arXiv:2305.00623v1 [cs.LG] 1 May 2023

用于节点表示对比学习的简化框架

Ilgee Hong[1,2], Huy Tran[1,3], 和 Claire Donnat[1,4]

[1]统计学系，芝加哥大学[2,3,4]{*ilgee,huydtran,cdonnat*}
@uchicago.edu

摘要

近年来，对比学习已成为一种强大的自监督学习框架，可用于提取丰富且多样的数据表示。广义而言，对比学习依靠数据增强方案生成输入数据的两个版本，并通过最大化归一化温度缩放交叉熵损失（NT-Xent），学习低维表示，以识别对应于同一原始实体的增强样本（Chen 等人，2020年）。本文研究将对比学习与图神经网络（Scarselli 等人，2008年）结合，用于学习图中节点的嵌入的潜力。具体而言，我们表明，只需对嵌入矩阵进行简单的逐列后处理，而非采用多数同类方法使用的通过多层感知机（MLP）进行按行后处理，就能显著提升所得嵌入的质量并缩短训练时间。这一改动使下游分类任务的表现最高提升了1.5%，并且在8个不同基准数据集中的6个上优于现有最先进的方法。我们通过重新审视Wang 和 Isola（2020年）提出的“对齐与均匀性范式”，说明按列后处理能够同时改善嵌入的“对齐”和“均匀性”。

1 引言

对比学习已成为学习丰富且通用的数据表示最受欢迎的自监督方法之一，这一成功很大程度上归功于其在多个领域展现出的出色性能，从图像分类（Chen 等人，2020年；Grill 等人，2020年；He 等人，2020年；Hjelm，2018年；Oord 等人，2018年a；Tian 等人，2020年）到序列建模（Kong 等人，2019年；Oord 等人，2018年b；Wu 等人，2020年a）以及对抗学习（Ho 和 Nvasconcelos，2020年；Grill 等人，2020年）。总的来说，对比学习通常包括生成数据的两个（或更多）扰动版本，随后学习识别“正样本对”中的表示（即对应于同一原始实体的嵌入），并将它们嵌入到嵌入空间中彼此接近的位置。相反，“负样本对”（即对应于不同原始实体的嵌入）则被鼓励彼此远离。值得注意的是，这种在数据视图之间进行差异学习的简单“无模型”方法（Gutmann 等人，2022年）已被证明，在与简单的线性分类器结合时，能够在许多下游监督任务中取得最先进的性能（Chen 等人，2020年）。这一结果相当令人惊讶，因为该方法不仅相较于其他无监督嵌入算法表现出色，而且相较于有监督

表 1：自监督节点表示学习方法的架构比较

Method	Target	MI-Estimator	Proj/Pred head	Asymmetric	Negative samples
DGI	N-G	✓	-	-	✓
MVGRL	N-G	✓	✓	✓	✓
BGRL	N-N	-	✓	✓	-
CCA-SSG	F-F	-	-	-	-
GRACE	N-N	-	✓	-	✓
CLNR（我们的方法）	N-N	-	-	-	✓

表 1 是 Zhang 等人（2021年）的扩展。目标：比较对；N/G/F 分别表示节点/图/特征。互信息估计器：参数化互信息估计器。投影/预测头：图神经网络编码器之后的参数化投影头或预测头。非对称：非对称架构，例如针对两个不同数据增强的两个独立编码器，或指数移动平均和停止梯度。负样本：计算损失时需要负样本。

端到端方　　法，其主要目标是学习针对特定任务定制的嵌入　　.

图对比学习。 受其在各种机器学习应用中的成功启发，近年来，对比学习逐渐被图表示学习领域所采用。在此情况下，对比学习通常与图神经网络（GNN）（Scarselli 等人，2008年；Kipf 和 Welling，2016年a；Zhou 等人，2020年；Wu 等人，2020年（b））相结合，以学习新类别的节点嵌入。当前的图对比学习方法可根据所要最大化的互信息类型分为两类。

1. 局部到全局互信息最大化，例如 Deep Graph InfoMax（DGI）（Velickovic 等人，2019年）和多视图图表示学习（MV-GRL）（Hassani 和 Khasahmadi，2020年）。这些方法通过最大化不同尺度下原始数据表征之间的互信息来学习节点嵌入：DGI 将局部（节点）表征与全局图嵌入之间的互信息作为目标，其中，局部（节点）表征是将原始图输入图神经网络编码器后获得的，全局图嵌入则通过对所有局部表征求和获得。负样本由扰动函数单独生成。另一方面，MVGRL 利用扩散过程生成额外的图，并获得相应的额外节点表征和图表征对。来自不同图的一对节点和图表征被视为负样本对。DGI 和 MVGRL 都使用由神经网络参数化的互信息估计器，为每个局部和全局~表征对获取概率分数。更具体地说，令 G 和 G 分别表示原始图和扰动图，则 DGI 的损失函数写为：

$$\mathcal{L}=\frac{1}{N+M}\left(\sum_{i=1}^{N}\mathbb{E}_{G}\left[\log\mathcal{D}\left(\boldsymbol{u}_{i},\boldsymbol{s}\right)\right]+\sum_{j=1}^{M}\mathbb{E}_{\widetilde{G}}\left[\log\left(1-\mathcal{D}(\widetilde{\boldsymbol{u}}_{j},\boldsymbol{s})\right)\right]\right),$$

其中，$\mathcal{D}$：$\mathbb{R}^{F}\times\mathbb{R}^{F}\rightarrow\mathbb{R}$ 是由多层感知机参数化的互信息估计器，$\boldsymbol{s}\in\mathbb{R}^{F}$ 是原始图的摘要向量，而 $\boldsymbol{u}_{i}$、$\widetilde{\boldsymbol{u}}_{j}\in\mathbb{R}^{F}$ 是来自原始图的节点表示

图及其损坏版本，分别为 $i = 1$、.....、 N 和 $j = 1$、.....、 M。

2. 局部到局部互信息最大化，例如图对比表征学习（GRACE）（Zhu 等人，2020年）及其扩展方法——自适应增强图对比表示学习（GCA）（Zhu 等人，2021年）。与之前的方法不同，这些方法直接使用节点对之间的点积定义相似度分数，无需参数化互信息估计器。这些方法通常将节点表示投影到单位超球面上，这已知有助于提高训练稳定性，而且更容易被下游线性分类器分离。GRACE 和 GCA 通过随机扰动边和特征（通常结合随机边删除与特征掩码）来生成原始图的两个视图。请注意，这里的扰动用于生成同一数据的不同版本，而不是生成负例。随后，通过最大化节点对之间的归一化温度缩放交叉熵损失（NT-Xent）（Chen 等人，2020年）来学习节点表示，这相当于解决一个分类问题：对于每个节点，在每个数据版本中，算法都要从图的两个版本中的所有其他嵌入里识别出来自同一原始节点的嵌入。目标函数见公式（2）和（3）。尽管 GRACE 和 GCA 表现出色，但它们仍存在一些缺点。这两种方法都依赖于通过带有非线性激活函数的多层感知机对原始嵌入矩阵进行按行后处理。这一选择最初受到该方法在视觉领域对应方法（Chen 等人，2020年）的启发；该方法的实证结果表明，与采用线性激活函数或不进行后处理步骤相比，在 ResNet（He 等人，2016年）之后使用带非线性激活函数的两层多层感知机，可以提升表征的表达能力和质量。然而，额外使用多层感知机进行后处理会显著增加模型架构的复杂度，而且其对最终表征的影响仍未得到充分理解。

