深度图表 INFOMAX

佩塔尔-韦利奇ˇ科维奇*
剑桥大学计算机科学与技术系
petar.velickovic@cst.cam.ac.uk

威廉-费杜斯
米拉 - 魁北克省人工智能研究所谷歌大脑
liamfedus@google.com

威廉-汉密尔顿
米拉 - 魁北克省人工智能研究所 麦吉尔大学
wlh@cs.mcgill.ca

彼得-利奥
剑桥大学计算机科学与技术系
pietro.lio@cst.cam.ac.uk

Yoshua Bengio†（本吉奥
Mila - 魁北克省人工智能研究所 蒙特里尔大学 yoshua.bengio@mila.quebec

R Devon Hjelm
微软研究院
米拉 - 魁北克省人工智能研究所
devon.hjelm@microsoft.com

摘要

我们提出了 *Deep Graph Infomax*（DGI），这是一种以无监督方式学习图结构数据中节点表示的通用方法。DGI 的关键在于最大化补丁表示与相应的图高层摘要之间的互信息--两者都是利用成熟的图卷积网络架构得出的。学习到的补丁表示法总结了以感兴趣的节点为中心的子图，因此可重新用于下流节点学习任务。与之前使用 GCN 进行非超视距学习的大多数方法相比，DGI 不依赖随机漫步目标，可轻松应用于传导式和归纳式学习设置。我们在各种节点分类基准上展示了极具竞争力的性能，有时甚至超过了监督学习的性能。

1 导言

将神经网络泛化到图结构输入是当前机器学习的主要挑战之一（Bronstein 等人，2017；Hamilton 等人，2017b；Battaglia 等人，2018）。虽然最近已经取得了长足的进步，特别是*在图卷积网络方面*（Kipf & Welling, 2016a；Gilmer 等人，2017；Velicˇkovic´ 等人，2018），但大多数成功的方法都使用了*监督学习*，这往往是不可能的，因为野生的大多数图数据都是无标记的。此外，从大规模图中发现新颖或有趣的结构往往是人们所希望的，因此，无监督图学习对于许多重要任务来说是必不可少的。

目前，对图结构数据进行无监督表示学习的主流算法依赖于基于随机游走的目标（Grover & Leskovec, 2016; Perozzi 等人, 2014; Tang 等人, 2015; Hamilton 等人, 2017a），有时会进一步简化以重建邻接信息（Kipf & Welling, 2016b; Duran & Niepert, 2017）。其基本原理是训

练编码器网络，使输入图中 "接近 "的节点在表示空间中也 "接近"。

随机漫步方法虽然功能强大，而且与个性化 PageRank 分数（Jeh & Widom，2003 年）等传统指标相关，但也存在已知的局限性。最突出的是，众所周知，随机漫步目标会过度强调邻近性信息而牺牲结构信息（Ribeiro 等人，2017 年），而且性能高度依赖于超参数的选择（Grover & Leskovec，2016 年；Perozzi 等人，2014 年）。此外，随着更强大的

*提交人在米拉工作期间完成的工作。
†CIFAR 研究员

基于图卷积的编码器模型（Gilmer 等人，2017 年），目前还不清楚随机漫步目标是否真的能提供任何有用的信号，因为这些编码器已经强制执行了一种归纳偏差，即相邻节点具有相似的表征。

在这项工作中，我们提出了一种基于*互信息*而非随机漫步的无监督图学习替代目标。最近，通过互信息神经估计（MINE，Bel- ghazi 等人，2018 年），互信息的可扩展估计变得既可能又实用，它依赖于训练一个*统计网络，将其*作为来自两个随机变量的联合分布及其边际乘积的样本的分类器。继 MINE 之后，Hjelm 等人（2018 年）又推出了用于学习高维数据表示的 Deep InfoMax（DIM）。DIM 训练编码器模型，使高级 "全局 "表示与输入的 "局部 "部分（如图像的补丁）之间的互信息最大化。这鼓励编码器携带存在于所有位置（因此与*全局相关*）的信息类型，例如类标签。

DIM 在很大程度上依赖于图像数据背景下的卷积神经网络结构，而据我们所知，还没有研究将互信息最大化应用于图结构输入。在这里，我们将 DIM 的思想应用到图领域，可以认为图领域的结构比卷积神经网络捕捉到的结构类型更普遍。在接下来的章节中，我们将介绍我们的方法，即 *Deep Graph Infomax*（DGI）。我们的实验证明，DGI 学习到的表示在转导和归纳分类任务中都具有持续的竞争力，常常优于有监督和无监督的强基线。

2 相关工作

对比方法。对表征进行无监督学习的一个重要方法是训练编码器，使其在能捕捉到相关统计依赖关系的表征和不能捕捉到相关统计依赖关系的表征之间形成*对比*。例如，对比方法可以使用*评分函数*，训练编码器在 "真实 "输入（又称正面示例）上提高评分，在 "虚假 "输入（又称负面示例）上降低评分。对比方法是许多流行的词嵌入方法的核心（Collobert & Weston, 2008; Mnih & Kavukcuoglu, 2013; Mikolov et al.对表示进行评分的方法有很多，但在图文献中，最常见的技术是使用分类（Perozzi 等人，2014；Grover & Leskovec，2016；Kipf & Welling，2016b；Hamilton 等人，2017b），不过也有使用其他评分函数的（Duran & Niepert，2017；Bojchevski & Gu¨nnemann，2018）。在这方面，DGI 也具有对比性，因为我们的目标是对局部-全局对和负采样对进行分类。

