

# 深度图谱对比性表征学习

朱彦桥<sup>1,2</sup> 徐一辰<sup>3</sup>† 于峰<sup>1,2</sup> 刘强<sup>4,5</sup> 吴姝<sup>1,2</sup> 王亮<sup>1,2</sup>

<sup>1</sup>中国科学院自动化研究所智能感知与计算研究中心

<sup>2</sup>中国科学院大学人工智能学院

<sup>3</sup>北京邮电大学计算机科学学院

<sup>4</sup>RealAI <sup>5</sup>清华大学

yanqiao.zhu@cripac.ia.ac.cn, linyxus@bupt.edu.cn

{feng.yu, shu.wu, wangliang}@nlpr.ia.ac.cn, qiang.liu@realai.ai

## 摘要

如今，图表示学习已经成为分析图结构数据的基础。受最近成功的对比方法的启发，在本文中，我们提出了一个新的框架，通过利用节点层面的对比目标进行无监督的图表示学习。具体来说，我们通过腐败产生两个图形视图，并通过最大化这两个视图中的节点表征的一致性来学习节点表征。为了给对比性目标提供不同的节点背景，我们提出了一个混合方案，在结构和属性层面上生成图的视图。此外，我们从相互信息和经典的三联体损失两个角度提供了我们动机背后的理论依据。我们使用各种真实世界的数据集对归纳和归纳学习任务进行了实证实验。实验表明，尽管我们提出的方法很简单，但它始终以很大的幅度超过了现有的最先进的方法。此外，我们的无监督方法甚至在归纳任务上超过了有监督的同类方法，显示了其在现实世界应用中的巨大潜力。

## 1 简介

在过去的几年里，图表示学习已经成为分析图结构数据的一个强大策略。图表示学习的目的是学习一个编码函数，将节点转换为低维密集嵌入，保留图的属性和结构特征。传统的无监督图表示学习方法，如DeepWalk[1]和node2vec[2]，遵循起源于skip-gram模型[3]的对比框架。具体来说，它们首先对短的随机行走进行采样，然后通过与其他节点的对比，强制同一行走上的相邻节点共享相似的嵌入。然而，基于DeepWalk的方法可以被视为重建图的接近矩阵，如高阶相邻矩阵[4]，它过分强调了定义在网络结构上的接近信息[5]。

最近，使用图形神经网络（GNN）的图形表示学习受到了极大的关注。然而，伴随着它的蓬勃发展，人们对训练模型时的标签可用性越来越关注。然而，现有的GNN模型大多是以监督的方式建立的[6-8]，它需要大量的标签节点进行训练。尽管有一些尝试将以前的无监督目标（即矩阵重建）连接到GNN模型[9, 10]，但这些方法仍然严重依赖预设的图接近矩阵。

†前两位作者对这项工作的贡献相同。

†这项工作是在CRIPAC, CASIA实习期间完成的。

预印本。正在审查中。

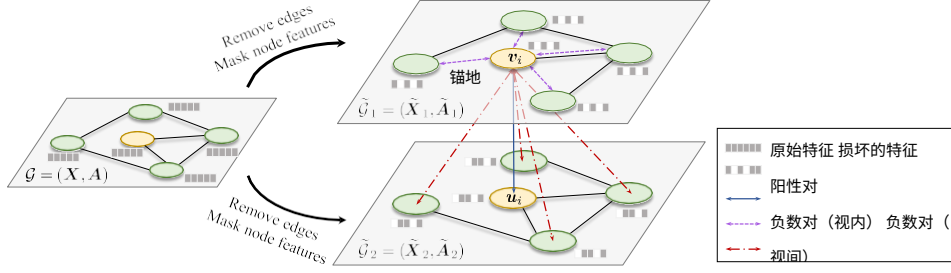


图1：我们提出的深度GRaPh Contrastive rEpresentation学习（GRACE）模型。

视觉表征学习不是优化重建目标，而是导致了经典的信息最大化（InfoMax）原则的复兴[11]。到目前为止，已经提出了一系列的对比学习方法[12-17]，这些方法通过将正向对与负向采样的对应物进行对比，寻求最大化输入（即图像）和其表示（即图像嵌入）之间的互信息（MI）。受之前Deep InfoMax（DIM）方法[15]在视觉表征学习中的成功启发，Deep Graph InfoMax（DGI）[18]提出了一个基于图域中MI最大化的替代目标。DGI首先采用GNN来学习节点嵌入，并通过读出函数获得全局概要嵌入（即图嵌入）。然后，DGI的目标是通过区分原始图中的节点和被破坏的图中的节点，使节点嵌入和图嵌入之间的MI最大化。

然而，我们认为，DGI中的局部-全局MI最大化框架仍处于起步阶段。它的目标被证明等同于在某些条件下使输入节点特征和高层节点嵌入之间的MI最大化。具体来说，为了实现InfoMax的目标，DGI需要一个注入式的读出函数来产生全局图嵌入，而注入式属性的限制性太强，无法实现。对于DGI采用的均值池读出函数，不能保证图嵌入能够从节点中提炼出有用的信息，因为它不足以保留节点级嵌入的独特特征。此外，DGI建议使用特征洗牌来生成图形的损坏视图。尽管如此，这个方案在生成负数节点样本时，考虑了在粗粒度层面上破坏节点特征。当特征矩阵是稀疏的，只进行特征洗牌不足以为被破坏的图中的节点生成不同的邻域（即上下文），导致对比性目标的学习困难。

在本文中，我们为无监督的图表示学习引入了一个简单而强大的对比框架（图1），我们将其称为深度GRaPh对比学习（GRACE）。<sup>3</sup>它是由传统的自组织网络[19]及其最近在视觉表征学习中的复兴[17]激发的。我们主要关注的不是对比节点级嵌入和全局嵌入，而是对比节点级的嵌入，而且我们的工作对生成图嵌入的注入性读出函数不做假设。在GRACE中，我们首先通过随机地进行破坏来生成两个相关的图视图。然后，我们使用对比损失来训练模型，使这两个视图中的节点嵌入之间的一致性最大化。与视觉数据不同，在那里有丰富的图像转换技术，如何执行腐败以生成图的视图仍然是一个开放的问题。在我们的工作中，我们共同考虑了拓扑结构和节点属性两个层面的破坏，即去除边缘和掩盖特征，为不同视图中的节点提供不同的背景，从而促进对比目标的优化。最后，我们提供了理论分析，揭示了我们的对比性目标与互信息和经典的三联体损失的联系。

我们的贡献总结如下。首先，我们提出了一个用于无监督的图表示学习的一般对比框架。提出的GRACE框架简化了以前的工作，并通过最大化两个图视图之间的节点嵌入的一致性来工作。其次，我们提出了两个具体的方案，即去除边缘和屏蔽特征，以生成图的视图。最后，我们使用六个流行的公共基准数据集，对常用的过渡性和归纳性节点分类进行了全面的实证研究。

<sup>3</sup>代码在<https://github.com/CRIPAC-DIG/GRACE> 公开。

线性评估协议。GRACE的性能一直优于现有的方法，我们的无监督方法甚至超过了有监督任务的同类方法，这表明它在现实世界的应用中具有巨大的潜力。

## 2 相关工作

**视觉表征的对比性学习。**对比法在自监督的视觉表征学习中很流行，它的目的是通过对比正反两方面的样本来学习鉴别性的表征。对于视觉数据，负面样本可以通过图像增强技术产生，如裁剪、旋转[20]、颜色扭曲[21]等。现有的工作[12-14]采用了一个存储库来存储负样本。其他工作[15-17]探讨了批量负样本。对于作为锚的图像补丁，这些方法通常会找到一个全局总结向量[22, 15]或邻近视图中的补丁[23, 24]作为正面样本，并将其与负面采样的对应物进行对比，例如同一批次中其他图像的补丁[22]。

理论分析揭示了它们成功背后的原因[25]。这些方法中使用的目标可以被看作是最大限度地提高输入特征和其表征之间的MI下限[11]。然而，最近的工作[26]显示，评估表征质量的下游性能可能在很大程度上取决于不仅在卷积结构中，而且在InfoMax目标的具体估计器中编码的偏差。

