基于 GCN 的图数据分类问题

辛未 郑钥方

Peking University

2019 年 4 月 23 日

Overview

问题背景

卷积神经网络 (CNN) 是一种重要的深度学习模型，能很好地处理图像、矩阵等数据类型，在众多领域有广泛的应用。然而，许多现实中重要的数据集以图形或网络的形式出现，例如社交多媒体网络数据，化学成分结构数据，生物基因蛋白数据以及知识图谱数据等。我们通常将以上的数据结构统称为非欧几里得数据。对这种数据找到合适的处理方式，是极具现实意义的话题。

图卷积神经网络 GCN

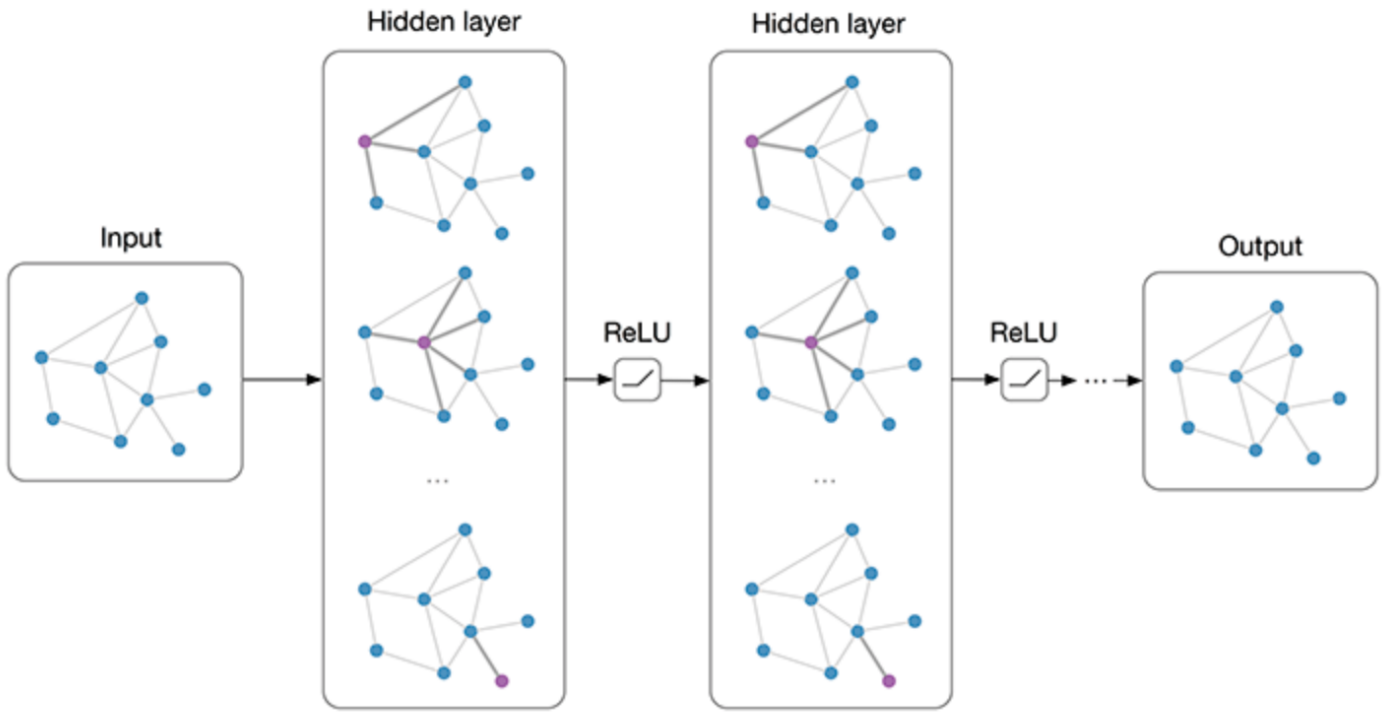
图卷积神经网络（Graph Convolutional Network, GCN）处理的对象是指数学中的拓扑图，对于图模型问题有着很好的解决能力。同时，社交网络、城市交通、关联结构等等很多问题也可以在图结构框架下得到很好的解释。所以图结构问题是很多实际问题的基础。

图卷积神经网络 GCN

与传统的 CNN 相比，GCN 能在以下方面发挥优势：

- 能处理非欧几里得数据，适应范围广
- GCN 模型的参数规模小很多，训练速度快

图卷积神经网络的经典结构



Multi-layer Graph Convolutional Network (GCN) with first-order filters.

图上的神经网络

- 在任意结构的图上推广 CNN/RNN 是一个很困难的问题。但是 2015 年到现在，陆续出现了一些基于具体问题的算法和基于谱图理论的算法。
- convolutional, because filter parameters are typically shared over all locations in the graph. ——Kipf

Graphical Convolutional Networks: Defination

- **目标：从信号和特征中学出一个函数模型。**
- **输入：** A feature description xi for every node i ; summarized in a $N \times D$ feature matrix X (N : number of nodes, D : number of input features)
- A representative description of the graph structure in matrix form; typically in the form of an adjacency matrix A (or some function thereof)
- **每个神经网络层可以写成非线性函数**

$$H^{(l+1)} = f(H^{(l)}, A)$$

简单例子

假设层之间的更新规则：

$$f(H^{(l)}, A) = \sigma\left(AH^{(l)}W^{(l)}\right),$$

则这个模型虽然很简单，但已经很强大了。其中 A 是邻接矩阵。我们还需要将模型归一化为

$$f(H^{(l)}, A) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right),$$

其中 $\hat{A} = A + I$

图数据集

- Cora Dataset
  - 引文网络，每篇文章作为一个结点，每篇文章的关键词是结点的一个特征。结点之间的连接关系是由引用来刻画的。
- Nashville Meetup Network
  - 每个人可以作为若干个 social group 中的一员；
  - 每个 group 会举办若干个 events

