屏蔽标签预测：半监督分类的统一信息传递模型

Yunsheng Shi , **Zhengjie Huang** , **Shikun Feng** , **Hui Zhong , Wenjing Wang** , **Yu Sun**
百度公司，中国
{shiyunsheng01, huangzhengjie, fengshikun01, zhonghui03, wangwenjin02,
sunyu02}@baidu.com

摘要

图神经网络（GNN）和标签传播算法（LPA）都是消息传递算法，在半监督分类中取得了优异的性能。GNN 通过神经网络的*特征传播*来进行预测，而 LPA 则通过图邻接矩阵的*标签传播*来获得结果。然而，目前仍没有有效的方法将这两种算法直接结合起来。为了解决这个问题，我们提出了一种新颖的统一消息传递模型（UniMP），它可以在训练和推理时结合*特征*和*标签传播*。首先，UniMP 采用图传网络，将特征嵌入和标签嵌入作为传播的输入信息。其次，为了在不过度拟合自循环输入标签信息的情况下训练网络，UniMP 引入了屏蔽标签预测策略，即随机屏蔽一定比例的输入标签信息，然后进行预测。UniMP 在概念上统一了特征传播和标签传播，在经验上非常强大。它在开放图基准（Open Graph Benchmark，OGB）中获得了最新的半监督分类结果。

1 导言

世界上有各种各样的场景，例如推荐相关新闻、发现新药或预测社会关系，这些都可以用图结构来描述。人们提出了许多方法来优化这些基于图的问题，并在许多相关领域取得了重大成功，如预测节点属性[Yang *等人*，2016； Kipf 和 Welling， 2016]、关系链接[Grover 和 Leskovec， 2016； Battaglia *等人*， 2018] 和图分类[Duvenaud *等人*，2015；Niepert *等人*，2016]。

在半监督节点分类任务中，我们需要利用有标签的示例进行学习，然后对这些无标签的示例进行预判。为了更好地对图中的节点标签进行分类，基于拉普拉斯平滑

	培训		推理	
模型	**特点**	**标签**	**特征**	**标签**

假设[Li *等人*，2018；Xu *等人*，2018b]，提出了消息传递模型来聚合图中其相连邻居的信息，获取足够的

LPA			C		C	
GCN	C		C			
APPNP	C		C			
GCN-LPA	C	C	C			
统一管理计划（我们的）		C	C	C	C	

表 1：比较信息传递模型在训练和推理中使用的输入信息。

事实，从而对无标记节点进行更稳健的预测。一般来说，实现消息传递模型的方法主要有两种，图神经网络（GNNs）[Kipf and Welling, 2016; Hamilton *et al.,* 2017; Xu et *al.*, 2018b; Liao et *al*、2019；Xu *等人，*2018a] 和标签传播算法（LPA）[Zhu *等人，*2003；Zhang 和 Lee，2007；Wang 和 Zhang，2007；Karasuyama 和 Mamitsuka，2013；Gong *等人，*2016；Liu *等人，*2019]。GNN 通过在多个神经层中传播和聚合节点特征来结合图结构，并从*特征传播*中获得预测结果。而 LPA 通过*标签传播*反复对未标记的实例进行预测。

由于 GNN 和 LPA 基于相同的假设，即通过信息提取进行半监督分类，因此直觉上将它们结合在一起可以提高性能。一些卓越的研究已经在此基础上提出了自己的图模型。例如，APPNP [Klicpera *等人，*2018] 和 TPN [Liu *等人*，2019] 使用 GNN 预测软标签，然后传播软标签；GCN- LPA [Wang 和 Leskovec，2019] 使用 LPA 对其 GNN 模型进行正则化。然而，如表 1 所示，上述方法仍无法直接将 GNN 和 LPA 纳入消息传递模型，在训练和推理过程中*传播特征*和*标签*。

在这项工作中，我们提出了统一消息传递模型（UniMP），通过两个简单但有效的想法来解决上述问题：(a) 将节点特征传播与标签相结合；(b) 屏蔽标签预测。以前基于 GNN 的方法只将节点特征作为输入，并使用部分观察到的节点标签进行监督训练。这些方法在推理过程中会丢弃观察到的标签。UniMP 在训练和推理阶段同时使用节点特征和标签。它使用嵌入技术将部分节点标签从单点转换为喜欢节点特征的稠密向量。

图。而多层图转换器网络则将其作为输入，在节点之间进行贴心的信息传播。因此，每个节点都可以从其邻居那里收集特征和标签信息。由于我们将节点标签作为输入，使用它进行监督训练会导致标签泄漏问题。模型会过度拟合自循环输入标签，而推理能力却很差。为了解决这个问题，我们提出了一种掩码标签预测策略，即随机掩码一些训练状态的标签，然后对其进行预测，以克服标签泄漏问题。这种简单有效的训练方法借鉴了 BERT 中屏蔽词预测的经验 [Devlin 等人，2018]，模拟了图中从有标签实例到无标签实例的标签转换过程。

