计算机科学
COMPUTER SCIENCE

融合对比学习的掩码图自编码器

王新喻, 宋小民, 郑慧明, 彭德中, 陈杰

引用本文

王新喻, 宋小民, 郑慧明, 彭德中, 陈杰. 融合对比学习的掩码图自编码器[J]. 计算机科学, 2026, 53(2): 145-151.

WANG Xinyu, SONG Xiaomin, ZHENG Huiming, PENG Dezhong, CHEN Jie. Contrastive Learning-based Masked Graph Autoencoder [J]. Computer Science, 2026, 53(2): 145-151.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100155

融合对比学习的掩码图自编码器

王新喻　宋小民　郑慧明　彭德中　陈　杰
四川大学计算机学院　成都 610065
(wangxinyu11@stu.scu.edu.cn)

摘　要　掩码图自编码器(Masked Graph Autoencoders,MGAEs)因能够有效处理图结构数据的节点分类任务而受到广泛关注。现有的掩码图自编码器模型在预训练编码器的过程中,存在语义信息损失和掩码节点嵌入相似两方面的不足。针对上述问题,提出一种融合对比学习的掩码图自编码器模型(CMGAE)。首先,将掩码图和原图分别输入在线编码器和目标编码器,生成在线嵌入和目标嵌入。然后,通过信息补充模块将在线嵌入和目标嵌入进行相似度对比,补充损失的语义信息。同时,将在线嵌入输入辨别函数和解码器,前者适当扩大掩码节点嵌入之间的方差,缓解掩码节点嵌入相似的问题,后者得到重构节点特征,用于训练在线编码器。最后,将预训练结束的在线编码器用于节点分类任务。在5个转导公共数据集和1个归纳数据集上进行节点分类实验,CMGAE的转导数据集准确率分别达到85.0%,73.6%,60.0%,50.5%,71.8%,归纳数据集的Micro-F1分数达到74.8%,相较于现有模型有着更好的性能。

关键词:图神经网络;节点分类;掩码图自编码器;图自监督学习;图对比学习
中图分类号　TP391

Contrastive Learning-based Masked Graph Autoencoder

WANG Xinyu,SONG Xiaomin,ZHENG Huiming,PENG Dezhong and CHEN Jie
College of Computer Science,Sichuan University,Chengdu 610065,China

Abstract　MGAEs have gained significant attention due to their effectiveness in handling node classification tasks on graph-structured data. However,existing MGAE models face two main limitations during the pretraining of the encoder:semantic information loss,and similarity of embeddings for masked nodes. To mitigate these issues,this paper proposes a Contrastive Masked Graph Autoencoder model(CMGAE). Firstly,the masked graph and the original graph are separately fed into the online encoder and the target encoder to generate online embeddings and target embeddings,respectively. Then,an information supplementation module is employed to compare the similarity between the online embeddings and target embeddings,thereby recovering the lost semantic information. Simultaneously,the online embeddings are passed through a discriminator function and decoder. The discriminator function helps increase the variance of the embeddings for masked nodes,mitigating the issue of similar embeddings for masked nodes. The decoder reconstructs node features that are used to train the online encoder. Finally,the pretrained online encoder is utilized for node classification tasks. Node classification experiments are conducted on five transductive benchmark datasets and one inductive dataset. The results show that CMGAE achieves a transductive accuracy of 85.0%,73.6%,60.0%,50.5%,and 71.8% on the respective datasets,while the Micro-F1 score on the inductive dataset reaches 74.8%. These results demonstrate that CMGAE outperforms existing models.

Keywords　Graph neural network,Node classification,Masked graph autoencoder,Graph self-supervised learning,Graph contrastive learning

1　引言

图是一种常见的数据结构,由节点、节点属性特征以及连接节点的边组成。图结构可以描述现实世界中的复杂系统,如引文网络[1-2]、蛋白质-蛋白质相互作用网络[3]、社交网络等。节点分类[4-7]是图数据分析中的重要任务,有助于揭示复杂网络的功能。例如,引文网络中的节点分类可以帮助研究人员快速找到相关领域的高质量文献;蛋白质-蛋白质相互作用网络中的节点分类有助于预测蛋白质功能,揭示生物系统的规律性等。

到稿日期:2025-01-24　返修日期:2025-04-27
基金项目:国家自然科学基金(62372315);四川省科技计划(2024NSFTD0049,2024ZDZX0004)
This work was supported by the National Natural Science Foundation of China(62372315) and Sichuan Provincial Science and Technology Program(2024NSFTD0049,2024ZDZX0004).
通信作者:陈杰(chenjie2010@scu.edu.cn)

对节点分类任务来说，如何使用现有方法有效地利用节点的属性特征与图的结构信息对节点进行正确的分类是一个关键的问题。基于传统机器学习的方法依赖人工设计的图特征，难以捕捉复杂拓扑关系，而基于大语言模型的新兴方法虽能利用预训练语义知识，却受限于大规模文本语料和高昂计算成本。相较之下，图自监督学习通过自监督训练挖掘图结构数据内在规律，逐渐成为主流范式，其有两种主流方法：掩码图自编码器方法和对比学习方法。掩码图自编码器通常带有编码器和解码器，两者都由图神经网络模型构成，例如GCN[8]，GAT[9]，GIN[10]等。编码器从图的掩码视图计算节点嵌入，解码器使用节点嵌入以重构掩码的部分作为训练过程。对比学习方法通常通过数据增强生成图的多个增强视图，并训练模型以最大化视图之间的互信息。具体来说，模型最大化正对之间的相似性，同时最小化负对之间的相似性。

