图数据增强方法比较

张鹤潇
2018011365
计84
731931282@qq.com

乐阳
2018011359
计84
ley18@mails.tsinghua.edu.cn

谢云桐
2018011334
计83
xieyt18@mails.tsinghua.edu.cn

1 简介

基于图神经网络 (GNN) 的深度学习方法在图机器学习任务上取得了巨大的成功，它已被广泛应用于社会网络、分子结构、知识图谱等图结构数据的研究中。然而，与图像、文字相比，图数据的标注更加困难，高昂的数据标注成本限制了有监督图机器学习进一步发展。为了让 GNN 用更少的数据学到更多的知识，各种自监督学习方法不断涌现。这些方法从无标签的图数据中构造监督信号，预训练深度图模型，进而获得输入数据的低维表示，供下游任务使用。

现有的图自监督学习方法大致可分为两类[1]：生成 (Generative) 学习和对比 (Contrastive) 学习。其中，基于对比学习的方法从输入数据定向地构造正负样本，让模型在表示空间中对正负样本进行判别，以此完成模型的预训练工作。所谓数据增强 (Data Augmentation)，即以输入数据生成正负样本的过程。

在 CV 和 NLP 领域，数据增强的技巧已经得到了广泛而深入的研究，而在图网络领域，相关探索还很少。已有的图对比学习模型已经包含了一些图数据增强方法，但这些方法与各自模型的耦合度很高，难以比较它们之间的优劣。为此，我们设计了一套基于对比学习的实验框架，旨在公平比较各种图数据增强方法的优劣。

本文的主要工作概述如下：

- 设计了一套基于对比学习的实验框架，以预训练、下游任务 (pretrain, fine-tune) 两阶段的模式测试图数据增强方法的表现。
- 在 3 个经典的节点分类数据集和 4 个图分类数据集上，实验对比多种数据增强方法的性能。
- 实验发现，在节点分类任务中， Personalized PageRank 效果最佳；而在图分类任务中，对于不同数据集，最优的数据增强方法不同。这表明数据增强方法的优劣与数据集和任务是相关的。

2 相关工作

2.1 图神经网络

图神经网络将深度学习与各行各业普遍存在的图结构数据嫁接在一起。它通常以带节点或边属性的图结构为输入，输出取决于具体任务。以节点分类为例， GNN 根据图结构和输

入节点属性学习图中每个节点的隐式向量表示，其目标是让该向量表示包含足够的信息，使不同类别的节点更容易区分。

图神经网络前向传播的主要过程是迭代地对邻居信息进行聚合和更新[2]。在一次迭代中，每个节点聚合其邻居节点的信息，并对聚合后的信息进行非线性变换，从而更新本节点的信息。通过堆叠多层网络，节点可以聚合多跳内邻居的信息。经典的图卷积网络[3] (GCN)，图注意力网络[4] (GAT) 和图异构网络[5] (GIN) 都遵循这个过程： GCN 对节点和邻居特征求平均； GAT 引入注意力机制对邻居信息进行加权；而 GIN 则从理论上指出什么样的聚合和更新策略可能是合理的。

2.2 自监督学习和预训练

很多时候，有标签数据的获取是昂贵而困难的，而无标签数据则很容易取得。从大规模无标注数据中生成伪标签作为监督信号，就是自监督学习。以 BERT[6] 为例，对于输入的无标签语句，随机掩盖 15% 的单词，以此为监督信号训练模型。

由于自监督学习作用于无标签数据，通常情况下，其产出模型为通用的预训练模型，可根据具体下游任务对其进行微调。预训练模型能从大规模无标签数据中学习通用规律，在 NLP、 CV 等领域已经取得了突破。

主流的自监督学习方法可分为两类。一类对输入数据进行生成重建，如 BERT、GPT[7] 系列；另一类从输入数据中构造正负样本，在表示空间对正负样本进行判别，如 Deep Infomax[8] 和 Momentum Contrast[9]. 本文讨论的数据增强方法主要应用于第二类，即对比学习模型中。通常情况下，对某个样本而言，其本身增广出的数据被视作正例，而其它样本及其增广出的数据被视为负例。

2.3 对比学习中的图数据增强

Graph Diffusion[10] 是一类将节点结构信息传播到全图的方法。论文[11]提出了一套基于最大化邻接矩阵和 Diffusion 矩阵互信息的对比学习方法，并通过实验得出结论，在各种 Graph Diffusion 方法中， Personalized PageRank 的效果最佳，且增加正样本的数量无助于提高深度图模型的性能。本文借鉴了该文提出的实验框架，去除了其中难以解释的技巧，并根据其实验结果，将 Personalized PageRank 作为候选数据增强方法之一。