**图表示学习。**许多关于无监督图表示学习的传统方法也采用了对比性模式[1, 2, 9, 27]。之前关于无监督图表示学习的工作集中在局部对比模式上，这迫使相邻的节点具有相似的嵌入。这种情况下的阳性样本是出现在同一随机行走中的节点[1, 2]。例如，开创性的工作DeepWalk[1]使用噪声对比性估计[28]对节点共现对的概率进行建模。这些基于随机漫步的方法被证明等同于对某些形式的图近似性进行因子化（例如，相邻矩阵的变换）[4]，这些方法过度强调了这些图近似性中编码的结构信息，而且在大规模数据集中也面临着严重的扩展问题。另外，这些方法在不恰当的超参数调整下也容易出错[1, 2]。

最近关于图神经网络（GNN）的工作采用了比传统方法更强大的图卷积编码器。其中，围绕有监督的GNN[6-8, 29]这一主题，已经形成了相当多的文献，这需要标记的数据集，而在现实世界的应用中可能无法获得。沿着另一条发展路线，无监督的GNNs很少受到关注。代表性的方法包括GraphSAGE[10]，它也包含了类似DeepWalk的目标。最近的工作DGI[18]结合了GNN和对比学习的力量，其重点是最大化全局图嵌入和局部节点嵌入之间的MI。然而，它很难满足图形读出函数的注入性要求，从而使图形嵌入可能恶化。与DGI相比，我们的工作并不依赖于明确的图嵌入。相反，我们专注于最大限度地提高节点嵌入在两个被破坏的图形视图中的一致性。

## 3 深度图谱对比性表征学习

在本节中，我们详细介绍了我们提出的GRACE框架，首先是对比性目标的整体框架，然后是具体的图视图生成方法。在这一节的最后，我们从两个角度提供了我们框架背后的理论依据，即与InfoMax原则和经典的三联体损失的联系。

### 3.1 预备工作

在无监督的图表示学习中，让 $G=(V, E)$ 表示一个图，其中 $V=\{v_1, v_2, \dots, v_N\}$ ,  $E \subseteq V \times V$ 分别代表节点集和边集。我们将特征矩阵和邻接矩阵表示为 $X \in \mathbb{R}^{N \times F}$ 和 $A \in \{0, 1\}^{N \times N}$ ，其中 $\mathbf{x}_i \in \mathbb{R}^F$ 是 $v_i$ 的特征， $A_{ij} = 1$  iff  $(v_i, v_j) \in E$ 。在训练中

没有给定 $G$ 中节点的类信息。我们的目标是学习一个GNN编码器 $f(X, A) \in \mathbb{R}^{N \times F}$ ，接收图形特征和结构作为输入，产生低维度的节点嵌入，即 $F$ 。

我们把  $H = f(X, A)$  表示为节点的学习表征，其中  $h_i$  是节点  $v$  的嵌入  $i$ 。这些表征可用于下游任务，如节点分类。

### 3.2 节点表征的对比性学习

#### 3.2.1 对比性学习框架

与之前通过利用局部-整体关系学习表征的工作相反，在GRACE中，我们通过直接最大化嵌入之间的节点级协议来学习嵌入。具体来说，我们首先通过随机破坏原始图生成两个图的视图。然后，我们采用一个对比性目标，强制要求两个不同视图中每个节点的编码嵌入彼此一致，并能与其他节点的嵌入区分开来。

在我们的GRACE模型中，在每个迭代中，我们生成两个图视图，表示为  $\tilde{G}_1$  和  $\tilde{G}_2$ ，并将两个生成的视图中的节点嵌入表示为  $U = f(X_1, A_1)$  和  $V = f(X_2, A_2)$ ，其中  $X_i$  和  $A_i$  是视图的特征矩阵和相邻矩阵。关于图形视图的生成细节将在后面的3.2.2节中讨论。

然后，我们采用一个对比性目标（即判别器），将同一节点在这两个不同视图中的嵌入与其他节点的嵌入区分开。对于任何节点  $v_i$ ，它在一个视图中产生的嵌入， $u_i$ ，被视为锚，它在另一个视图中产生的嵌入， $v_i$ ，构成正样本，而  $v_i$  以外的节点在两个视图中的嵌入自然被看作是负样本。形式上，我们定义批评家  $\vartheta(u, v) = s(g(u), g(v))$ ，其中  $s$  是余弦相似度， $g$  是一个非线性投影，以增强批评家的表达能力[17, 26]。投射  $g$  是用两层多层感知器（MLP）实现的。

我们为每一个正数对  $(u_i, v_i)$  定义成一对目标为

$$l(u_i, v_i) = \log \frac{e^{\vartheta(u_i, v_i)/\tau}}{\underbrace{\sum_{k=1}^N \mathbf{1}_{[k \neq i]} e^{\vartheta(u_i, v_k)/\tau}}_{\text{视图间负数对}} + \underbrace{\sum_{k=1}^N \mathbf{1}_{[k \neq i]} e^{\vartheta(u_i, v_k)/\tau}}_{\text{视觉内的负数对}}}, \quad (1)$$

其中  $\mathbf{1}_{[k \neq i]} \in \{0, 1\}$  是一个指示函数，等于1 iff  $k \neq i$ ， $\tau$  是一个温度参数。请注意，在我们的工作中，我们不对负数节点进行明确的采样。相反，给定一个正数对，我们自然地负数样本定义为两个视图中的所有其他节点。因此，负样本来自两个来源，即视图间或视图内的节点，分别对应于分母中的第二个和第三个项。由于两个视图是对称的，另一个视图的损失定义与  $l(v_i, u_i)$  类似。然后，要最大化的总体目标被定义为所有正数对的平均数，正式的定义为

$$J = \frac{1}{2N} \sum_{i=1}^N [l(u_i, v_i) + l(v_i, u_i)] \circ \quad (2)$$

总之，在每个训练周期，GRACE首先生成图  $G$  的两个图视图  $G_1$  和  $G_2$ 。然后，我们使用GNN编码器  $f$  获得  $G_1$  和  $G_2$  的节点表示  $U$  和  $V$ 。最后，通过最大化公式 (2) 中的目标来更新  $f$  和  $g$  的参数。该学习算法总结于算法1。

---

#### 算法1：GRACE训练算法

---

```

1 for epoch ← 1, 2, ... do
2   通过对  $G$  进行腐败处理，生成两个图视图  $G_1$  和  $G_2$ 
3   使用编码器  $f$  获得  $G$  的节点嵌入  $U$ 
4   使用编码器  $f$  获得  $G$  的节点嵌入  $V_2$ 

```

---

- 5 用公式 (2) 计算对比性目标J
  - 6 通过应用随机梯度上升法更新参数，使J最大化
-

### 3.2.2 图形视图的生成

生成视图是对比学习方法的一个关键组成部分。在图域中，图的不同视图为每个节点提供不同的背景。考虑到对比性方法依赖于不同视图中的节点嵌入之间的对比，我们建议在结构和属性层面上破坏原始图，这为模型构建了不同的节点上下文来进行对比。在GRACE中，我们设计了两种图形破坏的方法，去除拓扑结构的边缘和掩盖节点属性的特征。如何进行图的破坏仍然是一个开放的问题[18]。在我们的框架中，可以灵活地采用其他替代机制的腐败方法。

**移除边缘 (RE)。** 我们随机地删除原图中的一部分边。形式上，由于我们只删除现有的边缘，我们首先抽出一个随机掩蔽矩阵  $\mathbf{R} \in \{0, 1\}^{N \times N}$ ，其条目来自伯努利分布  $\mathbf{R}_{ij} \sim \text{B}(1-p_r)$ 。如果原图的  $\mathbf{A}_{ij} = 1$ ，否则  $\mathbf{R}_{ij} = 0$ 。这里  $p_r$  是每条边被移除的概率。由此产生的邻接矩阵可以计算为

$$\tilde{\mathbf{A}} = \mathbf{A} \circ \mathbf{R}, \quad (3)$$

其中  $(\mathbf{x} \circ \mathbf{y})_i = x_i y_i$  是Hadamard积。

**屏蔽节点特征 (MF)。** 除了去除边缘外，我们还随机屏蔽一部分维度，在节点特征中为零。形式上，我们首先对一个随机向量  $\mathbf{m} \in \{0, 1\}^F$  进行抽样调查。其中，它的每个维度都是独立地从伯努利分布中抽取的，概率为