掩码图自编码器方法避免了对比学习方法中所需的复杂训练策略，例如对比学习方法需要设计复杂的借口任务或需要依赖大量负对进行训练，而掩码图自编码器方法通过简单的策略即可达到较好的性能。尽管掩码图自编码器方法已在节点分类任务中取得了一定的成果，其仍存在两个问题亟须解决。1)过大的掩码比率可能导致丢失有效的语义信息。掩码图自编码器是视觉领域的掩码自编码器[11](Masked Autoencoder，MAE)在图领域的应用。对MAE来说，由于图像(或文本)属于欧几里德数据，其领域信息强相关，即基于周围非掩码图像块，掩码图像块的实际内容易于获取，对此类数据提供自监督指导的质量很高。而图结构数据属于非欧几里德数据，相较于欧几里德数据，其领域信息弱相关。具体来说，对猫的图像进行掩码时，人们可以很容易基于不完全的猫图像想象掩码图像块中的内容，而对社交网络这类图结构数据来说，对某个人进行掩码后，通过领域信息推断此人的信息相对更为困难。因此，尽管有重构策略帮助学习原图的语义信息，但对原图的过度掩码仍会使模型丢失大量可能有效的语义信息。2)掩码图自编码器在对节点特征掩码时选择使用掩码标记替换原特征，所有掩码节点被赋予相同的值，这可能导致编码器无法很好地区分这些节点并使掩码节点间的在线嵌入相似。

针对上述问题，本文提出一种融合对比学习的掩码图自编码器模型(Contrastive Learning-based Masked Graph Autoencoder，CMGAE)。首先，将掩码图和原图分别输入在线编码器和目标编码器生成在线嵌入和目标嵌入。然后，通过信息补充模块将在线嵌入和目标嵌入进行相似度对比，此模块通过两个嵌入的对比，将原图中的语义信息补入掩码图中，缓解了掩码导致的语义信息损失。同时，将在线嵌入输入辨别函数和解码器，辨别函数适当扩大掩码节点嵌入之间的方差，从而降低掩码节点在线嵌入之间的相似性，解码器输出重构节点特征，用于训练在线编码器。通过信息补充模块和新颖的辨别函数，缓解了掩码图自编码器存在的两个问题，进而提升了节点分类的性能。

本文的主要贡献如下：

1)提出了一个CMGAE模型，该模型补充了掩码图自编码器训练过程中损失的语义信息，并缓解了掩码节点嵌入相似的问题，可用于节点分类任务；

2)设计了一种语义信息补充模块，通过最大化两个视图之间的嵌入相似度来训练CMGAE模型；

3)在6个基准数据集上进行了节点分类实验，验证了该算法的有效性，并进行了CMGAE各个模块的消融实验。

2 相关工作

随着图神经网络在节点分类任务中的广泛应用，如何通过图自监督学习提升节点分类任务的性能成为研究热点。现有的图自监督学习根据模型结构和目标设计可以自然地分为生成式方法和对比学习方法。其中生成式方法中的主流方法为掩码图自编码器方法。

2.1 掩码图自编码器方法

掩码图自编码器方法一般专注于单独对特征或图结构进行重建，其会将图掩码后作为模型输入，并根据重建的对象分为两个子类别，即重建图特征信息和重建图结构信息。特征重建方法基于掩码特征回归，例如GraphMAE[12]掩码输入图中特定节点的特征并最后重建这部分特征，其虽能捕捉局部特征模式，但因忽略图结构语义关联，可能导致全局信息损失。GraphMAE2[13]设计了多视图随机重新掩码解码和潜在表示预测的策略，以规范特征重建，但多次掩码可能引入噪声干扰，降低潜在空间的一致性。RARE[14]通过进一步掩码和重建高阶潜在特征空间中的节点样本来提高推断掩码数据的确定性和自我监督机制的可靠性，然而高阶特征抽象易模糊细粒度语义。AUG-MAE[15]提出了一种对抗性掩码策略来提供难以对齐的样本，并引入了一个显式均匀性正则化器来确保学习表示的一致性，但其对抗训练可能加剧掩码节点嵌入的相似性。结构生成方法通常对图结构进行增强，例如MaskGAE[16]使用随机游走[17]对边进行掩码并重构，但存在对节点特征语义利用不足的缺点。S2GAE[18]采用方向感知的边掩码策略，但其离散掩码操作难以捕捉边权重的连续语义。Bandana[19]探索了非离散边缘掩码，使用带宽掩码和分层带宽预测目标。SeeGera[20]引入动态图结构掩码机制，通过自适应学习边重要性权重生成增强拓扑，但其掩码策略对节点特征的动态变化响应不足，可能在高阶语义推理任务中产生时序漂移误差。