论文[12]提出了基于随机游走的图采样方法，论文[13]提出用邻居特征替代节点特征的 DropNode 方法，本文将这两种方法从对应框架中剥离出来。

Deep Graph Infomax (DGI)[14] 对比了定向构造负样本的各种手段，并通过实验证明，打乱特征矩阵 (feature permutation) 构造的负例效果最好。本文根据其实验结果，将 feature permutation 作为构造负样本的方法。

3 方法

3.1 数据增强策略

对于图 $G(A, X)$[1]，记其邻接矩阵$A \in \mathbb{R}^{n\times n}$，特征矩阵$X \in \mathbb{R}^{n\times d_x}$，由该节点增广新数据$\mathcal{D}(A, X)$，其中 $\mathcal{D}(.,.) : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times d_x} \longmapsto \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times d_x}$ 是数据增强策略。下面给出不同 $\mathcal{D}$ 的定义。

[1]本文只考虑无边权的无向图。

3.1.1 Personalised PageRank

Graph Diffusion[10][11] 定义为(1), 其中 $T \in \mathbb{R}^{n\times n}$ 是广义传输矩阵，权重系数 θ 决定了全局信息和局部信息的比例。约束条件 $\sum_{k=0}^{\infty}\theta_k = 1$，$\theta_k \in [0,1]$，$T$ 的特征值 $\lambda_i \in [0,1]$.

$$S = \sum_{k=0}^{\infty}\theta_k T^k \in \mathbb{R}^{n\times n} \tag{1}$$

对于 Personalized PageRank，记图的节点度对角阵为 D，则 $T = AD^{-1}$，$\theta_k = \alpha(1-\alpha)^k$，公式(1)有闭式解

$$S = \alpha(I_n - (1-\alpha)D^{-1/2}AD^{-1/2})^{-1} \tag{2}$$

据此可得 $\mathcal{D}_{ppr}(A,X) = (S,X)$.

3.1.2 Random walk with restart

Random walk with restart[12] 以随机游走的方式在输入图上采样，得到不同子图作为样本的数据增强。

先给出 r-ego network 的定义。对于图 $G = (V,E)$，记其节点集合为 V，边集合为 $E \subseteq V \times V$. 对于节点 $v \in V$，其 r-neighbors 定义为 $S_v = \{u : d(u,v) \leq r\}$，其中 $d(u,v)$ 是节点 u 到 v 的最短路长度。节点 v 的 r-ego network 是由 S_v 导出的子图。

本文采用的 Random walk with restart 算法伪代码详述于 Algo. 1.

Algorithm 1: Random walk with restart

输入 : 图 G，起始节点 v，重启概率 p

输出 : 子图 $\tilde{G}$

令 v 为当前节点;
repeat
  if 当前节点不在子图中 **then**
    将当前节点加入子图 $\tilde{G}$;
  从当前节点开始在 G 上随机游走;
  以概率 p 回到节点v;
until 迭代步数或 $\tilde{G}$ 超过对应阈值;
重新标记 $\tilde{G}$ 的节点序号;

在 Random walk with restart 中，重启概率 (restart probability) 控制了提取子图的半径。重启概率越大，子图对应的 r-ego network 的半径就越大。重新标注节点序号的作用是防止模型简单地根据两个子图序号的对应关系进行判别，影响性能。

3.1.3 DropNode

前两种数据增强策略都聚焦于图结构层面，而 DropNode[13] 则从图的特征矩阵入手。大体上说，DropNode 以节点邻居的特征加权和代替其节点本身的特征，其算法详述于 Algo. 2.

DropNode 算法分为两步。首先，随机将 X 中的行向量置零，得到扰动后的特征矩阵 $\hat{X}$；然后，利用 $\hat{X}$ 进行特征传播，生成数据增强后的特征矩阵 $\tilde{X}$。这样就得到$\mathcal{D}_{DropNode}(A,X) = (A,\tilde{X})$.