抽样策略。对比方法的一个关键实施细节是如何提取正样本和负样本。上文提到的无监督图表征学习工作依赖于局部对比损失（强制近似节点具有相似的嵌入）。正样本通常对应于在图中*短随机漫步中*一起出现的节点对--从语言建模的角度来看，这实际上是将节点视为*单词*，将随机漫步视为*句子*。Bojchevski & Gu¨nnemann （2018）的最新研究使用节点锚定采样作为替代方法。这些方法的否定抽样主要基于随机对的抽样，最近的工作将这种方法调整为使用基于课程的否定抽样方案（具有逐步 "接近 "的否定示例；Ying 等人，2018a）或引入对手来选择否定示例（Bose 等人，2018）。

预测编码。对比预测编码（CPC，Oord 等人，2018 年）是另一种基于互信息最大化的深度

表征学习方法。与上述模型一样，CPC 也是对比性的，在这种情况下，它使用条件密度的估计值（以噪声对比估计的形式，Gutmann & Hyva¨rinen, 2010）作为评分函数。然而，与我们的方法不同的是，CPC 和上述图形方法都是*预测性的*：对比目标有效地在输入的结构指定部分之间（例如，相邻节点对之间或节点与其邻域之间）训练预测器。我们的方法不同之处在于，我们同时对比图的全局/局部部分，其中全局变量由所有局部变量计算得出。

据我们所知，之前唯一关注图的 "全局 "和 "局部 "表征对比的著作是通过邻接矩阵的（自动）编码目标来实现的。

(Wang等人，2016）和将群落级约束纳入节点嵌入（Wang等人，2017）。这两种方法都依赖于矩阵因式分解式损失，因此无法扩展到更大的图。

3 DGI 方法

在本节中，我们将以自上而下的方式介绍 Deep Graph Infomax 方法：首先对我们特定的无监督学习设置进行抽象概述，然后阐述我们方法所优化的目标函数，最后列举我们在单图设置中的所有步骤。

3.1 基于图形的无监督学习

我们假设了一个基于图的通用无监督机器学习设置：我们获得了一组 *节点特征*，$\mathbf{X} = \{\vec{x}_1, \vec{x}_2, ...X = \{\vec{x}_N, \vec{x}^F, ... \vec{x}\}$，其中 N 的节点数，$\vec{x}_i \in R$ 表示节点 i 的特征。我们还以 *邻接矩阵* $A \in R^{N\times N}$ 的形式获得了这些节点之间的关系信息。虽然 **A** 可以由任意实数（甚至任意边的特征）组成，但在我们的所有实验中，我们将假设图是*无权的，*即如果图中存在边 $i \to j$，*则* $A_{ij} = 1$，否则 $A_{ij} = 0$。

我们的目标是学习一个*编码器* $E: R^{N\times F} \times R^{N\times N} \to R^{N\times F'}$，这样 $E(\mathbf{X}, \mathbf{A}) = \mathbf{H} = \{\vec{h}_1, \vec{h}_2, ..., \vec{h}_N\}$ 表示高级表示 $\vec{h}_i \in R^{F'}$ 为每个节点 i。
然后可以检索这些数据，并将其用于节点分类等下游任务。

在这里，我们将重点讨论*图卷积*编码器--一类灵活的节点嵌入架构，它通过对局部节点邻域的重复聚合来生成节点表示（Gilmer 等人，2017 年）。一个关键的结果是，生成的节点嵌入 $\vec{h}_i$ *总结了*以节点 i *为*中心的图*片段，*而不仅仅是节点本身。在下文中，为了强调这一点，我们经常将 $\vec{h}_i$ 称为*补丁表示法。*

3.2 局部-全局互信息最大化

我们学习编码器的方法依赖于*局部互信息的最大化，*也就是说，我们寻求获得节点（即局部）表示，以捕捉整个图的全局信息内容，这些信息内容由*摘要向量* $\vec{s}$ 表示。

为了获得图级摘要向量 $\vec{s}$，我们利用一个*读出函数* $R: R^{N\times F} \to R^F$，并用它将获得的补丁表示总结为*图级表示*；即 $\vec{s} = R(E(\mathbf{X}, \mathbf{A}))$。

作为局部互信息最大化的替代方法，我们采用了一个*判别器* $D: R^F \times R^F \to R$，这样 $D(\vec{h}_i, \vec{s})$ 代表了分配给这一斑块-摘要对的概率分数（对于包含在摘要中的斑块来说应该更高）。

D 的负样本是通过将（**X**，**A**）中的摘要 $\vec{s}$ 与补丁代表（**X**，**A**）配对而得到的。

h_j 。在多图环境中，这种图可以作为训练集的其他元素获得。然而，对于单个图，需要一个明确的（随机）*破坏函数* $\mathcal{C} : \mathbb{R}^{N\times F} \times \mathbb{R}^{N\times N} \to \mathbb{R}^{M\times F} \times \mathbb{R}^{M\times M}$ 才能从以下图中获得一个反例 $\tilde{\mathbf{X}}, \tilde{\mathbf{A}}$

即 $(\mathbf{X}, \mathbf{A}) = \mathcal{C}(\mathbf{X}, \mathbf{A})$。负抽样程序的选择将决定作为最大化副产品的结构信息的具体类型。

