【动手做】利用可微池化完成图分类任务

爱科技勤折腾的苹果二 2024-1-22

讨论内容

0 工具准备

- Python和Anaconda
- DGL-Deep Graph Library

1 知识准备:处理多粒度的图表示学习

2 案例演示

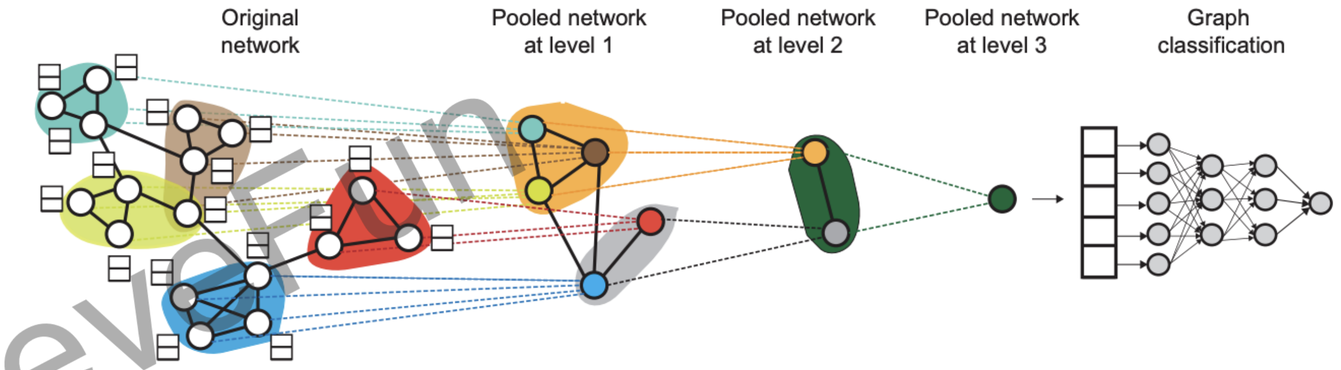
- 代码运行流程
- 代码结构
- 代码演示

3 参考资料

4 回顾【动手做】图神经网络模型

1 知识准备：处理多粒度的图表示学习

- 对图的不同粒度层次的信息建模，需要层次化模型，例如：需要对整个图进行分类。
- 层次化的模型：图的粗化过程，即不断地将相似节点聚类在一起，形成一个新的超节点。
- 超节点的表示由聚类一起的相似节点共同作用得到
- 超节点之间的连接边由聚类分配矩阵和原邻接矩阵共同作用得到



在每个层次层，运行一个 GNN 模型获取节点的嵌入。
使用这些学习到的嵌入聚类节点，并在此粗化图上运行另一个 GNN 层。
整个过程重复运行 L 层，使用最终的输出表示分类图。

图片来源： https://arxiv.org/pdf/1806.08804v4.pdf Hierarchical Graph Representation Learning with Differentiable Pooling

利用可微分池化进行分层图表示学习

- Recently, graph neural networks (GNNs) have revolutionized the field of graph representation learning through effectively learned node embeddings, and achieved state-of-the-art results in tasks such as node classification and link prediction. However, current GNN methods are inherently flat and do not learn hierarchical representations of graphs—a limitation that is especially problematic for the task of graph classification, where the goal is to predict the label associated with an entire graph. Here we propose DIFFPOOL, a differentiable graph pooling module that can generate hierarchical representations of graphs and can be combined with various graph neural network architectures in an end-to-end fashion. DIFFPOOL learns a differentiable soft cluster assignment for nodes at each layer of a deep GNN, mapping nodes to a set of clusters, which then form the coarsened input for the next GNN layer. Our experimental results show that combining existing GNN methods with DIFFPOOL yields an average improvement of 5–10% accuracy on graph classification benchmarks, compared to all existing pooling approaches, achieving a new state-of-the-art on four out of five benchmark data sets.
- 来源： https://arxiv.org/pdf/1806.08804v4.pdf

机器翻译：

最近，图神经网络（GNN）通过有效学习节点嵌入彻底改变了图表示学习领域，并在节点分类和链接预测等任务中取得了最先进的成果。然而，目前的 GNN 方法本质上是扁平的，不能学习图的分层表示--这种局限性对于图分类任务来说尤其成问题，因为图分类的目标是预测与整个图相关的标签。在这里，我们提出了 DIFFPOOL，它是一种可微分图池模块，可以生成图的层次化表示，并能以端到端的方式与各种图神经网络架构相结合。DIFFPOOL 为深度图神经网络每一层的节点学习可微分的软集群分配，将节点映射到一组集群，然后形成下一层图神经网络的粗化输入。我们的实验结果表明，将现有的 GNN 方法与 DIFFPOOL 相结合，与所有现有的池化方法相比，在图分类基准上平均提高了 5-10% 的准确率，在五个基准数据集中的四个数据集上达到了新的一流水平。

