【动手做】使用图卷积神经网络和半监督方法分类

爱科技勤折腾的苹果二 2023-11-27

讨论内容

0 工具准备

- Python和Anaconda
- DGL-Deep Graph Library

1 知识准备

- 卷积
- 半监督原理

2 案例演示

- 代码运行流程
- 代码演示

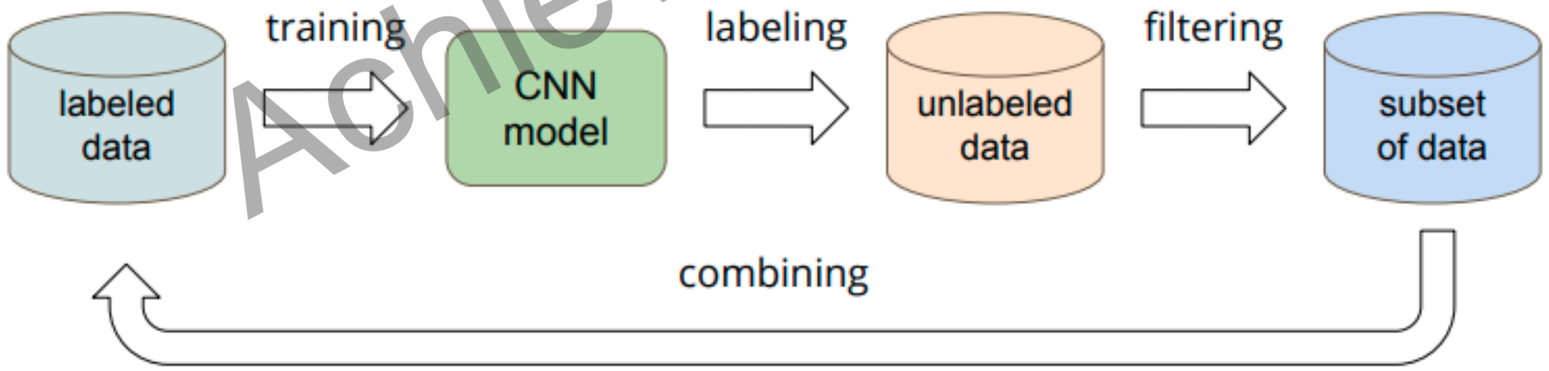
3 参考资料

1 知识准备：卷积

- 擅于抓取特征
- 参考幻灯片 https://github.com/bettermorn/IAICourse/blob/main/courseware/%E5%B7%A5%E4%B8%9A%E6%99%BA%E8%83%BD%E5%AE%9E%E6%88%983-%E8%AE%A1%E7%AE%97%E6%9C%BA%E8%A7%86%E8%A7%89%E6%8A%80%E6%9C%AF%E7%9A%84%E5%BA%94%E7%94%A8.pdf
- 视频：https://www.bilibili.com/video/BV17M411G7JG/ 【专题3:计算机视觉技术应用】1，2计算机视觉技术应用和CNN、对象检测基础知识和技术