尽管特征重建方法取得了进展，但这些研究仍同时存在两个缺陷，即语义信息损失和掩码节点嵌入相似，而本文模型能在一定程度上缓解这两个问题。

2.2 对比学习方法

对比学习方法通过最大化不同增广视图之间的互信息进行学习。经典工作如DGI[21]提出通过全局-局部互信息最大化学习图表示，其通过对比节点表示与全图池化向量的相似性捕捉图级语义，但受限于单视图对比策略，难以建模细粒度节点关系，且对图结构扰动的敏感性较高。MVGRL[22]进一步探索多视图图对比学习，利用个性化PageRank、热扩散等机制生成结构增强视图，并通过跨视图对比强化拓扑语义一致性。该方法虽通过谱域增广准则优化了视图生成过程，但多视图编码的计算复杂度限制了其在大规模图上的应用。CCA-SSG[23]提出谱图对比框架，利用典型相关分析约束增强

视图间的语义对齐，减少了冗余信息干扰，然而其线性投影设计可能削弱非线性语义的捕捉能力。GRACE[24]通过损坏生成两个视图，并通过最大化这两个视图中节点表示的一致性来学习节点表示。GCA[25]设计了基于节点中心性度量的增强方案来突出重要的连接结构。此类模型虽有效，但依赖于大量负对进行训练，带来了高昂的成本。BGRL[26]改进了原有的策略而不需要对比负样本，使得计算更有效，却也因对称性损失限制了模型的表达能力。近期研究进一步探索了消息传递机制的优化与生成式增强策略。例如，ReGCL[27]重新审视图对比学习中的消息传递过程，提出通过随机删除边生成不同的消息传播图，以增加输入数据的多样性并延缓过平滑现象。此方法虽缓解了传统图神经网络的过平滑问题，但未能有效增加同类节点间连边的数量，导致图拓扑结构优化能力有限。GACN[28]尝试结合生成对抗网络与对比学习，通过对抗训练生成具有判别性的增强视图，利用生成器构造高质量负样本，同时以判别器强化对比目标。尽管该方法在复杂图数据上展现了更强的表征鲁棒性，但其训练过程的不稳定性及生成样本的潜在偏差仍需进一步解决。

3 CMGAE 模型

3.1 问题描述

令 $G=(V,\boldsymbol{A},\boldsymbol{X})$ 表示给定的图，其中：$V=\{v_i\}_{i=1}^{N}$ 表示节点集，N 是节点总数；$\boldsymbol{A}\in\{0,1\}^{N\times N}$ 代表邻接矩阵，当 v_i 和 v_j 之间存在边时 $A_{ij}=1$；$\boldsymbol{X}\in\mathbb{R}^{N\times d}$ 是节点特征矩阵，其中 d 代表特征维度。大多数掩码图自编码器的目标是学习一个编码器 $f_\theta(\cdot)$，用于生成低维节点嵌入，令 $\boldsymbol{H}=f_\theta(\boldsymbol{X},\boldsymbol{A})\in\mathbb{R}^{N\times d'}$ 表示学习到的节点嵌入，其中 $\boldsymbol{h}_i\in\mathbb{R}^{d'}$ 代表节点 v_i 学习到的嵌入，d' 表示节点嵌入的维度。

3.2 CMGAE 具体结构

本节介绍 CMGAE 模型，其由 3 部分组成，分别是掩码图自编码器模块、信息补充模块和辨别函数。模型的框架图如图 1 所示。

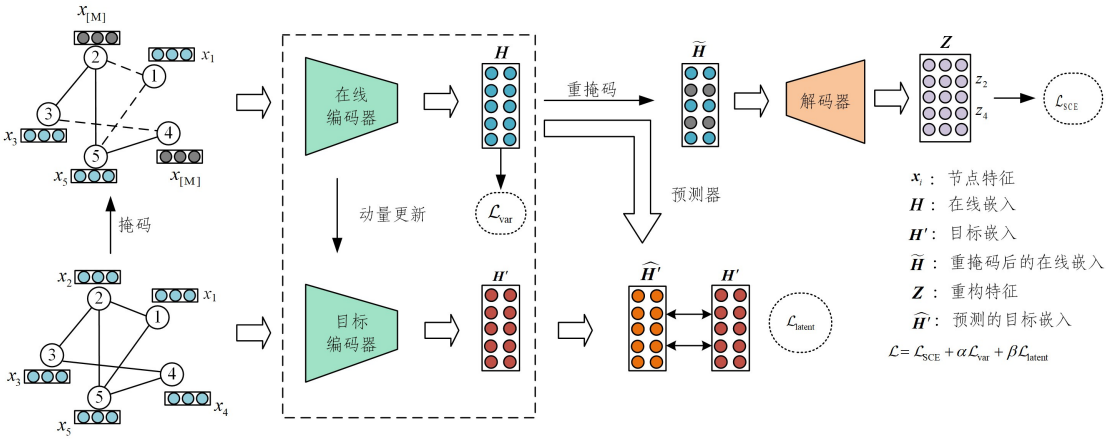

图 1 CMGAE 模型结构

Fig. 1 Structure of CMGAE

3.2.1 掩码图自编码器模块

此模块基于掩码重构策略，并以 GraphMAE 作为主干，为让模型学习到图的完整信息，选择同时对节点特征和图结构进行掩码。对于节点特征，首先随机采样部分节点 $\widetilde{V}\subset V$，并使用掩码标记对节点特征掩码，实践中掩码标记选择一个可学习的向量 $\boldsymbol{x}_{[\mathrm{M}]}$。掩码后的节点特征矩阵 $\widetilde{\boldsymbol{X}}$ 被定义为：

$$\widetilde{\boldsymbol{x}}_i=\begin{cases}\boldsymbol{x}_{[\mathrm{M}]}, & v_i\in\widetilde{V}\\ \boldsymbol{x}_i, & v_i\notin\widetilde{V}\end{cases}$$