图的非对比学习。 前一组方法在处理大型图时存在一些计算困难，因为评估每个节点的归一化因子需要计算所有负样本对之间的成对余弦相似度。因此，诸如自举图潜变量（BGRL）（Thakoor 等人，2021年）和受典型相关分析启发的图自监督学习（CCA-SSG）（Zhang 等人，2021年）等较新的方法提出采用不包含负样本的非对比目标。例如，在 CCA-SSG 中，目标是直接最小化正样本对之间的距离，同时使不同维度上的特征去相关。目标函数表示为：

$$\mathcal{L} = \|U - V\|_F^2 + \lambda \left(\left\|U^T U - I\right\|_F^2 + \left\|V^T V - I\right\|_F^2 \right)$$

其中， U， $V \in \mathbb{R}^{N \times F}$ 是原始图两个扰动版本的节点表示矩阵，$\lambda > 0$ 是超参数， $\|\cdot\|_F$ 表示弗罗贝尼乌斯范数。因此，BGRL 和 CCA-SSG 都通过忽略与负样本相关的计算，降低了计算复杂度。不过，BGRL 需要指数移动平均（EMA）和停止梯度等额外组件，这使得该方法成本更高，也更难解释。尽管 CCA-SSG 采用了简单且对称的架构，不需要额外组件，但这种特征级对比方法通常需要更大的嵌入维度，才能提升其表征质量。

贡献。本文的贡献有三点。(1) 我们提出了一个用于节点表示对比学习的简化框架，并将我们的方法称为节点表示对比学习（***CLNR***）。CLNR 沿用了 GRACE（Zhu 等人，2020年）和 GCA（Zhu 等人，2021年）的核心结构：通过随机数据增强生成输入图的两个视图，使用共享 GNN 编码器，并通过最大化 NT-Xent 损失（Chen 等人，2020年）获得节点表示。与 GRACE 和 GCA 不同，CLNR 通过将使用多层感知机对节点表示矩阵进行的按行后处理替换为简单的按列标准化，简化了模型架构。我们证明，这一替换显著提升了表征质量并缩短了模型训练时间。(2) 我们进一步阐释了各种后处理步骤的作用，并通过重新审视Wang 和 Isola（2020年）提出的"对齐与均匀性范式"，论证了我们对后处理（按列标准化）的选择。该范式旨在分析单位超球面上表征的质量。(3) 我们通过将生成的节点嵌入用作下游分类的输入，提供了其质量的实验依据。结果表明，与包括 CCA-SSG 和 GRACE 在内的最先进方法相比，CLNR 表现更优。

符号说明。对于给定的图 $G = (X, A)$，$X \in \mathbb{R}^{N \times D}$ 是特征矩阵，$A \in \{0,1\}^{N \times N}$ 是邻接矩阵。这里，N 表示图 G中的节点数，而 D 表示特征维度。

论文结构。我们在第2节介绍用于节点表征的一般对比学习框架，并提出我们的简化框架。第3 节给出实验结果，第4节讨论各种后处理步骤。第5 节给出补充实验，第6节总结全文。

2 用于节点表征的对比学习

我们首先介绍用于节点嵌入的通用对比学习框架，并重点研究嵌入后处理的选择。

2.1 模型框架

图增强。我们遵循Zhu 等人（2020年）提出的标准随机图增强流程，该流程包含两个部分：*(1)* 边丢弃 和*(2)* 特征掩码。因此，该过程依赖于两个不同的超参数："边丢弃率"（即每条边被丢弃的伯努利概率，各条边相互独立）和"特征掩蔽率"（即特征矩阵中每个元素被掩蔽的概率，各元素相互独立）。这组扰动允许在每个训练轮次生成同一输入图的两个不同版本 G 。将这两个版本分别记为 G_1 （X_1，A_1）和 G_2（X_2，A_2）。

共享 GNN 编码器。我们使用共享的图神经网络（GNN）（ G_1 和 G_2 将两者中的节点嵌入 Scarselli 等人，2008年）。令 f_θ 为共享的 GNN 编码器。我们将原始嵌入矩阵记为 $f_\theta(X_1, A_1) = Z_1$ 和 $f_\theta(X_2, A_2)$，其中 Z_1、$Z_2 \in \mathbb{R}^{N \times F}$ 和 F 表示嵌入维度。

后处理步骤。 前一个卷积步骤的结果是针对输入图两个版本的一组原始嵌入矩阵 Z_1 和 Z_2 。经验表明，对这些原始嵌入进行进一步后处理可以显著提升其表达能力。常见选项包括：

（1）按行后处理，在这种情况下，每个节点嵌入都独立于其他节点进行处理。令$g_{\boldsymbol{\eta}}$ 为一个具有非线性激活函数的两层多层感知机。对于两个版本的数据，最终的节点嵌入矩阵可表示为$g_{\boldsymbol{\eta}}(Z_1) = U$ 和 $g_{\boldsymbol{\eta}}(Z_2) = V$。例如，GRACE （Zhu 等人，2020 年）采用的正是这种设置，其架构与 SimCLR（Chen 等人，2020年）非常相似。后者表明，在 ResNet（He 等人，2016年）中添加具有非线性激活函数的两层多层感知机，在视觉任务上的表现优于其他方案，例如使用线性激活函数的多层感知机，或不进行后处理（Chen 等人，2020年；Bachman 等人，2019年）。

（2）按列后处理——更具体地说，是按列标准化，即对原始嵌入的每个维度进行中心化和重新缩放。因此，每个节点嵌入都与同一批次中的其他嵌入一同处理，而不再是独立处理。这种处理机制在深度学习文献中称为“批归一化”（BN）（Ioffe 和 Szegedy，2015年；Santurkar 等人，2018年；Bjorck 等人，2018年）。训练深度神经网络（DNNs）时，通常会在层间的激活函数之前使用 BN 对输入进行标准化。BN 最初旨在缓解内部协变量偏移，即训练过程中网络参数变化所导致的网络激活值分布变化（Ioffe 和 Szegedy，2015年）。随后研究发现，BN 通过平滑优化景观（Santurkar 等人，2018年），并允许采用更大的学习率（Bjorck 等人，2018年），能够提升模型训练的速度与稳定性，以及泛化准确率。在此，我们提出一种在自监督对比学习中将 BN 用作后处理步骤的新方法。我们认为，常用方法中的按行多层感知机后处理使算法架构过于复杂，却没有带来明显的性能提升。相反，在第3节和第4节中，我们将展示，按列 BN 后处理能够得到更好的表征；对比学习文献中有关对齐与均匀性范式的近期研究结果也支持这一结论（Wang 和 Isola，2020年）。我们将采用这种后处理方式的对比学习框架称为节点表征对比学习方法（*CLNR*）。最终的节点嵌入可表示如下：

$$U = \frac{Z_1 - \text{mean}(Z_1)}{\text{std}(Z_1)} \quad \text{and} \quad V = \frac{Z_2 - \text{mean}(Z_2)}{\text{std}(Z_2)}, \tag{1}$$

