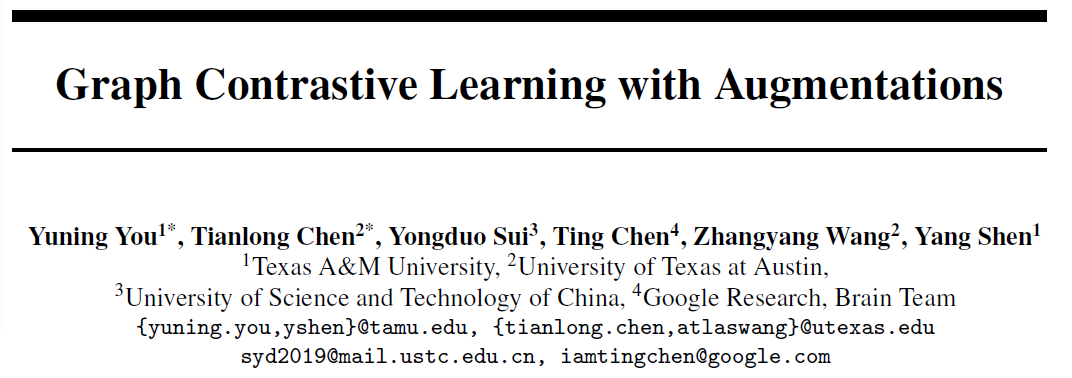
**论文分享：Graph Contrastive Learning with Augmentations**



**Abstract**

本篇论文发表于NeurIPS2020。获取具有泛化性、迁移性和鲁棒性的图表示是一个亟待解决的问题。现有的图表示通常采用监督学习的方式，这种方式高度依赖于标注信息。针对该问题，这篇文章提出了利用图对比学习框架学习（GraphCL），利用四种针对图的数据增强方式，通过无监督（自监督）方式作为预训练模型获取图表示。在半监督、无监督、迁移学习和对抗攻击四类实验设定中，GraphCL框架表现出了良好的泛化性、迁移性和鲁棒性。

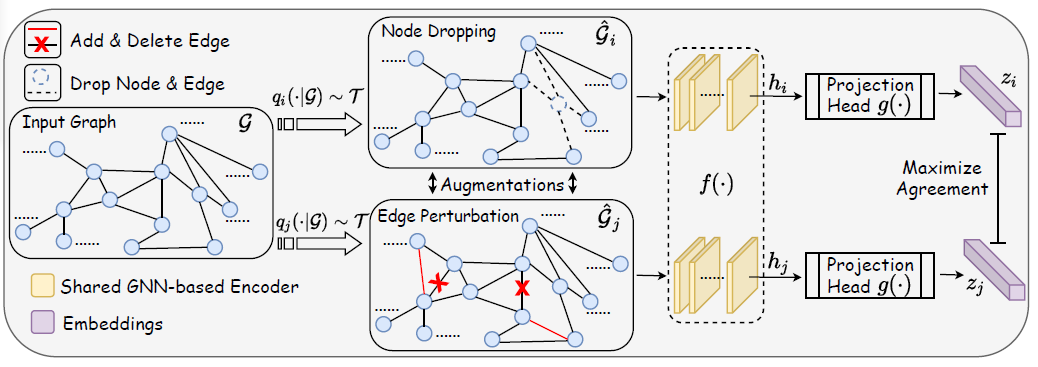
**Introduction**

GNN作为图表示方法，广泛应用于节点分类、链路预测等图任务上。在这些图任务中，GNN通常是采用有监督的方式端到端进行学习。这种方式高度依赖于任务相关的标签信息，然而这类标签信息的获取成本可能较高。而预训练的方法是该问题的解决手段之一，采用预训练方法的GNN将具有更好的泛化性。图表示预训练的难点在于：由于图表示与下游任务紧密相关，因此很难设计通用的GNN预训练方法。一种朴素的针对图数据的预训练方法是重建其节点邻接信息（如GAE和GraphSAGE）。但是这种方法过于关注邻居信息，并不总是对下游任务有效。因此，需要提出一种预训练方法需要准确捕捉到图数据的高度异质信息.