$1 - p_m$ ，即  $\mathbf{m}_i \sim \text{B}(1 - p_m)$ 。然后，生成的节点特征  $\tilde{\mathbf{x}}$  的计算方法是

$$\mathbf{X} \leftarrow [\mathbf{x}_1 \circ \mathbf{m}; \mathbf{x}_2 \circ \mathbf{m}; \dots; \mathbf{x}_N \circ \mathbf{m}]^T. \quad (4)$$

这里  $[-; -]$  是串联运算符。

请注意，尽管我们提出的RE和MF方案在技术上与Dropout[30]和DropEdge[31]相似，但我们的GRACE模型和这两种参考方法的目的根本不同。Dropout是一种通用技术，在训练期间随机掩盖神经元以防止大规模模型的过度拟合。在图域中，DropEdge的提出是为了防止过度拟合，并缓解GNN架构过深时的过度平滑。然而，我们的GRACE框架随机应用RE和MF来产生不同的图视图，以便在图拓扑和节点特征层面进行对比学习。此外，GRACE中采用的GNN编码器是一个相当浅的模型，通常只由两层或三层组成。

在我们的实现中，我们共同利用这两种方法来生成图视图。  $\mathbf{G}_1$  和  $\mathbf{G}_2$  的生成由两个超参数  $p_r$  和  $p_m$  控制。为了在两个视图中提供不同的语境，两个视图的生成过程使用了两套不同的超参数

$p_{r,1}, p_{m,1}$  和  $p_{r,2}, p_{m,2}$ 。实验证明，在温和的条件下，我们的模型对  $p_r$  和  $p_m$  的选择并不敏感，例如，  $p_r \leq 0.8$  和  $p_m \leq 0.8$ ，这样原始图形就不会被过度破坏。我们请读者参考附录C.1中的敏感性分析，了解经验结果。

### 3.3 理论上的理由

在这一节中，我们从两个角度提供了我们模型背后的理论依据，即相互信息最大化和三联体损失。详细的证明可以在附录D中找到。

**与互信息的联系。** 首先，我们揭示了我们的损失与两个视图中的节点特征和嵌入之间的相互信息最大化之间的联系，这在表征学习文献中被广泛地应用[13, 15, 25, 26]。相互信息量化了通过观察另一个随机变量获得的关于一个随机变量的信息量。

**定理1.** 让  $\mathbf{x}_i = \{\mathbf{x}_{ik}\}_{k \in \mathbf{N}(i)}$  是节点  $v_i$  的邻域，集体映射到其输出嵌入，其中  $\mathbf{N}(i)$  表示GNN架构指定的节点  $v_i$  的邻域集合， $\mathbf{x}$  是相应的随机变量，具有均匀分布  $p(\mathbf{x}_i) = \frac{1}{K}$ 。给定两个随机变量  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^F$  是两个视图中的嵌入，它们的联合分布表示为  $p(\mathbf{u}, \mathbf{v})$ ，我们的目标是编码器输入  $\mathbf{x}$  和两个图视图  $\mathbf{u}, \mathbf{v}$  中的节点表示之间的MI的下限。形式上，



$$j \leq i \quad (\mathbf{x}; \mathbf{u}, \mathbf{v}) \circ \quad (5)$$

**证明简述。**我们首先观察到，我们的目标J是InfoNCE目标的下界[23]、

其定义为  $I_{\text{NCE}}(\mathbf{U}; \mathbf{V})$ ，E 
$$\frac{Q}{P(u_i, v_i)} = \frac{1}{N} \sum_{i=1}^N \log \frac{e^{g(u_i, v_i)}}{\sum_{j=1}^N e^{g(u_i, v_j)}} \quad [25].$$
 根据 [23]，InfoNCE估计器是真实MI的下界。因此，该定理直接来源于数据处理不等式的应用，即  $I(\mathbf{U}; \mathbf{V}) \leq I(\mathbf{X}; \mathbf{U}, \mathbf{V})$ 。□

**备注。**从定理1可以看出，最大化J相当于最大化输入节点特征和学习节点表征之间的相互信息  $I(\mathbf{X}; \mathbf{U}, \mathbf{V})$  的下限。与此相反的是，最近的工作进一步提供了经验证据，即优化更严格的MI约束可能不会导致视觉表征学习的更好下游性能[26]，这

突出了编码器设计的重要性。在附录C.3中，我们还将我们的目标与InfoNCE损失进行了比较，它是一个更严格的MI估算器，以进一步证明GRACE模型的优越性。

**与三联体损失的联系。**另外，我们可以把公式(2)中的优化问题看作是经典的三联体损失，常用于深度度量学习。

**定理2.**当投影函数  $g$  是身份函数，并且我们通过简单地取内积来衡量嵌入相似性，即  $s(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T \mathbf{v}$ ，并进一步假设正数对远比负数对更对齐，最小化成对目标  $(\mathbf{u}_i, \mathbf{v}_i)$  与最大化三联体损失相吻合，如后文所述

$$-I(\mathbf{u}_i, \mathbf{v}_i) \propto 4Nt + \sum_{j=1}^N \mathbf{1}_{ij|J \neq I} \|\mathbf{u}_i - \mathbf{v}_i\|^2 - \|\mathbf{u}_i - \mathbf{v}_j\|^2 + \|\mathbf{u}_i - \mathbf{v}_j\|^2 - \|\mathbf{u}_j - \mathbf{v}_i\|^2 \quad (6)$$

**备注。**定理2在目标和经典的三联体损失之间建立了联系。换句话说，我们可以将公式(2)中的问题视为学习图卷积编码器，以鼓励正样本在嵌入空间中离负样本更远。此外，通过从度量学习的角度看目标，我们强调了适当选择负样本的重要性，这在以前基于InfoMax的方法中经常被忽视。最后，对比性目标的优化成本很低，因为我们不需要明确地生成负样本，所有的计算都可以并行进行。与此相反，已知三联体损失的计算成本很高[32]。

## 4 实验

在这一节中，我们使用六个公共基准数据集对产生的节点嵌入在节点分类上的质量进行了经验评估。我们请感兴趣的读者参考补充材料中的实验细节，包括数据集的配置（附录A），实施和超参数（附录B），以及额外的实验（附录C）。

### 4.1 数据集

为了进行综合比较，我们使用了六个广泛使用的数据集来研究过渡性和归纳性节点分类的性能。具体来说，我们使用三种数据集：（1）引文网络，包括Cora、Citeseer、Pubmed和DBLP[33, 34]，用于过渡性节点分类、

（2）来自Reddit帖子的社交网络，用于大规模图上的归纳学习[10]，以及（3）生物蛋白质-蛋白质相互作用（PPI）网络[35]，用于多个图上的归纳节点分类。这些数据集的细节可以在附录A中找到。

### 4.2 实验设置

对于每个实验，我们都遵循[18]中的线性评估方案，其中每个模型首先以无监督的方式进行训练。所得的嵌入被用来训练和测试一个简单的  $l_2$ -正则化逻辑回归分类器。我们对该模型进行了20次训练，并报告了每个数据集的平均性能。此外，我们使用归纳任务的微观平均

F1分数和归纳任务的准确性来衡量性能。请注意，对于归纳学习任务，测试是在未见过或未训练过的节点和图形上进行的，而对于归纳学习任务，我们使用所有数据的特征，但测试集的标签在训练期间被屏蔽。

**归纳学习。** 在归纳学习任务中，我们采用两层的GCN[6]作为编码器。我们的编码器结构的形式是这样的

$$GC(X, A) = \sigma(D^{-1} A D^{-1} X W), \quad (7)$$

$$f(X, A) = GC_2(GC_1(X, A), A). \quad (8)$$

其中， $\hat{A} = A + I$ 是带有自循环的邻接矩阵， $D_{\hat{A}}$ 是度矩阵， $\sigma(\cdot)$ 是一个非线性激活函数，例如， $\text{ReLU}(\cdot) = \max(0, \cdot)$ ， $W_i$ 是一个可训练的权重矩阵。

