置为 4，根据文献 [31] 的经验结论，设置 $\tau^{\mathrm{prior}} = 2/5, \tau = 2/3$. 1D 卷积层数量为 2 层，分别插入基于 Transformer 的主干网络的中间层和最后一层，考虑到计算资源限制，经初步实验结果表明，设置 6 种尺寸卷积核最佳，分别为 1、2、5、6、7 和 30，先验概率分别为 0.4、0.3、0.2、0.04、0.03 和 0.03，在卷积核数量变多、尺寸变大的情况下，预测结果会明显下降，可能原因是数据噪声比较大的情况下，参数量增多，导致模型出现了过拟合现象. 对超参数 β 的选择进行了初步实验，结果显示在取 0.5 时效果最好，在取 0 时，即只用局部结构信息时效果较差.

4.3 实验结果与分析

4.3.1 OGBG-MolHIV 数据集

为了自适应卷积模块的实验分析能够更加明确，不受其他因素干扰，简单地采用 Transformer 编码器网络作为本文的主干模型，进行消融实验以验证各模块的有效性，即与 4 种方法进行比较：

(1) 本文方法的主干模型，Transformer 编码器网络.

(2) 本文方法去掉自适应选择卷积核尺寸部分，尺寸统一设为 3，即局部编码（Local encoding）.

(3) 本文方法在局部编码时使用多卷积核计算，将计算结果取平均，即多尺度局部编码（Multi-scale local encoding）.

(4) 本文方法在局部编码时使用多尺度卷积核计算，使用自注意力机制，直接使用 $\boldsymbol{\pi}_j$ 对卷积核加权，即软分配局部编码（Soft assignment local encoding）.

同时，也与目前最先进方法进行比较，由于本文使用的 Transformer 模型没有考虑节点之间边的

信息, 为公平起见, 此处也只与未考虑边信息的方法[36]进行比较, 结果如表 1 所示. 从表中结果可以看出:

表 1　不同方法在 OGBG-MolHIV 的 ROC-AUC 值比较结果

Table 1　Test results of ROC-AUC on OGBG-MolHIV

Method	ROC-AUC
EGC-M	0.7818
Transformer	0.7058
Local encoding	0.7249
Multiscale local encoding	0.7535
Soft-assignment local encoding	0.7519
Ours	**0.7839**

(1)使用局部编码的 4 种方法与主干模型相比较预测结果普遍升高, 可以证明本文方法中使用卷积提取局部邻域结构的有效性.

(2)使用局部编码方法比使用单一尺度卷积核的编码方法比较结果有所提升, 说明使用多卷积核尺寸是必要的.

(3)软分配多尺度卷积核的方法与平均分配多尺度卷积核的方法结果相比, 结果下降, 可能的原因是软分配局部编码方法参数较多, 导致对噪声的过学习, 因此证明变分推断的必要性.

(4)本文方法与其他方法相比效果提升较为明显, 表明自适应地挖掘局部拓扑信息可以有效提高图网络预测的性能, 而且变分推断可以有效防止过拟合.

另外, 为了与其他最先进的方法比较, 将本文方法的 Transformer 模型替换为 Graphormer[10], 结果如表 2 所示, 证明了本方法性能优于最先进方法. 值得注意的是, Graphormer 也使用了局部结构信息, 但是本方法仍然在结果上有所提升, 说明本文提出的自适应卷积模块对局部拓扑信息的表达更加有效.

表 2　本文与最先进方法在 OGBG-MolHIV 的比较结果

Table 2　Comparison of our method and the state-of-the-art on OGBG-MolHIV

Ref.	ROC-AUC
Zhang, et al.[37]	0.7799
Bouritsas, et al.[14]	0.8039
Wijesinghe, et al.[8]	0.7972
Ying, et al.[10]	0.8051
Ours (Graphormer)	**0.8189**

4.3.2　USPTO 数据集

分别在 USPTO 数据集上的克级和亚克级数据上进行了随机划分(Random)、时间划分(Time)、随机划分(数据平滑)(Smoothed), 进行消融实验以验证各模块的有效性. 实验结果如表 3 所示. 与 OGBG-MolHIV 数据集上的实验结果类似, 本文方法在两个级别的多种数据划分情况下均取得了最好的实验结果. 消融实验中, 本文方法优于局部编码、多尺度局部编码、软分配局部编码方法的结果表明, 所提出的自适应卷积模块中学习卷积核以及变分优化在挖掘图局部拓扑结构方面的重要作用. 与 Schwaller 等[32]的实验结果比较显示, 本文方法能够取得优于最先进方法的性能.