其中，$\boldsymbol{x}_i$ 表示节点 v_i 原本的节点特征。与 GraphMAE 不同，对于图结构，CMGAE 令掩码边满足伯努利分布，具体如式(1)所示：

$$\varepsilon_{\text{mask}}\sim\text{Bernoulli}(q),0<q<1 \tag{1}$$

其中：$\varepsilon_{\text{mask}}$ 表示掩码边集，Bernoulli(·)表示伯努利分布，q 表示采样比率。然后将掩码图 $\widetilde{G}=(\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{A}})$ 输入在线编码器 $f_\theta(\cdot)$ 得到在线嵌入 $\boldsymbol{H}$，具体如式(2)所示：

$$\boldsymbol{H}=f_\theta(\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{A}}) \tag{2}$$

其中，$\boldsymbol{H}\in\mathbb{R}^{|V|\times d}$，$|V|$ 表示节点数，d 表示在线嵌入的维度。

为了鼓励在线编码器学习更有表现力的嵌入，将在线嵌入 $\boldsymbol{H}$ 重掩码，重掩码后的嵌入 $\widetilde{\boldsymbol{H}}$ 被定义为：

$$\widetilde{\boldsymbol{h}}_i=\begin{cases}\boldsymbol{h}_{[\mathrm{M}]}, & v_i\in\widetilde{V}\\ \boldsymbol{h}_i, & v_i\notin\widetilde{V}\end{cases}$$

其中，$\boldsymbol{h}_{[\mathrm{M}]}$ 表示另一个可学习向量。然后将重掩码后的在线嵌入 $\widetilde{\boldsymbol{H}}$ 输入解码器 $g_\varphi(\cdot)$ 中，得到重构节点特征 $\boldsymbol{Z}$，如式(3)所示：

$$\boldsymbol{Z}=g_\varphi(\widetilde{\boldsymbol{H}},\boldsymbol{A}) \tag{3}$$

最后选择缩放余弦误差作为掩码图自编码器模块损失函数 $\mathcal{L}_{\text{SCE}}$，如式(4)所示：

$$\mathcal{L}_{\text{SCE}}=\frac{1}{|\widetilde{V}|}\sum_{v_i\in\widetilde{V}}\left(1-\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{z}_i}{\|\boldsymbol{x}_i\|\cdot\|\boldsymbol{z}_i\|}\right)^{\gamma},\gamma\geqslant1 \tag{4}$$

其中，γ 表示缩放因子，用于控制简单样本对训练的影响程度。

3.2.2 信息补充模块

掩码图自编码器的训练基于对原图大幅度的掩码，这会丢失大量语义信息；同时，对比学习方法中两个视图的对比自然地补充了视图之间的语义信息。因此，CMGAE 通过将对比学习方法融入掩码图自编码器中，以创新性的方式补充后者损失的语义信息。

对比方法一般使用共享编码器生成两个视图的嵌入，然而这可能使得编码器学习到较为混乱的语义信息，导致生成低质量的嵌入。因此本节使用两个不同的编码器生成两个视图的嵌入。具体来说，首先重用掩码图自编码器的掩码图 $\widetilde{G}=(\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{A}})$，$\widetilde{G}$ 保留了较少语义信息，并选择保留丰富语义信息的原图 $G=(\boldsymbol{X},\boldsymbol{A})$ 作为对比视图。然后将掩码图 $\widetilde{G}$ 和原图 G 分别输入在线编码器 $f_\theta(\cdot)$ 和目标编码器 $f_\phi(\cdot)$，得到在线嵌入 $\boldsymbol{H}$ 和目标嵌入 $\boldsymbol{H}'$，具体如式(5)所示：

$$\begin{aligned}\boldsymbol{H}&=f_\theta(\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{A}})\\ \boldsymbol{H}'&=f_\phi(\boldsymbol{X},\boldsymbol{A})\end{aligned}\tag{5}$$

接着将在线嵌入 $\boldsymbol{H}$ 进一步输入预测器 $p_\theta(\cdot)$ 中，得到目标嵌入的预测值 $\hat{\boldsymbol{H}}'$，具体如式(6)所示：

$$\hat{\boldsymbol{H}}'=p_\theta(\boldsymbol{H})\tag{6}$$

其中，$p_\theta(\cdot)$ 由 3 层 MLP 组成。最后选择缩放余弦误差作为该模块的损失函数，具体如式(7)所示：

$$\mathcal{L}_{\text{Latent}}=\frac{1}{|V|}\sum_{v_i\in V}\left(1-\frac{\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{h}_i'}{\|\boldsymbol{h}_i\|\cdot\|\boldsymbol{h}_i'\|}\right)^{\gamma},\gamma\geqslant 1\tag{7}$$