在目标方面，我们遵循 Deep InfoMax（DIM，Hjelm 等人，2018 年）的直觉，使用噪声对比型目标，并在联合样本（正面例子）和边际乘积（负面例子）之间采用标准二元交叉熵（BCE）损失。以下内容

我们使用以下目标[1]：

$$L = \frac{1}{N+M} \sum_{i=1}^{N} E_{(\mathbf{x},\mathbf{A})} \log D(\vec{h}_i, \vec{s}) + \sum_{j=1}^{M} E_{(\tilde{\mathbf{x}}_j, \tilde{\mathbf{A}}_j)} \log\left(1 - D(\vec{\tilde{h}}_j, \vec{s})\right) \quad (1)$$

根据联合边际和乘积边际之间的詹森-香农发散[2]，这种方法能有效地最大化 $\vec{h}_i$ 和 $\vec{s}$ 之间的互信息。

由于所有派生的斑块表征都是为了保持与全局图摘要的互信息，这就允许在斑块层面发现并保持相似性--例如，具有相似结构作用的远处节点（众所周知，这对于许多节点分类任务来说都是一个强有力的预测因素；Donnat 等人，2018 年）。请注意，这是 Hjelm 等人（2018 年）提出的论点的 "反向 "版本：对于节点分类，我们的目标是让*补丁*与整个图中的相似补丁建立链接，而不是强制要求摘要包含所有这些相似性（不过，原则上这两种效果应该同时出现）。

3.3 理论动机

现在，我们提供一些直观的方法，将我们的判别器的分类误差与图表示的互信息最大化联系起来。

定理 1. *让* $\{\mathbf{X}^{(k)}\}_{k=1}^{|\mathbf{X}|}$ *是一组从经验概率分布中提取的节点表示。让* $\mathsf{R}(\cdot)$ *是图的确定性读出函数，* $\vec{s}^{(k)} = \mathsf{R}(\mathbf{X}^{(k)})$ *是第 k 个图的摘要向量，边际分布为* $p(\vec{s})$*。联合分布之间的最优分类器是* $p(\mathbf{X}, \vec{s})$ *与边际值* $p(\mathbf{X})p(\vec{s})$ *的乘积，假定类平衡，误差率为上界为* $\mathrm{Err}^* = \frac{1}{2}\sum_{k=1}^{|\mathbf{X}|} p(\vec{s}^{(k)})^2$*。如果* R *是注入式的，就可以达到这个上界。*

*证明。*用 $\mathcal{Q}^{(k)}$ 表示输入集中由 R 映射到 $\vec{s}^{(k)}$ 的所有图的集合，即 $\mathcal{Q}^{(k)} = \{\mathbf{X}^{(j)} \mid \mathsf{R}(\mathbf{X}^{(j)}) = \vec{s}^{(k)}\}$。由于 $\mathsf{R}(\cdot)$ 是确定的，从联合 $(\mathbf{X}^{(k)}, \vec{s}^{(k)})$ 中抽取样本的概率为 $p(\vec{s}^{(k)})p(\mathbf{X}^{(k)})$，分解为：

$$p(\vec{s}^{(k)}) \sum_{\mathbf{s}}^{\vec{s}} p(\mathbf{X}^{(k)}, \vec{s}) = p(\vec{s}^{(k)}) p(\mathbf{X}^{(k)} \mid \vec{s}^{(k)}) p(\vec{s}^{(k)}) = \frac{p(\mathbf{X}^{(k)})}{\sum_{\mathbf{X}' \in \mathcal{Q}^{(k)}} p(\mathbf{X}')} p(\vec{s}^{(k)})^2 \quad (2)$$

为方便起见，让 $\rho^{(k)} = \frac{p(\mathbf{X}^{(k)})}{\sum_{\mathbf{X}' \in \mathcal{Q}^{(k)}} p(\mathbf{X}')}$。根据定义，$\mathbf{X}^{(k)} \in \mathcal{Q}^{(k)}$，因此 $\rho^{(k)} \le 1$。

当 $\mathcal{Q}^{(k)} = \{\mathbf{X}^{(k)}\}$，即当 R 对 $\mathbf{X}^{(k)}$ 是注入式时，这个概率比值最大为 1。从边际乘积中抽取任何联合样本的概率都有界

上式中 $\sum_{k=1}^{|\mathbf{X}|} p(\vec{s}^{(k)})^2$。由于从联合中抽取 $(\mathbf{X}^{(k)}, \vec{s}^{(k)})$ 的概率为 $\rho^{(k)} p(\vec{s}^{(k)}) \ge \rho^{(k)} p(\vec{s}^{(k)})^2$，我们知道将这些样本归类为来自联合样本的误差较小，而不是把它们归类为来自边际积的样本。因此，这种分类器的误差率就是在下列条件下从联合样本中抽取样本作为边际乘积样本的概率

混合概率，我们可以用 $\mathrm{Err} \le \frac{1}{2}\sum_{k=1}^{|\mathbf{X}|} p(\vec{s}^{(k)})^2$ 对其进行约束，其上限为
如上所述，当 $\mathsf{R}(\cdot)$ 对 $\{\mathbf{X}^{(k)}\}$ 的所有元素都是注入式时，就实现了。 □

注意到 $\frac{1}{2} \le \mathrm{Err}^* \le \frac{1}{2|\mathbf{X}|}$ 可能是有用的。第一个结果是通过一个微不足道的应用得到的
而只有在恒定读出函数的边缘情况下，才会达到另一个极端（此时，来自联合的每个示例

