MGAE：用于图形自监督学习的掩蔽自动编码器

谭巧玉
德克萨斯 A&M 大学
qytan@tamu.edu

刘宁浩
佐治亚大学
ninghao.liu@uga.edu

小黄
香港理工大学
xiaohuang@comp.polyu.edu.hk

陈睿
三星美国研究中心
rui.chen1@samsung.com

崔秀贤
三星美国研究中心
soohyunc@gmail.com

夏虎
莱斯大学
xia.hu@rice.edu

抽象的

我们引入了一种新颖的掩蔽图自动编码器 (MGAE) 框架，用于对图结构数据进行有效学习。借鉴自监督学习的经验，我们随机掩蔽了大部分边，并尝试在训练期间重建这些缺失的边。MGAE 有两种核心设计。首先，我们发现掩蔽输入图结构的高比例，例如，70%，产生了一个非平凡且有意义的自监督任务，有益于下游应用。其次，我们使用图神经网络（GNN）作为编码器在部分掩蔽的图上执行消息传播。为了重建大量被掩蔽的边，提出了一种定制的互相关解码器。它可以多粒度地捕获锚边的头节点和尾节点之间的互相关。将这两种设计结合起来使 MGAE 能够高效、有效地进行训练。在多个开放数据集（Planetoid 和 OGB 基准）上进行的大量实验表明，MGAE 在链接预测和节点分类方面的表现通常优于最先进的无监督学习竞争对手。

1 简介

图结构数据在现实世界系统中无处不在 [1, 2]，例如社交网络、学术图谱和生物相互作用网络。鉴于标签通常不可用，无监督图表征学习在学术界和工业界都引起了相当大的关注。目标是学习节点表征以保留输入图结构 [3–6]。基于学习到的表征，我们可以执行各种无监督任务，例如链接预测和异常检测。此外，我们可以直接将现成的学习算法应用于学习到的表征来执行监督任务，例如节点分类。

为了以无监督的方式学习节点表示，有两条研究路线。首先，图自编码器，例如基于 GNN 编码器的自编码器，在许多图域 [7, 8] 中被证明在节点分类和链接预测任务上是有效的。它旨在重建原始网络结构（即观察到的边）以进行模型训练。人们一直致力于探索有效的编码器网络。例如，GAE [4] 采用经典的图卷积网络 (GCN) [9] 模型作为编码器，而 GraphSage [10] 为图引入了 GCN 的归纳变体

预印本。正在审核中。

编码。其次，图自监督学习（GSSL）专注于设计用于自监督训练的高级借口任务[11，12]。例如，DGI[13]和GIC[14]旨在通过最大化节点级表示和锚节点所在的图级表示之间的互信息[15]来训练GNN模型。GSSL方法可以学习用于图编码的稳健而强大的GNN模型[16]。

虽然边重建和边丢弃是图自编码和 GSSL 中的常用技术，但掩蔽自编码从未在图上进行过探索。其核心思想是删除一定比例的输入数据，并使用删除的内容来指导训练。掩蔽自编码已被证明在文本 [17] 和图像 [18] 建模中有效且高效。图自编码将完整图作为输入，目标是重建整个边。在 GSSL 中，图增强方法之一是删除一些边 [11]。实验表明，边丢弃在许多场景中可能效果不佳。其目标是学习鲁棒表示，而不是预测被删除的边。此外，在实践中，其有效掩蔽比值小于 30%。因此，我们请求：*如何为图设计适当的掩码自动编码？重建输入图并学习有效的节点表示真正需要多少百分比的边*？

在本文中，我们对这些悬而未决的问题给出了肯定的答案。我们提出了一个简单而有效的图自动编码器框架，称为掩蔽图自动编码器（MGAE），用于无监督图表示学习。MGAE 的目标是随机屏蔽输入图结构的大部分，然后恢复被屏蔽的边。与传统的图自动编码器不同，I）我们的 GNN 编码器仅对部分网络结构（没有掩蔽边）进行卷积，II）我们的解码器旨在捕获锚边的头节点和尾节点之间的互相关性，以有效地从它们的潜在表示中重建链接（见图 1）。在这种设计下，我们的 MGAE 模型可以实现双赢的局面，具有显著的高掩蔽率（例如，70%：它优化了模型性能，同时允许 GNN 编码器仅处理一小部分（30%原始图结构。我们总结了我们的主要贡献：

- 我们针对图结构数据引入了一种新颖的图自动编码器替代方案，称为掩蔽图自动编码器 (MGAE)。它受到自监督学习的启发，不仅稳健有效，而且非常适合链接预测和节点分类。
- 我们提出了一种定制的互相关解码器，以有效利用掩蔽图结构产生的噪声隐藏表示进行边缘重建。这种细致的设计使我们能够对边缘进行非常高的掩蔽率（例如 70%），这也提高了有效性和效率。
- 大量实验表明，在链接预测和节点分类任务方面，MGAE 在 Planetoid 和 OGB 基准测试中的表现更佳，有时甚至与最先进的竞争对手相当。