我们在开放图基准（Open Graph Benchmark，OGB）中的三个半监督分类数据集上评估了我们的 UniMP 模型，我们的新方法在所有任务中都取得了新的一流结果，在 *ogbn-products* 中获得了 82.56% 的 ACC，在 *ogbn-proteins* 中获得了 86.42% 的 ROC-AUC，在 *ogbn-arxiv* 中获得了 73.11% 的 ACC。我们还对 UniMP 模型进行了消融研究，以评估我们的统一方法的有效性。此外，我们还对 la- bel 传播如何提高模型性能进行了最透彻的分析。

2 序言

在本节中，我们将简要回顾相关工作，并介绍我们的术语。我们将图表示为 $G = (V, E)$，其中 V 表示图中的节点，E 表示图中的节点。

$|V| = n$，E 表示边。节点由特征矩阵 $X \in \mathbb{R}^{n \times m}$ 和目标类矩阵 $Y \in \mathbb{R}^{n \times c}$ 描述，特征矩阵通常是维数为 m 的稠密向量，目标类矩阵 Y R 的类数为 c。邻接矩阵 $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ 用于描述图 G，对角阶数矩阵用 $D = \text{diag}(d_1, d_2, ..., d_n)$ 表示，其中 $d_i = \sum_j a_{ij}$ 是节点 i 的阶数。

矩阵定义为 D^{-1} A 或 $D^{-} AD^{-}$，我们采用的是本文中的第一个定义。

图神经网络。在半监督节点分类中，GCN [Kipf and Welling, 2016] 是基于拉普拉斯平滑假设的最经典模型之一。GCN 通过多个层（包括线性层和非线性激活层）在图中转换和传播节点特征 X，以建立近似映射：$X \mapsto Y$。第 l 层 GCN 的特征传播方案是

$$H^{(l+1)} = \sigma(D^{-1} A H^{(l)} W^{(l)})$$
$$Y = f_{out}(H)^{(L)} \quad (1)$$

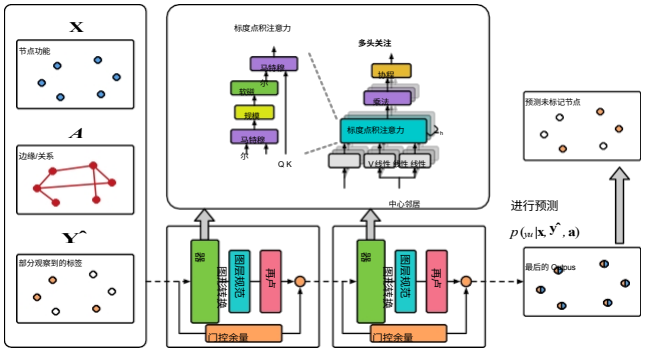

图 1：UniMP 的结构。

在图中反复传播标签。给定初始标签矩阵 $\hat{Y}^{(0)}$，该矩阵由贴有标签的节点的一热标签指示向量 $\hat{y}^0$ 或未贴标签的零向量组成。LPA 的简单迭代方程如下：

$$\hat{Y}^{(l+1)} = D^{-1} A \hat{Y}^{(l)} \quad (2)$$

标签通过标准化邻接矩阵 D^{-1} A 从其他节点传播。

结合 GNN 和 LPA。最近，在半分类任务中结合 GNN 和 LPA 成为一种趋势。APPNP [Klicpera 等人，2018] 和 TPN [Liu 等人，2019] 提出使用 GCN 预测软标签，然后用个性化 Pagerank 传播软标签。不过，这些研究仍然只考虑了部分节点标签作为监督训练信号。GCN-LPA 与我们的工作最为相关，因为他们也将部分节点标签作为监督信号。不过，他们以一种更间接的方式将 GCNN 和 LPA 结合起来，只是在训练中使用 LPA 来规范其 GAT 模型的权重边。而我们的 UniMP

GCN-LPA 将 GNN 和 LPA 正确地结合在一个网络中，在训练和预测中传播节点特征和标签。此外，GCN-LPA 的正则化策略只能用于 GAT [Velic˘kovic´ *et al.*, 2017]、GAAN [Zhang *et al.*, 2018]等具有可训练权重边的 GNN，而我们的训练策略则不同，可以很容易地扩展到 GCN 和 GAT 等各种 GNN 中，以进一步提高其性能。我们将在下一节更具体地介绍我们的方法。

3 统一信息传递模型

如图 1 所示，给定节点特征 X 和部分

其中，σ 是激活函数，$W^{(l)}$ *是第* 1 层的可训练权重，$H^{(l)}$ 是*第* 1 层的节点代表。$H^{(0)}$ 等于节点输入特征 X 。最后，f_{out} 输出层应用于最终表示，对 Y 进行预测 。

标签传播算法。标签传播算法（LPA）等传统算法仅利用节点间的标签和关系进行预测。LPA 认为连接节点之间的标签相似且

观察到的标签 $\hat{Y}$ ，我们采用图形变换器，共同
利用标签嵌入将上述功能和标签传播结合在一起，构建出我们的 UniMP 模型。此外，我们还引入了一种屏蔽标签预测策略来训练我们的模型，以防止标签泄漏问题 。

3.1 图形转换器

由于 Transformer [Vaswani *等人*，2017 年；Devlin *等人* ，2018 年] 已被证明在 NLP 中功能强大，因此我们采用其 vanilla