Algorithm 2: DropNode

输入： 邻接矩阵 A，特征矩阵 X, DropNode 概率 δ，随机传播次数 K

输出： 数据增强后的特征矩阵 $\tilde{X}$

for $X_i \in X$ **do**

  sample $\epsilon \sim Bernoulli(1-\delta)$;

  $\hat{X}_i := \epsilon \cdot X_i$;

$\hat{X} := \frac{1}{1-\delta}\hat{X}$;

$\tilde{X} := \sum_{k=0}^{K} \frac{1}{1+K} A^k \hat{X}$;

在此过程中，每个结点的特征随机地与来自其邻居的信号相混合。根据同质性假设[15]，图上相邻节点倾向于具有相似的特征和标签，因此节点在扰动中丢失的信息可以通过其邻居节点来补偿。这样，我们就能为每个节点生成多个增广表示。

3.2 网络架构与预训练

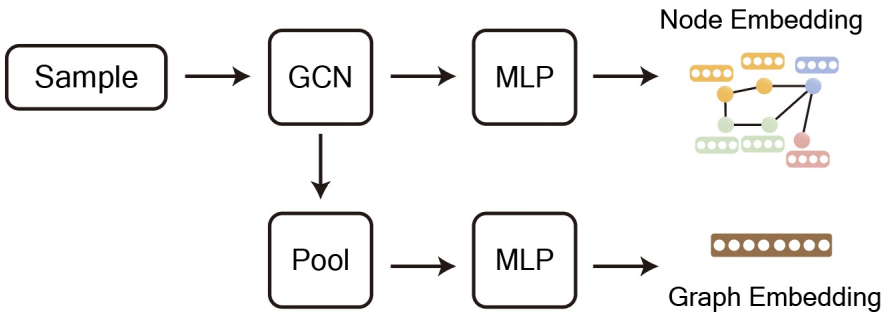

Figure 1: 网络结构图

本实验框架参考了论文[11]的设计，支持以多种 GNN 网络作为图编码器，为简单起见，我们选择了 GCN[3]。模型架构如图1所示。GCN $g(.) : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times d_x} \longmapsto \mathbb{R}^{n\times d_h}$ 对输入数据进行编码，再经过一个 MLP，得到节点在隐向量空间上的表示矩阵 H，即 Node embedding.

对于 GCN 输出的节点特征，用一个 graph pooling layer $\mathcal{P}(.) : \mathbb{R}^{n\times d_h} \longmapsto \mathbb{R}^{d_h}$ 将其聚合为全图的特征向量。$\mathcal{P}(.)$ 公式如下，

$$\mathcal{P}(H) = \sigma\left(\Big\|_{i=1}^{L}\left[\sum_{i=1}^{n} h_i^{(l)}\right] W\right) \tag{3}$$

其中 L 是 GCN 的层数，$h_i^{(i)}$ 是第 i 层 GCN 输出的节点特征向量，$W \in \mathbb{R}^{(L\times d_h)\times d_h}$ 是权重矩阵，$\|$ 是向量拼接运算符，σ 是激活函数 PReLU[16]. pooling layer 的输出经过 MLP，得到隐空间上的图表示向量 $\vec{h}_g$，即 Graph embedding.

网络中的两个 MLP 均由两层隐藏层构成，以 PReLU 为激活函数。

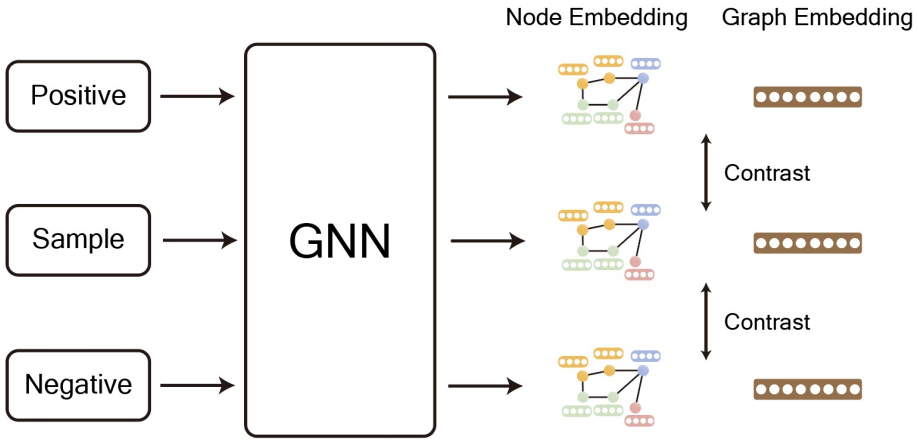

Figure 2: 预训练示意图

预训练框架如图2所示。目标函数

$$\mathcal{O} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(\frac{1}{g}\sum_{i=1}^{|g|}\mathbf{Dis}(\vec{h}_i,\vec{h}_i^{pos}) + \mathbf{Dis}(\vec{h}_g,\vec{h}_g^{pos}) - \frac{1}{g}\sum_{i=1}^{|g|}\mathbf{Dis}(\vec{h}_i,\vec{h}_i^{neg}) - \mathbf{Dis}(\vec{h}_g,\vec{h}_g^{neg})) \tag{4}$$

其中 $\mathcal{G}$ 表示图数据集，$\vec{h}_i^{pos}$ 和 $\vec{h}_g^{pos}$ 表示正样本的节点、图表示向量，$\vec{h}_i^{neg}$ 和 $\vec{h}_g^{neg}$ 表示负样本的节点、图表示向量。

$\mathbf{Dis}(.,.) : \mathbb{R}^{d_h} \times \mathbb{R}^{d_h} \longmapsto \mathbb{R}$ 度量了两个向量的相似程度，实现为一层双线性层 (Bilinear Layer).