2 问题陈述

符号：我们使用粗体小写字母（例如，x）表示矢量，以及粗体大写字母（例如，十）表示矩阵。*五*⊟排十表示为十*五*.我们假设一个无向图*格*= (*五*、）和*n*节点数为，其中*五*和*埃*分别表示节点和边的集合。每个节点*五*∈*V*有一个*弗*-维属性向量十*五*∈R*弗*描述其属性。当节点属性不可用时，十*五*可以初始化为独热索引向量或可学习参数。为了研究自动编码器框架下的图表示学习，我们遵循以下介绍的文献[4,5,19]。

图形自动编码器（GAE）。给定一个图表*格*= (*五*、）的目标是学习一个编码器网络
f: *能量* × *能量* ⇉ 赫映射每个节点*五*∈*V*变成*d*维嵌入向量时长*五*∈R*d*，以及
解码器网络*问*: 赫→ −*E*重建网络结构（例如，*埃*）来自潜在的
空间H.它是一个无监督框架，其关键是保留潜在空间中的拓扑结构和节点属性赫通过准确恢复观察到的边缘*埃*然后通过链接预测和节点分类任务来评估图自动编码器的性能。

编码器。编码器网络被实例化为 GNN 模型 [9, 10]，它们是图上成熟的表示学习器。遵循消息传递策略 [20]，GNN 的核心思想是通过聚合自身及其邻居的表示来更新每个节点的表示。

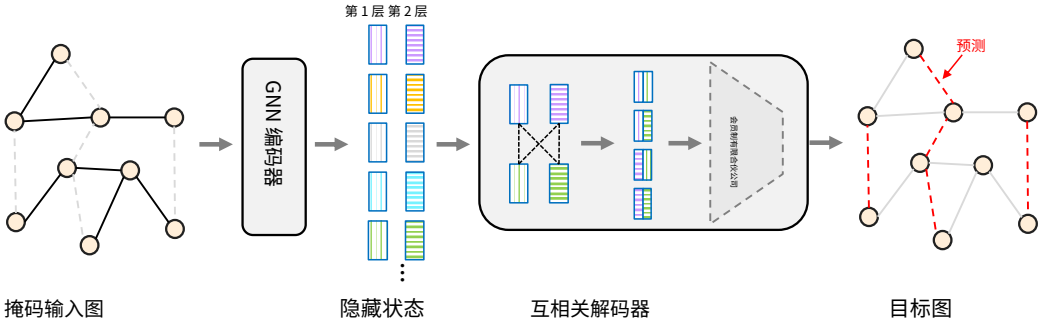


图 1：提出的 MGAE 架构。给定一个图，一大组随机的边被屏蔽。GNN 编码器应用于剩余的边以产生隐藏的表示。然后将这些表示输入到定制的解码器中，以在训练期间重建被屏蔽的边，从而捕获具有多粒度特征的端节点之间的互相关性。

正式地，在*钾*第层，我们有，

$$\text{时长}_{\text{五}}^{\text{钾}}=\text{通信}(\text{时长}_{(k-1)\text{五}}\ ,\text{AGG}(\{\text{时长}_{\text{你}^{(k-1)}}\ :\ \text{你}\in N_{\text{五}}\}))\ , \tag{1}$$

在哪里时长（五 表示节点的嵌入*五*在*钾*第层，以及*否*五={*u*|*u*|*e*，你∈*E*}是节点的直接邻居*五*.我们经常初始化时长（0） 五 =十五在实践中。函数 AGG 表示邻域聚合器。它通过可学习的注意力权重 [21, 22] 或由图拓扑确定的固定组合权重 [9, 23] 选择性地聚合来自邻居的特征。要更新节点*五*，另一个函数 COM 用于将聚合的邻居信息与来自前一层的自身节点嵌入相结合。对于具有*钾*层，有*钾*

