《人工神经网络》大作业开题报告

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 任务定义

数据增强(Data Augmentation)在深度学习任务中应用广泛，它作为自监督学习不可或缺的一部分，帮助神经网络用更少的数据学到更多了的知识。在 CV 和 NLP 领域，各种数据增强技巧已经得到了广泛而深入的研究，而在图网络领域，相关的探索还很少[1]。

现有的图自监督学习方法已经包含了许多数据增强的技巧，但它们与各自的模型紧紧耦合，难以公平比较其优劣。我们希望提炼出这些方法，设计模型，进行实验，对比它们的优劣，总结其中的规律。

形式化描述任务如下，对于图$G(A, X)$，记其邻接矩阵为A，特征矩阵为X，数据增强正样本$G_p(A_p, X_p)$，负样本$G_n(A_n, X_n)$；预训练图网络模型GNN，使$GNN(G_p)$ 和$GNN(G)$ 间的信息差尽可能小，$GNN(G_n)$ 和$GNN(G)$ 间的信息差尽可能大。调整G_p和G_n，对比GNN 在下游任务（节点分类和图分类）中的性能。

2 数据集

节点分类数据集包括 Cora, Citeseer, Pubmed[2]. [1]

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: 节点分类数据集

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

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

- MUTAG是188个突变芳香族和杂芳香族硝基化合物的数据集。
- 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包含了雄性大鼠可致癌化合物。

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

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

3 挑战和基线

3.1 问题和挑战

- 本项目所涉及的领域处于前沿状态，没有过往成熟的工作可以借鉴。我们需要从图自监督学习相关论文中尝试提炼出图数据增强的方法，并设计合适的模型进行实验。
- 图数据的体量较大，对于内存的需求很高。我们希望以稀疏矩阵的形式对图的邻接矩阵和特征矩阵进行读写，以解决内存瓶颈。
- 图网络coding本身也有难度。
- 本组成员均为大三同学，其它专业课的压力很重。

3.2 基线

以图卷积网络(GCN)[2]和图注意力网络(GAT)[3]为基线[2]。

Table 3: baseline

Model	Cora	Citeseer	Pubmed
GCN	81.0	70.7	79.0
GAT	84.0	70.9	78.6

图分类的baseline暂未实现。

需要注意的是，本项工作立足于提炼和对比已有方法，模型性能的绝对值或SOTA并不是追求的目标。

4 研究计划

对于图的数据增强可以从特征矩阵和邻接矩阵两个方面入手，

- 对邻接矩阵的数据增强可以包括子图采样[4]，加边减边[5].
- 对特征矩阵的数据增强可以包括Diffusion[6] [7]，Relabeling[8]，以及用n-hop neighbor的特征替代节点特征[9]。

本项目计划时间安排如下:

- 第八周结束前(-11.8)，完成相关论文的调研。
- 九到十二周(11.9-12.6)，设计模型，进行实验。
- 十三周之后(12.7-)，解题，撰写实验报告。

5 可行性

- 研究方向可行性见上一节。
- 本组准备使用所在实验室的服务器进行实验，配备多张V100。

[2]https://github.com/dmlc/dgl/tree/master/examples/pytorch

- 本组同学其它专业课压力非常大，能投入的时间有限，因此工作会以提炼和对比已有方法为主。
- 本组同学有丰富的PyTorch编程经验，能熟练使用图神经网络库DGL(Deep Graph Library)进行编程。

6 申请MegStudio的额外算力

本组不申请MegStudio的额外算力。

References

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

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

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

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

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

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

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

[8] Hu, Z., Y. Dong, K. Wang, et al. Gpt-gnn: Generative pre-training of graph neural networks, 2020.

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