使用此损失函数避免了对比学习方法中普遍存在的模型依赖于大量负对训练的问题，使此模块专注于补充两个视图之间的语义信息。

3.2.3 辨别函数

掩码图自编码器在对节点特征掩码时选择使用掩码标记替换原特征，所有掩码节点被赋予相同的值，这可能导致在线编码器无法区分这些节点并使掩码节点间的在线嵌入相似。为缓解这个问题，设计了一个新颖的辨别函数，作用于掩码节点的在线嵌入，具体如式(8)所示：

$$\mathcal{L}_{\text{var}}=-\sqrt{Var(\boldsymbol{H}_{\widetilde{V}})+\epsilon}\tag{8}$$

其中，ϵ 是一个小标量，防止数值为 0 时数据不稳定，实践中可以取 0.0001；$\boldsymbol{H}_{\widetilde{V}}\in\mathbb{R}^{|\widetilde{V}|\times d}$ 由掩码节点集 $\widetilde{V}$ 的在线嵌入组成，掩码节点的在线嵌入都经过正则化处理；$Var(\boldsymbol{H}_{\widetilde{V}})$ 表示在线嵌入的方差。

辨别函数强制让掩码节点在线嵌入在特定方差范围内映射到不同空间，这种方法自然地扩大了在线嵌入之间的方差，导致在线嵌入之间的相似度降低，提高了模型的分辨性能。

3.2.4 训练步骤

结合以上模块的优势，使 CMGAE 能缓解现存模型的局限性。模型总的损失函数如式(9)所示：

$$\mathcal{L}=\mathcal{L}_{\text{SCE}}+\alpha\mathcal{L}_{\text{var}}+\beta\mathcal{L}_{\text{Latent}}\tag{9}$$

其中，α 和 β 是非负超参数。

在线编码器的参数更新遵循梯度反向传播，具体如式(10)所示：

$$\theta\leftarrow optimize(\theta,\eta,\partial_\theta\mathcal{L})\tag{10}$$

其中，η 是学习率，最终更新仅从损失函数 $\mathcal{L}$ 相对于 θ 的梯度计算。

目标编码器参数更新与梯度反向传播分离，使用衰减率 τ 更新为在线参数 θ 的指数移动平均值，具体如式(11)所示：

$$\phi\leftarrow\tau\phi+(1-\tau)\theta\tag{11}$$

3.2.5 计算复杂度分析

对一个具有 N 个节点和 M 条边的图，类似 GAT 的简单编码器 f 所需的计算时间为 $O(N+M)$。CMGAE 对每个更新步骤执行 3 个编码器计算(目标和在线编码器两次，每次数据增强一次)，且由于 CMGAE 中的掩码图自编码器模块也由 GAT 构成，因此执行一个解码器计算，辨别函数的计算时间为 $O(N)$，信息补充模块中预测器的计算时间为 $O(N)$。假设反向传播的计算成本与前向传播类似，可以得出计算复杂度为 $7O(N+M)+3O(N)$。CMGAE 模型预训练算法如算法 1 所示。

算法 1 CMGAE 模型预训练算法

输入：特征矩阵 $\mathbf{X}\in\mathbb{R}^{N\times F}$，邻接矩阵 $\mathbf{A}\in\mathbb{R}^{N\times N}$，最大迭代次数 E，衰减率 τ，学习率 η，超参数 α 和 β，节点掩码比率 γ，边掩码比率 q，模型参数 θ 和 φ

输出：预训练结束的在线编码器参数 θ

1. 初始化模型参数 θ 和 φ
2. for e←1 to E do
3. 对特征矩阵 $\mathbf{X}$ 以及邻接矩阵 $\mathbf{A}$ 掩码，得到掩码后的特征矩阵 $\widetilde{\mathbf{X}}$ 和邻接矩阵 $\widetilde{\mathbf{A}}$
4. 通过式(5)得到在线嵌入 $\mathbf{H}$ 以及目标嵌入 $\mathbf{H}'$
5. 重掩码在线嵌入 $\mathbf{H}$，得到 $\widetilde{\mathbf{H}}$
6. 通过式(3)得到重构节点特征 $\mathbf{Z}$
7. 通过式(4)得到掩码图自编码器模块的损失函数 $\mathcal{L}_{\text{SCE}}$
8. 通过式(6)得到目标嵌入的预测值 $\mathbf{H}'$
9. 通过式(7)得到信息补充模块的损失函数 $\mathcal{L}_{\text{Latent}}$
10. 通过式(8)得到辨别损失函数 $\mathcal{L}_{\text{var}}$
11. 通过式(9)得到总的损失函数 $\mathcal{L}$
12. 通过式(10)和式(11)更新模型参数
13. End for
14. 返回在线编码器参数 θ

4 实验结果和分析

将 CMGAE 模型在多个数据集上进行节点分类实验，并与其他先进的模型进行性能比较。

4.1 数据集

在 6 个公开的数据集上进行节点分类任务，表 1 列出了数据集的基本信息。

表 1 数据集基本信息

Table 1 Basic information of datasets

数据集	节点数/个	边数/条	特征维度/个	类数/个
Cora	2708	5278	1433	7
CiteSeer	3327	4732	3703	6
PPI	56944	8187636	50	121
Corafull	19793	126842	8710	70
Flickr	89250	899756	500	7
Ogbn-arxiv	169343	1166243	128	40