节点表示{时长（1） 五，时长 五) · · · ，时长（五 }正在生成，其中时长（*钾*） 五 占领邻里结构内部*钾*酒花。

解码器。重建观察到的边缘*格*来自隐藏空间H，解码器网络被定义为边相似度函数。我们想要预测概率*是*，你那个边缘 *埃*，你存在。例如，一些研究[4]通过内积来估计似然值，即*是*，你=

$$\text{时长}_{\text{五}}^{(\text{钾})}\cdot\text{时长}_{\text{你}}^{(\text{钾})}$$

而其他一些工作[5, 24]使用多层参数化相似度函数感知器，即*是*，你=MLP（时长（钾）五，时长（钾你）。

3 掩蔽图自动编码器

在图 1 中，我们详细阐述了所提出的掩蔽图自动编码器 (MGAE) 框架的核心思想。这是一种简单的方法，只需给出部分观察到的边即可重建原始网络结构。与传统的图自动编码器一样，MGAE 有一个 GNN 编码器，用于将每个节点映射到潜在嵌入中，还有一个解码器，用于从潜在空间恢复输入图中的原始链接。不同之处在于：（1）我们允 许 GNN 编码器仅在部分网络结构上运行（即，一部分边被掩盖）；（2）设计了一种新颖的互相关解码器，通过以不同的粒度捕获边的头节点和尾节点之间的互相关，从潜在表示中重建被掩盖的边。我们的方法有四个组成部分：网络掩蔽、GNN 编码器、互相关解码器、和重建目标。下面我们将详细介绍它们。

网络掩蔽。我们扰乱了原始图*格*通过随机屏蔽边缘子集。正式地，我们使用*埃面具*和*埃预订*分别表示被掩盖和保留的边集，其中*埃面具*∪*E预订*=*埃*。为了提高效率，我们采用高掩蔽率的随机采样来获取掩蔽边缘集*埃面具*。未来我们将探索更复杂的采样策略。具体来说，我们进一步考虑两种类型的随机采样方案来生成掩码输入图。

我。*无向掩蔽*. 我们将节点之间的链接视为*五*和*五*无向的。也就是说，*埃，你*和*埃紫外线*是等价的，并且只包含一个副本*埃*。因此，进行随机抽样*埃*，蒙版边缘集*埃面具*也是无向的。

二.*定向掩蔽*背后的关键假设是图中的链接是有向的。因此，*埃，你*和*埃紫外线*是不同的，并且它们都包含在*埃*. 删除*埃，你*并不意味着*埃紫外线*也被删除。因此，在对*埃*，结果的掩蔽边集*埃面具*是导演。

两种掩蔽方案的主要区别在于，无向掩蔽会比有向掩蔽使重建更具挑战性。根据自监督原理，更困难的任务（例如，掩蔽高比例的边）在理论上有望取得更好的性能。然而，在网络掩蔽场景中，我们发现最佳采样策略应该与网络结构有关。例如，如果网络密集，则最好采用更严格的策略（即无向 掩蔽）。相反，如果原始图稀疏，则有向掩蔽可能是更好的选择，因为无向掩蔽可能会显著破坏节点的邻域结构，导致 GNN 编码器的表达能力大幅下降。例如，在实验中，我们发现有向掩蔽对 OGB 数据集效果更好，而无向掩蔽对 Planetoid 数据集效果更好。

GNN 编码器。我们遵循标准的 GAE 方法，采用公式 (1) 中所示的成熟 GNN 模型作为我们的编码器。具体来说，我们将 GCN [9] 和 GraphSage [10] 架构视为主干。然而，我们的编码器只对整个边缘集的一小部分（例如 30%）进行操作 *埃*用于消息传播。*埃面具*在训练过程中被移除。这使我们能够以更少的计算和内存成本训练 GNN 编码器。被屏蔽的边缘由定制的解码器恢复，这将在后面介绍。值得注意的是，在相同的屏蔽率下，用于训练的边总数对于上述两种采样策略是相同的。这是因为 GNN 中的消息传播默认是双向的。

互相关解码器。所提出的 MGAE 解码器接收 *钾*隐藏表示矩阵*{*赫（1），赫（2），・・・，赫（钾）*}* 由 GNN 编码器生成。参见图 1。每个节点*五*有 *钾* 嵌入向量表示为*{*时长（钾）∈R*d* 五 *}*钾 钾=1，在哪里时长（钾五 是通过聚合来自邻居 *五*之内 *钾*跃点定义为*埃预订*。假设我们掩盖了原始图中高比例的边，则剩余的网络结构（邻接矩阵）由剩余的边组成
放*埃预订*不可避免地是不完整的。直接使用最终层表示赫（钾） 重建被遮盖的边相当困难。为了应对这一挑战，我们采用不同粒度对两个节点之间的相关性进行建模。具体来说，给定一条边*埃，你*和他们的 *钾*隐
陈述*{*时长钾 五，时长 五*}*钾钾=1，我们的目标是捕捉它们的互相关性，如下所示。

$$\text{时长}_{\text{埃}_{\text{五，你}}} = //_{\text{钾}_{\text{千}=1}} \text{时长（钾五} \quad \overset{\text{时长}(\text{杰})}{\text{你}}, \tag{2}$$