多头关注图学习，同时考虑到边缘特征的情况。具体来说，给定节点特征
图 $H^{(l)} = \{h_1, h_2^{(l)(l)}, ..., h_n^{(l)}\}$，我们计算多头在...
从 j 到 i 的每条边的注意事项 如下

$$
\begin{aligned}
q_{c,i}^{(l)} &= W_{c,qhi}^{(l)(l)} + b_{c,q}^{(l)} \\
k_{c,j}^{(L)} &= W^{(L)(L)} h + b \\
e_{c,ij} &= W_{c,eeij} + b_{c,e} \\
\alpha_{c,ij}^{(l)} &= \frac{\{q_{c,i}^{(l)} k_{c,u}^{(l)} + e_{c,ij}\}}{\sum_{u \in N(i)} \langle q_{c,i}^{(l)} k_{c,u}^{(l)} + e_{c,iu} \rangle}
\end{aligned}
\tag{3}
$$

其中，$\langle q, k \rangle = \exp(\frac{q^T k}{\sqrt{d}})$ 为指数级点积函数，d 为每个头部的隐藏大小。对于第 c 个头的注意力，我们首先将源特征 $h_i^{(l)}$ 和
远距离特征 $h_j^{(l)}$ 变为查询向量 $q_{c,i}^{(l)} \in \mathbb{R}^d$ 和关键向量 $k_{c,j}^{(l)} \in \mathbb{R}^d$，分别使用不同的可训练参数 $W_{c,q}^{(l)}, W_{c,k}^{(l)}, b_{c,q}^{(l)}, b_{c,k}^{(l)}$。提供的边缘特征 e_{ij} 将编码并添加到关键矢量中，作为每一层的附加信息。

在获得图多头关注后，我们会从远方的 j 向来源地 i 进行信息聚合：

$$
\begin{aligned}
v_{c,j}^{(l)} &= W_{c,v}^{(l)} h_j^{(l)} + b_{c,v}^{(l)} \\
\hat{h}_i^{(l+1)} &= \sum_{c=1}^{C} \sum_{j \in N(i)} \alpha_{c,ij}^{(l)} (v_{c,j}^{(l)} + e_{c,ij})
\end{aligned}
\tag{4}
$$

其中$\|$是 C head atten- tion 的连接操作。与公式 1 相比，多头注意力矩阵取代了原来的归一化邻接矩阵，成为信息传递的过渡矩阵。远端特征 h_j
转化为 $v_{c,j} \in \mathbb{R}^d$ 进行加权求和。

此外，受李[2019]和陈[2020]的启发，我们
如公式 5 所示，我们建议在层与层之间使用门控残差连接，以防止模型过度平滑。

$$
\begin{aligned}
r_i^{(L)} &= W^{(L)(L)} h_i + b^{(L)} \\
\beta_i^{(l)} &= \text{sigmoid}(W_g^{(l)} [\hat{h}_i^{(l+1)}; r_i^{(l)}; \hat{h}_i^{(l+1)} - r_i^{(l)}]) \\
h_i^{(l+1)} &= \text{ReLU}(\text{LayerNorm}((1 - \beta_i^{(l)}) \hat{h}_i^{(l+1)} + \beta_i r_i))^{(l)(l)}
\end{aligned}
\tag{5}
$$

特别是，与 GAT 类似，如果我们在最后一个输出层应用图形变换，我们将对多头输出进行平均，并去除非线性变换，如下所示：

和零向量。然后，我们将标签传播与 Graph Transformer 结合起来，简单地将节点特征和标签向量加在一起，即

传播信息（$H^0 = X + \hat{Y}$）$\in \mathbb{R}^{n \times m}$。我们可以
证明通过映射部分标记的 $\hat{Y}$ 和节点特征
我们的模型是将标签传播和特征传播统一在同一空间内
。
共享消息传递框架。让我们把 $\hat{Y}_d = \hat{Y} W_d$
和 A^* 是归一化邻接矩阵 D^{-1} A 或图形转换器中的保留矩阵，如等式 3。
那我们就能找到：

$$
\begin{aligned}
H^{(0)} &= X + \hat{Y} \\
H^{(l+1)} &= \sigma(((1-\beta)A^* + \beta I) H^{(l)(l)} W_d)
\end{aligned}
\tag{7}
$$

其中，β 可以是等式 5 这样的门控函数，也可以是 APPNP 这样预先定义的超参数 [Klicpera 等人，2018]。我们可以
为简化起见，我们将 σ 设为标识函数，那么
得到：

$$
\begin{aligned}
H^{(l)} &= ((1-\beta)A^* + \beta I)^l (X + \hat{Y} W_d) W^{(1)(2)} \ldots W^{(l)} \\
&= ((1-\beta)A^* + \beta I)^l X W + ((1-\beta)A^* + \beta I)^l \hat{Y} W W_d
\end{aligned}
\tag{8}
$$

其中 $W = W^{(1)(2)} \ldots W^{(l)}$。然后我们可以发现，我们的模型可以近似分解为特征传播（$((1-\beta)A^* + \beta I)^l X W$ 和标签传播
$((1-\beta)A^* + \beta I)^l \hat{Y} W_d W$。