其中，对于任意矩阵 $A \in \mathbb{R}^{n \times q}$，$\text{mean}(A) \in \mathbb{R}^{n \times q}$ 和 $[\text{mean}(A)]_{i,\,j} = \sum_{k=1}^{n} a_{k,\,j}/n$ 以及 $\text{std}(A) \in \mathbb{R}^{n \times q}$ 和 $[\text{std}(A)]_{i,\,j} = \sqrt{\sum_{k=1}^{n}(a_{k,\,j} - [\text{mean}(A)]_{i,\,j})/n}$。

投影到超球面。无论选择哪种后处理方式，最后都会将嵌入投影到单位超球面上。这最后一步源自大量深度学习文献（Bojanowski 和 Joulin，2017；Chen 等人，2020年；Davidson 等人，2018年；Meng 等人，2019年；Xu 和 Durrett，2018年）。对最终嵌入矩阵（例如，ℓ_2 以及U 和 V）进行归一化，已被证实能够提高学习过程的稳定性（Wang 和 Isola，2020年；Xu 和 Durrett，2018年），并使下游线性分类器更容易分离这些嵌入。

对比损失函数。我们采用 Zhu 等人（2020年）提出的对比目标函数，称为“NT-Xent 损失”；这一术语最早由 Chen 等人（2020年）使用。与 GRACE 不同，GRACE 会计算图中所有节点对的余弦相似度；我们在每个训练轮次随机抽取m 个节点，并仅计算这些节点之间的余弦相似度。当 m 足够大时，这种方法能够很好地扩展到大型图，同时不会牺牲准确率。对于随机小批次中的每个节点 $i = 1$，...，m，我们令$\boldsymbol{u}_i = U_i$、:和$\boldsymbol{v}_i = V_i$、:分别为 U和V的第 i行。然后，对于每个 $i = 1$，...，m，若 $\boldsymbol{u}_i$是目标节点，则将（$\boldsymbol{u}_i$，$\boldsymbol{v}_i$）视为正样本对，而（$\boldsymbol{u}_i$，$\boldsymbol{v}_k$）和（$\boldsymbol{u}_i$，$\boldsymbol{u}_k$）——其中 $k \neq i$——自然地视为负样本对。请注意，第一组负样本对来自跨视图节点，第二组负样本对来自视图内节点。令 s： $\mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}$ 为余弦相似度函数（定义为 s（$\boldsymbol{u}$，$\boldsymbol{v}$）$= \boldsymbol{u} \cdot \boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$）。我们的目标函数具有如下明确形式：

$$\ell(\boldsymbol{u}_i, \boldsymbol{v}_i) = \log \frac{\overbrace{e^{s(\boldsymbol{u}_i, \boldsymbol{v}_i)/\tau}}^{positive\ pair}}{\underbrace{e^{s(\boldsymbol{u}_i, \boldsymbol{v}_i)/\tau}}_{positive\ pair} + \underbrace{\sum_{k=1}^{m} \mathbf{1}_{[\mathrm{k}\neq\mathrm{i}]} \left[e^{s(\boldsymbol{u}_i, \boldsymbol{u}_k)/\tau} + e^{s(\boldsymbol{u}_i, \boldsymbol{v}_k)/\tau} \right]}_{negative\ pairs}}, \tag{2}$$

其中，τ表示温度参数。值得注意的是，一般而言，$\ell(\boldsymbol{u}_i, \boldsymbol{v}_i) \neq \ell(\boldsymbol{v}_i, \boldsymbol{u}_i)$。因此，对所有可能的配对取平均后，待最大化的最终目标函数为

$$\mathcal{L} = \frac{1}{2m} \left(\sum_{i=1}^{m} [\ell(\boldsymbol{u}_i, \boldsymbol{v}_i) + \ell(\boldsymbol{v}_i, \boldsymbol{u}_i)] \right). \tag{3}$$

3 实验

本节将我们的简化版节点对比学习框架应用于多个真实数据集，并与现有的同类方法进行比较，以展示该框架的潜力。

真实数据集。 我们使用真实数据上的**节点分类任务**来评估节点表征的质量。我们考虑了八个节点分类基准数据集：*Cora*、*CiteSeer*、*Pubmed*、*Amazon Computers*、*Amazon Photo*、*Coauthor CS*、*Coauthor Physics*、*ogbn-arxiv*。前七个基准数据集由Yang 等人(2016年)和Shchur 等人(2018年)使用，而*ogbn-arxiv* 来自开放图基准数据集（Hu 等人，2020年）。我们采用引文网络数据集（*Cora*、*Citeseer*、*Pubmed*）的公开划分，以及 Amazon 和合著网络数据集（*Amazon Computer*、*Amazon Photo*、*Coauthor CS*、*Coauthor Physics*）的 1:1:8 训练集/验证集/测试集划分。数据集的详细信息见附录A.1。

实验设置。 我们遵循 Velickovic 等人 (2019年）提出的评估协议，该协议包括两个阶段：1）预训练和 2）线性评估。首先，我们使用公式（2）中描述的 NT-Xent 损失训练图神经网络编码器，从而获得整个图的节点嵌入。随后，我们在部分节点（训练集）上拟合一个简单的线性分类器。然后，我们在剩余节点（测试集）上评估分类准确率。所有实验均使用 PyTorch 和 PyTorch Geometric 模块，在具有 16 GB 内存的图形处理器上完成。

表2：节点分类准确率

METHODS	INPUT	CORA	CITESEER	PUBMED	COMPUTERS	PHOTO	CS	PHYSICS	OGBN-ARXIV
GAE	X, A	71.5±0.4	65.8±0.4	72.1±0.5	85.3±0.2	91.6±0.1	90.0±0.7	94.9±0.1	52.57±0.1
DGI	X, A	82.3±0.6	71.8±0.7	76.8±0.6	84.0±0.5	91.6±0.2	92.2±0.6	94.5±0.5	67.95±0.6
MVGRL	X, A, S	83.5±0.4	**73.3±0.5**	80.1±0.7	87.5±0.1	91.7±0.1	92.1±0.1	95.3±0.1	OOM
CCA-SSG	X, A	84.2±0.4	73.1±0.3	81.6±0.4	88.7±0.3	93.1±0.1	93.3±0.2	95.4±0.1	60.5 ± 0.2
BGRL	X, A	82.1±1.0	71.8±0.5	80.6±1.0	89.0±0.4	92.7±0.4	92.0±0.2	94.5±0.3	70.3±0.1
GRACE	X, A	83.6±0.8	71.8±1.1	81.6±0.9	89.1±0.4	**93.3±0.4**	93.6±0.3	95.3±0.1	69.8 ± 0.1
CLNR	X, A	**84.3±0.6**	73.1±0.7	**83.0 ± 0.6**	**89.5±0.4**	93.2±0.3	**93.8±0.2**	**95.5±0.1**	**70.4 ± 0.3**
GCN	X, A, Y	81.5±0.7	70.3±0.7	79.0±0.3	86.5±0.5	92.4±0.2	93.0±0.3	95.7±0.2	71.7±0.3
GAT	X, A, Y	83.0±0.7	72.5±0.7	79.0±0.3	86.9±0.3	92.6±0.4	92.3±0.2	95.5±0.2	71.7±0.3