在哪里//表示连接和 表示元素乘法。这里时长*埃* ，你 ∈R氪度2 是边缘的最终表示*埃，你*。时长（钾） 五 时长你表示节点之间的交叉表示*五*和节点*你*，考虑到他们的*钾*-阶邻域和*杰*阶邻域。交叉表示背后的关键见解是突出不同粒度特征中两个节点的共同模式。此操作对于 MGAE 至关重要，因为*钾*潜在嵌入向量本身并不完整，含有噪声，因此我们必须识别它们的共享模式作为边缘重建的有效特征。此外，我们想指出的是*钾*通常是一个较小的数字（例如，*钾*=2）在图自动编码器中。因此，额外的计算和内存成本是有限的。

重建目标。与传统的图自动编码器不同，我们的 MGAE 通过学习预测被掩盖的边缘来恢复原始的图结构*埃面具*而不是观察到的*埃预订*。此设置的灵感来自最近计算机视觉领域的自监督学习[18]。通过强制 MGAE 解码器仅重建被遮盖的边缘，我们的 GNN 编码器对网络噪声具有鲁棒性，并且可以更有效地编码图形信息。
我们采用标准
基于图的损失函数来训练我们的模型。

$$\text{大号} = -\sum_{(\text{，你}) \in \text{E面具}} \text{日志}\,\Sigma \frac{\text{指数（}\gamma\text{）}}{\text{是}\in\text{V指数（是规则）}}, \tag{3}$$

在哪里是，你=MLP（时长*埃，你* ）是边缘的重建分数*埃，你*，而 MLP 是一个多层具有 ReLU 激活函数的感知器。在实验中，我们采用负采样[25, 26]

算法1：掩蔽图自动编码器 (MGAE)

输入：图形*格*= (*五*、）、GNN编码器深度*钾*、掩蔽比ω，嵌入维数d;

尽管未收敛 做

  随机掩蔽*格*按比例ω分为两个边集：*埃面具*和*埃预订*;对保留边集进行GNN编码

  *埃预订*根据公式（1）；获得边的交叉表示*埃面具*根据公式（2）；

  通过最小化重建损失来更新模型*埃面具*根据公式（3）；

返回训练好的 MGAE 模型。

加速优化，因为公式 (3) 中分母的求和运算计算量过大。我们的 MGAE 模型通过算法 1 中概述的随机梯度下降进行优化。

在训练 MGAE 模型后，与其他图自动编码器架构类似，我们可以使用解码器的输出对未见边执行链接预测任务。对于节点分类，我们使用 GNN 编码器为图中的节点生成表示。然后，我们将*钾*每个节点的表示作为其最终表示输入到分类器中。

表 1：数据集统计数据。

数据	# 节点	# 边 # 特征	分割率 # 类		
科拉	2，708	5，429	1，433	85/5/15	7
引用者	3，312	4，660	3，703	85/5/15	6
PubMed	19，717	四十四，338	500	85/5/15	3
ogbl-ddi	4，267	1，334，889	-	80/10/10	–
ogbl 合作	235，868	1，285，465	128	92/4/4	–
ogbl-ppa	576，289	三十，326，273	-	70/20/10	–
ogn-arxiv 文档	169，343	三十，326，273	128	–	40
ogn蛋白	132，534	三十九，561，252	8	–	112

4 实验

我们在各种开放图数据集上评估了 MGAE 的性能。具体来说，我们试图回答两个问题。问题 1：与最先进的链接预测和节点分类模型相比，MGAE 的效果如何？问题2：我们的模型在不同的掩蔽率下表现如何？

数据集和实验设置。我们在六个基准图数据集上评估了 MGAE 的预测性能，其中包括三个 Planetoid 数据集 [27]（Cora、Citeseer 和 Pubmed）和三个 OGB 数据集 [24]（ogbl-ddi、ogbl-collab 和 ogbl-ppa）。对于节点分类，我们在五个数据集（包括 Cora、Citeseer 和 Pubmed）和两个 OGB 节点分类数据集（ogbn-arxiv 和 ogbn-proteins）上进行了评估。数据统计汇总在表 1 中。

我们的模型基于 Pytorch [28] 和 PyG (PyTorch Geometric) 库 [29] 构建。我们使用 Adam [30] 优化器对 MGAE 进行了 200 次训练，并以 50 次训练周期的耐心进行了提前停止。我们的模型中有三个超参数，即掩蔽率ω，嵌入维数d以及编码器层*钾*.我们设置*钾*=2和ω=0。7如果没有指定，则默认。对于嵌入维度，我们固定 d=128和d=256分别用于 Planetoid 和 OGB 数据集。此外，我们应用*无向掩蔽*对于行星数据集和*定向掩蔽*默认用于 OGB 数据集。