Cora 数据集：属于引文网络数据集，其中节点表示文献，边表示文献之间的引用关系，按研究方向分为 7 类。

CiteSeer 数据集[29]：属于引文网络数据集，其中节点表示科学出版物，边表示科学出版物之间的引用关系，按出版物种类分为 6 类。

PPI 数据集：蛋白质相互作用数据集，其中节点表示蛋白质，边表示蛋白质之间的相互作用。此数据集包含 24 个图，

图中平均节点数为2372个，每个节点的特征维度为50，节点被分为121类。其中训练集包含20个图，验证集包含2个图，测试集包含2个图。

Corafull数据集[30]：扩展的Cora数据集，属于引文网络数据集，其中节点表示文献，边表示文献之间的引用关系，节点被分为70类。

Flickr数据集[31]：基于图像数据，其中节点表示图片，边表示图片之间的联系，按图片种类分为7类。

Ogbn-arxiv数据集[32]：属于引文网络数据集，其中节点表示arXiv平台的学术论文，边表示论文之间的引用关系，节点按学科领域划分为40类。

4.2 对比模型

选取以下8种用于节点分类的基线方法进行比较。

1）DGI：经典的图对比学习模型，通过对比学习最大化全局图表示和局部节点表示之间的互信息。

2）MVGRL：通过最大化图的不同结构视图之间的互信息学习节点和图的表示，采用多视图对比学习框架。

3）BGRL：通过两个不同的图编码器对一个图的两个增强视图编码来学习节点表示。

4）CCA-SSG：通过数据增强生成输入图的两个视图，旨在通过学习不变表示来丢弃增强变量信息。

5）SeeGera：生成图自监督学习模型，基于自监督变分图自动编码器家族，通过重构输入图数据来学习节点表示。

6）MaskGAE：生成式图自监督学习模型，采用掩码图建模作为借口任务，通过掩码一部分边并对其重构来学习节点表示。

7）GraphMAE：生成式图自监督学习模型，通过掩码一部分节点特征并重构来学习节点表示。

8）AUG-MAE：生成式图自监督学习模型，通过动态掩码部分节点特征和图结构以学习鲁棒表示。

4.3 评价指标

实验关注节点分类这个下游任务，对PPI数据集采用Micro-F1分数作为评价指标，其余数据集采用准确率作为评价指标。

4.4 实验设置

实验对上述数据集进行节点分类任务。节点分类任务是预测网络中的未知节点标签。其中PPI数据集用于归纳学习，并遵循GraphSage[33]中的归纳设置，其余数据集用于转导学习。

对于评估协议，首先，在无监督条件下，通过提出的CMGAE训练一个在线编码器。然后冻结在线编码器的参数并生成所有节点的嵌入。为了评估，进一步训练一个线性分类器，并通过20次随机初始化报告测试节点的平均准确度。所有数据集都采用标准数据拆分。

CMGAE模型是基于GraphMAE的改进模型。在Cora，CiteSeer，PPI数据集上，对于共有超参数，CMGAE在原有超参数基础上进行微调。在Corafull，Flickr数据集上，超参数经过网格化搜索进行调优。最终学习率设置为0.001，0.0001，优化器选择Adam，优化器中的衰减率设置为0，0.00002，0.0001，0.0002，正则化方法选择L2正则化。CMGAE独有的参数α和β设置均为0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9，1.0，边掩码比率q设置为0.5，0.6，0.7，动量更新中的超参数衰减率设置为0.996。对于基线，如果数据集可用，则参考原文献中的最佳设置；否则，根据官方代码实现并进行超参数搜索。

实验均在配置NVIDIA 4090 GPU的Linux服务器上运行，PyTorch版本为2.1.1，DGL版本为2.1.0，CUDA版本为12.2，Python版本为3.9。

4.5 结果分析

4.5.1 节点分类结果分析

表2列出了实验结果，其中加粗字体表示最优指标值，下划线字体表示次优指标值。结果显示，CMGAE在大部分数据集上的性能都优于其余图自监督基线。其中GraphMAE是最近提出的生成式方法，且CMGAE模型是GraphMAE的改进模型，因此重点关注两者的性能差距。结果证明，CMGAE在所有数据集上都优于GraphMAE。由此可得，利用信息补充模块补充原图语义信息，并通过辨别函数扩大掩码节点嵌入之间的方差，可以有效提升模型性能，验证了CMGAE策略的有效性。

表2 节点分类结果

Table 2 Results of the node classification

Method	Cora	CiteSeer	PPI	Corafull	Flickr	Ogbn-arxiv
DGI	82.3±0.6	71.8±0.7	63.8±0.2	48.2±0.5	45.0±0.2	70.3±0.2
MVGRL	83.5±0.4	73.3±0.5	—	52.6±0.5	—	—
BGRL	82.7±0.6	71.1±0.8	68.8±0.2	47.4±0.5	39.4±0.1	71.6±0.1
CCA-SSG	83.4±0.4	<u>73.3±0.3</u>	73.3±0.2	53.5±0.4	49.1±0.1	71.2±0.2
SeeGera	83.0±0.5	71.4±0.9	73.4±0.3	52.0±0.4	49.4±0.5	71.2±0.2
MaskGAE	82.9±0.3	72.4±0.6	73.9±0.3	52.7±0.1	50.1±0.2	71.2±0.3
GraphMAE	84.2±0.4	73.0±0.4	74.2±0.3	55.5±0.1	50.2±0.2	71.3±0.6
AUG-MAE	<u>84.3±0.4</u>	73.2±0.4	<u>74.3±0.1</u>	<u>57.6±0.3</u>	<u>50.3±0.2</u>	**71.9±0.2**
CMGAE	**85.0±0.6**	**73.6±0.4**	**74.8±0.3**	**60.0±0.2**	**50.5±0.5**	<u>71.8±0.3</u>