3. 3 掩码标签预测

以往关于 GNN 的研究很少考虑在训练和推理阶段使用部分观察到的标签 $\hat{Y}$。它们
只将这些标签信息作为地面实况目标，在给定 X 和 A 的情况下对其模型参数 θ 进行监督训练：

$$
\arg\max_{\theta} \log p_{\theta}(\hat{Y}|X, A) = \sum_{i=1}^{\hat{V}} \log p_{\theta i}(\hat{y}_i|X, A)
\tag{9}
$$

其中，$\hat{V}$ 表示带有标签的部分节点。然而
我们的 UniMP 模型通过传播节点特征和标签来进行预测：$p(y|X, \hat{Y}, A)$。在我们的模型中简单地使用上述目标，将使训练中的标签泄漏成为可能。
阶段，导致推理效果不佳。从
BERT 可屏蔽输入字词，并对输入字词进行预测。

他们对模型进行预训练（屏蔽词预测），我们

我们提出了一种掩码标签预测策略来训练我们的模型。在训练过程中，每一步我们都会通过将部分节点标签屏蔽为零，将 $\hat{Y}$ 破坏为 $\tilde{Y}$，并保留
其他标签则保留，这由一个称为标签率的超参数控制。

假设这些被屏蔽的标签为 $\bar{Y}$，我们的目标是

$$\hat{h}_i^{(l+1)} = \frac{1}{C} \sum_{c=1} \sum_{j \in N(i)} \alpha_{c,ij}^{(l)} (y_{c,j}^{(l)} + e_{c,ij}^{(l)}) \tag{6}$$

$$h_i^{(l+1)} = (1 - \beta_i^{(l)}) \hat{h}_i^{(l+1)} + \beta_i^{(l)} r_i^{(l)}$$

3.2 标签嵌入和传播

我们建议将部分观测到的标签嵌入到与节点特征相同的空间中：$\hat{y} \in r^{n \times c} \rightarrow \hat{y}_d \in r^{n \times m}$、

其中包括已标注节点的标签嵌入向量 $\hat{Y}$ 作为输入标签，用于预测其余未标注节点。

功能是在给定 X、$\tilde{Y}$ 和 A *的情况*下预测 $\bar{Y}$：

$$\arg\max_{\theta} \log p_{\theta}(\bar{Y} | X, \tilde{Y}, A) = \sum_{i=1}^{\bar{}} \log p_{\theta}(\bar{y}_i | X, \tilde{Y}, A) \tag{10}$$

其中

V 代表那些带有屏蔽标签的节点。这样，我们就可以在不泄露自循环标签信息的情况下训练模型。在推理过程中，我们将使用所有

4 实验

我们提出了一种用于半监督节点分类的统一消息传递模型（UniMP），该模型通过图变换器将特征和标签传播联合起来，并采用屏蔽标签预测策略对其进行优化。我们在开放图基准（Open Graph Benchmark，OGBN）的节点属性预测（Node Property Pre- diction of Open Graph Benchmark）上进行了实验，该基准包括多个具有挑战性的大型数据集，用于半监督分类，并在与实际应用密切相关的程序中进行了拆分[Hu *等人*，2020]。为了验证我们的模型是否有效，我们在*ogbn-products*、*ogbn-proteins* 和*ogbn-arxiv* 三个 OGBN 数据集中将我们的模型与其他最先进（SOTA）模型进行了比较。我们还提供了更多的实验和全面的消融研究，以更直观地展示我们的动机，以及 LPA 如何改进我们的模型以获得更好的结果。

4.1 数据集和实验设置

名称	节点	边	任务	任务类型	公制
产品	2,449,029	61,859,140	1	多级类	准确性
ogbn 蛋白	132,534	39,561,252	112	二进制类	ROC-AUC
ogbn-arxiv	169,343	1 ,166,243	1	多级类	准确性

表 2：OGB 节点属性预测的数据集统计

数据集。与实际应用中的图形相比，大多数常用图形数据集的规模都非常小。由于这些数据集的小规模性、不可忽略的重复率或泄漏率、不真实的数据分割等问题，GNN 在这些数据集上的性能非常不稳定[Hu *等人*，2020]。因此，我们在最近发布的开放图基准（Open Graph Benchmark，OGB）数据集上进行了实验[Hu *等人*，2020]，这些数据集克服了常用数据集的主要缺点，因此更加真实和具有挑战性。OGB 数据集涵盖了现实世界中的各种应用，跨越了从社交和信息网络到生物网络、分子图和知识边缘图等多个重要领域。这些数据集还涵盖了节点、图和链接/边层面的各种预测任务。如表 2 所示，在这项工作中，我们在三个不同规模和任务的 OGBN 数据集上进行了实验，以获得可信的结果，这三个数据集包括：*ogbn-products*（关于 47 个产品类别的分类）和给定的 100 维节点特征；ogbn-*proteins*（关于 112 种蛋白质功能的预测）和给定的 8 维边缘特征；*ogbn-arxiv*（关于 40 类主题的分类）和给定的 128 维节点特征。有关这些数据集的更多详情，请参阅补充文件中的附录 A。

	ogbn-products	ogbn-proteins	ogbn-arxiv
取样方法	邻域取样	随机分区	全批取样
隐藏尺寸	128	64	128
辍学者	0.3	0.1	0.3
体重衰减	0.001	0.001	0.001
标签率	0.625	0.5	0.625