4.3.3　Buchwald-Hartwig 数据集

USPTO 数据集提出时间较短、任务较难, 因此该数据集上的对比方法较少. 为了方便与已有方法进行较为广泛的比较, 同时验证本方法在面向一类特定反应(胺化反应)的收率预测任务上的有效性, 在 Buchwald-Hartwig 数据集上进行了实验. 表 4 展示了本文方法在该数据集上与已有方法的对比实验结果, 本方法在特定反应收率预测的实验条件下同样取得了最好的结果, 且显著优于对比方法, 展示了本方法的优越性能.

5　结论

本文提出了一种利用变分卷积推断局部拓扑

表 3　不同方法在 USPTO 数据集的 R^2 值比较结果

Table 3　Comparison of R^2 scores on the USPTO dataset

Dataset	Data split	Schwaller, et al.[32]	Local encoding	Multiscale local encoding	Soft-assignment local encoding	Ours
Subgram	Random	0.195	0.198	0.196	0.195	**0.199**
	Time	0.142	0.146	0.147	0.145	**0.150**
	Smoothed	0.388	0.390	0.396	0.397	**0.435**
Gram	Random	0.117	0.118	0.119	0.118	**0.121**
	Time	0.095	0.096	0.096	0.095	**0.098**
	Smoothed	0.277	0.279	0.285	0.284	**0.311**

表 4 Buchwald-Hartwig 数据集上与已有方法的平均 R^2 值比较结果

Table 4 Comparison of average R^2 scores on the Buchwald-Hartwig dataset

Methods	Average R^2
Ahneman, et al.[18]	0.69
Chuang, et al.[38]	0.59
Granda, et al.[39]	0.60
Schwaller, et al.[32]	0.73
Ours	**0.76**

结构的图表示方法. 该方法在自注意力图网络的基础上加入自适应局部结构表示模块，不但可以全局地学习图节点之间的关系推理与信息传递，还能够自适应地编码图节点局部拓扑结构信息，通过充分挖掘多种结构信息来优化图网络的表达能力. 在三个公共数据集上的实验结果均表明了该方法能够有效地挖掘图数据节点间的局部结构信息并进行表示，从而提高自注意力图网络的预测性能.

虽然本方法能够自动学习选择合适的卷积模板，以自适应地学习不同尺寸和结构的局部子结构信息表示. 但是根据初步实验可看出，供选择的卷积模板过多时仍然会出现较多冗余，影响优化性能. 因此，未来的工作是设计一种生成式模型，通过学习自动张成局部子结构的表示模块，使得模型更加普适于更复杂的应用场景.

参 考 文 献

[1] Ying R, He R N, Chen K F, et al. Graph convolutional neural networks for web-scale recommender systems // *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London, 2018: 974

[2] Dai H J, Li C T, Coley C W, et al. Retrosynthesis prediction with conditional graph logic network // *Advances in Neural Information Processing Systems*. Vancouver, 2019: 8870

[3] Han K, Wang Y, Guo J, et al. Vision GNN: An image is worth graph of nodes // *Advances in Neural Information Processing Systems*. New Orleans, 2022

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // *Advances in Neural Information Processing Systems*. Long Beach, 2017: 5998

[5] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder –decoder for statistical machine translation // *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*). Doha, 2014: 1724

[6] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs // *Advances in Neural Information Processing Systems*. Long Beach, 2017: 1024

[7] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks // *International Conference on Learning Representations*. Toulon, 2018

[8] Wijesinghe A, Wang Q. A New Perspective on "How graph neural networks go beyond Weisfeiler-Lehman?" // *International Conference on Learning Representations*. Online, 2022

[9] Liu J W, Liu J W, Luo X L. Research progress in attention mechanism in deep learning. *Chin J Eng*, 2021, 43(11): 1499
(刘建伟, 刘俊文, 罗雄麟. 深度学习中注意力机制研究进展. 工程科学学报, 2021, 43(11): 1499)

[10] Ying C X, Cai T L, Luo S J, et al. Do transformers really perform badly for graph representation? // *Advances in Neural Information Processing Systems*. Online, 2021: 28877

[11] Alon U, Yahav E. On the bottleneck of graph neural networks and its practical implications[J/OL]. *arXiv preprint* (2020-6-9) [2022-7-24].https://arxiv.org/abs/2006.05205