预训练结束后，将模型输出的节点表示矩阵 H 和图表示向量 $\vec{h}_g$ 作为下游任务模型的输入，如图3所示。

4 实验

4.1 实验设置

我们的实验在节点分类和图分类两类下游任务中进行。其中，节点分类数据集包括[2] Cora, Citeseer, Pubmed. 这三个数据集均为引文网络，即由论文及其引用关系构成的网络。

Table 1: 点分类数据集

Dataset	Nodes	Edges	Features	Classes
Cora	2708	5429	1433	7
Citeseer	3327	4732	3703	6
Pubmed	19717	44338	500	3

[2]https://github.com/kimiyoung/planetoid/tree/master/data

图分类数据集包括[3] MUTAG, IMDB-B, IMDB-M, PTR-MC.

- MUTAG 包含了 188 个突变芳香族和杂芳香族硝基化合物。
- IMDB-B 是一个电影协作数据集，收集了 IMDB 上不同电影的演员和类型信息，IMDB-M 与之类似。
- PTR-MC 是雄性大鼠可致癌化合物数据集。

Table 2: 图分类数据集

Dataset	Graphs	Classes	Avg Nodes
MUTAG	188	2	17.93
IMDB-B	1000	2	19.77
IMDB-M	1500	3	13.00
PTC-MR	344	2	14.29

图分类数据集均不自带节点特征向量。为了解决这个问题，我们将节点标签和度的 one-hot 编码拼接在一起，作为节点特征输入 GNN.

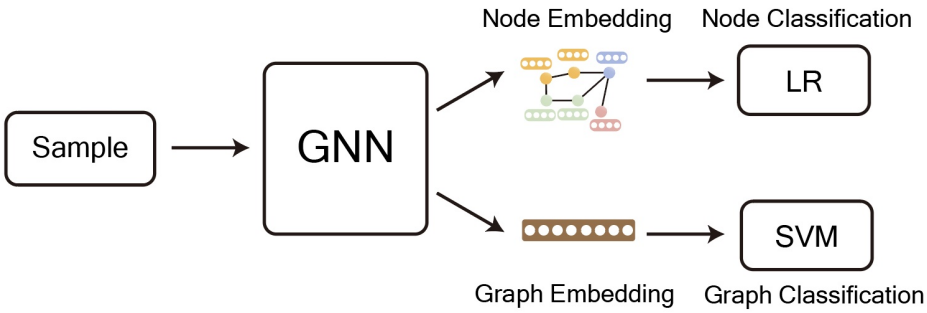

Figure 3: 下游任务示意图

我们遵照图对比学习领域先前的 state-of-the-art 来设计下游任务。对于节点分类，我们参照 DGI[14] ，将预训练模型输出的节点表示作为逻辑回归分类器的输入，训练 300 个epoch 之后，测试预测准确率。对于图分类，我们参照 InfoGraph[17]，将预训练得到的图表示输入线性 SVM, 测试其在 10 折交叉验证中的平均预测准确率。SVM 的参数 C 在$[10^{-3}, 10^{-2}, \cdots, 10^{3}]$ 中选取。

我们用深度学习框架 PyTorch 和图网络编程库 DGL 搭建模型，用 optuna 库进行超参数调优；设置 GNN 隐藏层维数为 512，用 Adam 优化器优化网络，并取学习率为 0.001，当预训练模型损失连续 20 个 epoch 不下降时结束预训练。在 citeseer 数据集上，我们设置L2 weight decay 为 0.01，而在其它数据集上该超参数均设为 0. Personalized PageRank 的系数 α 设置为 0.2.