在 7 个基准数据集上进行节点分类任务时的平均分类准确率及标准差。输入： X：节点特征， A：邻接矩阵， S：扩散矩阵，以及 Y：节点标签。对于 GRACE 和 CLNR，我们在公式（ $m = 1024$ ）所示的损失中使用 $m = 1024$ 2）。OOM 表示使用 16GB 图形处理器时发生内存溢出。

实现细节。对于第一组实验，我们采用 f_θ标准的两层图卷积网络模型作为图编码器（Kipf 和 Welling，2016年a），在除*CiteSeer*、*Coauthor CS*和*ogbn-arxiv*之外的所有数据集上，隐藏维度和输出维度均相同。与Zhang 等人（2021年）一样，我们发现，对于*CiteSeer*，一层图卷积网络的效果更好；对于*Coautor CS*，两层多层感知机比两层图卷积网络效果更好。由于任务难度较高（*ogbn-arxiv* 包含40个类别，而其他数据集的类别数均少于15个），我们为*ogbn-arxiv*使用三层图卷积网络。第二阶段中，我们使用简单的逻辑回归作为线性分类器。我们采用 Adam 优化器（Kingma 和 Ba，2014年）。我们对模型参数和数据划分分别进行20次随机初始化，并报告均值准确率及其标准差。有关超参数调优的更多细节见附录A.2。

结果。为确保公平比较，我们为所有方法采用一致的训练设置：在所有方法和所有基准数据集上，均使用 Adam 优化器和不带 ℓ_2 正则化的简单逻辑回归。表 2 通过报告分类准确率的均值（和标准差），比较了 CLNR 与现有的同类方法。我们首先关注 CLNR（按列后处理）与 GRACE（按行后处理）之间的比较。可以看到，在除 Photo 之外的所有基准数据集上，CLNR 的表现均优于 GRACE，最高提升 1.5%。这表明，通常按列后处理比按行后处理能获得更好的表示。我们还将 CLNR 与其他现有方法进行了比较。具体而言，我们将 CLNR 与一种经典的无监督模型：图自编码器（Kipf 和 Welling，2016年b）、五种自监督模型：DGI（Velickovic 等人，2019年）、MVGRL（Hassani 和 Khasahmadi，2020年）、BGRL（Thakoor 等人，2021年）、CCA-SSG（Zhang 等，2021年），以及两种监督基线：GCN（Kipf 和 Welling，2016年a）和 GAT（Veli ˇckovi ´c 等人，2017年）进行了比较。从表 2 可以看出，在除 CiteSeer（MVGRL 表现更优）之外的所有基准数据集上，CLNR 的表现均优于其他所有无监督和自监督方法。除 Coauthor Physics 和 ogbn-arxiv 外，在所有基准数据集上，CLNR 的表现甚至优于 GCN 和 GAT 这两种监督基线，提升幅度最高达 1 到 4%。另一方面，在使用非对比损失的方法中，

CCA-SSG 在除*Computer* 和*ogbn-arxi*之外的所有基准数据集上均优于 BGRL $v.$

训练时间。 还值得强调的是，CLNR 通常只需较少的训练轮次即可生成最佳节点表示。这得益于批归一化的另一项优势：它允许使用更高的学习率，并使优化景观更加平滑，从而加快模型训练。此外，CLNR 无需在多层感知机中学习额外参数。表 3 展示了 GRACE 与 CLNR 在模型训练时间方面的对比。我们观察到，CLNR 的训练速度最高可比 GRACE 快 20 倍。

表3：CLNR 和 GRACE 的训练时间。

Model	Dataset	Training time
CLNR	ogbn-arxiv	**2879.75s**
GRACE	ogbn-arxiv	7286.71s
CLNR	Photo	**7.87s**
GRACE	Photo	143.25s
CLNR	Computers	**25.62s**
GRACE	Computers	104.59s
CLNR	PubMed	**134.51秒**
GRACE	PubMed	538.75s

4 重新审视“对齐与均匀性范式”

在本节中，我们将深入比较 CLNR 和 GRACE，并尝试解释 CLNR 所获得的嵌入为何更优。为此，我们重新审视Wang 和 Isola (2020年提出的“对齐与均匀性范式”。

对齐与均匀性。 鉴于将特征映射到单位超球面并仅对其进行分析取得了成功，Wang 和 Isola (2020年) 提出从两个量的角度分析对比损失： ***(a)*** 对齐，用于量化正样本对的表征（即对应于同一实体的嵌入）彼此映射得有多接近：

$$\ell_{\text{align}} = \mathbb{E}_{(x,y)\sim p_{\text{pos}}} \|f(x) - f(y)\|_2^{\alpha}. \tag{4}$$

在前一个表达式中， $f(x)$ 表示节点的嵌入， $\alpha > 0$ 是任意正指数（通常取 2）； 以及 ***(b)*** 均匀性，定义为平均成对高斯势的对数：

$$\ell_{\text{uniform}}(f,t) = \log \mathbb{E}_{x,y \overset{\text{i.i.d}}{\sim} p_{\text{data}}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right], \quad t > 0. \tag{5}$$

均匀性本质上衡量特征在单位球面上的分布是否均匀。 Wang 和 Isola (2020年) 确实指出，在超球面上分布得越均匀的数据分布保留的信息越多。随后，他们证明 NT-Xent 损失渐近地可以表示为

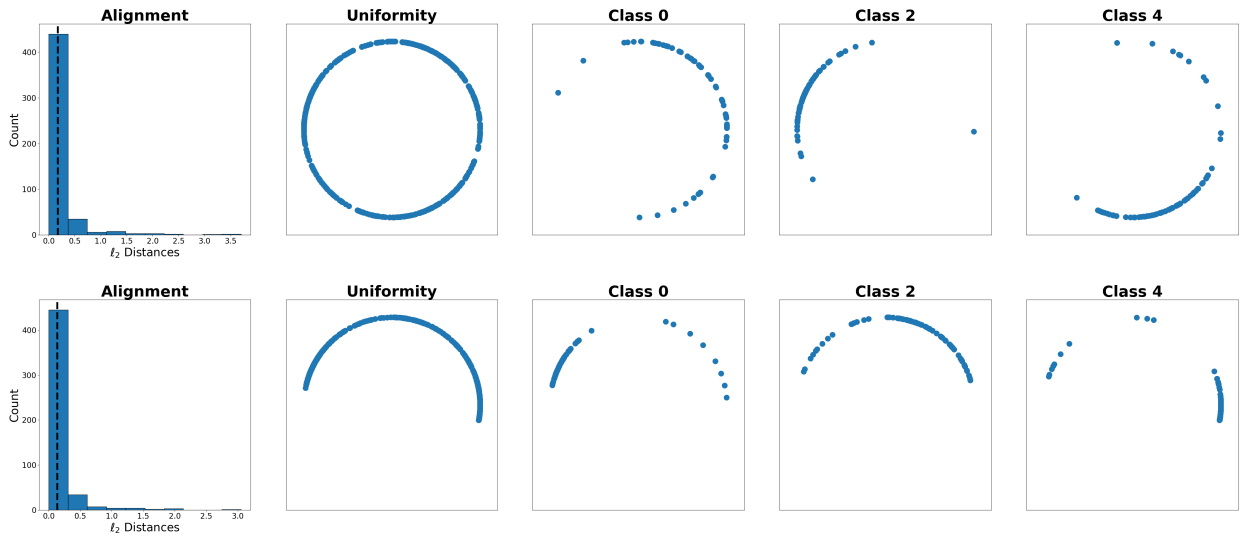

图 1：使用 CLNR（上）和 GRACE（Zhu 等人，2020年）（下）在二维单位超球面上呈现的 Cora 验证集节点表示。最左侧图中的虚线表示平均对齐度指标。请注意，与 GRACE 相比，CLNR 的均匀性有了显著提升。