4.1 链接预测

我们首先回答问题 1基于链接预测任务。

基线。我们考虑以下基线方法。三种代表性的 GAE 模型（GAE [9]、GraphSAGE [10] 和 ARGE [5]）和三种自监督方法（DGI [13]、GIC [14] 和 SelfTask-GNN [11]）。为了公平比较，我们在实验中采用 EdgeMask 作为 SelfTask-GNN 的自监督信号。我们的目标是通过使用相同的数据集分割和训练程序，在每个数据集上对不同模型进行严格而公平的比较。具体来说，对于

表 2：具有掩蔽比的行星数据的链接预测结果0。7.最好的结果被突出显示。

	科拉		引用者		PubMed	
	曲线下面积	美联社	曲线下面积	美联社	曲线下面积	美联社
直肠癌	90。02±0。80	90。61±1。00	95。53±0。40	95。72±0。10	91。24±0。60	93。33
政府投资公司	93。54±0。60		±0。7097。04±0。	50 96。80±0。	5093。71±0。	30 93。23±0。00
阿尔及利亚	92。40±0。00	92。83±0。03	88 91。24±0。	87 91。	95。±03	00
盖亚	91。09±0。01		。52±0。04 85。		91。68±0。05	
圣人	86。33±1。06		65±2。56 89。98		87。90±2。54	
自任务GNN	90。65±0。50		±0。四十五		91。52±0。52	
MGAE至GCN	93。52±0。2394。四十六±0。2493。二十九±0。49				93。81±0。4098。四十五±0。03 98。22±0。05	
MGAE一SAGE95。05±0。76 94。50±0。8694。85±0。49					94。68±0。34 97。三十八±0。17 97。11±0。19	

	96。81±0。00 96	11±0。00 96。50
	。40±0。01 89。	±0。02 89。四十
	22±0。87 97。四	四±0。82 97。11
	十八±0。12	±0。16

表 3：具有掩蔽率的 OGB 数据集上的链接预测性能0。7、除 ogbl-ppa 外，还需使用掩蔽率0。4.最佳结果已突出显示。“OOM”表示 GeForce RTX 3090 GPU 设备 (24GB) 内存不足。

	ogbl-ddi		ogbl 合作		ogbl-ppa	
	点击数@20	点击数@30	点击数@50	点击数@100	点击数@10	点击数@50
直肠癌	十三。87±4。81	15。31±5。52 13	对象对象模型	对象对象模型	对象对象模型	对象对象模型
政府投资公司	10。56±6。77 37	。二十九±7。44	对象对象模型	对象对象模型	对象对象模型	对象对象模型
地理信息网络	。07±5。07 53。	51。56±4。19	四十四。75±1。07	52。三十±1。01	2。52±0。四十七	10。82±1。04 8。
图形管理软件	90±4。74 20。43	65。80±6。9454。	63±1。12 60。	23±1。201。87±0。	67 23。86	92±2。二十八
阿尔及利亚	±4。66 42。二十	±6。53 50。67±6。	02 三十九±2。	三十七。66±1。	0。41±0。二十六	3。83±0。84
自任务GNN	六±4。85		51 33。03±5。11	98 42。22±7。14	0。65±0。三十	4。二十六±1。二十七
MGAE至GCN	65。91±3。50 75。02±2。26 54。74±1。06 61。01±1。18 3。98±1。33					9。97±1。55
MGAE一SAGE66。00±9。49 75。18±6。5749。二十七±0。96				55。四十四±0。82 1。三十七±0。三十八		4。79±0。16

对于 Planetoid 数据集（Cora、CiteSeer 和 PubMed），我们按照 [4] 的方法将所有边随机分成三组，即训练集（85%）、验证集（5%）和测试集（10%），并根据 AUC 和平均精度 (AP) 得分评估性能。对于 OGB 数据集（ogbl-ddi、ogbl-collab 和 ogbl-ppa），我们按照 [24] 的方法将数据集按照表 1 中总结的分割比例分成三组，并使用命中率 (Hits@N) 评估其性能，其中 *否* 是被召回的节点数。对于我们的模型，我们考虑两种变体：MGAE-GCN 和 MGAE-SAGE，这意味着我们使用 GCN 和 GraphSage 架构来实现我们的 GNN 编码器。