我们认为以下两类代表性算法是基线，包括（1）传统方法DeepWalk[1]和node2vec[2]，以及（2）深度学习方法GAE、VGAE[9]和DGI[18]。此外，我们还报告了使用逻辑回归分类器对原始节点特征和DeepWalk与输入节点特征相连接的嵌入所获得的性能。为了与有监督的同行进行直接比较，我们还报告了两个有代表性的模型SGC[29]和GCN[6]的性能，它们是以端到端方式训练的。

**大图上的归纳学习。** 考虑到Reddit数据的大规模，我们紧跟[18]，采用三层GraphSAGE-GCN[10]与剩余连接[36]作为编码器，其表述为

$$MP_i(X, A) = \sigma([D^{-1} A^{-1} X; X] W_i), \quad (9)$$

$$f(X, A) = MP_3(MP_2(MP_1(X, A), A), A). \quad (10)$$

这里我们使用了均值传播规则，因为 $D^{-1}$ 是对节点特征的平均数。由于Reddit的规模很大，它不能完全装入GPU内存。因此，我们应用了[10]中提出的子抽样方法，我们首先随机选择一个小批的节点，然后通过对节点邻居进行替换抽样，得到以每个选定节点为中心的子图。具体来说，我们在第一跳、第二跳和第三跳分别抽取30、25、20个邻居。在这种基于抽样的设置下生成图视图，RE和MF都可以毫不费力地适应抽样子图。

**多图上的归纳学习。** 对于多图的归纳学习PPI，我们用跳过的连接堆叠了三个均值池层，类似于DGI[18]。图卷积编码器可以被表述为

$$h_1 = MP_1(x, a), \quad (11)$$

$$h_2 = MP(x w_{\text{skip}} + h_1, a), \quad (12)$$

$$f(X, A) = H_3 = MP(x_{\text{跳}} + h_1 + h_2, a), \quad (13)$$

其中 $w_{\text{skip}}$ 和 $w_{\text{跳}}$ 是两个投影矩阵， $|$ 在公式(9)中定义。尽管PPI数据集由多个图组成，出于效率的考虑，我们只计算一个锚节点的负样本，作为同一图内的其他节点。

大图和多图设置中的基线都是以类似于过渡性任务的方式选择。我们考虑（1）传统方法DeepWalk[1]，以及（2）深度学习方法GraphSAGE[10]和DGI[18]。此外，我们还报告了在与过渡性任务相同的设置下，使用原始特征和DeepWalk+特征的性能。我们进一步提供了两个有代表性的监督方法的性能，包括FastGCN[37]和GaAN-mean[38]作为参考。在表格中，基线的结果是按照其原始论文的性能报告的。对于GraphSAGE，我们重新使用了无监督的结果以进行公平比较。

#### 4.3 结果和分析

表1中总结了经验性能。总的来说，从表中我们可以看出，我们提出的模型在所有六个数据集上都表现出很强的性能。GRACE在过渡性和归纳性任务上的表现一直比无监督的基线好

很多。强大的性能验证了所提出的对比性学习框架的优越性。我们

---

<sup>4</sup>DeepWalk不适用于多图实验，因为DeepWalk产生的嵌入空间可能相对于不同的不相交图任意旋转[10]。

表1: 节点分类的性能摘要, 以百分比 (在归纳任务上) 或微观平均F1得分 (在归纳任务上) 的标准偏差表示。每种方法在训练阶段的可用数据显示在第二列, 其中 $X$ 、 $A$ 、 $Y$ 分别对应于节点特征、邻接矩阵和标签。无监督模型的最高性能以黑体字突出显示。

(a) 传导式

方法	训练数据	Cora	Citeseer	公共医学杂志	DBLP
原始特征	$X$	64.8	64.6	84.8	71.6
node2vec	$A$	74.8	52.3	80.3	78.8
DeepWalk	$A$	75.7	50.5	80.5	75.9
DeepWalk + 功能	$X, A$	73.1	47.6	83.7	78.1
GAE	$X, A$	76.9	60.6	82.9	81.2
VGAE	$X, A$	78.9	61.2	83.0	81.7
DGI	$X, A$	82.6±0.4	68.8±0.7	86.0±0.1	83.2±0.1
<b>格雷斯</b>	$X, A$	<b>83.3±0.4</b>	<b>72.1±0.5</b>	<b>86.7±0.1</b>	<b>84.2±0.1</b>
SGC	$X, A, Y$	80.6	69.1	84.8	81.7
GCN	$X, A, Y$	82.8	72.0	84.9	82.7

(b) 感应式

方法	训练数据	睿迪特 (Reddit 公司)	PPI
原始特征	$X$	58.5	42.2
DeepWalk	$A$	32.4	-
DeepWalk + 功能	$X, A$	69.1	-
GraphSAGE-GCN	$X, A$	90.8	46.5
图表SAGE-mean	$X, A$	89.7	48.6
GraphSAGE-LSTM	$X, A$	90.7	48.2
图形AGE-池	$X, A$	89.2	50.2
DGI	$X, A$	94.0±0.1	63.8±0.2
<b>格雷斯</b>	$X, A$	<b>94.2±0.0</b>	<b>66.2±0.1</b>
快速GCN	$X, A, Y$	93.7	-
平均值 (GaAN-mean)	$X, A, Y$	95.8±0.1	96.9±0.2

特别要注意的是, GRACE在所有四个过渡性数据集和归纳性数据集Reddit上与用标签监督训练的模型有竞争性。

我们还观察到以下几点。首先, GRACE比另一种有竞争力的对比学习方法DGI在PPI上取得了相当大的改进。我们认为这是由于节点特征的极度稀少 (超过40%的节点的特征为零 [10]), 这强调了在选择负面样本时考虑拓扑信息的重要性。对于像PPI这样的数据集, 极端的特征稀疏性使得DGI无法区分原始图和通过洗牌节点特征产生的破坏图中的样本, 因为洗牌节点特征对对比目标没有影响。相反, GRACE中使用的RE方案不依赖于节点特征, 在这种情况下起到了补救作用, 这可以解释GRACE与DGI相比在PPI上的优势。另外, 我们注意到, 我们的方法与监督模型之间仍然存在巨大的差距。这些监督模型的另一个优点是受益于标签, 它为模型学习提供了其他辅助信息。考虑到现实世界数据集的稀疏性, 我们进

行了另一个实验来验证我们的方法对稀疏的节点特征是稳健的（附录C.4）。结果表明，通过随机删除节点特征，我们的方法仍然优于现有的基线。

其次，像DeepWalk这样的传统对比学习方法的表现不如在一些数据集（Citeseer、Pubmed和Reddit）上只使用原始特征的天真分类器，这表明这些方法在利用节点特征方面可能是无效的。与传统工作不同，我们看到基于GCN的方法，如GraphSAGE和GAE，能够在学习嵌入时纳入节点特征。然而，我们注意到，在某些数据集（Pubmed）上，它们的性能

我们认为这可以归因于他们选择负面样本的天真方法，即简单地根据边缘来选择对比对。这一事实进一步证明了在对比性表征学习中选择负样本的重要作用。与GAEs相比，GRACE的优越性能也再次验证了我们提出的GRACE框架的有效性，该框架在图的视图中对节点进行对比。

此外，我们对关键的超参数 $p_r$ 和 $p_m$ （附录C.1）进行了敏感性分析，并对我们的混合方案进行了生成图视图的消融研究（附录C.2）。结果表明，我们的方法对这些参数的扰动是稳定的，并验证了在图的拓扑结构和节点特征层面上的破坏的必要性。我们还比较了经典的InfoNCE损失（附录C.3），验证了我们设计选择的有效性。这些额外实验的细节可以在补充材料中找到。

## 5 总结

在本文中，我们开发了一个新的图对比表征学习框架，该框架基于最大化节点水平上的一致。我们的模型通过首先使用两种建议的方案生成图的视图，去除边缘和掩盖节点特征，然后应用对比损失来最大化这两种视图中节点嵌入的一致性来学习表示。理论分析揭示了我们的对比性目标与互信息最大化和经典的三重损失的联系，这证明了我们的动机。我们使用各种真实世界的数据集，在归纳和归纳设置下进行了综合实验。实验结果表明，我们提出的方法可以持续地以较大的幅度超过现有的最先进的方法，甚至在过渡性任务上超过有监督的同行。