均匀性项与对齐项之和，因此它同时优化这两者。基于这些观察，作者认为，与使用经典的负交叉熵损失相比，显式地同时优化均匀性和对齐可能会取得更好的结果。然而，他们提出的加权损失 $\ell = \ell_{\text{uniform}} + \lambda \ell_{\text{align}}$ 有一个缺点：它引入了另一个需要通过交叉验证来确定的参数 λ 。

通过批归一化增强均匀性。Wang 和 Isola（2020年）通过实证表明，更好的表征具有更低的对齐度和均匀性指标值。借助这些发现，我们将批归一化的成功归因于它增强表征均匀性的能力。事实上，批归一化会缩放嵌入空间的所有维度，从而使所有嵌入方向具有可比性，并使嵌入空间呈各向同性。对于包括主成分分析在内的其他无监督学习技术而言，这种标准化实际上是常规做法，旨在防止数据的某个维度凌驾于其他维度之上（Hastie 等人，2009）。同样，我们认为，批归一化 BN 能够稳定各个维度的贡献，从而更容易优化均匀性。我们认为，这种批归一化足以促进均匀性，而*Wang 和 Isola （2020年）*已证明，均匀性有益于嵌入质量；此外，它还有一个显著优势，即无需进行超参数调优。

为这一现象提供实证依据，并清晰观察每种后处理的效果，我们还考虑了以下变体：

nCLNR： 我们首先将 CLNR 和 GRACE 与一种使用恒等映射作为后处理（即不进行后处理）的对应方法（称为 $nCLNR$）进行比较。这将作为我们的消融基准，用于验证后处理步骤的重要性。

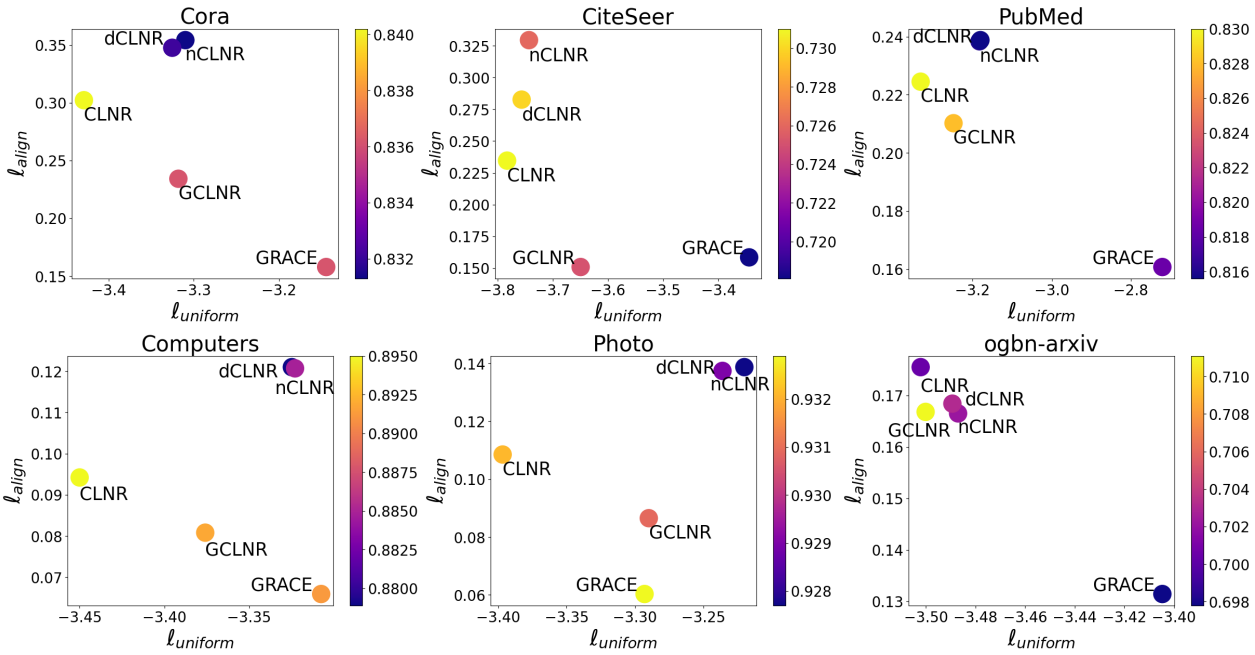

图 2：比较 nCLNR、CLNR、GRACE 及其变体 dCLNR 和 GCLNR 在各自的对齐度量和均匀性度量方面的表现（公式（4）和（5））。每个点代表一种方法，颜色渐变表示相应的节点分类准确率。

dCLNR： 借鉴深度神经网络领域关于这一主题的丰富文献，按列后处理的另一种选择是使用去相关批归一化（dBN）（Huang 等人，2018年）。dBN 在先前的标准化技术基础上进一步发展，并进一步促使不同维度上的特征彼此不相关。这有望进一步改善 Hessian 矩阵的条件数，但在标准文献中，通常被认为实现成本过高。为丰富我们的讨论，我们也考虑 dBN 后处理步骤，并将采用这种后处理方式的对比学习框架称为 $dCLNR$。

GCLNR： 这一最后的变体对应一种后处理方法：先使用带有非线性激活函数的 2 层多层感知机，然后进行逐列标准化（即同时采用逐行和逐列后处理）。

我们首先在图1中绘制了使用 GRACE 和我们的方法 CLNR 在 Cora 数据集上学到的表征。如预期所示，我们基于批次归一化的方法显著提高了所学表征的均匀性。

图 2 展示了在 6 个不同数据集上，各方法达到的对齐度（公式（4））与均匀性（公式（5））指标。我们观察到，在 6 个基准数据集中，有 5 个数据集上的单独 CLNR 位于 nCLNR（无后处理）的左下方。这表明，与不进行后处理相比，按列后处理同时提升了对齐度和均匀性。另一方面，GRACE 始终位于图的右下角。这表明，按行后处理以牺牲大量均匀性为代价，着重提升表征的对齐度