表 2 和表 3 分别报告了 Planetoid 和 OGB 数据集上 10 次运行的平均结果。结合两个表的结果，我们得出以下主要观察结果。

- 我们的模型 MGAE 在几乎所有情况下在六个数据集上的表现都优于基于图自动编码器的基线（AGRE、GAE 和 GraphSage）。具体而言，MGAE 在 ogbl-collab 和 ogbl-ppa 数据集上取得了与最佳图自动编码器基线相当的结果，而在其他四个数据集上的表现则明显优于它们。在六个数据集中，MGAE 在 Cora、PubMed 和 ogb-ddi 数据集上获得了新的最佳性能。
- 与自监督学习基线（DGI、GIC 和 SelfTask-GNN）相比，我们的模型 MGAE 在六个数据集中的五个上实现了显着的性能提升。特别是，MGAE 仅在 CiteSeer 数据集上输给 GIC，而在其余五个数据集上表现出色。此外，我们还观察到我们的模型与三个自监督基线之间的性能差距在 OGB 数据集上有所增加。该结果表明图自动编码器架构更适合大规模数据集上的链接预测任务。
- 另一个重要的观察结果是*没有任何*我们的两个变体 MGAE-GCN 和 MGAE-SAGE 在六个数据集上的表现始终优于另一个。例如，尽管 GAE 在 Cora 和 CiteSeer 数据集上的表现优于 GraphSage，但 MGAE-SAGE 在这两个数据集上的表现优于 MGAE-GCN。这表明 MGAE 的最佳 GNN 编码器在不同的图场景下有所不同。

表 4：基于所有数据集的节点分类性能70%随机边缘遮蔽。

方法	科拉 ACC。	引用者 ACC。	PubMed ACC。	ogn-arxiv 文档 ACC。	ogn蛋白 曲线下面积
地理信息网络 图形管理软件 澳大利亚国立大学 自任务GNN 直肠癌 政府投资公司	83。60±0。52 74 。三十±1。84 85 。86±0。72 84。 69±0。09 85。41 ±0。三十四 87。 70±0。01	63。三十七±1。 21 60。20±2。15 73。10±0。86 71 。82±0。13 74。 51±0。51 76。三 十九±0。02	78。23±1。63 81 。96±0。74 81。 85±1。01 83。92 ±0。18 85。95± 0。66 85。99±0 。十三	66。01±0。37 64 。79±2。91 50。 06±1。21 68。三 十±0。02 67。08 ±0。43 64。00± 0。22	61。67±0。35 55 。三十九±0。79 40。73±0。68 60 。93±0。44 50。 31±0。55 48。55 ±0。四十七
镁基合金	86。15±0。二十五	74。60±0。06	86。91±0。二十八	72。02±0。05	63。33±0。12

- 我们要指出的 MGAE 的另一个有希望的特性是表 2 和表 3 中的结果是在高掩蔽比下获得的（ω=0。7）、ogbl-ppa 除外。也就是说，我们只将原始图的 30% 原始边馈送到 GNN 编码器。因此，MGAE 自然比传统的图自动编码器模型更高效，因为消息传播是 GNN 模型中最耗时的过程。此外，它还表明图数据中的许多节点连接是冗余的。这一观察结果与结构学习 [31, 32] 或图稀疏化 [33, 34] 的动机一致。

4.2 节点分类

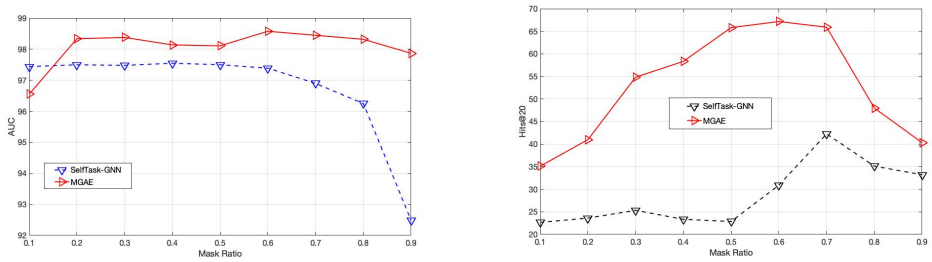
除了链接预测之外，为了进一步回答问1，我们在 Planetoid（Cora、CiteSeer 和 PubMed）和 OGB（ogbn-arxiv 和 ogbn-protiens）数据集上的节点分类上评估我们的模型。我们随机将所有边的 10% 分成验证集，并使用剩余的 90% 的边作为训练集。验证集用于调整超参数。模型训练完成后，我们使用全集边作为输入来生成节点表示以供下游评估。具体来说，我们在所有模型的学习到的节点表示上训练一个 SVM 分类器，并应用 5 倍交叉验证来估计性能。为了避免随机性，我们重复该过程 10 次，并按照 [26, 24] 报告 Core、CiteSeer、PubMed 和 ogbn-arxiv 的准确度 (ACC) 和 ogbn-proteins 的 AUC 的平均结果。我们采用相同的基线作为链接预测设置，并在表 4 中报告结果。主要观察结果如下。