也是来自边际乘积的示例，因此没有分类器的表现比偶然性更好）。

推论 1.*从现在起，假设所使用的读出函数* R *是注入式的。假设* $\rightarrow_s$ *空间中允许的状态数* $|\rightarrow_s|$ *大于或等于* |X|*。那么，对于* $\rightarrow_s^{\wedge}$ *，*

[1]请注意，Hjelm 等人（2018 年）使用的是二元交叉熵的软加版本。

[2]这里定义的 "GAN "距离--根据 Goodfellow 等人（2014 年）和 Nowozin 等人（2016 年）--与詹森-香农分歧的关系是 $_{DGAN} = {}_{2DJS}$ - log 4。因此，任何参数在优化一个参数的同时，也会优化另一个参数。

在联合边际和乘积边际之间的最优分类器的分类误差下的最优汇总，即 $|\rightarrow s^{\wedge}| = |\mathsf{X}|$。

*证明。*根据 R 的注入性，我们知道 $\rightarrow s^{\wedge} = \text{argmin}_{\rightarrow s}\, \text{Err}^*$。由于误差上界 Err* 是一个简单的几何和，我们知道当 $p(\rightarrow s^{(\mathrm{k})})$ 是均匀的时候，误差上界是最小的。由于 R(-) 是确定性的，这意味着每个潜在的摘要状态至少需要使用一次。结合条件 $|\rightarrow \mathsf{s}| \geq |\mathsf{X}|$，我们得出结论：最优值为 $|\rightarrow s^{\wedge}| = |\mathsf{X}|$。□

定理 1. $\rightarrow s^{\wedge} = \text{argmax}_{\rightarrow s}\, \text{MI}(\mathbf{X}; \rightarrow s)$，*其中* MI *是互信息。*

*证明*这是因为互信息在可逆反形式下是不变的。由于 $|\rightarrow s^{\wedge}| = |\mathsf{X}|$，而 R 是注入式的，因此它有一个反函数，$\mathsf{R}^{-1}$。由此可见，对于任意 $\rightarrow s$，$\text{MI}(X; \rightarrow s) \leq H(\mathbf{X}) = \text{MI}(\mathbf{X}; \mathbf{X}) = \text{MI}(X; \mathsf{R}(\mathbf{X})) = \text{MI}(X; \rightarrow s^{\wedge})$，其中 H 是熵。 □

定理 1 表明，对于有限的输入集和合适的确定性函数，最小化判别器中的分类误差可用于最大化输入和输出之间的互信息。然而，正如 Hjelm 等人（2018）的研究所示，仅凭这一目标还不足以学习有用的表征。与他们的研究一样，我们在全局摘要向量和局部高级表征之间进行判别。

定理 2. *设* $\mathbf{X}^{(\mathrm{k})} = \{\rightarrow x\}_{\mathrm{j} \in \mathrm{n}(\mathbf{X}^{(k)}, \mathrm{i})}$ *是第 k 个图中节点 i 的邻域，即集合映射到其高级特征，即* $\rightarrow h_{\mathrm{i}} = \mathsf{E}(\mathbf{X}^{(\mathrm{k})})$，*其中* n_{i} *是邻域函数，返回图* $\mathbf{X}^{(\mathrm{k})}$ *中节点 i 的邻域索引集，*E *是确定性编码器函数。假设* $|\mathbf{X}_{\mathrm{i}}| = |\mathsf{X}| = |\rightarrow \mathsf{s}| \geq |\rightarrow h_{\mathrm{i}}|$。*那么，最小化分类的* $p(\rightarrow h_{\mathrm{i}}, \rightarrow s)$ *与* $p(\rightarrow h_{\mathrm{i}})p(\rightarrow s)$ *之间的误差也会使* $\text{MI}(\mathbf{X}^{(\mathrm{k})}_{\mathrm{i}}; \rightarrow h_{\mathrm{i}})$ *最大化。*

*证明。*鉴于我们假设 $|\mathbf{X}_{\mathrm{i}}| = |\rightarrow \mathsf{s}|$，存在一个逆 $\mathbf{X}_{\mathrm{i}} = \mathsf{R}^{-1}(\rightarrow s)$，因此 $\rightarrow h_{\mathrm{i}} = \mathsf{E}(\mathsf{R}^{-1}(\rightarrow s))$，即存在一个将 $\rightarrow s$ 映射到 $\rightarrow h_{\mathrm{i}}$ 的确定性函数 $(\mathsf{E} \circ \mathsf{R}^{-1})$。那么，联合 $p(\rightarrow h_{\mathrm{i}}, \rightarrow s)$ 与边际值 $p(\rightarrow h_{\mathrm{i}})p(\rightarrow s)$ 的乘积之间的最优分类器有（根据 Lemma 1）错误率上限为 $\text{Err}^* = \frac{1}{2}\sum^{|\mathsf{X}|} p(\rightarrow h)^{(\mathrm{k})}$。因此（如推论 1 所示），对于最优的 $\rightarrow h_{\mathrm{i}}$，$|\rightarrow h_{\mathrm{i}}| = |\mathbf{X}_{\mathrm{i}}|$，根据与定理 1 相同的论证，互信息最大化邻域特征与高层特征之间的关系，即 $\text{MI}(\mathbf{X}^{(\mathrm{k})}; \rightarrow h)_{\mathrm{i}}$。 □