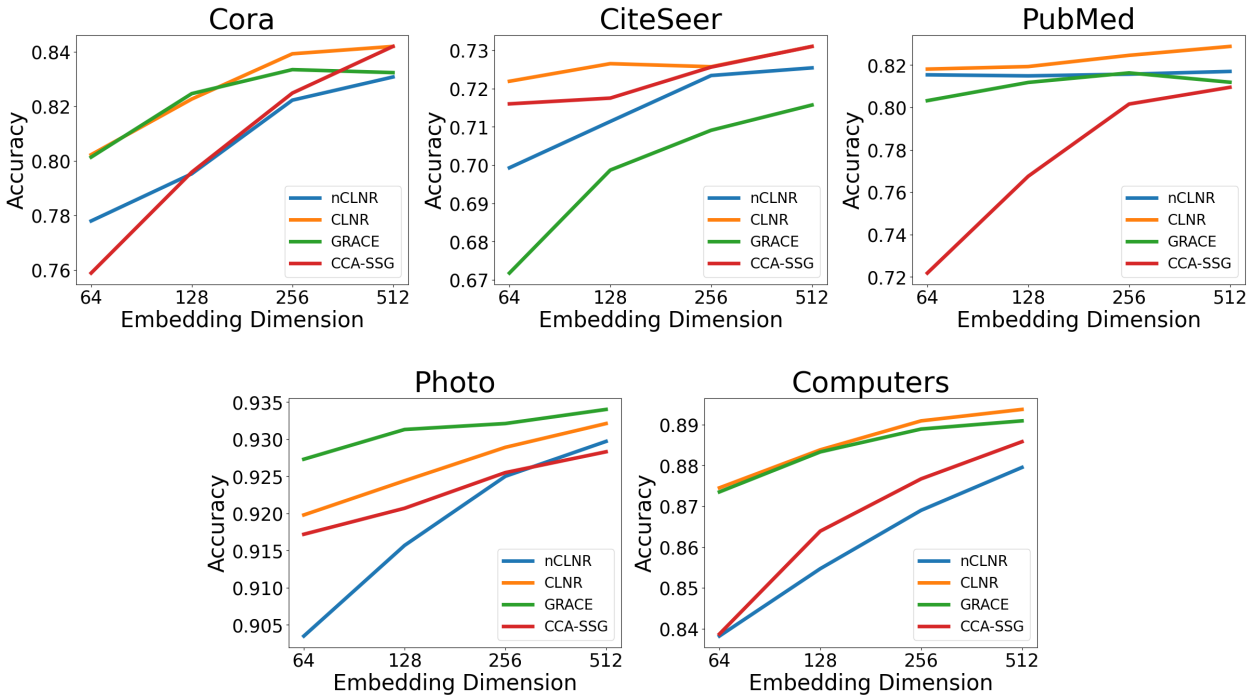

图 3：nCLNR、CLNR、GRACE 和 CCA-SSG 随嵌入维度变化的对比。

。唯一的例外是*Photo*，在该数据集上，GRACE 也位于 nCLNR 的左下方，并且节点分类准确率优于 CLNR。我们可能会预期，同时采用按行和按列后处理的 GCLNR，相比只采用其中一种方法的模型，能够提升表征质量。有趣的是，我们观察到 GCLNR 始终位于 CLNR 和 GRACE 之间，这表明一旦完成按行后处理，就无法通过按列后处理完全恢复均匀性。令人惊讶的是，额外对不同特征进行去相关处理的 dCLNR，其对齐度和均匀性指标与 nCLNR 相近。

5 其他实验

本文最后讨论 CLNR 相较于现有方法的其他优势。讨论重点有两方面：（a）嵌入方法的效率（以嵌入大小变化时下游节点分类器的准确率衡量）；（b）对扰动的鲁棒性。在整个讨论中，我们将非对比式竞争方法 CCA-SSG（Zhang 等人，2021年）纳入对比，因为它具有出色的性能和可扩展性。

嵌入效率。现在，我们研究嵌入维度对节点表示质量的影响。从图 3可以看出，CLNR 在所有基准数据集的各个嵌入维度上都表现出稳定且具有竞争力的性能。GRACE 也表现出

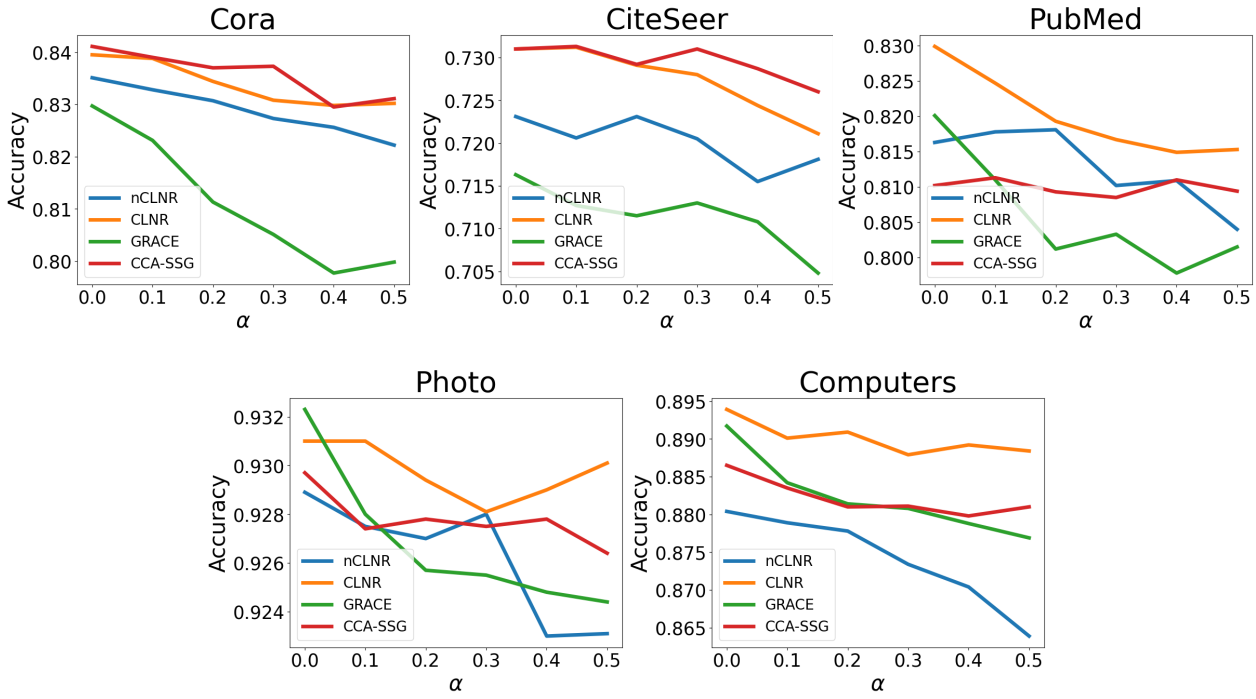

图4：nCLNR、CLNR、GRACE和CCA-SSG在边扰动鲁棒性方面的比较。

在除 *CiteSeer* 之外的所有基准数据集和嵌入维度上均表现良好。相反，我们发现，不对原始嵌入进行后处理的 nCLNR，以及采用非对比损失的 CCA-SSG，对维度的变化都更加敏感。

方法的鲁棒性。最后，我们比较nCLNR、CLNR、GRACE和CCA-SSG在边扰动方面的鲁棒性。设 $|A|$ 为输入图 G中的边数。对于给定的 $p \in (0,1)$，我们生成包含 $\Delta_A \in \mathbb{R}^{N\times N}$ $|A| \times p$ 条~随机无向边的邻接矩阵噪声。随后，我们生成扰动后的邻接矩阵 $A = A + \Delta_A$，并使用扰动后的图 ~ ~训练GCN编码器 $G = (X, A)$，而非使用 G。结果汇总于图4中。图4 展示了随着 p值增大，四种模型在5个基准数据集上的节点分类准确率。我们观察到，在除*CiteSeer*之外的所有基准数据集上，CLNR始终优于其他所有方法。