对比学习能够通过最大化不同增强视图（augmented view)下的特征一致性进行表示学习，其中通过与数据或任务相关的增强视图来得到表示的不变性。将对比学习方法扩展至GNN的预训练上，将可能有利于克服上述基于邻居的预训练方法的局限性，其中的关键是如何针对图数据进行相应的数据增强并且构建对比学习目标。

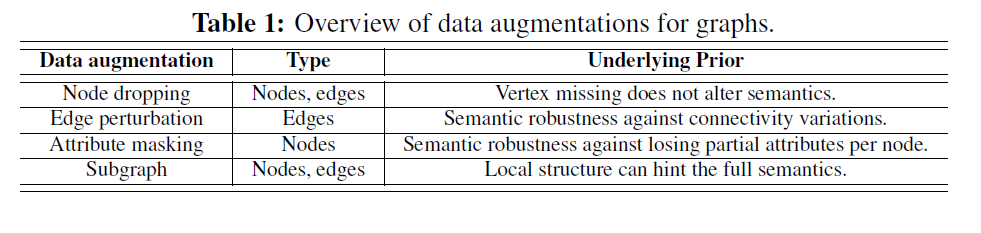
**Method**

文章提出的图对比学习框架如下图所示，共分为四部分：



（1） Graph data augmentation

数据增强的目的是在不影响语义标签的情况下，通过某种转换来创建新颖的、合理的数据。针对图的数据增强可表示为：给定一个图，可将增强图表示为：，其中代表预定义的数据增强分布。数据增强分布代表了对于数据分布的先验信息。文章提出了四种针对图的数据增强方法：



* Node dropping：从图中随机去除一定比例的节点及其连边，使得学习的表示在节点扰动下具有一致性。代表的先验信息是：缺失部分节点不影响图的语义。
* Edge perturbation：随机增加或删除一定比例的边，使学习的表示在边扰动下具有一致性。代表的先验信息是：增减部分连边不影响图的语义。
* Attribute masking：随机去除部分节点的属性信息，促使模型使用其他信息来重建被屏蔽的节点属性。
* Subgraph：使用随机游走的方式从原图中提取子图。

在整个GraphCL框架中，给定一个图数据，通过数据增强生成两个相关的增强图,作为正样本对。

（2）GNN-based encoder

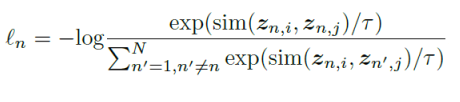
通过GNN-based encoder从增强图,中得到Graph-level的初步表示。

（3）Projection head

将非线性函数作为projection head将增强图的表示映射到另一隐空间，用于计算对比误差。文章采用MLP作为投影头。

（4）Contrastive loss function

对比损失函数采用NT-Xent，计算过程如下：在GNN预训练过程中，从数据集中随机采样得到包含N个图的Minibatch，通过数据增强得到2N个增强图以及相应的对比误差进行优化。其中，增强图的表示与同一张图的另一增强图表示作为正样本对，相应的该表示与其他N-1个增强图表示作为负样本对。采用余弦相似度函数作为相似度度量，第n个图数据的NT-Xent误差可以定义为：