这就促使我们使用联合样本和边际乘积样本之间的分类器，而使用二元交叉熵（BCE）损失来优化该分类器在神经网络优化中已广为人知。

3.4 总干事办公室概况

假设是单图设置（即输入为 $(\mathbf{X}, \mathbf{A})$），我们现在总结一下深度图 Infomax 程序的步骤：

1. 使用腐败函数：$(\mathbf{X}, \mathbf{A}) \sim \mathsf{C}(\mathbf{X}, \mathbf{A})$，对负面示例进行采样。
2. 通过编码器获取输入图形的补丁表示 $\rightarrow h_{\mathrm{i}}$：
   $\mathbf{H} = \mathsf{E}(\mathbf{X}, \mathbf{A}) = \{\rightarrow h_1, \rightarrow h_2, \ldots, \rightarrow h_{\mathrm{N}}\}$。

3.通过编码器获取负示例的补丁表示，$\grave{h}_j$ ：

$\mathbf{H} = \mathsf{E}(\mathbf{X}, \tilde{\mathbf{A}}) = \{\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_M\}$。

4. 通过读出函数对输入图形的补丁表示进行汇总：$\rightarrow s = \mathsf{R}(\mathbf{H})$。

5. 通过梯度下降法更新 E、R 和 D 的参数，使公式 1 最大化。图 1 全面总结了这一算法。

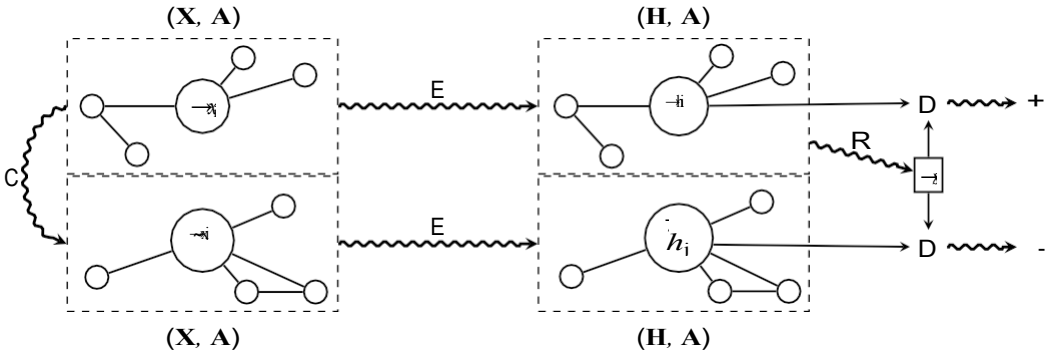

图 1：Deep Graph Infomax 高级概览。详情请参见第 3.4 节。

表 1：实验中使用的数据集摘要。

数据集	任务	节点	边缘	特点	班级	训练/评估/测试节点
科拉	传导式	2,708	5,429	1,433	7	140/500/1,000
Citeseer	传导式	3,327	4,732	3,703	6	120/500/1,000
发表于	传导式	19,717	44,338	500	3	60/500/1,000
Reddit	感应式	231,443	11,606,919	602	41	151,708/23,699/55,334
PPI	感应式	56,944 (24 幅图表)	818,716	50	121 (多栏)	44,906/6,514/5,524 (20/2/2 图表)

4 分类性能

我们评估了 DGI 编码器在各种节点分类任务（传导式和归纳式）中学习到的表示法的优势，并获得了有竞争力的结果。在每种情况下，我们都使用 DGI 以完全无监督的方式学习补丁表示，然后评估这些表示的节点级分类效用。具体方法是直接使用这些表征来训练和测试简单的线性（逻辑回归）分类器。

4.1 数据集

我们按照 Kipf & Welling (2016a) 和 Hamilton 等人 (2017a) 所描述的实验设置完成了以下基准任务：(1) 在 Cora、Cite-seer 和 Pubmed 引用网络中将研究论文分类为主题（Sen 等人，2008 年）；(2) 预测以 Reddit 帖子为模型的社交网络的社区结构；(3) 在蛋白质-蛋白质相互作用（PPI）网络中对蛋白质角色进行分类（Zitnik & Leskovec，2017 年），这需要对未见过的网络进行泛化。

有关数据集的更多信息，请参见表 1 和附录 A。

4.2 实验装置

在三种实验环境（转导学习、大型图形的归纳学习和多图形）中，我们分别采用了不同的编码器和适合该环境的腐败函数（如下所述）。

迁移学习。对于归纳学习任务（Cora、Citeseer 和 Pubmed），我们的编码器是一个单层图卷积网络（GCN）模型（Kipf & Welling, 2016a），传播规则如下：

$$E(\mathbf{X}, \mathbf{A}) = \sigma\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\boldsymbol{\Theta}\right) \tag{3}$$

其中，$\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ 是插入自循环的邻接矩阵，$\hat{\mathbf{D}}$是其相应的度矩阵，即 $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$。对于非线性系数 σ，我们采用了参数 ReLU