## 关于更广泛影响的讨论

我们提出的自监督图表示学习技术有助于缓解在现实世界中部署机器学习应用时的标签稀缺问题，从而节省了大量的人力注释工作。例如，我们的GRACE框架可以插入到现有的推荐系统中，为用户和商品产生高质量的嵌入，解决冷启动问题。此外，从实证结果来看，我们的工作蛋白质功能预测方面比现有的基线有明显的优势，这表明它在药物发现和治疗方面的巨大潜力，因为在这个关键时刻，COVID-19危机。请注意，我们的工作主要是作为现有机器学习模型的插件，它并没有带来新的伦理问题。然而，GRACE模型仍然可能给出有偏见的输出（例如，性别偏见、种族偏见），因为在数据收集、图形构建等过程中，所提供的数据本身可能有强烈的偏见。

## 鸣谢

作者要感谢孙涛和吕思锐有见地的讨论。这项工作由国家重点研发计划（2018YFB1402600，2016YFB1001000）和国家自然科学基金（U19B2038，61772528）共同支持。



## A 数据集详情

**归纳学习。**我们利用四个广泛使用的引文网络，Cora、Citeseer、Pubmed和DBLP，来预测文章的主题类别。在这些数据集中，图是由计算机科学文章的引文链接构建的。具体来说，节点对应于文章，无向边对应于文章之间的引文链接。此外，每个节点都有一个稀疏的词包特征和一个相应的文章类型标签。前三个网络是由[33, 39]提供的，后一个DBLP数据集是由[34]提供的。在这些引文网络中，我们随机选择10%的节点作为训练集，10%的节点作为验证集，剩下的节点作为测试集。

**大图上的归纳学习。**然后我们预测一个大规模社会网络的社区结构，该网络收集自Reddit。该数据集由[10]预处理，包含2014年9月创建的Reddit帖子，这些帖子属于不同的社区（subreddit）。在数据集中，节点对应于帖子，如果同一用户在两个帖子上都有评论，则边连接着帖子。节点特征由帖子的标题、内容和评论构成，使用现成的GloVe词嵌入[40]，以及其他指标，如帖子得分和评论数量。按照[10, 18]的归纳设置，在Reddit数据集上，我们选择前20天的帖子进行训练，包括151,708个节点，其余的用于测试（30%的数据包括23,699个节点用于验证）。

**对多个图的归纳学习。**最后，我们根据基因本体的细胞功能，在蛋白质-蛋白质相互作用（PPI）网络[35]中预测蛋白质的作用，以评估所提出的方法在多个图中的泛化能力。PPI数据集包含多个图，每个图对应一个人体组织。图由[10]构建，其中每个节点有多个标签，是基因本体集的子集（共121个），节点特征包括位置基因集、主题基因集和免疫学特征（共50个）。按照[10]，我们选择了20个由44,906个节点组成的图作为训练集，两个包含6,514个节点的图作为验证集，其余四个包含12,038个节点的图作为测试集。

表2汇总了数据集的统计数据；表3包括了下载链接。对于归纳性任务，与[6]类似，在训练阶段，所有的节点特征都是可见的，但节点标签被掩盖了。在归纳式设置中，我们密切关注[10]；在训练期间，用于评估的节点是完全不可见的；然后在不可见的或未训练的节点和图上进行评估。

表2：实验中使用的数据集的统计数据。

数据集	类型	#节点	# 边缘	#特点	#Classes
Cora	横向的	2,708	5,429	1,433	7
Citeseer	横向的	3,327	4,732	3,703	6
公共医学杂志	横向的	19,717	44,338	500	3
DBLP	横向的	17,716	105,734	1,639	4
睿迪特 (24张图)	感应式	231,443	11,606,919	602	41

表3：数据集下载链接。

数据集	下载链接
Cora	<a href="https://github.com/kimiyoung/planetoid/raw/master/data">https://github.com/kimiyoung/planetoid/raw/master/data</a>
Citeseer	<a href="https://github.com/kimiyoung/planetoid/raw/master/data">https://github.com/kimiyoung/planetoid/raw/master/data</a>
Pubmed	<a href="https://github.com/kimiyoung/planetoid/raw/master/data">https://github.com/kimiyoung/planetoid/raw/master/data</a>

DBLP	<a href="https://github.com/abojchevski/graph2gauss/raw/master/data/dblp.npz">https://github.com/abojchevski/graph2gauss/raw/master/data/dblp.npz</a>
红豆网	<a href="https://s3.us-east-2.amazonaws.com/dgl.ai/dataset/reddit.zip">https://s3.us-east-2.amazonaws.com/dgl.ai/dataset/reddit.zip</a>
PPI	<a href="https://s3.us-east-2.amazonaws.com/dgl.ai/dataset/ppi.zip">https://s3.us-east-2.amazonaws.com/dgl.ai/dataset/ppi.zip</a>

## B 实施

**计算基础设施。**所有的模型都是用PyTorch Geometric 1.5.0[41]和PyTorch 1.4.0[42]实现的。整个实验中使用的数据集都可以在PyTorch Geometric库中找到。对于节点分类，我们使用Scikit-learn[43]中现有的带有 $l_2$ 正则化的逻辑回归实现。所有的实验都是在一台带有8个NVIDIA Titan Xp GPU（每个12GB内存）和14个Intel Xeon E5-2660 v4 CPU的计算机服务器上进行的。

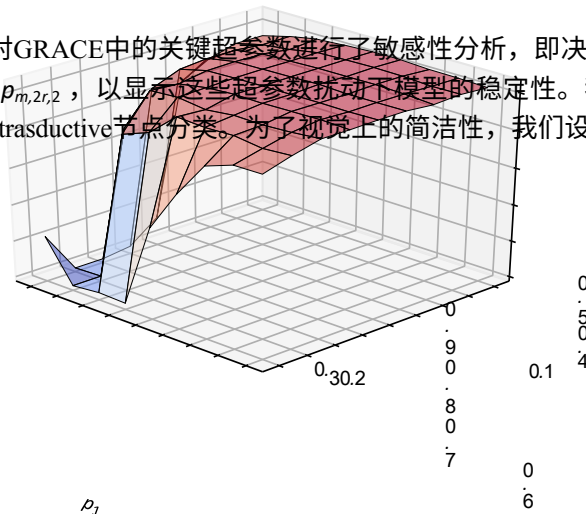
**超参数。**所有模型都用Glorot初始化[44]，并在所有数据集上用Adam SGD优化器[45]进行训练。初始学习率设置为0.001，但在Cora上例外为0.0005，在Reddit上为 $10^{-5}$ 。在所有数据集上， $l_2$ 权重衰减因子被设置为 $10^{-5}$ 。在过渡性和归纳性任务中，我们对模型进行固定次数的训练，具体来说，Cora、Citeseer、Pubmed和DBLP分别为200、200、1500、1000次，Reddit为40次，PPI为200次。控制抽样过程的概率参数，第一视角的 $p_{r,1}$ ， $p_{m,1}$ 和第二视角的 $p_{r,2}$ ， $p_{m,2}$ ，都选择在0.0和0.4之间，因为原始图会当概率设置得过大时，会被过度破坏。请注意，为了给两个视图中的节点生成不同的上下文， $p_{r,1}$ 和 $p_{r,2}$ 应该是不同的， $p_{m,1}$ 和 $p_{m,2}$ 也是如此。所有针对数据集的超参数都在表4中总结。

数据集	$p_{m,1}$	$p_{m,2}$	价值 ,1	价值 ,2	超参数规格。 率水	训练纪 元	隐藏的 维度	激活功能
Cora	0.3	0.4	0.2	0.4	0.005	10	128	リング
Citeseer	0.3	0.2	0.2	0.0	0.001	$10^{-5}$	200	ループ
公共医学杂志	0.0	0.2	0.4	0.1	0.001	$10^{-5}$	256	プレゼ
DBLP	0.1	0.0	0.1	0.4	0.001	$10^{-5}$	1,500	ル
表油柱	0.2	0.2	0.1	0.2	0.00001	$10^{-5}$	1,000	リング