2 案例演示

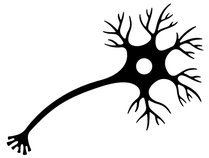
- 案例来源：
  - https://paperswithcode.com/method/diffpool

- 代码
  - https://github.com/dmlc/dgl/tree/master/examples/pytorch/diffpool

学习目标

- 来源： https://github.com/dmlc/dgl/tree/master/examples/pytorch/diffpool
- 理解DiffPool的原理，如何实现图分类任务。

环境准备



1. 安装Anaconda https://www.anaconda.com/download/
2. 在Anaconda建立环境dgl，根据dgl文档(https://www.dgl.ai/pages/start.html)确定**Python版本**
3. 根据代码，使用pip install -r requirements.txt 安装依赖库并验证库的存在pip list
    - torch
4. 安装dgl https://www.dgl.ai/pages/start.html

例如：conda install -c dglteam dgl

代码结构

文件名	作用	方法
data_utils.py	数据工具	one_hotify pre_process
模型的dgl_layers aggregator.py bundler.py gnn.py		Aggregator，MeanAggregator， MaxPoolAggregator， LSTMAggregator，Bundler
模型的tensorized_layers assignment.py diffpool.py graphsage.gy		DiffPoolAssignment， BatchedDiffPool，BatchedGraphSAGE

代码结构

文件名	作用	方法
encoder.py	编码器	DiffPool gcn_forward,gcn_forward_tensorized, forwad,loss
loss.py	计算损失	EntropyLoss， LinkPredLoss
model_utils.py	模型工具类	batch2tensor masked_softmax
train.py	训练模型	prepare_data,**graph_classify_task,** train, evaluate
model_param	模型参数model.iter-0	不适用

2.1 代码运行流程

- 问题：图分类任务
- 训练流程：
    1. 装载数据集
    2. 初始化模型
    3. 训练模型
    4. 评估模型

2.2 代码演示

- https://github.com/dmlc/dgl/tree/master/examples/pytorch/diffpool
- python train.py --dataset ENZYMES --pool_ratio 0.10 --num_pool 1 --epochs 1000
- python train.py --dataset DD --pool_ratio 0.15 --num_pool 1 --batch-size 10
- 数据集 ENZYMES：一个从 BRENDA 酶数据库中获得的包含 600 个蛋白质三级结构的数据集。ENZYMES 数据集包含 6 种酶。https://github.com/snap-stanford/GraphRNN/tree/master/dataset/ENZYMES
- 数据集 DD：一个包含1178个蛋白质结构的数据集。每个蛋白质被表示为一个图，其中节点是氨基酸,如果两个节点之间的距离小于6埃，那么他们之间会有一条边相连。预测任务是将蛋白质分类为酶和非酶。https://github.com/snap-stanford/GraphRNN/tree/master/dataset/DD

3 参考资料

- https://github.com/dmlc/dgl/tree/master/examples/pytorch/diffpool
- https://arxiv.org/pdf/1806.08804v4.pdf

4 回顾【动手做】图神经网络模型

- GraphSAGE：链接预测
- GCN图卷积网络：分类
- GAT图注意力网络：分类
- VGAE 图变分自编码模型：链接预测
- DiffPool：Hierarchical Graph Representation Learning with Differentiable Pooling 利用可微池进行层次图表示学习
- GCC: Graph Contrastive Coding for Graph Neural Network Pre-Training 图对比学习模型 自监督学习
- MVGRL: Contrastive Multi-View Representation Learning on Graphs 图上的对比多视图表征学习
- HGT: Heterogeneous Graph Transformer 异构图转换器
- Text Generation from Knowledge Graphs with Graph Transformers 利用图转换器从知识图谱生成文本
- GPT-GNN: Generative Pre-Training of Graph Neural Networks 生成式图网络预训练框架

总结

0 工具准备

- Python和Anaconda
- DGL-Deep Graph Library

1 知识准备:处理多粒度的图表示学习

2 案例演示

- 代码运行流程
- 代码结构
- 代码演示

3 参考资料

4 回顾【动手做】图神经网络模型