6 结论

我们介绍了用于节点表示的自监督对比学习（CLNR），这是一种简化而有效的节点表示学习方法。我们用简单的批归一化取代 GRACE（Zhu 等人，2020年）中计算开销较大的基于多层感知机的按行后处理，从而简化了现有基于 NT-Xent 损失的对比学习方法的架构，并显著提升了其性能、可扩展性和鲁棒性。为了解释这一成功，我们重新审视了对齐与均匀性范式。我们提供了实证证据，表明逐列

后处理足以同时改善嵌入的对齐性和均匀性，而按行后处理则以显著损失均匀性为代价，改善了节点表示的对齐性。

参考文献

P. Bachman、R. D. Hjelm 和 W. Buchwalter。通过最大化视图间的互信息来学习表征。神经信息处理系统进展，32，2019年。N. Bjorck、C. P. Gomes、B. Selman 和 K. Q. Weinberger。理解批归一化。神经信息处理系统进展，31，2018年。P. Bojanowski 和 A. Joulin。通过预测噪声进行无监督学习。载于国际机器学习会议，第 517–526 页。PMLR，2017年。T. Cali´nski 和 J. Harabasz。用于聚类分析的树突法。《统计学通讯——理论与方法》，3(1):1–27，1974年。T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton。用于视觉表示对比学习的简明框架。载于国际机器学习会议，第 1597–1607 页。PMLR，2020年。T. R. Davidson、L. Falorsi、N. De Cao、T. Kipf 和 J. M. Tomczak。超球面变分自编码器。*arXiv*预印本 *arXiv:1804.00891*，2018年。D. L. Davies 和 D. W. Bouldin。一种聚类分离度量。*IEEE*模式分析与机器智能汇刊，(2):224–227，1979年。J.-B. Grill、F. Strub、F. Altch´e、C. Tallec、P. Richemond、E. Buchatskaya、C. Doersch、B. Avila Pires、Z. Guo、M. Gheshlaghi Azar 等。自举式潜在表示学习：一种自监督学习的新方法。神经信息处理系统进展，33:21271–21284，2020年。M. U. Gutmann、S. Kleinegesse 和 B. Rhodes。对比学习的统计学应用。《行为计量学》，第 1–25 页，2022年。K. Hassani 和 A. H. Khasahmadi。图上的对比式多视图表示学习。载于国际机器学习会议，第 4116–4126 页。PMLR，2020年。T. Hastie、R. Tibshirani、J. H. Friedman 和 J. H. Friedman。《统计学习要素：数据挖掘、推断与预测》，第 2 卷。施普林格，2009年。K. He、X. Zhang、S. Ren 和 J. Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016年。K. He、H. Fan、Y. Wu、S. Xie 和 R. Girshick。用于无监督视觉表示学习的动量对比。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9729–9738 页，2020年。R. D. Hjelm、A. Fedorov、S. Lavoie-Marchildon、K. Grewal、P. Bachman、A. Trischler 和 Y. Bengio。通过互信息估计与最大化学习深度表征。*arXiv*预印本 *arXiv:1808.06670*，2018年。

C.-H. Ho 和 N. Nvasconcelos。利用对抗样本进行对比学习。神经信息处理系统进展，33:17081–17093，2020年。W. Hu、M. Fey、M. Zitnik、Y. Dong、H. Ren、B. Liu、M. Catasta 和 J. Leskovec。开放图基准：用于图上机器学习的数据集。神经信息处理系统进展，33:22118–22133，2020年。L. Huang、D. Yang、B. Lang 和 J. Deng。去相关批归一化。载于*IEEE*计算机视觉与模式识别会议论文集，第 791–800 页，2018年。S. Ioffe 和 C. Szegedy。批归一化：通过减少内部协变量偏移加速深度网络训练。载于国际机器学习会议，第 448–456 页。PMLR，2015年。D. P. Kingma 和 J. Ba。Adam：一种随机优化方法。*arXiv*预印本 *arXiv:1412.6980*，2014年。T. N. Kipf 和 M. Welling。基于图卷积网络的半监督分类。*arXiv*预印本 *arXiv:1609.02907*，2016年a。T. N. Kipf 和 M. Welling。变分图自编码器。*arXiv*预印本 *arXiv:1611.07308*，2016年b。L. Kong、C. d. M. d'Autume、W. Ling、L. Yu、Z. Dai 和 D. Yogatama。语言表示学习的互信息最大化视角。*arXiv*预印本 *arXiv:1910.08350*，2019年。J. McAuley、C. Targett、Q. Shi 和 A. Van Den Hengel。基于图像的风格与替代品推荐。载于第 *38* 届 *ACM SIGIR*信息检索研究与发展国际会议论文集，第 43–52 页，2015年。Y. Meng、J. Huang、G. Wang、C. Zhang、H. Zhuang、L. Kaplan 和 J. Han。球面文本嵌入。神经信息处理系统进展，32，2019年。A. v. d. Oord、Y. Li 和 O. Vinyals。基于对比预测编码的表示学习。*arXiv*预印本 *arXiv:1807.03748*，2018年a。A. v. d. Oord、Y. Li 和 O. Vinyals。基于对比预测编码的表示学习。*arXiv*预印本 *arXiv:1807.03748*，2018年b。P. J. Rousseeuw。轮廓系数：用于解释和验证聚类分析的图形化辅助工具。计算与应用数学杂志，20:53–65，1987年。S. Santurkar、D. Tsipras、A. Ilyas 和 A. Madry。批归一化如何帮助优化？神经信息处理系统进展，31，2018年。F. Scarselli、M. Gori、A. C. Tsoi、M. Hagenbuchner 和 G. Monfardini。图神经网络模型。*IEEE*神经网络汇刊，20(1):61–80，2008年。P. Sen、G. Namata、M. Bilgic、L. Getoor、B. Galligher 和 T. Eliassi-Rad。网络数据中的集体分类。《*AI Magazine*》，29(3):93–93，2008年。