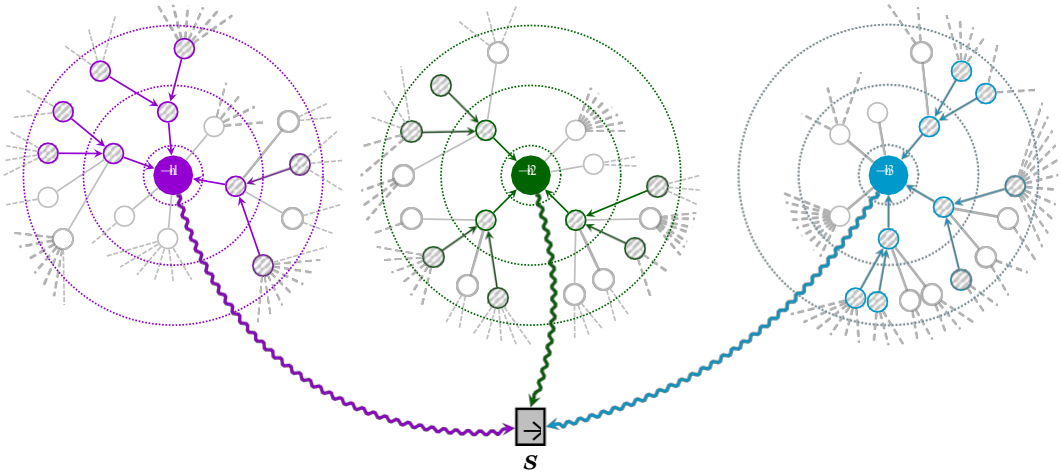

图 2：大型图（如 Reddit）上的 DGI 设置。摘要向量 $\rightarrow s$ 是通过组合多个子采样斑块表示 $\rightarrow h_i$ （此处分别通过对第一层和第二层的三个和两个邻居进行采样获得）而得到的。

(PReLU）函数（He 等人，2015 年），$\mathbf{\Theta} \in \mathbb{R}^{F \times F'}$ 是应用于每个节点的可学习线性变换，计算的 $F' = 512$ 个特征（由于内存限制，Pubmed 上的 $F' = 256$ 个）。

在这种情况下使用的破坏函数旨在鼓励表示法对图中不同节点的结构相似性进行适当编码；为此，C 保留了
原始邻接矩阵（$\widetilde{\mathbf{A}} \eqqcolon \mathbf{A}$），而被破坏的特征 $\widetilde{\mathbf{X}}$ 是通过行向
也就是说，被破坏的图由与原始图完全相同的节点组成，但这些节点位于图中的不同位置，因此将获得不同的补丁表示。我们在附录 C 中证明了 DGI 对其他损坏函数选择的稳定性，但我们发现那些保留图结构的函数会产生最强的特征。

大型图上的归纳学习对于归纳学习，我们可能不再在编码器中使用 GCN 更新规则（因为学习到的过滤器依赖于固定的已知邻接矩阵）；取而代之的是，我们应用 GraphSAGE-GCN 所使用的*均值池*传播规则（Hamilton 等人，2017a）：

$$\mathrm{mp}(\mathbf{x}, \mathbf{a}) = \mathbf{d}\hat{}\mathbf{-1}\,\mathbf{a}\hat{}\,\mathbf{x}\,\mathbf{\theta} \tag{4}$$

参数定义如公式 3。请注意，乘以 $\mathbf{D}\hat{}\mathbf{-1}$ 实际上是执行归一化和（因此是均值池化）。虽然等式 4 明确指定了邻接矩阵和程度矩阵，但*它们并不是必需的*：通过对节点邻居的*持续*关注机制可以观察到相同的归纳行为，正如 Const-GAT 模型所使用的那样（Velic˘kovic´ et al. ）

对于 Reddit，我们的编码器是一个具有跳接连接的三层均值池模型（He 等人，2016 年）：

$$\widehat{\mathrm{MP}}(\mathbf{X}, \mathbf{A}) = \sigma\left(\mathbf{X}\mathbf{\Theta} \,\|\, \mathrm{MP}(\mathbf{X}, \mathbf{A})\right) \qquad \mathrm{E}(\mathbf{X}, \mathbf{A}) = \widehat{\mathrm{MP}}_3\left(\widehat{\mathrm{MP}}_2\left(\widehat{\mathrm{MP}}_1\left(\mathbf{X}, \mathbf{A}\right), \mathbf{A}\right), \mathbf{A}\right) \tag{5}$$

其中$\|$为特征串联（即中心节点及其邻近节点分别处理）。我们在每个 MP 层计算 $F' = 512$ 个特征，对 σ *采用* PReLU 激活。

由于数据集规模庞大，无法完全容纳在 GPU 内存中。因此，我们采用了 Hamilton 等人（2017a）的子采样方法，即首先选取一小批节点，*然后通过对节点邻域进行替换采样，*得到

以每个节点为中心的子图。具体来说，我们在第一、第二和第三层分别抽取 10、10 和 25 个邻域，因此每个子抽样斑块有 1 + 10 + 100 + 2500 = 2611 个节点。我们只对中心节点 i 的补丁表示 $\rightarrow h_i$ 进行必要的计算。然后使用这些表示法得出迷你批的摘要向量 $\rightarrow s$（图 2）。我们在整个训练过程中使用了 256 个节点的迷你批。

为了在这种情况下定义破坏函数，我们采用了与转导任务类似的方法，但将每个子采样片段视为一个单独的待破坏图（即，我们对每一个子采样片段进行逐行洗牌）。

的特征矩阵）。需要注意的是，这很可能会导致中心节点的特征被替换为采样邻居的特征，从而进一步促进负样本的多样性。然后，中央节点获得的片段表示将提交给判别器。

多图归纳学习。对于 PPI 数据集，受之前成功的超级可视化架构（Velic˘kovic´ 等人，2018 年）的启发，我们的编码器是一个具有密集跳转连接的三层均值池模型（He 等人，2016 年；Huang 等人，2017 年）：

$$\mathbf{H}_1 = \sigma\left(\mathrm{MP}_1\left(\mathbf{X}, \mathbf{A}\right)\right) \tag{6}$$

$$\mathbf{H}_2 = \sigma\left(\mathrm{MP}_2\left(\mathbf{H}_1 + \mathbf{X}\mathbf{W}_{\text{skip}}, \mathbf{A}\right)\right) \tag{7}$$

$$\mathsf{E}(\mathbf{X}, \mathbf{A}) = \sigma\left(\mathrm{MP}_3\left(\mathbf{H}_2 + \mathbf{H}_1 + \mathbf{X}\mathbf{W}_{\text{skip}}, \mathbf{A}\right)\right) \tag{8}$$