表 3：模型的超参数设置

实施细节。如上所述，这些数据集在规模或任务上各不相同。因此，我们按照之前的研究[Li 等人，2020]，用不同的采样方法对我们的模型进行了评估，得到了可信的比较结果。在 *ogbn-products* 数据集中，我们使用 Neigh- borSampling（每层大小=10）在训练期间对子图进行采样，并使用全批次进行推理。在 *ogbn 蛋白*数据集中，我们使用随机分割法将密集图分割成子图来训练和测试模型。至于小规模的 *ogbn-arxiv* 数据集，我们只在训练和测试中使用全批处理。我们在表 3 中为每个数据集设置了模型的超参数，标签率是指在应用屏蔽标签预测策略时保留标签的百分比。我们使用 Adam 优化器，其 lr
= 0.001 来训练我们的模型。特别地，我们将模型在小规模 *ogbn-arxiv* 数据集中的权重设置为 0.0005，以防止过拟合。有关调整后超参数的更多详情，请参阅补充文件中的附录 B。

4.2 与 SOTA 模型的比较

OGB leaderboard 提供基线和其他 SOTA 比较模型。所有这些结果都保证可以通过开放源代码进行复制。按照 OGB 的要求，我们对每个数据集运行了 10 次，并报告了平均值和标准偏差。如表 4、表 5 和表 6 所示，在三个 OGBN 数据集中，我们的统一模型优于所有其他比较模型。由于大多数比较模型只考虑了优化其模型的特征传播，这些结果表明，将标签传播纳入 GNN 模型可以带来显著的改进。具体而言，与 DeeperGCN 等新的 SOTA 方法相比，我们在 *ogbn-products* 中获得了 82.56% 的 ACC，在 *ogbn-proteins* 中获得了 86.42% 的 ROC- AUC，实现了约 0.6-1.6% 的 ab- solute 改进 [Li 等人，2020]。在*ogbn-arxiv中*，我们的方法获得了73.11%的ACC，与参数比我们大四倍的GCNII[Chen等人，2020]相比实现了0.37%的绝对改进。

表 4：ogbn-产品的结果

模型	测试 ROC-AUC	验证 ROC-AUC	参数
GaAN [Zhang *et al*.]	0.7803 ± 0.0073	-	-
GeniePath-BS [Liu 等人，2020b]	0.7825 ± 0.0035	-	316,754
MWE-DGCN	0.8436 ± 0.0065	0.8973±	538,544
DeepGCN [Li *et al*.]	0.8496 ± 0.0028	0.0057	2,374,456
DeeperGCN [Li *et* al.]	0.8580 ± 0.0017	0.8921 ± 0.0011	2,374,568
		0.9106 ± 0.0016	
统一管理计划	**0.8642 ±0.0008**	**0.9175 ± 0.0007**	1,909,104

表 5：ogbn 蛋白的结果

模型	测试精度	验证精度	参数
DeeperGCN [Li *et* al.]	0.7192 ± 0.0016	0.7262 ± 0.0014	1,471,506
GaAN [Zhang *et al*.]	0.7197 ± 0.0024	-	1,471,506
DAGNN [Liu *et* al.]	0.7209 ± 0.0025	-	1,751,574
JKNet [Xu *等人*，2018b]	0.7219 ± 0.0021	0.7335 ± 0.0007	331,661
GCNII [Chen *et al*.]	0.7274 ± 0.0016	-	2,148,648
统一管理计划	**0.7311 ± 0.0021**	**0.7450 ± 0.0005**	473,489

表 6：ogbn-arxiv 的结果

-

模型

	测试精度	验证精度	参数
GCN-Cluster [Chiang *et al*.]	0.7897 ± 0.0036	0.9212 ± 0.0009	206,895
GAT 集群	0.7923 ± 0.0078	0.8985 ± 0.0022	1,540,848
GAT-NeighborSampling	0.7945 ± 0.0059	-	1,751,574
GraphSAINT [Zeng *et al*.]	0.8027 ± 0.0026	-	331,661
DeeperGCN [Li *et* al.]	0.8090 ± 0.0020	0.9238 ± 0.0009	253,743
统一管理计划	**0.8256 ± 0.0031**	**0.9308 ± 0.0017**	1,475,605