- 我们的模型 MGAE 在五个数据集上的表现始终优于三个基于图自动编码器的基线（GCN、GraphSage 和 ARGVA）。鉴于 MGAE 在链接预测任务中可以获得至少与经典图自动编码器相当的结果，这表明所提出的 MGAE 是一种强大的图自动编码器替代方案。
- 与基于自监督学习的模型（DGI、GIC 和 SelfTask-GNN）相比，我们的模型在两个小型数据集（Cora 和 CiteSeer）上的表现不如它们中的最佳模型。然而，MGAE 在三个大型数据集（PubMed、obgn-arxiv 和 ogbn-proteins）上的表现却远远胜过它们。

- SelfTask-GNN 是一项与 MGAE 密切相关的工作，因为它们都专注于随机屏蔽一些边缘作为自监督训练任务。然而，根据表 2、3 和 4 中的结果，由于解码器设计和屏蔽策略有限，SelfTask-GNN 与最先进的基线相比表现不佳。这些结果表明，我们的 MGAE 模型是第一个能够成功采用自监督学习来提升经典图自动编码器性能的工作。

4.3 敏感性分析

在本节中，我们进行实验来验证掩蔽率的影响ω在我们的模型 MGAE 上（问2)。为了对 MGAE 进行全面评估，我们引入了 SelfTask-GNN 进行比较，因为它可以被视为我们模型的变体，通过用简单的 MLP 网络替换互相关解码器。具体来说，我们改变ω从 0.1 到 0.9，步长为 0.1。图 2 展示了 MGAE-GCN 和 SelfTask-GNN 在 PubMed 和 ogbl-ddi 数据集上的结果。在其他数据集上也观察到了类似的曲线。从图中我们有两个主要观察结果。



（一）PubMed　　（b）ogbl-ddi

图 2：MGAE 与 SelfTask-GNN 在链接预测方面的表现ω。

- MGAE 的性能首先随着掩蔽比的增加而提高ω直到达到0。7、然后它下降ω进一步增加。此外，我们的 MGAE 模型在ω约为0.5到0.7。这些结果表明了我们的模型在高掩蔽率下的稳健性。

- MGAE 在两个数据集上的表现在不同情况下始终优于 SelfTask-GNN ω值，除非ω <0。2在 PubMed 数据集上。此外，当ω约为 0.5 和 0.7。这些观察结果验证了我们提出的互相关解码器对于自监督图自动编码器训练的有效性。

5 结论

我们探索了一种新颖的掩蔽图自动编码器 (MGAE) 框架，用于图上的无监督表示学习。它从自监督学习中汲取灵感，可以看作是自监督图自动编码器的替代品。与原始图自动编码器模型不同，MGAE 建议随机掩蔽大部分（即 70%）原始图结构作为输入，并仅重建被掩蔽的边进行模型训练。具体来说，我们引入了两种边掩蔽策略：无向掩蔽和有向掩蔽，以生成有效的自监督任务。此外，我们还提出了一种定制的互相关解码器，通过捕获其头节点和尾节点之间的交叉表示来有效地恢复缺失的边。在多个开放图基准上的大量实验结果验证了 MGAE 在链接预测和节点分类任务方面优于最先进的基线。

参考

[1] William L Hamilton、Rex Ying 和 Jure Leskovec。图上的表示学习：方法和应用。arXiv 预印本 arXiv:1709.05584，2017年。

[2] Fenxiao Chen, Yun-Cheng Wang, Bin Wang, and CC Jay Kuo. 图表示学习：一项调查。APSIPA 信号与信息处理学报，2020年9月9日。

[3] Bryan Perozzi、Rami Al-Rfou 和 Steven Skiena。Deepwalk：社交表征的在线学习。第 20 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第 701–710 页，2014 年。

[4] Thomas N Kipf 和 Max Welling。变分图自动编码器。arXiv 预印本 arXiv:1611.07308，2016 年。

[5] 潘诗睿，胡瑞琪，龙国栋，蒋静，姚丽娜，张成琪。用于图嵌入的对抗性正则化图自动编码器。arXiv 预印本 arXiv:1802.04407，2018 年。

[6] 谭乔宇, 刘宁浩, 胡侠. 深度表示学习在社会网络分析中的应用。大数据前沿，2：2，2019。

[7] 张道琨，尹杰，朱兴全，张承其。网络表征学习：一项调查。IEEE 大数据交易 ,6(1):3–28, 2018。

[8] 张慕涵、李攀、夏英龙、王凯、金龙，重新审视用于链接预测的图神经网络。arXiv 预印本 arXiv:2010.16103，2020年。

[9] Thomas N Kipf 和 Max Welling。使用图卷积网络进行半监督分类。arXiv 预印本 arXiv:1609.02907，2016 年。

[10] William L Hamilton、Rex Ying 和 Jure Leskovec。大型图上的归纳表示学习。第 31 届神经信息处理系统国际会议论文集，第 1025–1035 页，2017 年。