O. Shchur、M. Mumme、A. Bojchevski 和 S. G¨unnemann。图神经网络评估中的陷阱。2018年。A. Sinha、Z. Shen、Y. Song、H. Ma、D. Eide、B.-J. Hsu 和 K. Wang。Microsoft Academic Service（MAS）及其应用概述。载于第*24*届万维网国际会议论文集，第243–246页，2015年。S. Thakoor、C. Tallec、M. G. Azar、M. Azabou、E. L. Dyer、R. Munos、P. Veli˘ckovi´c 和 M. Valko。通过自举在图上进行大规模表示学习。*arXiv*预印本 *arXiv:2102.06514*，2021年。Y. Tian、D. Krishnan 和 P. Isola。对比式多视图编码。载于欧洲计算机视觉会议，第776–794页。施普林格，2020年。P. Veli˘ckovi´c、G. Cucurull、A. Casanova、A. Romero、P. Lio 和 Y. Bengio。图注意力网络。*arXiv*预印本 *arXiv:1710.10903*，2017年。P. Velickovic、W. Fedus、W. L. Hamilton、P. Li`o、Y. Bengio 和 R. D. Hjelm。深度图信息最大化。国际学习表征会议（海报），2(3):4，2019年。K. Wang、Z. Shen、C. Huang、C.-H. Wu、Y. Dong 和 A. Kanakia。Microsoft Academic Graph：专家不足时的应对之道。定量科学研究，1(1):396–413，2020年。T. Wang 和 P. Isola。通过超球面上的对齐与均匀性理解对比式表征学习。载于国际机器学习会议，第9929–9939页。PMLR，2020年。Z. Wu、S. Wang、J. Gu、M. Khabsa、F. Sun 和 H. Ma。CLEAR：用于句子表征的对比学习。*arXiv*预印本 *arXiv:2012.15466*，2020年（a）。Z. Wu、S. Pan、F. Chen、G. Long、C. Zhang 和 S. Y. Philip。图神经网络综合综述。*IEEE* 神经网络与学习系统汇刊，32(1):4–24，2020年（b）。J. Xu 和 G. Durrett。用于稳定变分自编码器的球形潜在空间。*arXiv*预印本 *arXiv:1808.10805*，2018年。Z. Yang、W. W. Cohen 和 R. Salakhutdinov。借助图嵌入重新审视半监督学习。2016年。H. Zhang、Q. Wu、J. Yan、D. Wipf 和 P. S. Yu。从典型相关分析到自监督图神经网络。神经信息处理系统进展，第34卷：76–89，2021年。J. Zhou、G. Cui、S. Hu、Z. Zhang、C. Yang、Z. Liu、L. Wang、C. Li 和 M. Sun。图神经网络：方法与应用综述。*AI Open*，1:57–81，2020年。Y. Zhu、Y. Xu、F. Yu、Q. Liu、S. Wu 和 L. Wang。深度图对比表征学习。*arXiv*预印本 *arXiv:2006.04131*，2020年。Y. Zhu、Y. Xu、F. Yu、Q. Liu、S. Wu 和 L. Wang。采用自适应增强的图对比学习。载于*2021*年万维网大会论文集，第2069–2080页，2021年。

A 实现细节

A.1 数据集

我们采用八个节点分类基准数据集来评估模型：*Cora*、*CiteSeer*、*PubMed*、*Amazon Computers*、*Amazon Photo*、*Coauthor CS*、*Coauthor Physics*、*ogbn-arxiv*。在所有实验中，我们使用 PyTorch Geometric 提供的预处理版本数据集。数据集简要介绍如下：

Cora、CiteSeer、PubMed。 一种引文网络，其中节点代表科学论文，边表示论文对之间的引用关系（Sen 等人，2008年）。

Amazon Computers、Amazon Photo。 Amazon 商品共同购买网络，其中节点代表商品，边表示经常被一起购买的商品对（McAuley 等人，2015年）。

Coauthor CS、Coauthor Physics。 Microsoft Academic 网络，其中节点代表作者，边表示共同撰写过论文的作者之间的关系（Sinha 等人，2015年）。

Ogbn-arxiv。 MAG 收录的所有计算机科学（CS）arXiv 论文之间的引文网络，其中节点代表 arXiv 论文，有向边表示一篇论文引用了另一篇论文（Wang 等人，2020年；Hu 等人，2020年）。

数据集的更多详细信息见表4。

表 4：数据集统计信息。

Dataset	# Nodes	# Edges	Feature dimension	# Classes
Cora	2,708	10,556	1,433	7
CiteSeer	3,327	9,104	3,703	6
PubMed	19,717	88,648	500	3
Computers	13,752	491,722	767	10
Photo	7,650	238,162	745	8
CS	18,333	163,788	6,805	15
Physics	34,493	495,924	8,415	5
ogbn-arxiv	169,343	1,166,243	128	40

A.2 超参数

对于所有数据集，所有超参数均通过小规模网格搜索确定。具体而言，由于搜索空间较大，我们固定温度参数 $\tau = 0.5$ 和 $\mathrm{wd1} = 0$。表格列出了我们在八个基准数据集上的实验中所用超参数的全部细节。5。

表5：超参数的详细信息。

Dataset	Epoch	# layers	embedding dimension	τ	lr1	wd1	p_f	p_e	lr2	wd2
Cora	50	2	512	0.5	1e-3	0	0.2	0.5	5e-3	1e-4
CiteSeer	50	1	512	0.5	1e-3	0	0.2	0.5	1e-2	1e-2
PubMed	600	2	512	0.5	1e-3	0	0.3	0.5	1e-2	1e-4
Computers	200	2	512	0.5	1e-3	0	0.0	0.5	1e-2	1e-4
Photo	100	2	512	0.5	1e-3	0	0.0	0.5	1e-2	1e-4
CS	600	2	512	0.5	1e-3	0	0.3	0.1	5e-3	1e-2
Physics	50	2	512	0.5	1e-3	0	0.2	0.4	5e-3	1e-4
ogbn-arxiv	5000	3	512	0.5	1e-2	0	0.0	0.5	1e-2	1e-4

B 进一步的节点嵌入比较 S

B.1 聚类评分

我们进一步通过考察三种聚类评分来评估节点嵌入的质量： 轮廓系数、戴维斯-博尔丁指数、卡林斯基-哈拉巴斯指数 (Rousseeuw，1987年；Davies 和 Bouldin，1979年；Cali´nski 和 Harabasz，1974年）。我们考察三种方法：CLNR、GRACE 和 CCA-SSG。以下简要介绍这三种聚类评分：

轮廓系数。 设第 i 个节点的轮廓系数为 SC_i ，所有节点的轮廓系数为 $SC = \frac{1}{N}\sum_{i=1}^{N} SC_i$ ，其中 N 为节点数量。对于每个节点 i，轮廓系数计算为 $SC_i = \frac{Y_i - X_i}{max\ X_i}$ （，Y_i），其中 X_i 为簇内距离的均值， Y_i 为第 i 个节点不属于的最近簇距离的均值。 SC 所能达到的最佳值和最差值分别为 1 和 -1。 $SC \approx 0$ 表示簇之间存在重叠。

Davies-Boulding 指数。 设 Davies-Boulding 指数为 DB。Davies-Boulding 指数计算为 $DB = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} R_{i,\ j}$ ，其中 K 为类别数量，且 $R_{i,\ j} = \frac{S_i + S_j}{M_i}$ ，，其中 S_i 为每个节点与第 i 个簇质心之间距离的均值，$M_{i,\ j}$ 为第 i 个簇与第 j 个簇质心之间的距离。 DB 的值较小表示簇之间分离良好。

卡林斯基-哈拉巴斯指数。 我们将卡林斯基-哈拉巴斯指数记为 CH。卡林斯基-哈拉巴斯指数的计算公式为 $CH = \frac{tr(B)}{tr(W)} \times \frac{N-K}{K-1}$ 其中， $tr(B)$ 是组间离散矩阵的迹， $tr(W)$ 是簇内离散矩阵的迹， K 是类别数，N 是节点数。指数越高，表明簇越紧密且分离得越好。

CLNR、GRACE 和 CCA-SSG 在 Cora 数据集上的聚类评分汇总于表 6中。我们观察到，CLNR 达到了最佳的轮廓系数，而 CCA-SSG 达到了

最佳的 Davies-Boulding 指数和卡林斯基-哈拉巴斯指数。此外，CLNR 的所有聚类评分均高于 GRACE。

表6：Cora 数据集上的聚类评分。

METHODS	SC	DB	CH
CLNR	**0.105**	2.248	63.066
GRACE	0.041	2.686	49.503
CCA-SSG	0.089	**2.002**	**70.760**