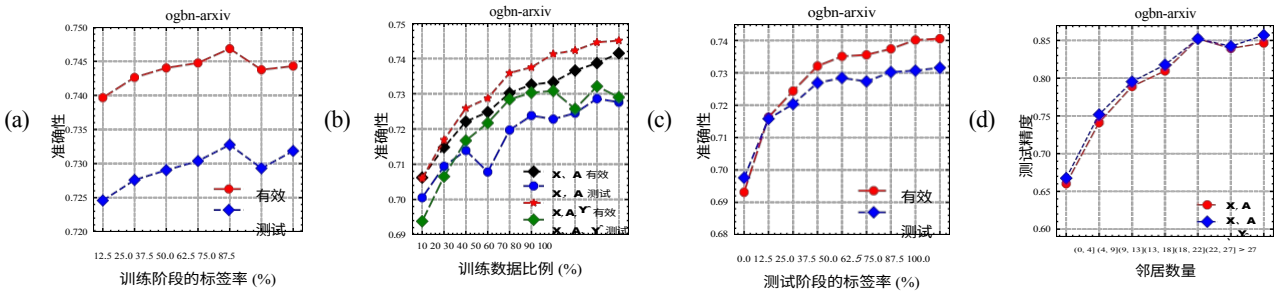

图 2：标签覆盖率如何影响标签传播的探索：（a）不同标签_率下的训练；（b）不同标签数据比例下的训练；（c）不同标签_率下的测试；（d）不同邻域下的测试准确率。

输入	模型	数据集		
		产品 测试 ACC	ogbn 蛋白 测试 ROC-AUC	ogbn-arxiv 测试 ACC
X	多层感知器	0.6106 ± 0.0008	0.7204 ± 0.0048	0.5765 ± 0.0012
X、A	GCN	0.7851 ± 0.0011	0.8265 ± 0.0008	0.7218 ± 0.0014
	GAT	0.8002 ± 0.0063	0.8376 ± 0.0007	0.7246 ± 0.0013
	图形转换器	0.8137 ± 0.0047	0.8347 ± 0.0014	0.7292 ± 0.0010
A、$\hat{Y}$	GCN	0.7832 ± 0.0013	0.8083 ± 0.0021	0.7018 ± 0.0009
	GAT	0.7751 ± 0.0054	0.8247 ± 0.0033	0.7055 ± 0.0012
	图形转换器	0.7987 ± 0.0104	0.8160 ± 0.0007	0.7090 ± 0.0007
X、A、$\hat{Y}$	GCN	0.7987 ± 0.0104	0.8247 ± 0.0032	0.7264 ± 0.0003
	GAT	0.8193 ± 0.0017	0.8556 ± 0.0009	0.7278 ± 0.0009
	图形转换器	**0.8256 ± 0.0031**	0.8560 ± 0.0003	**0.7311 ± 0.0021**
	1 w/ Edge Feature	*	**0.8642 ± 0.0008**	*

表 7：这是对不同输入的模型进行的消融研究，其中 **X** 表示节点特征，**A** 是图邻接矩阵，$\hat{\mathbf{Y}}$ 是观察到的标签。在 *ogbn 蛋白*中，节点特征最初并不提供。在本实验中，我们将边缘特征作为其节点特征的平均值，并提供不带边缘特征的 Transformer 的实验结果，以进行公平比较，该结果与表 5 略有不同。

4.3 消融研究

在本节中，为了更好地确定我们提出的模型的不同组成部分所带来的改进，我们从以下四个方面进行了扩展研究：

- 首先，我们在各种 GNNS 上应用了掩蔽标签预测策略，以显示结合 LPA 和 GNN 的有效性和鲁棒性，如表 7 所示。
- 为了获得更实用、更有效的屏蔽标签预测策略，我们在训练和推理过程中调整了标签_率，以探索标签覆盖率与 GNN 性能之间的关系，如图 2 所示。
- 我们还分析了 LPA 如何影响 GNN，使其表现更好，如图 3 所示。
- 此外，在表 8 中，与 GAT 相比，我们提供了更多关于 UniMP 的消融研究，显示了我们模型的超强性。

不同输入的图神经网络

在表 7 中，我们对各种 GNN 应用了屏蔽标签预测，以提高它们的性能。首先，我们重新实现了经典的 GNN 方法，如 GCN 和 GAT，采样方法和模型设置如表 3 所示。GCN 的隐藏大小为头部_num*hidden_size，因为它没有头部注意力。其次，我们改变了这些模型的不同输入，以研究 GCN 和 GAT 的有效性。

特征和标签传播，使用我们的**屏蔽标签预设词典，**以部分节点标签 $\mathbf{Y}^{\hat{}}$ 作为输入来训练模型。

表 7 中的第 4 行显示，只有在 $\mathbf{Y}$和 **A 的**情况下，GNN 在所有三个数据集中的表现仍然很好，优于只给定 **X 的** MLP 模型。比较表 7 中的第 3 行和第 5 行，包含 **X**、**A** 和 $\mathbf{Y}^{\hat{}}$**的**模型优于包含 **X** 和 **A 的**模型，这表明在半监督分类中，GNN 在不包含地面实况训练标签 $\mathbf{Y}^{\hat{}}$ 的情况下进行预测是一种信息浪费。表 7 中的第 3-5 行还表明，在不同的输入设置下，我们的图变换器的性能优于 GAT 和 GCN。

标签覆盖率与性能之间的关系

虽然我们已经验证了使用这种策略结合 LPA 和 GNN 的有效性，但 LPA 覆盖率与它对 GNN 性能的影响之间的关系仍不确定。因此，如图 2 所示，我们在 *ogbn-arxiv* 中进行了更多实验，以研究它们在以下不同情况下的关系：

- 在图 2a 中，我们使用 **X**、$\mathbf{Y}^{\hat{}}$、**A** 作为输入来训练 UniMP。我们调整了输入标签_率，这是屏蔽标签预测任务的超参数，并显示了验证和测试的准确率。当标签_率约为 0.625 时，我们的模型能取得更好的性能。