## C 额外的实验

### C.1 敏感度分析

在这一节中，我们对GRACE中的关键超参数进行了敏感性分析，即决定图形视图生成的四个概率 $p_{m,1}$ ， $p_{r,1}$ ， $p$ ， $p_{m,2}$ ， $p_{r,2}$ ，以显示这些超参数扰动下模型的稳定性。我们通过改变这些参数从0.1到0.9来进行transductive节点分类。为了视觉上的简洁性，我们设定 $p_1 = p_{r,1} = p_{m,1}$



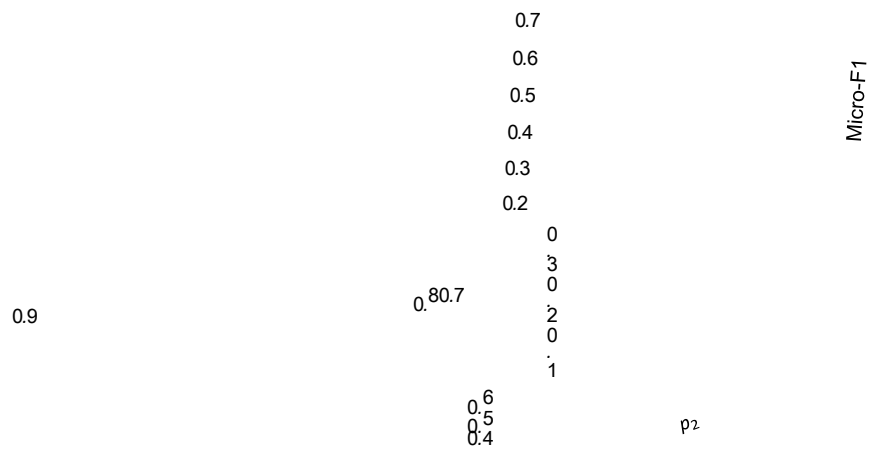


图2：在Citeseer数据集上，GRACE在不同的超参数下，以Micro-F1的方式进行过渡性节点分类的性能。

和 $p_2 = p_{r,2} = p_{m,2}$ 。换句话说， $p_1$  和  $p_2$  控制着两个图形视图的生成。请注意，我们在敏感性分析中只改变了这四个参数，其他参数与之前描述的保持一致。

Citeseer数据集上的结果如图2所示。从图中可以看出，当参数不太大时，Micro-F1的节点分类性能是相对稳定的，如图中的高原所示。因此我们得出结论，总体而言，我们的模型对这些概率不敏感，证明了对超参数调整的稳健性。如果概率设置得太大（例如， $>0.5$ ），原图将被严重破坏。比如说，当 $p_r = 0.9$ 时，几乎每条现有的边都被删除了，导致孤立的节点在生成的图视图。那么，在这种情况下，图卷积网络就很难从节点邻域学习到有用的信息。因此，在两个图视图中学习到的节点嵌入是不够独特的，这将导致难以优化对比目标。

## C.2 消融研究

在本节中，我们对生成图视图的两种方案，即去除边缘（RE）和屏蔽节点特征（MF）进行消融研究，以验证所提出的混合方案的有效性。我们将GRACE（-RE）表示为不去除边缘的模型，GRACE（-MF）表示为不掩盖节点特征的模型。我们报告了GRACE（-RE）、GRACE（-MF）和原始模型GRACE在过渡性节点分类上的表现，除了不同的启用方案外，设置与之前相同。结果列于表5。

可以看出，我们联合应用RE和MF的混合方法明显优于只使用一种独立的RE或MF方法的两个降级模型。这些结果验证了我们提出的方案对图的破坏的有效性，并进一步显示了在图的拓扑结构和节点特征两个层面上共同考虑破坏的必要性。

表5：在消融研究中，模型变体与原始GRACE模型在节点分类准确性方面的表现。GRACE(-RE)和GRACE(-MF)分别表示没有去除边缘和屏蔽节点特征的模型。

方法	科拉	馨予	医学论文	DBLP
遗传学	<b>83.2±0.5</b>	<b>72.1±0.5</b>	<b>86.7±0.1</b>	<b>84.2±0.1</b>
遗传因子 (-RE)	82.3±0.4	72.0±0.4	84.8±0.2	83.6±0.2
GRACE (-MF)	81.6±0.4	69.9±0.6	85.7±0.1	83.5±0.1

## C.3 与InfoNCE损失的比较

在本节中，我们考虑另一个广泛使用的目标，即InfoNCE损失[23]，在对比方法中。为了公平比较，我们使用InfoNCE目标来衡量两个图视图之间的节点相似性，其定义为

$$J_{\text{NCE}} = \frac{1}{2} [I_{\text{NCE}}(\mathbf{V}, \mathbf{U}) + I_{\text{NCE}}(\mathbf{U}, \mathbf{V})], \quad (14)$$

其中，成对的目标由 $I_{\text{NCE}}(\mathbf{U}, \mathbf{V})$ ， $N$ 定义。
$$\frac{1}{N} \sum_{i=1}^N \log \frac{e^{g(\mathbf{u}_i, \mathbf{v}_i)}}{\sum_{j=1}^N e^{g(\mathbf{u}_i, \mathbf{v}_j)}} \cdot \text{INCE}(\mathbf{V}, \mathbf{U})$$
可以被对称地定义。修改后的模型在下文中被称为GRACE-NCE。我们报告了GRACE-NCE在与原始模型GRACE相同的设置下的传导性节点分类的性能。结果总结在表6中。

从表中我们可以清楚地看到，在所有四个数据集上，变体模型GRACE-NCE的性能都不如原始模型GRACE。结果实证表明，虽然InfoNCE是一个更严格的互信息估计器，但我们的目标更有效，显示出更好的下游性能，这与之前在视觉表示学习中的观察结果一致[26]。我们认为，与InfoNCE相比，我们的目标的优越性能可以归因于包含了更多的负面样本。具体来

说，我们在目标中考虑了视图内的负数对，这可以被看作是对图卷积算子带来的平滑问题的一种正则化。

表6: GRACE和GRACE-NCE在四个引文数据集上的过渡性节点分类中的表现。

方法	科拉	馨予	医学论文	DBLP
遗传学	83.2±0.5	72.1±0.5	86.7±0.1	84.2±0.1
GRACE-NCE	82.1±0.4	70.9±0.6	85.0±0.1	82.1±0.1

#### C.4 对稀疏特征的稳健性

如前所述,对于现有的工作DGI,使用特征洗牌方案为具有密集特征的节点生成负样本是相对容易的。然而,当节点特征稀疏时,特征洗牌可能不足以为节点生成不同的邻域,这就促使我们的混合方案在拓扑和属性两个层面上对原始图进行破坏。

在这一节中,我们进行了随机污染训练数据的实验,将某一部分节点特征屏蔽为零。具体来说,我们在四个引文网络上将节点特征的污染率从0.5变化到0.9。我们在所有其他参数与之前描述的相同的情况下,进行了过渡性节点分类的实验。准确率方面的表现如图3所示。

从图中我们可以看出,在不同的污染率下,GRACE始终以较大的优势胜过DGI,这证明了我们提出的GRACE模型对稀疏特征的鲁棒性。我们将GRACE的鲁棒性归功于我们提出的RE方法在拓扑层面对图的破坏的优越性,因为RE能够为节点构建不同的拓扑环境而不依赖于节点特征。这些结果再次验证了在拓扑和属性两个层面上考虑图损坏的必要性。请注意,当很大一部分节点特征被屏蔽时,例如90%的特征被屏蔽,GRACE和DGI的表现都很差。这可以解释为,当节点特征被过度污染时,节点是高度稀疏的,这样GNN模型就不能有效地从节点中提取有用的信息,导致性能下降。