上述误差可重写为：



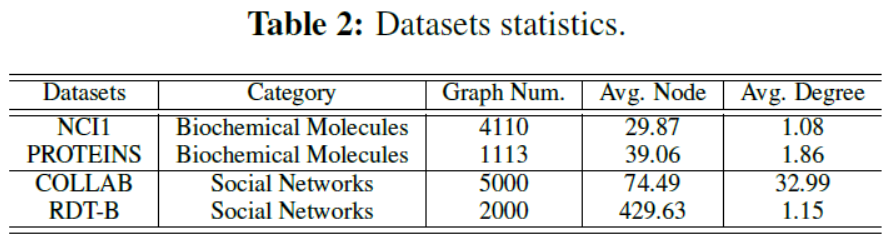
实际是最大化两类增强图的表示之间的互信息下界。

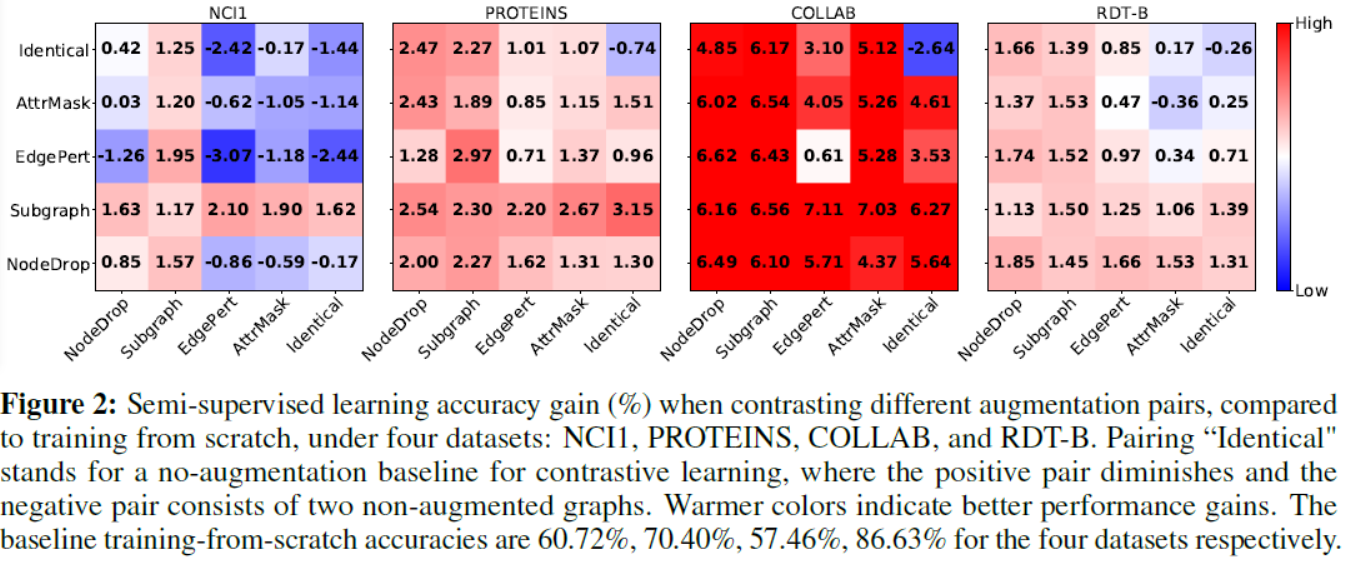
**Experiment**

文章的实验分为两部分，分别讨论数据增强对GraphCL效果的影响以及比较GraphCL与SOTA图表示学习方法的性能。

**数据增强在图对比学习中的作用**

这部分实验评估了采用之前提出的四种数据增强方法的图对比学习框架在半监督图分类任务上的效果。在半监督设定下，模型的训练采用pre-training加finetuning的方法，采用的数据集包括Biochemical Molecules以及Social Networks两类。通过实验得到了文章所提出的预训练方法相对于learn from scratch方法的性能提升（Fig2）。





实验主要讨论了三部分内容：

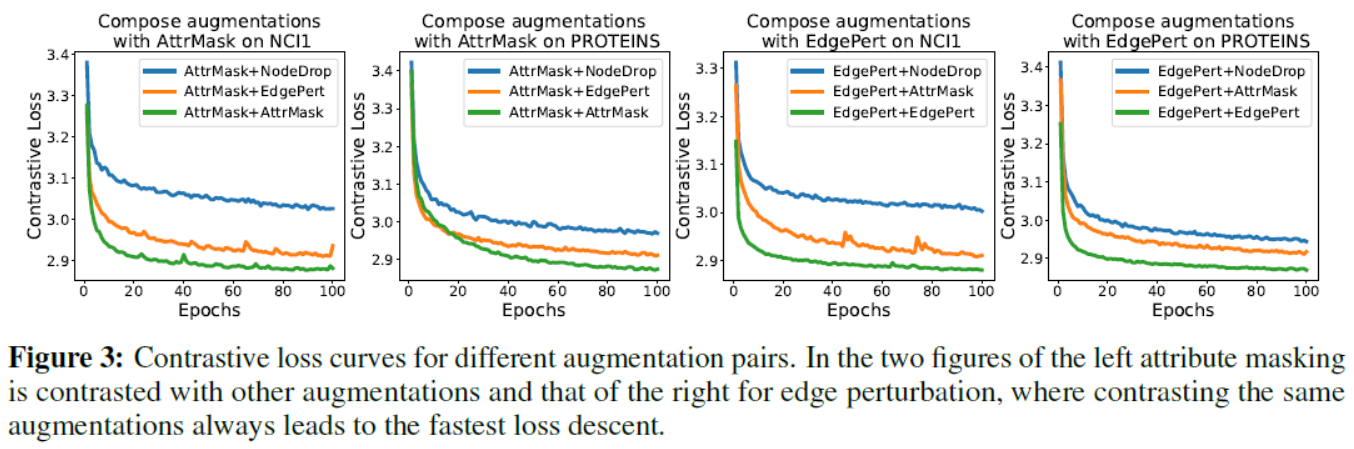
**1. 数据增强对于图对比学习的效果具有关键作用**

（1）加入数据增强有效提升了GraphCL的效果

通过观察Fig2中每个数据图实验结果中的最上一行与最右一列可以发现采用数据增强能有效提升GraphCL的分类准确度。这是由于适当的数据增强会对数据分布注入相应的先验，通过最大化图与其增强图之间的一致性，使模型学习得到的表示对扰动具有不变性。