- 图 2b 描述了训练数据比例与标签生成效果之间的相关性。我们将输入标签率固定为 0.625。唯一的变化是训练数据的比例。按照常理，随着训练数据量的增加，性能会逐渐提高。而采用标签传播 $\hat{Y}$ 的模型可以从以下方面获得更大的收益
增加标注数据的比例。
- 我们的统一模型总是掩盖部分训练标签，并试图恢复它们。但在推理阶段，我们的模型会利用所有训练标签进行预测，这与训练阶段略有不同。在图 2c 中，我们在训练阶段将输入标签率固定为 0.625，并在推断阶段执行不同的输入标签率。在训练阶段，我们发现在预测阶段降低标签率时，UniMP 的性能（低于 0.70）可能比基线（约 0.72）更差。但是，当标签率上升时，性能可以提高到 0.73。
- 在图 2d 中，我们计算了按邻居数量分组的未标记节点的准确率。前
实验结果表明，拥有更多邻居的节点具有更高的准确率。而且，即使训练邻居的数量不同，带有标签预估 $\hat{Y}$ 的模型也总能有所改进。

测量节点之间的连接

在图 3 中，我们分析了 LPA 如何影响 GNN，使其表现更好。Wang [2019] 曾指出，在训练过程中为 GCN 使用 LPA 可以使同一节点内的
类/标签的连接更紧密，从而提高了模型预测的准确性（ACC）。我们的模型可以看作是它们的升级版，在我们的图形转换器的训练和测试中都使用了 LPA。因此，我们尝试基于我们的模型对上述想法进行实验验证。

$$MSF = \frac{1}{N}\sum_{i=1}^{N} \log 1 + \sum_{j\in N(i)_{pos}} \sum_{k\in N(i)_{neg}} e^{\alpha i_{j,k}} - e \qquad (11)$$

我们使用等式 11 所示的边际相似度函数 (MSF) 来反映同一类节点之间的连接紧密度（分数越高，连接越紧密）。我们在 *ogbn-arxiv* 上进行了实验。如图 3 所示，模型预测的 ACC 与边际相似度成正比。统一特征和边际传播可以进一步加强它们之间的联系，从而提高它们的 ACC。此外，在不同的输入条件下，我们的 Graph Transformer 在连接紧密度和 ACC 方面都优于 GAT。

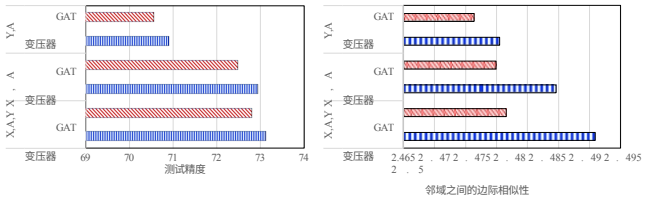

图 3：精确度与邻居之间的边距相似性之间的相关性。

节点之间的相互作用更多。此外，残差和门控残差也能加强浅层的 GNN。此外，我们的统一模型可以将额外的估值标签作为输入，进一步提高模型的性能，而无需更多的训练步骤。因此，当我们将模型应用于真实场景，并逐步积累标签数据时，无需从头开始训练我们的模型，非标签数据的准确率就能不断提高，而其他没有明确标签建模的 GNN 则无法充分利用附加标签的优势。

模型	OGBN-ARXIV	OGBN-PRDOUCTO
GAT（关注总和）	0.8002	0.7246
1 w/ 剩余	0.8033	0.7265
1 带门控余量 变压器（点积）	0.8050	0.7272
1 带余量	0.8091	0.7259
1 w/ gated residual	0.8125	0.7271
1 带列车标签（UniMP）	0.8137	0.7292
1 w/ 验证标签	0.8256	0.7311
	0.8312	**0.7377**

表 8：UniMP 与 GAT 的消融研究比较

5 结论

我们首先提出了一种统一的消息传递模型 UniMP，该模型可联合执行特征传播和标签传播。
在图形转换器中进行分析，从而实现半监督分类。此外，我们还提出了一种屏蔽
标签预测法对我们的模型进行监督训练，预

关于 UniMP 的更多消融研究

最后，我们从以下 4 个方面对 UniMP 模型与 GAT 进行了更多的消减研究：（1）带有点积注意力的香草变换器或带有总和注意力的 GAT；（2）简单残差或门控残差；（3）以训练标签为输入；（4）以训练标签和验证标签为输入。如表 8 所示，我们可以发现点积注意力优于总和注意力，因为点积提供了

以避免自循环标签信息的过度拟合。实验结果表明，UniMP 在三个主要 OGBN 数据集（*ogbn-products*、*ogbn-proteins* 和 *ogbn-arxiv*）上的表现远远优于之前的一流模型，消融研究证明了统一特征传播和标签传播的有效性。

参考资料

[Battaglia *et al.*, 2018] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *ArXiv preprint arXiv:1806.01261*, 2018.

[陈明等，2020] 陈明、魏哲伟、黄增锋、丁博林、李亚良。 简单和深度图卷积网络。*ArXiv 预印本 arXiv:2007.02133*, 2020.

[Chiang *et al.*, 2019] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh.Cluster-gcn：训练深度和大型图卷积网络的高效算法。In *SIGKDD*, pages 257-266, 2019.