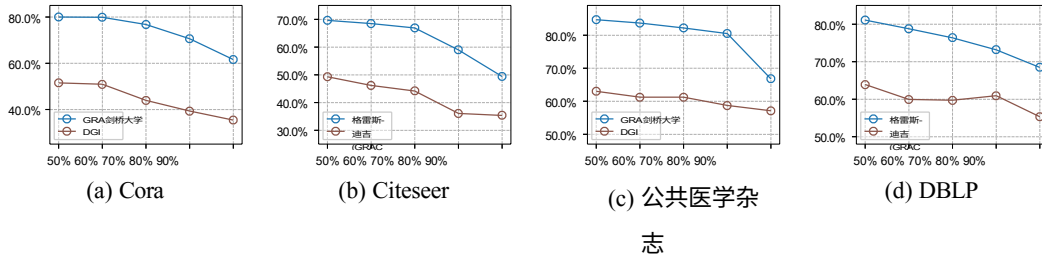


图3: GRACE和DGI在四个引文数据集上,在不同掩蔽率下掩蔽部分节点特征的情况下,以Micro-F1的方式进行传导节点分类的性能。

## D 详细证明

### D.1 定理1的证明

**定理1.** 让  $\mathbf{x}_i = \{\mathbf{x}_{kk} \}_{k \in \mathcal{N}(i)}$  是节点  $v_i$  的邻域, 集体映射到其输出嵌入, 其中  $\mathcal{N}(i)$  表示GNN架构指定的节点  $v_i$  的邻域集合,  $\mathbf{x}$  是相应的随机变量, 具有均匀分布  $p(\mathbf{x}_i) = \frac{1}{|\mathcal{N}(i)|}$ 。给定两个随机变量  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^F$  是两个视图中的嵌入, 它们的联合分布表示为  $p(\mathbf{u}, \mathbf{v})$ , 我们的目标是编码器输入  $\mathbf{x}$  和两个图视图  $\mathbf{u}, \mathbf{v}$  中的节点表示之间的MI的下限。正式地,

$$j \leq i \quad (\mathbf{x}; \mathbf{u}, \mathbf{v}) \circ \quad (15)$$



证明。我们首先说明我们的目标J和InfoNCE目标[23]之间的联系，它可以被定义为[25]

$$J_{\text{NCE}}(\mathbf{U}; \mathbf{V}) = \mathbb{E}_{\mathbf{Q}(\mathbf{u}, \mathbf{v})} \left[ \frac{1}{N} \sum_{i=1}^N \frac{\exp(\vartheta(\mathbf{u}_i, \mathbf{v}_i))}{\frac{1}{N} \sum_{j=1}^N \exp(\vartheta(\mathbf{u}_i, \mathbf{v}_j))} \right],$$

其中批判函数被定义为 $\vartheta(\mathbf{x}, \mathbf{y}) = s(g(\mathbf{x}), g(\mathbf{y}))$ 。我们进一步定义 $\rho_r(\mathbf{u}_i) = \frac{1}{N} \sum_{j=1, j \neq i}^N \exp(\vartheta(\mathbf{u}_i, \mathbf{u}_j))$ ,  $\rho_c(\mathbf{u}_i) = \frac{1}{N} \sum_{j=1}^N \exp(\vartheta(\mathbf{u}_i, \mathbf{v}_j))$ 。为方便记述， $\exp(\vartheta(\mathbf{u}_i, \mathbf{v}_j))$ 。

请注意， $\rho_r(\mathbf{v}_i)$ 和 $\rho_c(\mathbf{v}_i)$ 可以对称地定义。然后，我们的目标J可以被改写为

$$J = \mathbb{E}_{\mathbf{Q}(\mathbf{u}, \mathbf{v})} \left[ \frac{1}{N} \sum_{i=1}^N \frac{\exp(\vartheta(\mathbf{u}_i, \mathbf{v}_i))}{\log \sqrt{(\rho_c(\mathbf{u}_i) + \rho_r(\mathbf{u}_i))(\rho_r(\mathbf{v}_i) + \rho_c(\mathbf{v}_i))}} \right]. \quad (16)$$

使用 $\rho_c$ 的符号，InfoNCE估计器 $I_{\text{NCE}}$ 可以写成：

$$I_{\text{NCE}}(\mathbf{U}, \mathbf{V}) = \mathbb{E}_{\mathbf{Q}(\mathbf{u}, \mathbf{v})} \left[ \frac{1}{N} \sum_{i=1}^N \frac{\exp(\vartheta(\mathbf{u}_i, \mathbf{v}_i))}{\rho_c(\mathbf{u}_i)} \right]. \quad (17)$$

因此，

$$\begin{aligned} 2j &= I_{\text{NCE}}(\mathbf{u}, \mathbf{v}) - \mathbb{E}_{\mathbf{Q}(\mathbf{u}, \mathbf{v})} \left[ \frac{1}{N} \sum_{i=1}^N \frac{1 + \frac{\rho_r(\mathbf{u}_i)}{\rho_c(\mathbf{u}_i)}}{1 + \frac{\rho_r(\mathbf{v}_i)}{\rho_c(\mathbf{v}_i)}} \right] \\ &\quad + I_{\text{NCE}}(\mathbf{v}, \mathbf{u}) - \mathbb{E}_{\mathbf{Q}(\mathbf{u}, \mathbf{v})} \left[ \frac{1}{N} \sum_{i=1}^N \frac{1 + \frac{\rho_r(\mathbf{v}_i)}{\rho_c(\mathbf{v}_i)}}{1 + \frac{\rho_r(\mathbf{u}_i)}{\rho_c(\mathbf{u}_i)}} \right] \\ &\leq I_{\text{NCE}}(\mathbf{u}, \mathbf{v}) + I_{\text{NCE}}(\mathbf{v}, \mathbf{u}) \end{aligned} \quad (18)$$

根据[25]，InfoNCE估计器是真实MI的下限，即

$$I_{\text{NCE}}(\mathbf{u}, \mathbf{v}) \leq i(\mathbf{u}; \mathbf{v}). \quad (19)$$

因此，我们得出了

$$2j \leq i(\mathbf{u}; \mathbf{v}) + i(\mathbf{v}; \mathbf{u}) = 2i(\mathbf{u}; \mathbf{v}), \quad (20)$$

这就导致了不等式

$$J \leq I(\mathbf{U}; \mathbf{V}). \quad (21)$$

根据数据处理不等式，即对于所有满足马尔可夫关系 $\mathbf{X} \rightarrow \mathbf{Y} \rightarrow \mathbf{Z}$ 的随机变量，不等式 $I(\mathbf{X}; \mathbf{Z}) \leq I(\mathbf{X}; \mathbf{Y})$ 成立。然后，我们观察到， $\mathbf{X}, \mathbf{U}, \mathbf{V}$ 满足关系 $\mathbf{U} \leftarrow \mathbf{X} \rightarrow \mathbf{V}$ 。由于 $\mathbf{U}$ 和 $\mathbf{V}$ 在观察 $\mathbf{X}$ 之后是有条件独立的，所以该关系等同于 $\mathbf{U} \rightarrow \mathbf{X} \rightarrow \mathbf{V}$ 的马尔可夫关系，这导致 $I(\mathbf{U}; \mathbf{V}) \leq I(\mathbf{U}; \mathbf{X})$ 。我们进一步注意到，关系 $\mathbf{X} \rightarrow (\mathbf{U}, \mathbf{V}) \rightarrow \mathbf{U}$ 成立，因此可以得出 $I(\mathbf{X}; \mathbf{U}) \leq I(\mathbf{X}; \mathbf{U}, \mathbf{V})$ 。将这两个不等式结合起来，可以得到所需的不等式

$$i(\mathbf{u}; \mathbf{v}) \leq i(\mathbf{x}; \mathbf{u}, \mathbf{v}). \quad (22)$$

按照公式(21)和公式(22)，我们最终得出不等式

$$j \leq i(\mathbf{x}; \mathbf{u}, \mathbf{v}), \quad (23)$$

这就结束了这个证明。□

## D.2 定理2的证明

**定理2.** 当投影函数 $g$ 是身份函数，并且我们通过简单地取内积来衡量嵌入相似性，即 $s(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}$ ，并进一步假设正数对远比负数对更对齐，最小化成对目标 $(\mathbf{u}_i, \mathbf{v}_i)$ 与最大化三联体损失相吻合，如后文所述