其中，$\mathbf{W}_{\text{skip}}$ 是可学习的投影矩阵，MP 如公式 4 所定义。我们计算 $F' = 512$ 个特征，每个 MP 层使用 PReLU 激活σ。

在这种多图设置中，我们选择使用*随机采样的训练图*作为负样本（即我们的破坏函数只是从训练集中采样不同的图）。考虑到该数据集中有超过 40% 的节点特征为零，我们发现这种方法最为稳定。为了进一步扩大负面示例池，我们还对采样图的输入特征应用了 dropout（Srivastava 等人，2014 年）。我们发现，在向逻辑回归模型提供所学嵌入之前，对整个训练集进行标准化是有益的。

读出、判别和其他训练细节。在所有三种实验设置中，我们采用了相同的读出功能和判别器结构。

对于读出功能，我们使用的是所有节点特征的简单平均值：

$$\mathsf{R}(\mathbf{H}) = \sigma\left(\frac{1}{N}\sum_{i=1}^{N}\vec{h}_i\right) \tag{9}$$

其中 σ 是对数 sigmoid 非线性。虽然我们发现这种读出方式在所有实验中表现最佳，但我们认为它的威力会随着图大小的增加而减弱，在这种情况下，更复杂的读出架构，如 set2vec（Vinyals 等，2015 年）或 DiffPool（Ying 等，2018 年 b）可能更合适。

判别器通过应用简单的双线性评分函数（类似于 Oord 等人（2018 年）使用的评分方法）对摘要-补丁表示对进行评分：

$$\mathsf{D}(\vec{h}_i, \vec{s}) = \sigma\left(\vec{h}_i^{\mathrm{T}} W \vec{s}\right) \tag{10}$$

这里，$\mathbf{W}$ 是一个可学习的评分矩阵，σ 是 logistic sigmoid 非线性，用于将评分转换为 $(\vec{h}_i, \vec{s})$ 成为正面示例的概率。

所有模型均使用 Glorot 初始化（Glorot & Bengio，2010 年），并使用 Adam SGD 优化器（Kingma & Ba，2014 年）在可用节点上（转导式数据集为所有节点，归纳式数据集仅为训练节点）进行训练，以最大化等式 1 中提供的互信息，初始学习率为 0.001（特指 10^{-5} on Reddit）。在转发型数据集上，我们对观察到的*训练*损失采用早期停止策略，耐心等待 20 个历时[3]。在归纳数据集上，我们使用固定数量的历时进行训练（Reddit 为 150 个历时，

PPI 为 20 个历时）。

4.3 成果

表 2 总结了我们的对比评估实验结果。

对于转导任务，我们报告了我们的方法在经过 50 次运行训练（之后是逻辑回归）后在测试节点上的平均分类准确率（含标准偏差），并重新使用了 Kipf & Welling（2016a）中已报告的指标，以衡量 DeepWalk 和 GCN 以及标签传播（Label Propagation，LP）（Zhu 等人，2003 年）和 Planetoid（Yang 等人，2016 年）--一种代表性的监督随机行走方法--的性能。特别值得一提的是，我们提供了在原始输入特征上训练逻辑回归的结果，以及在输入特征串联后训练 DeepWalk 的结果。

[3]DGI 的参考实现可在 `https://github.com/PetarV-/DGI` 上找到。

表 2：分类准确率（在转导任务中）或微平均 F_1 分数（在归纳任务中）的结果汇总。在第一列中，我们强调了每种方法在训练过程中可用的数据类型（**X**：特征，**A**：邻接矩阵，**Y**：标签）。"GCN "对应于以监督方式训练的双层 DGI 编码器。

传导式

现有数据	方法	科拉	Citeseer	发表于
X	原始功能	47.9 ± 0.4%	49.3 ± 0.2%	69.1 ± 0.3%
A, **Y**	LP （Zhu 等人，2003 年）	68.0%	45.3%	63.0%
A	DeepWalk （Perozzi 等人，2014 年）	67.2%	43.2%	65.3%
X、**A**	DeepWalk + 功能	70.7 ± 0.6%	51.4 ± 0.5%	74.3 ± 0.9%
X、**A**	随机启动（我们的）	69.3 ± 1.4%	61.9 ± 1.6%	69.6 ± 1.9%
X、**A**	**DGI** （我们的）	**82.3** ± 0.6%	**71.8** ± 0.7%	**76.8** ± 0.6%
X、**A**、**Y**	GCN （Kipf 和 Welling，2016a）	81.5%	70.3%	79.0%
X、**A**、**Y**	平面类（Yang 等人，2016 年）	75.7%	64.7%	77.2%