[Devlin *等人*，2018] Jacob Devlin、Ming-Wei Chang、Ken-ton Lee 和 Kristina Toutanova。Bert: Pre-training of deep bidirectional transformers for language understand- ing. *arXiv preprint arXiv:1810.04805*, 2018.

[Duvenaud *et al.*, 2015] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Ala´n Aspuru-Guzik, and Ryan P Adams.用于学习分子指纹的图上卷积网络。In *NIPS*, pages 2224-2232, 2015.

[Gong *等人*，2016] Chen Gong、Dacheng Tao、Wei Liu、Liu Liu 和 Jie Yang。通过 "教到学 "和 "学到教 "进行标签传播。*IEEE TNNLS*，28（6）：1452-1465，2016.

[Grover and Leskovec, 2016] Aditya Grover and Jure Leskovec. node2vec：可扩展的网络特征学习。In *SIGKDD*, pages 855-864, 2016.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec.大型图上的归纳表示学习。In *NIPS*, pages 1024-1034, 2017.

[Hu *et al.*, 2020] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec.开放图基准：*ArXiv preprint arXiv:2005.00687*, 2020.

[Karasuyama and Mamitsuka, 2013] Masayuki Karasuyama and Hiroshi Mamitsuka.基于 Manifold 的标签传播相似性调整。In *NIPS*, pages 1547-1555, 2013.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. 用图卷积网络进行半监督分类》，*arXiv preprint arXiv:1609.02907*, 2016.

[克利*珀拉等人*，2018] 约翰内斯-克利珀拉、亚历山大-博-切夫斯基、斯蒂芬-古奈曼。Predict then propagate：*ArXiv preprint arXiv:1810.05997*, 2018.

[Li *et al.*, 2018] Qimai Li, Zhichao Han, and Xiaoming Wu. 半监督学习的图卷积网络深度解读。*神经计算*》，第 3538-3545 页，2018 年。

[李国豪、马蒂亚斯-穆勒、阿里-塔贝特、伯纳德-加内姆。Deepgcns：gcns 能否像 cnns 一样深入？In *ICCV*, pages 9267-9276, 2019.

[李国豪、熊晨鑫、Ali Thabet 和 Bernard Ghanem。Deepergcn：训练更深层次 GCN 所需的一切。*ArXiv 预印本 arXiv:2006.07739*, 2020。

[廖仁杰 等，2019] 廖仁杰、赵志珍、Raquel Ur- tasun 和 Richard S Zemel 。 Lanczosnet： *ArXiv preprint arXiv:1901.01484*, 2019.

[Liu 等人，2019] 刘彦斌、Juho Lee、Minseop Park、Sae- hoon Kim、Eunho Yang、Sung Ju Hwang 和 Yi Yang.学习传播标签：用于少量学习的传导式传播网络。*ArXiv：Learning*，2019。

[Liu 等人，2020a] Meng Liu、Hongyang Gao 和 Shuiwang Ji.走向更深入的图神经网络。在 *SIGKDD* 中，第 338-348 页，2020 年。

[Liu *et al.*, 2020b] Ziqi Liu, Zhengwei Wu, Zhiqiang Zhang, Jun Zhou, Shuang Yang, Le Song, and Yuan Qi. 用于训练图神经网络的Ban-dit采样器。*arXiv预印本 arXiv:2006.05806*，2020.

[Niepert 等人，2016 年] Mathias Niepert、Mohamed Ahmed 和 Konstantin Kutzkov。学习图的卷积神经网络。In *ICML*, pages 2014-2023, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.注意力就是你所需要的一切。In *NIPS*, pages 5998-6008, 2017.

[Velicˇkovic´ *et al.*, 2017] Petar Velicˇkovic´, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio.Graph attention networks. *ArXiv preprint arXiv:1710.10903*, 2017.

[Wang and Leskovec, 2019] 王宏伟和尤雷-莱斯科维奇。统一图卷积神经网络和标签传播》，*arXiv：Learning*，2019.

[Wang and Zhang, 2007] Fei Wang and Changshui Zhang. 通过线性邻域的标签传播。*IEEE TKDE*，20（1）：55-67，2007.

[Xu *et al.*, 2018a] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka.图神经网络有多强大？ *arXiv preprint arXiv:1810.00826*, 2018.

[Xu *et al.*, 2018b] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka.带有跳跃知识网络的图上表征学习。In *ICML*, pages 5453-5462, 2018.

[杨志林、William Cohen 和 Ruslan Salakhudinov.用图嵌入重新审视半监督学习。In *ICML*, pages 40-48.PMLR, 2016.

[Zeng *et al.*, 2019] 曾翰清、周宏宽、Ajitesh Srivastava、Rajgopal Kannan 和 Viktor Prasanna.图圣：基于图采样的归纳学习方法。*arXiv preprint arXiv:1907.04931*, 2019.

[Zhang and Lee, 2007] Xinhua Zhang and Wee S Lee.基于图的半监督学习算法的参数学习。In *NIPS*, pages 1585-1592, 2007.

[Zhang *et al.*, 2018] Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit-Yan Yeung.Gaan：用于大型时空图学习的门控注意力网络。*ArXiv 预印本 arXiv:1803.07294*, 2018.

[Zhu *et al.*, 2003] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty.使用高斯场和谐函数的半监督学习。*ICML*, pages 912-919, 2003.