之

$$-l(\mathbf{u}_i, \mathbf{v}_i) \propto 4N\tau + \sum_{j=1}^J \mathbf{1}_{\{j \neq i\}} \left( \|\mathbf{u} - \mathbf{v}_i\|^2 - \|\mathbf{u} - \mathbf{v}_j\|^2 + \|\mathbf{u}_i - \mathbf{v}\|^2 - \|\mathbf{u}_i - \mathbf{v}_j\|^2 \right) \quad \circ$$

(24)

证明。基于假设，我们可以将成对目标重新排列为

$$\begin{aligned}
 -l(u_i, v_i) &= -\log \frac{\exp(u_i^T v_i / \tau)}{\sum_{k=1}^N \exp(u_i^T v_k / \tau) + \sum_{k=1}^N \mathbf{1}_{[k \neq i]} \exp(u_i^T u_k / \tau)} \\
 &= \text{对数} \quad 1 + \sum_{k=1}^N \mathbf{1}_{[k \neq i]} \exp \frac{u_i^T v_k - u_i^T v_i}{\tau} + \sum_{k=1}^N \mathbf{1}_{[k \neq i]} \exp \frac{u_i^T u_k - u_i^T u_i}{\tau}.
 \end{aligned} \tag{25}$$

通过一阶泰勒扩展、

$$\begin{aligned}
 -l(u_i, v_i) &\approx \sum_{k=1}^N \mathbf{1}_{[k \neq i]} \exp \frac{u_i^T v_k - u_i^T v_i}{\tau} + \sum_{k=1}^N \mathbf{1}_{[k \neq i]} \exp \frac{u_i^T u_k - u_i^T u_i}{\tau} \\
 &\approx 2 + \frac{1}{\tau} \sum_{k=1}^N \mathbf{1}_{[k \neq i]} (u_i^T v_k - u_i^T v_i) + \sum_{k=1}^N \mathbf{1}_{[k \neq i]} (u_i^T u_k - u_i^T u_i) \\
 &= 2 - \frac{1}{2\tau} \sum_{k=1}^N \mathbf{1}_{[k \neq i]} \|u_i - v_k\|^2 - \|u_i - v_i\|^2 + \|u_i - u_k\|^2 - \|u_i - u_i\|^2 \\
 &\propto \sum_{k=1}^N \mathbf{1}_{[k \neq i]} \|u_i - v_k\|^2 - \|u_i - v_i\|^2 + \|u_i - u_k\|^2 - \|u_i - u_i\|^2,
 \end{aligned} \tag{26}$$

这就结束了这个证明。□

## 参考文献

- [1] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: 社会表征的在线学习。在 *KDD*, 2014.
- [2] Aditya Grover 和 Jure Leskovec. node2vec: 可扩展的网络特征学习。在 *KDD*, 2016.
- [3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 词和短语的分布式表示及其构成性。在 *NIPS*, 2013.
- [4] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 网络嵌入作为矩阵分解: 统一 DeepWalk、LINE、PTE 和 node2vec。In *WSDM*, 2018.
- [5] Leonardo Filipe Rodrigues Ribeiro, Pedro H. P. Saverese, and Daniel R. Figueiredo. Struc2vec: 从结构特征中学习节点表征。In *KDD*, 2017.
- [6] Thomas N. Kipf and Max Welling. 用图卷积网络进行半监督性分类。在 *ICLR*, 2017.
- [7] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 图注意网络。在 *ICLR*, 2018.
- [8] Fenyu Hu, Yanqiao Zhu, Shu Wu, Liang Wang, and Tieniu Tan. Hierarchical Graph Convolutional Networks for Semi-supervised Node Classification. In *IJCAI*, 2019.
- [9] Thomas N. Kipf and Max Welling. Variational Graph Auto-Encoders. 在 *BDL@NIPS*, 2016.
- [10] William L. Hamilton, Zhitaoying, and Jure Leskovec. 大图上的归纳表征学习。In *NIPS*, 2017.
- [11] Ralph Linsker. 感知网络中的自我组织。 *IEEE Computer*, 1988.
- [12] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 通过非参数实例识别进行无监督特征学习。In *CVPR*, 2018.

- [13] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. *arXiv.org*, June 2019.
- [14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 无监督的视觉表征学习的动量对比。在 *CVPR*, 2020.

- [15] Philip Bachman, R. Devon Hjelm, and William Buchwalter.通过最大限度地提高跨视图的相互信息来学习表征。In *NeurIPS*, 2019.
- [16] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang.通过不变和传播实例特征的无监督嵌入学习。In *CVPR*, 2019.
- [17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton.A Simple Framework for Contrastive Learning of Visual Representations. *arXiv.org*, February 2020.
- [18] Petar Velic`kovic`, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm.Deep Graph Infomax.在*ICLR*, 2019年。
- [19] Suzanna Becker和Geoffrey E. Hinton。在随机点立体图中发现表面的自组织神经网络。《自然》, 355(6356), 1992。
- [20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis.通过预测图像旋转进行无监督的表征学习。In *ICLR*, 2018.
- [21] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich.着色作为视觉理解的代理任务。在*CVPR*, 2017.
- [22] R.Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio.通过互感信息估计和最大化学习深度表征。在*ICLR*, 2019年。
- [23] Aäron van den Oord, Yazhe Li, and Oriol Vinyals.Representation Learning with Contrastive Predictive Coding. *arXiv.org*, 2018.
- [24] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord.Data-Efficient Image Recognition with Contrastive Predictive Coding. *arXiv.org*, May 2019.
- [25] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker.On Variational Bounds of Mutual Information.In *ICML*, 2019.
- [26] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic.关于表征学习的互信最大化。In *ICLR*, 2020.
- [27] William L. Hamilton, Rex Ying, and Jure Leskovec.Representation Learning on Graphs: 方法和应用 *IEEE Data Eng.Bull.*, 2017.
- [28] Michael Gutmann和Aapo Hyvärinen.非正态化统计模型的噪声对比估计, 并应用于自然图像统计。 *JMLR*, 2012.
- [29] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger.Simplifying Graph Convolutional Networks.在*ICML*, 2019年。
- [30] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov.辍学: 防止神经网络过拟合的简单方法。 *JMLR*, 15(1), 2014.
- [31] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang.DropEdge: Towards Deep Graph Convolutional Networks on Node Classification.在*ICLR*, 2020.
- [32] Florian Schroff, Dmitry Kalenichenko, and James Philbin.FaceNet: 用于人脸识别和聚类的统一嵌入。In *CVPR*, 2015.
- [33] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad.网络数据中的集体分类。 *AI杂志*, 2008年。
- [34] Aleksandar Bojchevski and Stephan Günnemann.Deep Gaussian Embedding of Graphs: 通过排名进行无监督的归纳学习。在*ICLR*, 2018.

- [35] Marinka Zitnik和Jure Leskovec.通过多层组织网络预测多细胞功能.  
*Bioinform.*, 2017.
- [36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.图像识别的深度残差学习.In *CVPR*, 2016.
- [37] 陈杰, 马腾飞, 和曹晓.FastGCN: 通过重要性采样实现图卷积网络的快速学习。In *ICLR*, 2018.

- [38] Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit-Yan Yeung. GaAN: 用于在大型和时空图上学习的门控注意力网络。In *UAI*, 2018.
- [39] Zhilin Yang, William W. Cohen, and Ruslan R. Salakhutdinov. 重新审视用图嵌入进行的半监督学习。在 *ICML*, 2016.
- [40] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: 词汇表征的全局矢量。In *EMNLP*, 2014.
- [41] Matthias Fey and Jan Eric Lenssen. 使用PyTorch Geometric的快速图形表示学习。在 *rlgm@iclr*, 2019.
- [42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: 一个强制性的风格, 高性能的深度学习库。In *NeurIPS*, 2019.
- [43] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Python 中的机器学习。 *JMLR*, 2011.
- [44] Xavier Glorot和Yoshua Bengio. 了解训练深度前馈神经网络的难度。在 *AISTATS*, 2010.
- [45] Diederik P. Kingma 和 Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.