（2）组合不同数据增强方式对算法效果提升更大

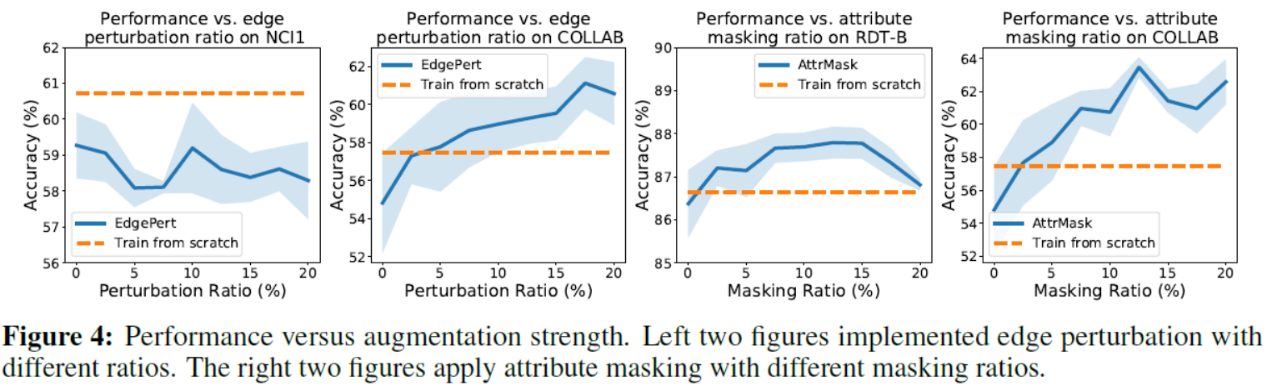
通过观察Fig2发现每个数据集上采用相同数据增强方式构建的样本对所对应的结果均不是该数据集上的最优结果，而每个数据集上的最优结果均采用不同数据增强组合的方式。文章给出的解释是，采用不同数据增强组合的方式避免了学习到的特征过于拟合低层次的“shortcut”，使特征更加具有泛化性。同时通过Fig3发现当采用不同数据增强方式组合时，相比于单一数据增强时的对比误差下降的更慢，说明不同数据增强组合的方式意味着“更难”的对比学习任务。



**2. 数据增强的类型，强度以及模式对GraphCL效果的影响**

（1）Edge perturbation的方式对于Social Network有效但在部分biochemical Molecules数据集上反而有负面效果

通过Fig2可以看出Edge perturbation的数据增强方式在除NCI1之外的三个数据集上均有较好的效果，但是在NCI1上的效果反而比baseline算法差。这是由于对NCI1中的网络的语义对于边的扰动更加敏感，对网络中边进行修改可能会改变分子的性质从而破坏网络语义，进而影响下游任务。针对Edge perturbation的强度，从Fig4中可以得出，在COLLAB数据集上，算法性能随Edge perturbation的强度增加而提升，但在NCI1数据集上，Edge perturbation强度对算法效果无明显影响。



（2）Attribute masking的方式在更“密集”的图数据上能取得更好效果

从Fig2中可以发现Attribute masking的增强方式在平均度更高的数据集上具有更好的性能增益（例如COLLAB），而在平均度较低的数据集上增益明显减小。文章对这个结果做出的假设是，当图数据越密集时，意味着Attribute masking之后模型仍然有足够的其他数据来重建被屏蔽的数据，而反之则难以重建。在强度方面，通过增加Attribute masking的强度可以在更“密集”的数据集上提升算法效果。