[11] Wei Jin, Tyler Derr, Haochen Liu, Yiqi Wang, Suhang Wang, Zitao Liu, Jiliang Tang. 图上的自监督学习：深刻见解和新方向。arXiv 预印本 arXiv:2006.10141，2020年。

[12] 徐东宽，程伟，罗东升，陈海峰，张翔。 Infogcl：信息感知图对比学习。神经信息处理系统的进展，34，2021年。

[13] Petar Veličković、William Fedus、William L Hamilton、Pietro Liò、Yoshua Bengio 和 R Devon Hjelm。深度图信息最大化。arXiv 预印本 arXiv:1809.10341，2018 年。

[14] Costas Mavromatis 和 George Karypis。图信息聚类：最大化图中的粗粒度互信息。在亚太知识发现与数据挖掘会议，第 541–553 页。Springer，2021 年。

[15] Philip Bachman、R Devon Hjelm 和 William Buchwalter。通过最大化各个视图之间的相互信息来学习表示。arXiv 预印本 arXiv:1906.00910，2019年。

[16] 刘晓，张凡进，侯振宇，李勉，王兆宇，张静，唐杰。自我监督学习：生成式学习或对比式学习。IEEE 知识与数据工程学报，2021年。

[17] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。Bert：用于语言理解的深度双向转换器的预训练。arXiv 预印本 arXiv:1810.04805，2018 年。

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár 和 Ross Girshick. 蒙版自动编码器是可扩展的视觉学习器。arXiv 预印本 arXiv:2111.06377，2021年。

[19] Guillaume Salha、Romain Hennequin、Viet Anh Tran 和 Michalis Vazirgiannis。可扩展图自动编码器的退化框架。arXiv 预印本 arXiv:1902.08813，2019年。

[20] Justin Gilmer、Samuel S Schoenholz、Patrick F Riley、Oriol Vinyals 和 George E Dahl。量子化学的神经信息传递。国际机器学习会议，第 1263–1272 页。PMLR，2017 年。

[21] Petar Veličković、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。图注意力网络。arXiv 预印本 arXiv:1710.10903，2017年。

[22] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。你只需要关注。神经信息处理系统的进展，第 5998–6008 页，2017 年。

[23] Felix Wu、Amauri Souza、Tianyi Zhang、Christopher Fifty、Tao Yu 和 Kilian Weinberger。简化图卷积网络。国际机器学习会议，第 6861-6871 页。PMLR，2019 年。

[24] Weihua Hu、Matthias Fey、Marinka Zitnik、Yuxiao Dong、Hongyu Ren、Bowen Liu、Michele Catasta 和 Jure Leskovec。开放图基准：图机器学习数据集。arXiv 预印本 arXiv:2005.00687，2020年。

[25] 杨振，丁明，周常，杨红霞，周景仁，唐杰。了解图表示学习中的负采样。在第 26 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第 1666–1676 页，2020 年。

[26] Weihua Hu、Matthias Fey、Hongyu Ren、Maho Nakata、Yuxiao Dong 和 Jure Leskovec。Ogblsc：基于图的机器学习的大规模挑战。arXiv 预印本 arXiv:2103.09430，2021年。

[27] Prithviraj Sen、Galileo Namata、Mustafa Bilgic、Lise Getoor、Brian Galligher 和 Tina Eliassi-Rad. 网络数据中的集体分类。人工智能杂志 , 29(3):93–93, 2008。

[28] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等人。Pytorch：命令式、高性能深度学习库。神经信息处理系统的进展 , 32: 8026–8037, 2019年。

[29] Matthias Fey 和 Jan Eric Lenssen。使用 Pytorch 几何进行快速图形表示学习。 arXiv 预印本 arXiv:1903.02428，2019年。

[30] Diederik P Kingma 和 Jimmy Ba. Adam：一种随机优化方法。 国际学习表征会议，2015年。

[31] Luca Franceschi、Mathias Niepert、Massimiliano Pontil 和 Xiao He。学习图神经网络的离散结构。国际机器学习会议，第 1972-1982 页。PMLR，2019 年。

[32] 金伟，马瑶，刘晓瑞，唐先锋，王苏航，唐继良。鲁棒图神经网络的图结构学习。第 26 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第 66–74 页，2020 年。

[33] 郑程，宗波，程伟，宋东进，倪景超，余文超，陈海峰，王伟。通过神经稀疏化进行鲁棒图表示学习。在国际机器学习会议，第 11458-11468 页。PMLR，2020 年。

[34] Guihong Wan 和 Haim Schweitzer。通过元学习实现图的边缘稀疏化。在2021 IEEE 第 37 届数据工程国际会议（ICDE），第 2733-2738 页。IEEE，2021 年。