4.5.2 消融实验

为验证CMGAE中各个模块（主要分为两个模块，即信息补充模块和辨别函数模块）的有效性，进行了对应的消融实验。选用了4个数据集进行实验，表3列出了对应的结果，其中，CMGAE表示原始CMGAE；Variant 1表示移除辨别函数模块，保留信息补充模块；Variant 2表示保留辨别函数模块，移除信息补充模块；GraphMAE表示原始GraphMAE，即移除辨别函数模块，移除信息补充模块。

表 3 消融实验

Table 3 Ablation study

消融实验	Cora	PPI	Corafull	CiteSeer
CMGAE	85.0	74.8	60.2	73.6
Variant 1	84.9	74.7	58.6	73.1
Variant 2	84.7	74.7	59.1	73.2
GraphMAE	84.2	74.2	55.5	73.0

结果表明，引入信息补充模块，能帮助模型获取更多有效语义信息。对于较小图，信息补充模块弥补了大规模掩码导致的语义信息缺失；对于较大图，其稳定了掩码图自编码器模块。辨别函数通过扩大嵌入方差，缓解了嵌入相似的问题。两者的结合，使模型达到了更好的性能，验证了两个模块的有效性。

4.5.3 参数分析

为了验证信息补充模块与辨别函数占比的影响，在Cora，Corafull，PPI 数据集上进行分析。

1)辨别函数。仅固定 α 值，并对模型进行调参，表 4 列出了对应的结果。

表 4 不同 α 设置的实验结果

Table 4 Experiment results of different settings of α

α 占比	准确率%		Micro-F1 分数
	Cora	Corafull	PPI
0.1	84.6	60.0	74.6
0.3	84.6	60.1	74.5
0.5	84.6	60.1	74.7
0.7	84.6	60.1	74.7
0.9	84.6	60.1	74.7

由表 4 可知，在 0.1 到 0.9 的占比中，3 个数据集的实验结果均较为稳定，并没有随着 α 占比提高而导致模型效果变差，这是因为辨别函数首先对嵌入正则化，再扩大方差。这一定程度上能帮助辨别函数模块稳定地提升模型的性能，而不会使 α 过高导致嵌入映射到不合理的范围内，说明辨别函数模块具有较强的鲁棒性。

2)信息补充模块。仅固定 β 值，并对模型进行调参，表 5 列出了对应的结果。

表 5 不同 β 设置的实验结果

Table 5 Experiment results of different settings of β

β 占比	准确率%		Micro-F1 分数
	Cora	Corafull	PPI
0.1	84.6	60.2	74.6
0.3	84.7	60.2	74.7
0.5	84.4	60.1	74.7
0.7	84.0	60.0	74.7
0.9	84.0	59.8	74.7

由表 5 可知，当 β 占比提高时，Cora 数据集的实验效果逐渐下降，这是因为 Cora 数据集较小，引入过量原图信息可能会导致模型过分关注图中低价值语义信息，无法学习到有用的嵌入。而 Corafull 数据集虽也随 β 占比提高而效果下降，但由于 Corafull 数据集较大，模型不易对过量语义信息过拟合，因此还能保持较稳定的性能。当 β 占比提高时，PPI 数据集的性能保持稳定，这是因为 PPI 数据集属于归纳设置，各个图之间的语义信息影响较小，能保持稳定的性能。

4.5.4 计算复杂度实验

为了研究 CMGAE 的计算需求，进行对应的计算复杂度实验，主要与 GraphMAE 进行对比，表 6 列出了对应的结果。

表 6 计算复杂度实验

Table 6 Computational complexity study

数据集	CMGAE 显存占用/MB	GraphMAE 显存占用/MB	CMGAE 时间	GraphMAE 时间
Cora	910	712	125 s	55 s
CiteSeer	1522	1030	35 s	15 s
PPI	2600	1868	77 min 45 s	67 min 5 s
Corafull	7502	6374	15 min	10 min 55 s
Flickr	4168	2904	11 min 10 s	7 min 35 s
Ogbn-arxiv	19654	12986	45 min	25 min

由表 6 可知，CMGAE 在各个数据集中的显存占用增长较明显，在不同数据集上的训练时间也有不同程度的增加，其以较长时间的训练，以及更多的参数量换取较高的性能。

结束语 为解决语义信息损失和掩码节点嵌入相似的问题，本文提出融合对比学习的掩码图自编码器模型(CMGAE)，通过引入信息补充模块以及新的辨别函数，缓解掩码图自编码器因过度掩码导致的有效语义信息损失，以及节点掩码策略导致的节点嵌入相似。实验结果表明，CMGAE 在节点分类任务中相较于其他模型性能更好。但本文的辨别函数只是间接缓解掩码策略导致的负面影响，因此未来考虑设计一种新的掩码策略来解决这类问题。

参 考 文 献

[1] WANG K, SHEN Z, HUANG C, et al. Microsoft academic graph: When experts are not enough[J]. Quantitative Science Studies, 2020, 1(1): 396-413.