（3）Node dropping和Subgraph的方式对所有数据集都有效

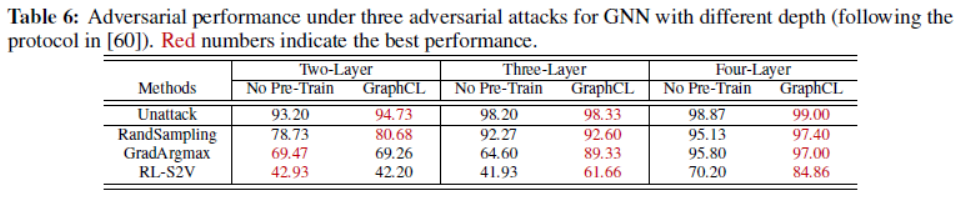
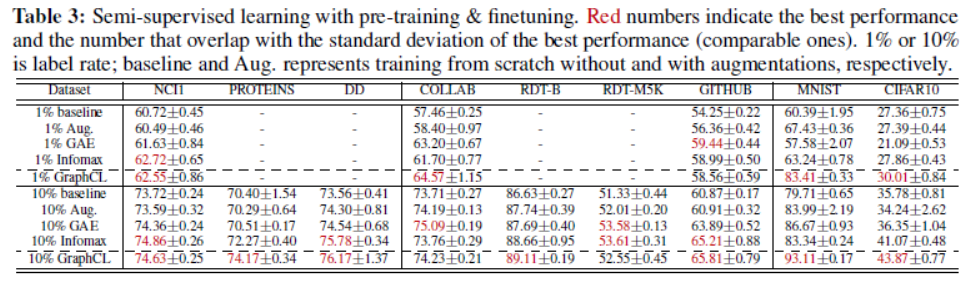
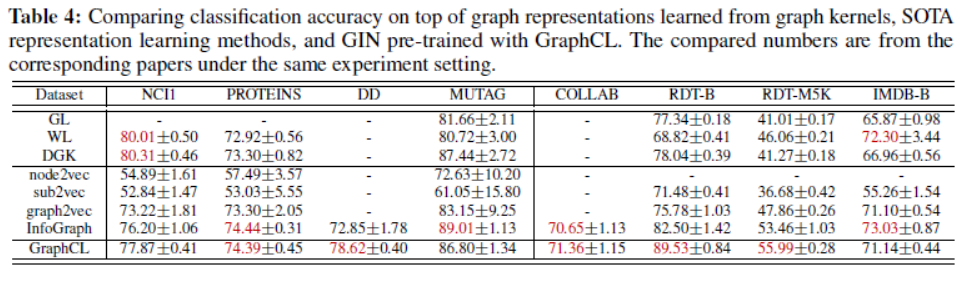
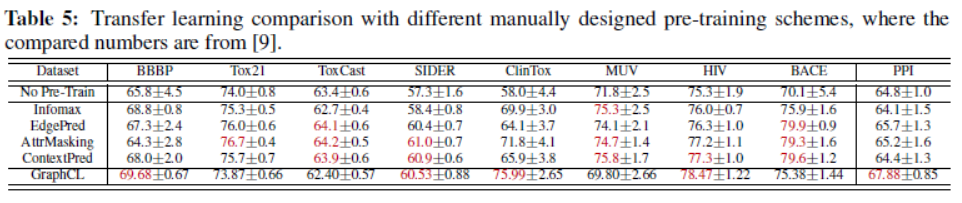
上述两种方式，尤其是Subgraph的数据增强方式在实验中的数据集上都能给图对比学习算法带来性能增益。Node dropping有效的原因是，在许多图数据中去掉部分节点并不影响整图的语义。而对于Subgraph的方式，之前的相关研究已经说明了采用Local-Global 的对比学习方式训练图表示是有效的。

**3. 相对于“更难”的任务，过于简单的对比任务对算法性能提升没有帮助**

“更难”的任务有利于提升GraphCL的效果，这里包含两种情况，一种是将不同的数据增强方法进行组合，另一种是增加数据增强的强度或者提高增强模式的难度，例如采用更高的概率进行Node dropping，或者采用均匀分布之外的复杂分布进行Node dropping。

**GraphCL与SOTA算法的性能对比**

在这部分实验中，文章对比了GraphCL与SOTA的图表示方法在四种setting下的图分类任务中的性能，包括半监督、无监督、迁移学习以及对抗攻击setting。具体实验设置详见原文。在半监督、无监督、迁移学习任务中，GraphCL在大部分数据集上的分类准确率都达到了SOTA，在对抗攻击setting下，实验结果表明GraphCL增强了模型的鲁棒性。



**Conclusion**

文章针对图表示问题，采用了自监督的图对比学习作为预训练方法，主要贡献有：

* 提出了针对Graph的数据增强方法；
* 提出了用于GNN预训练的图对比学习框架GraphCL；
* 评估了不同图数据增强方法的性能并分析了其基本原理；
* 提出的GraphCL在半监督、无监督表示和迁移学习setting下达到SOTA，并增强了对对抗攻击的鲁棒性。