其它超参数见表3。

我们以用 GCN, GAT 搭建的有监督模型作为baseline[4]，同时以样本本身作为增广正例 (No Augment) 进行对照，检验各种数据增强策略的实际效果。

[3]https://ls11-www.cs.tu-dortmund.de/people/morris/graphkerneldatasets/

[4]GCN 和 GAT 在图分类任务上的实验结果来源于[11]。

Table 3: 模型超参数

dataset	Cora	Citeseer	Pubmed	IMDB-B	IMDB-M	MUTAG	PTC-MR
batch size	4	4	4	256	128	256	32
GCN layers	1	1	1	2	4	4	4
Epochs	2000	2000	2000	20	40	20	100
DropNode propagation	1	0	0	2	1	1	1
restart probability	0.4	0.1	0.4	0.8	0.8	0.8	0.8

4.2 实验结果

实验结果见表4和表5。

Table 4: 节点分类实验结果

method	Cora	Citeseer	Pubmed
PPR	**86.8±0.3**	**73.3±1.1**	**78.3±1.1**
RWR	84.9±0.5	69.6±1.0	77.9±0.8
DropNode	85.0±0.7	72.0±0.8	76.9±0.5
No Augment	84.2±0.2	71.1±0.4	74.5±0.7
GCN[3]	81.5	70.3	79.0
GAT[4]	83.0	72.5	79.0

Table 5: 图分类实验结果

method	IMDB-B	IMDB-M	MUTAG	PTC-MR
PPR	74.3±0.8	**50.7±0.6**	88.6±0.7	60.8±0.7
RWR	73.5±0.3	50.2±0.1	**89.0±1.1**	60.6±1.4
DropNode	**75.2±0.1**	**50.7±0.3**	88.6±0.9	**62.7±1.6**
No Augment	73.6±0.9	50.7±0.8	88.9±0.9	60.5±1.6
GCN	74.0±3.4	51.9±3.8	85.6±5.8	64.2±4.3
GAT	70.5±2.3	47.8±3.1	89.4±6.1	66.7±5.1

实验表明，在节点分类任务上，Personalized PageRank 的效果最优，且自监督学习模型的性能普遍优于有监督模型。而在图分类任务上，对于不同的数据集，最优的数据增强策略不同，自监督学习模型的性能与有监督学习相当或略低。由此可见，数据增强策略的效果与下游任务类型和数据集是相关的。

作为消融实验，我们增设无数据增强的自监督模型作为对照组。由实验结果可知，在大多数数据集上，数据增强策略确实提高了模型的性能。这在一定程度上证明了实验设计和实现的一致性。

5 总结

本文设计了一套基于对比学习的实验框架，以预训练、下游任务两阶段的模式测试图数据增强方法的表现；在三个的节点分类数据集和四个图分类数据集上，实验对比了三种数据增强方法的性能；从实验结果中，我们发现，数据增强方法的效果与数据集和下游任务类型是相关的。

为简单起见，我们在实验中固定正例和负例的数量为 1，这可能限制了某些数据增强方法的性能。在进一步的研究中，应当将这个问题考虑在内，探索增广样本的数量对模型性能的影响。

References

[1] Liu, X., F. Zhang, Z. Hou, et al. Self-supervised learning: Generative or contrastive, 2020.

[2] Battaglia, P. W., J. B. Hamrick, V. Bapst, et al. Relational inductive biases, deep learning, and graph networks, 2018.

[3] Kipf, T. N., M. Welling. Semi-supervised classification with graph convolutional networks, 2017.

[4] Veličković, P., G. Cucurull, A. Casanova, et al. Graph attention networks, 2018.

[5] Xu, K., W. Hu, J. Leskovec, et al. How powerful are graph neural networks?, 2019.

[6] Devlin, J., M.-W. Chang, K. Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] Radford, A., I. Sutskever. Improving language understanding by generative pre-training. In *arxiv*. 2018.

[8] Hjelm, R. D., A. Fedorov, S. Lavoie-Marchildon, et al. Learning deep representations by mutual information estimation and maximization, 2019.

[9] He, K., H. Fan, Y. Wu, et al. Momentum contrast for unsupervised visual representation learning, 2020.

[10] Klicpera, J., S. Weißenberger, S. Günnemann. Diffusion improves graph learning, 2019.

[11] Hassani, K., A. H. Khasahmadi. Contrastive multi-view representation learning on graphs, 2020.

[12] Qiu, J., Q. Chen, Y. Dong, et al. Gcc. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery*, 2020.

[13] Feng, W., J. Zhang, Y. Dong, et al. Graph random neural network for semi-supervised learning on graphs, 2020.

[14] Veličković, P., W. Fedus, W. L. Hamilton, et al. Deep graph infomax, 2018.

[15] McPherson, M., L. Smith-Lovin, J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[16] He, K., X. Zhang, S. Ren, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

[17] Sun, F.-Y., J. Hoffmann, V. Verma, et al. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization, 2020.