[2] MERNYEI P, CANGEA C. Wiki-cs: A wikipedia-based benchmark for graph neural networks[J]. arXiv: 2007.02901, 2020.

[3] FOUT A, BYRD J, SHARIAT B, et al. Protein interface prediction using graph convolutional networks[C]// Advances in Neural Information Processing Systems. 2017.

[4] ZHANG L Y, SUN H H, SUN Y F, et al. Review of Node Classification Methods Based on Graph Convolutional Neural Networks [J]. Computer Science, 2024, 51(4): 95-105.

[5] LI X, LU W, MA Z Y, et al. A Node Classification Method Based on Graph Attention and Improved Transformer [J]. Chinese Journal of Electronics, 2024, 52(8): 2799-2810.

[6] LIU Y L, QIU R H, TANG Y R, et al. PUMA: Efficient Continual Graph Learning for Node Classification With Graph Condensation [J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(1): 449-461.

[7] DUAN L, CHEN X, LIU W, et al. Structural entropy based graph structure learning for node classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 8372-8379.

[8] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv: 1609.02907, 2016.

[9] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv: 1710.10903, 2017.

[10] XU K, HU W, LESKOVEC J, et al. How powerful are graph

neural networks? [J]. arXiv:1810.00826,2018.
[11] HE K,CHEN X,XIE S,et al. Masked autoencoders are scalable vision learners[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16000-16009.
[12] HOU Z,LIU X,CEN Y,et al. GraphMAE:Self-supervised masked graph autoencoders[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022:594-604.
[13] HOU Z, HE Y, CEN Y, et al. GraphMAE2: A decoding-enhanced masked self-supervised graph learner[C]// Proceedings of the ACM Web Conference 2023. 2023:737-746.
[14] TU W X, LIAO Q, ZHOU S H, et al. Rare: Robust masked graph autoencoder[J]. IEEE Transactions on Knowledge and Data Engineering,2024,36(10):5340-5353.
[15] WANG L,TAO X,LIU Q,et al. Rethinking Graph Masked Autoencoders through Alignment and Uniformity[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 15528-15536.
[16] LI J,WU R,SUN W,et al. What's Behind the Mask: Understanding Masked Graph Modeling for Graph Autoencoders [C]// Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023:1268-1279.
[17] PEROZZI B,AL-RFOU R,SKIENA S. Deepwalk:Online learning of social representations[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014:701-710.
[18] TAN Q,LIU N,HUANG X,et al. S2gae:Self-supervised graph autoencoders are generalizable learners with graph masking [C]// Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 2023:787-795.
[19] ZHAO Z,LI Y,ZOU Y,et al. Masked Graph Autoencoder with Non-discreteBandwidths[C]// Proceedings of the ACM on Web Conference. 2024. 2024:377-388.
[20] LI X, YE T, SHAN C, et al. Seegera: Self-supervised semi-implicit graph variational auto-encoders with masking[C]// Proceedings of the ACM Web Conference 2023. 2023:143-153.
[21] VELIČKOVIĆ P, FEDUS W, HAMILTON W L, et al. Deep graph infomax[J]. arXiv:1809.10341,2018.
[22] HASSANI K,KHASAHMADI A H. Contrastive multi-view representation learning on graphs[C]// International Conference on Machine Learning. PMLR,2020:4116-4126.
[23] ZHANG H,WU Q,YAN J,et al. From canonical correlation analysis to self-supervised graph neural networks[J]. Advances in Neural Information Processing Systems,2021,34:76-89.
[24] ZHU Y,XU Y,YU F,et al. Deep graph contrastive representation learning[J]. arXiv:2006.04131,2020.
[25] ZHU Y,XU Y,YU F,et al. Graph contrastive learning with adaptive augmentation[C]// Proceedings of the Web Conference 2021. 2021:2069-2080.
[26] THAKOOR S,TALLEC C,AZAR M G,et al. Large-scale representation learning on graphs via bootstrapping[J]. arXiv: 2102.06514,2021.
[27] JI C,HUANG Z,SUN Q,et al. ReGCL:rethinking message passing in graph contrastive learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024:8544-8552.
[28] WU C,WANG C,XU J,et al. Graph contrastive learning with generative adversarial network[C]// Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023:2721-2730.
[29] SEN P,NAMATA G,BILGIC M,et al. Collective classification in network data[J]. AI Magazine,2008,29(3):93-93.
[30] BOJCHEVSKI A,GÜNNEMANN S. Deep Gaussian embedding of graphs:Unsupervised inductive learning via ranking[J]. arXiv:1707.03815,2017.
[31] ZENG H,ZHOU H,SRIVASTAVA A,et al. Graphsaint:Graph sampling based inductive learning method[J]. arXiv: 1907.04931,2019.
[32] HU W H,FEY M,ZITNIK M,et al. Open Graph Benchmark: Datasets for Machine Learning on Graphs[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook,NY:Curran Associates Inc.,2020: 22118-22133.
[33] HAMILTON W,YING Z,LESKOVEC J. Inductive representation learning on large graphs[C]// Advances in Neural Information Processing Systems. 2017.

WANG Xinyu, born in 2001, postgraduate. His main research interest is graph neural network.

CHEN Jie, born in 1982, Ph.D, associate professor. His main research interests include machine learning, big data analysis and artificial intelligence.

(责任编辑:喻藜)