[12] Jin W G, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation // *International Conference on Machine Learning*. Stockholm, 2018: 2323

[13] Chen Z D, Chen L, Villar S, et al. Can graph neural networks count substructures? // *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*. Vancouver, 2020: 10383

[14] Bouritsas G, Frasca F, Zafeiriou S, et al. Improving graph neural network expressivity *via* subgraph isomorphism counting. *IEEE Trans Pattern Anal Mach Intell*, 2023, 45(1): 657

[15] Yu H, Zhao S Y, Shi J Y. STNN-DDI: A substructure-aware tensor neural network to predict drug-drug interactions. *Brief Bioinform*, 2022, 23(4): bbac209

[16] Hu W, Fey M, Zitnik M, et al. Open graph benchmark: Datasets for machine learning on graphs // *Advances in Neural Information Processing Systems*. Online, 2020

[17] Lowe D. Chemical reactions from US patents (1976-Sep2016) [J/OL]. *Figshare* (2017-6-14) [2022-7-24]. https://doi.org/10.6084/m9.figshare.5104873.v1

[18] Ahneman D T, Estrada J G, Lin S, et al. Predicting reaction performance in C-N cross-coupling using machine learning. *Science*, 2018, 360(6385): 186

[19] Wu F, Fan A, Baevski A, et al. Pay less attention with lightweight and dynamic convolutions // *International Conference on Learning Representations*. New Orleans, 2019

[20] Wu Z H, Liu Z J, Lin J, et al. Lite transformer with long-short range attention[J/OL]. *arXiv preprint* (2020-4-24) [2022-7-24].https://arxiv.org/abs/2004.11886

[21] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition // *Interspeech Conference*. Shanghai, 2020: 5036

[22] Wang Y Q, Xu Z L, Wang X L, et al. End-to-end video instance segmentation with transformers // *IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (*CVPR*). Nashville, 2021: 8737

[23] Wu H P, Xiao B, Codella N, et al. CvT: introducing convolutions to vision transformers // *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Montreal, 2022: 22

[24] Si C Y, Yu W H, Zhou P, et al. Inception transformer[J/OL]. *arXiv preprint* (2022-5-25) [2022-7-24].https://arxiv.org/abs/2205.12956

[25] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision // *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle, 2016: 2818

[26] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning // *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco, 2017: 4278

[27] Zhou B L, Andonian A, Oliva A, et al. Temporal relational reasoning in videos // *Proceedings of the European Conference on Computer Vision*. Munich, 2018: 803

[28] Kim Y. Convolutional neural networks for sentence classification[J/OL]. *arXiv preprint* (2014-8-25) [2022-7-24]. https://arxiv.org/abs/1408.5882

[29] Kingma D P, Welling M. Auto-encoding variational bayes // *International Conference on Learning Representations*. Banff, 2014: 1

[30] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models // *International Conference on Machine Learning*. Beijing, 2014: 1278

[31] Maddison C J, Mnih A, Teh Y W. The concrete distribution: A continuous relaxation of discrete random variables[J/OL]. *arXiv preprint* (2016-11-2) [2022-7-24].https://arxiv.org/abs/1611.00712

[32] Schwaller P, Vaucher A C, Laino T, et al. Prediction of chemical reaction yields using deep learning. *Mach Learn*:*Sci Technol*, 2021, 2(1): 015016

[33] Landrum G. Rdkit documentation[J/OL]. *Rdkit* (2012-12-1) [2022-7-24]. http://www.rdkit.org/RDKit_Docs.2012_12_1.pdf

[34] Schwaller P, Laino T, Gaudin T, et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent Sci*, 2019, 5(9): 1572

[35] Schwaller P, Probst D, Vaucher A C, et al. Mapping the space of chemical reactions using attention-based neural networks. *Nat Mach Intell*, 2021, 3(2): 144

[36] Tailor S A, Opolka F L, Liò P, et al. Do we need anisotropic graph neural networks? [J/OL]. *arXiv preprint* (2021-4-3) [2022-7-24].https://arxiv.org/abs/2104.01481

[37] Zhang M, Li P. Nested graph neural networks // *Advances in Neural Information Processing Systems*. Online, 2021: 15734

[38] Chuang K V, Keiser M J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science*, 2018, 362(6416): 186

[39] Sandfort F, Strieth-Kalthoff F, Kühnemund M, et al. A structure-based platform for predicting chemical reactivity. *Chem*, 2020, 6(6): 1379