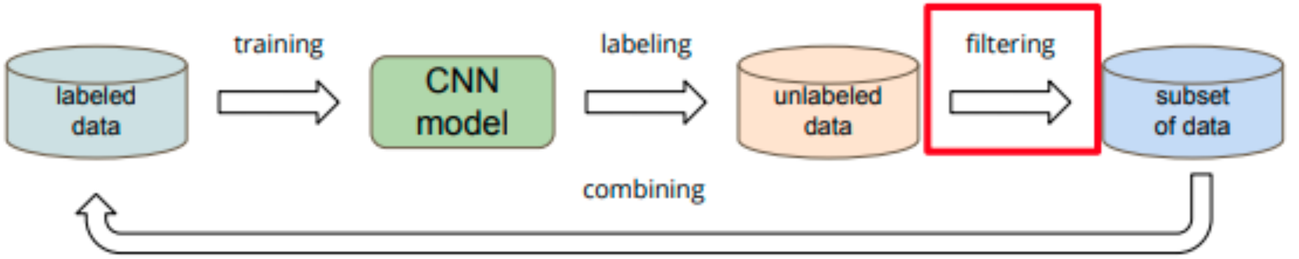
1 知识准备：半监督学习的原理

- Lee 2013年提出
- 使用一小部分有标签的数据和大量无标签的数据提高模型的性能
- 为未标记的标签生成伪标签，并训练

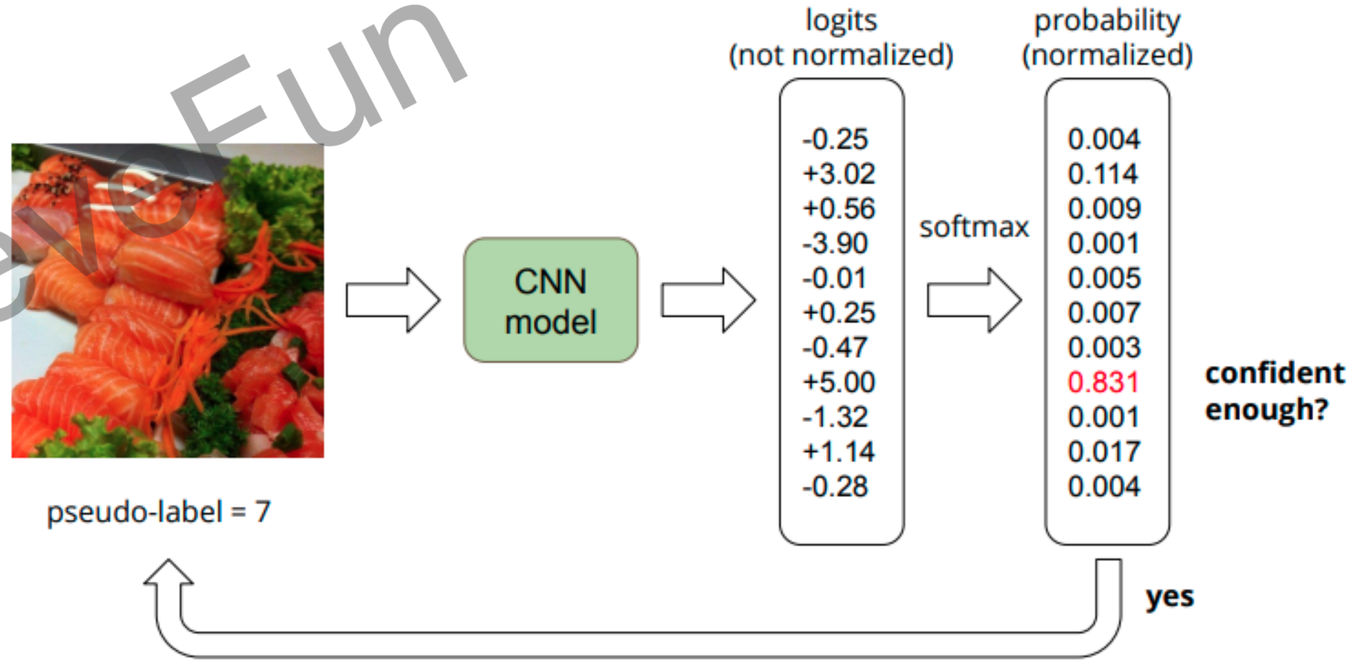


图片来源：https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/hw/HW03/HW03.pdf

伪标签方法的4个步骤



- 在一批有标签的数据上训练模型
- 使用训练好的模型预测一批无标签数据的标签
- 使用预测的标签计算无标签数据的损失
- 将有标签的损失与无标签的损失结合起来，进行反向传播



图片来源：https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/hw/HW03/HW03.pdf

3 案例演示

- 案例来源：
  - https://paperswithcode.com/paper/semi-supervised-classification-with-graph
  - https://docs.dgl.ai/tutorials/models/1_gnn/1_gcn.html
- 代码
  - https://github.com/dmlc/dgl/tree/master/examples/pytorch/gcn
  - https://docs.dgl.ai/_downloads/12e99dd8b30e32f2fe4cf6b5a3e27af3/1_gcn.py
  - https://docs.dgl.ai/_downloads/4a28323096e685201ab0a13483dfbaa3/1_gcn.ipynb

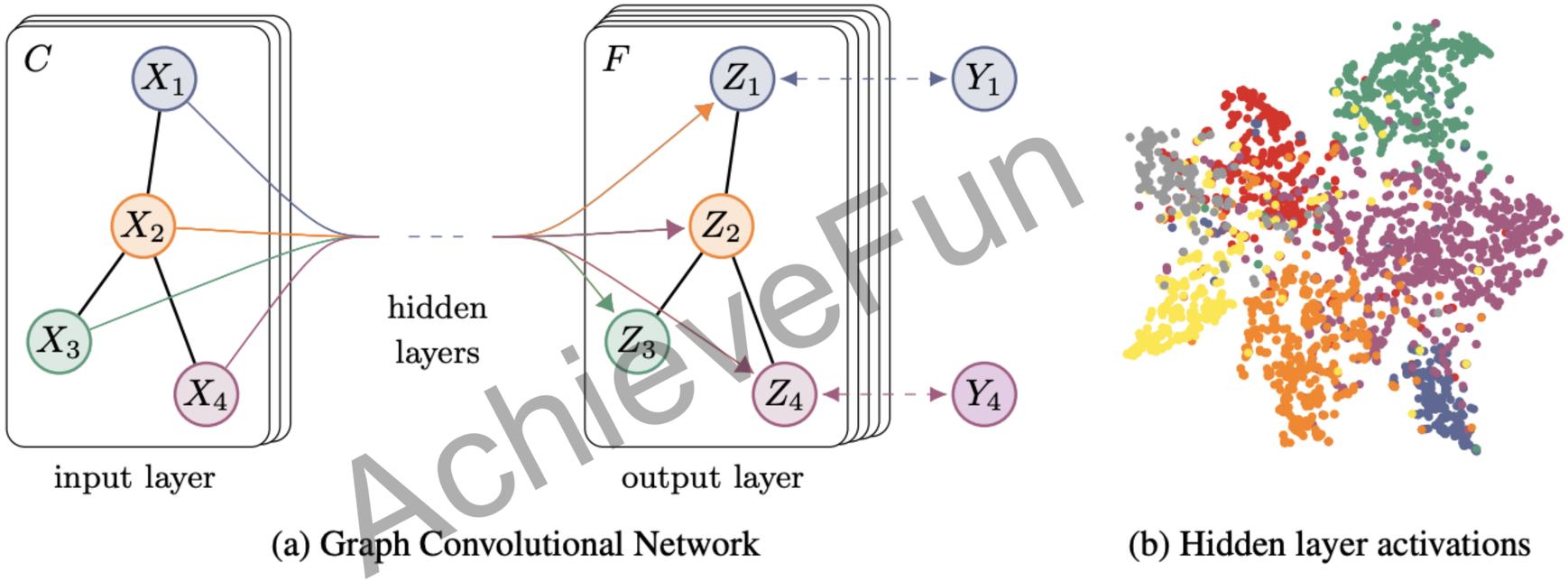
SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS

- We present a scalable approach for semi-supervised learning on graph-structured data that is based on an efficient variant of convolutional neural networks which operate directly on graphs. We motivate the choice of our convolutional architecture via a localized first-order approximation of spectral graph convolutions. Our model scales linearly in the number of graph edges and learns hidden layer representations that encode both local graph structure and features of nodes. In a number of experiments on citation networks and on a knowledge graph dataset we demonstrate that our approach outperforms related methods by a significant margin.

- 来源： https://arxiv.org/pdf/1609.02907v4.pdf

我们提出了一种对图结构数据进行半监督学习的可扩展方法，该方法基于直接在**图**上运行的**卷积神经网络**的高效变体。我们通过对**谱图卷积的局部一阶近似**来激励我们选择卷积架构。我们的模型与图边的数量成线性关系，并学习同时**编码局部图结构和节点特征的隐藏层表征**。在对引文网络和知识图谱数据集的大量实验中，我们证明我们的方法明显优于相关方法。

用于半监督学习的多层 GCN



(a) Graph Convolutional Network

(b) Hidden layer activations

https://arxiv.org/pdf/1609.02907v4.pdf

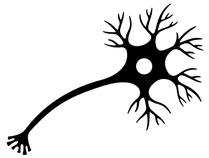
用于半监督学习的多层 GCN的优点

- 克服以下局限：
  - 基于图-拉普拉斯正则化的方法（Zhu 等人，2003 年；Belkin 等人，2006 年；Weston 等人，2012 年）很可能由于其假设边仅编码节点的相似性而受到限制。
  - 基于跳过图的方法由于基于难以优化的多步骤流水线而受到限制。
- 优点：
  - 在效率（measured in wall-clock time）方面仍然优于相关方法。
  - 与只聚合标签信息的 ICA（Lu & Getoor，2003 年）等方法相比，在每一层传播邻近节点的特征信息可提高分类性能
  - 进一步证明，所提出的重归一化传播模型（公式 8）既提高了效率（减少了参数和运算，如乘法或加法），又提高了对大量数据的预测性能。
- 来源： https://arxiv.org/pdf/1609.02907v4.pdf

学习目标

- 如何在半监督环境下对输入图的节点进行分类。
- 使用图卷积神经网络 (GCN) 作为图特征的嵌入机制。

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

代码运行流程

- 问题：在一个图（如引文网络）中对节点（如文档）进行分类，在这个图中，只有一小部分节点有标签。这个问题可以归结为基于图的半监督学习，即通过某种形式的显式基于图的正则化对标签信息进行平滑处理（Zhu 等人，2003 年；Zhou 等人，2004 年；Belkin 等人，2006 年；Weston 等人，2012）
- 流程：
  - 定义消息和还原 reduce 函数
  - 定义GCNLayer 模块
  - 装载数据集
  - 定义模型评估方法
  - 训练网络

代码演示

- https://docs.dgl.ai/_downloads/12e99dd8b30e32f2fe4cf6b5a3e27af3/1_gcn.py
- https://docs.dgl.ai/_downloads/4a28323096e685201ab0a13483dfbaa3/1_gcn.ipynb
- https://github.com/dmlc/dgl/tree/master/examples/pytorch/gcn

3 参考资料

- https://docs.dgl.ai/tutorials/models/1_gnn/1_gcn.html
- https://arxiv.org/pdf/1609.02907v4.pdf
- https://github.com/dmlc/dgl/tree/master/examples/pytorch/gcn
- https://docs.dgl.ai/_downloads/12e99dd8b30e32f2fe4cf6b5a3e27af3/1_gcn.py
- https://docs.dgl.ai/_downloads/4a28323096e685201ab0a13483dfbaa3/1_gcn.ipynb