感应式

现有数据	方法	Reddit	公众宣传局
X	原始功能	0.585	0.422
A	DeepWalk （Perozzi 等人，2014 年）	0.324	-
X, **A**	DeepWalk + 功能	0.691	-
X, **A**	GraphSAGE-GCN （Hamilton 等人，2017a）	0.908	0.465
X, **A**	GraphSAGE-mean （Hamilton 等人，2017a）	0.897	0.486
X, **A**	GraphSAGE-LSTM （Hamilton 等人，2017a）	0.907	0.482
X, **A**	GraphSAGE 池（Hamilton 等人，2017a）	0.892	0.502
X, **A**	随机启动（我们的）	0.933 ± 0.001	0.626 ± 0.002
X, **A**	**DGI** （我们的）	**0.940** ± 0.001	**0.638** ± 0.002
X、**A**、**Y**	FastGCN (Chen 等人，2018)	0.937	-
X、**A**、**Y**	平均集合（Zhang 等人，2018 年）	0.958 ± 0.001	0.969 ± 0.002

对于归纳任务，我们报告了（未见的）测试节点上的微平均 F_1 分数，这是训练 50 次运行后的平均值，并重新使用了 Hamilton 等人（2017a）为其他技术报告的指标。具体来说，由于我们的设置是无监督的，因此我们将与无监督的 GraphSAGE 方法进行比较。我们还提供了两个相关架构--FastGCN（Chen 等人，2018 年）和 Avg. pooling（Zhang 等人，2018 年）的监督结果。

我们的结果表明，在所有五个数据集上都取得了很好的性能。我们特别注意到，DGI 方法*与采用监督损失的* GCN 模型所报告的结果相比具有竞争力，在 Cora 和 Citeseer 数据集上甚至超过了其性能。我们认为这些优势源于这样一个事实，即 DGI 方法允许每个节点间接访问整个图的结构属性，而有监督的 GCN 则仅限于两层邻域（由于训练信号的极端稀疏性

和相应的过拟合威胁）。值得注意的是，虽然我们能够超越同等的监督编码器架构，但我们的性能仍然没有超越目前的监督转导技术（由 GraphSGAN（丁等人，2018 年）等方法保持）。我们进一步发现，在 Reddit 和 PPI 数据集上，DGI 方法成功超越了所有与之竞争的无监督 GraphSAGE 方法，从而验证了基于局部互信息最大化的方法在归纳节点分类领域的潜力。我们在 Reddit 数据集上的结果与有监督的技术水平相比具有竞争力，而在 PPI 数据集上的差距仍然很大--我们认为这归因于可用节点特征的极度稀疏性（超过 40% 的节点特征为零），而我们的编码器在很大程度上依赖于这种稀疏性。

我们注意到，*随机初始化的*图卷积网络可能已经提取出非常有用的特征，并代表了一个强大的基线--这是众所周知的事实，考虑到它与 Weisfeiler-

图 3：根据原始特征（**左**）、随机初始化的 DGI 模型特征（**中**）和学习的 DGI 模型特征（**右**）对 Cora 数据集中的节点进行 t-SNE 嵌入。学习的 DGI 模型嵌入的聚类非常清晰，Silhouette 得分为 0.234。

Kipf & Welling（2016a）和 Hamilton 等人（2017a）已经强调并分析了雷曼图同构测试（Weisfeiler & Lehman，1968）。因此，我们还提供了随机初始化编码器获得的嵌入式逻辑回归性能（*Random-Init*）。除了证明 DGI 能够在这一强大基线的基础上进一步提高之外，它还特别揭示出，在归纳数据集上，以前基于随机游走的负采样方法可能无法有效地学习分类任务所需的适当特征。

最后，需要指出的是，较深的编码器对应的是恢复的补丁表示之间更明显的*混合*，从而降低了正/负示例库的有效可变性。我们认为，这就是较浅的架构在某些数据集上表现更好的原因。虽然我们不能说这些趋势在一般情况下都会成立，但在使用 DGI 损失函数时，我们普遍发现采用*更宽*而非*更深的*模型更有优势。

5 定性分析

为了更好地理解 DGI 的特性，我们对 DGI 算法学习到的嵌入进行了一系列不同的分析。我们的分析完全集中在 Cora 数据集上（因为它的节点数量最少，大大提高了分析的清晰度）。

图 3 给出了嵌入式的一组标准 "演化 "t-SNE 图（Maaten 和 Hinton，2008 年）。正如定量结果所预期的那样，学习的嵌入式二维投影在二维投影空间（尤其是与原始特征和 Random-Init 相比）中产生了明显的聚类，这符合 Cora 的七个主题类别。该投影获得的 Silhouette 分数（Rousseeuw，1987 年）为 0.234，与之前报告的 0.234 分数相比毫不逊色。
嵌入式传播为 0.158（Duran 和 Niepert，2017 年）。

我们进行了进一步的分析，揭示了 DGI 的学习机制，分离出有*偏差的*em-bedding维度来降低负面示例的分数，并利用其余维度来编码关于正面示例的有用信息。我们利用这些洞察力，即使从编码器提供的补丁代表中移除*一半*维度，也能保持与有监督 GCN 的竞争性能。这些以及其他一些定性和消减研究见附录 B。

6 结论

我们提出了 Deep Graph Infomax (DGI)，这是一种在图结构数据上学习无监督表示的新方法。通过利用强大的图卷积架构获得的图补丁表示的局部互信息最大化，我们能够获得考虑到图的全局结构属性的节点嵌入。这使得我们在各种转导式和归纳式分类任务中都能获得极具竞争力的性能，有时甚至优于相关的*监督*式架构。