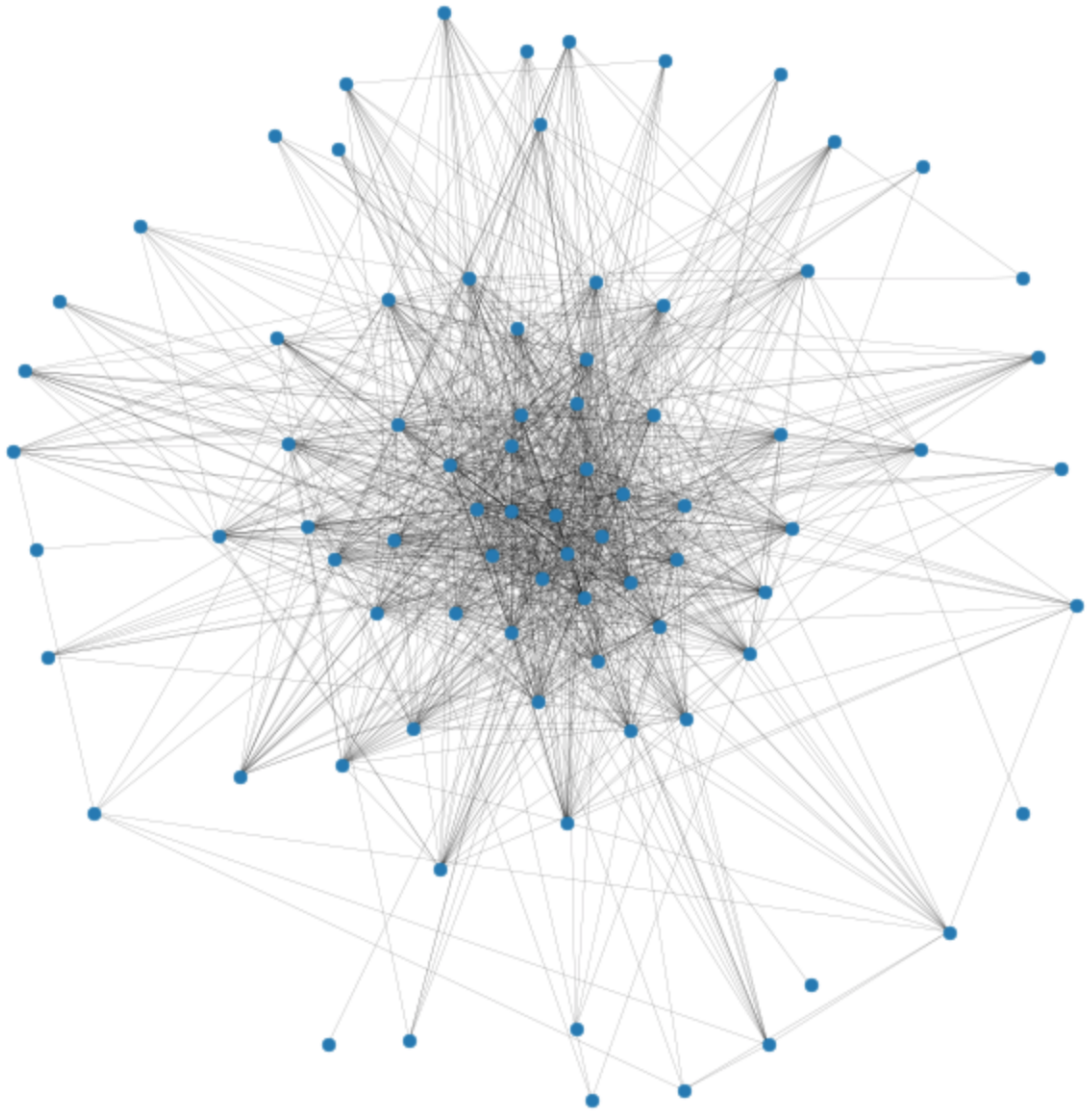
Cora Dataset

- cora.content 文件：
  - 2708 篇机器学习领域的论文，主题分为 7 类作为标签
  - 所有论文根据 1433 个关键词生成的 one-hot 向量
- cora.cites 文件：
  一个拓扑图，包含所有论文的引用关系，其中两个结点之间连一条边，当且仅当两篇论文之间有引用关系

Nashville Meetup Network Dataset

- Graph data
  - member-to-group-edges.csv：成员属于哪些团队，成员活动最多的团队作为标签
  - group-edges.csv：团队之间的关系
  - member-edges.csv：成员之间的关系
- Metadata
  - meta-groups.csv：团队的信息，包括组成人员
  - meta-members.csv：成员的信息，包括属于的团队和姓名等
  - meta-events.csv：活动的信息，包括参加人员和时间等

Nashvile Tech Meetups 的网络结构



GCN 模型

对 Cora Dataset 和 Nashvile Tech Meetups 两个图数据集，我们使用含有两个图卷积层的 GCN 模型进行训练。
我们使用的参数为 epoch=200，lr=0.01，其他参数的选取详见 train.py。

Cora Dataset 的训练结果

	训练集	验证集	测试集
loss	0.361	0.669	0.674
accuracy	0.964	0.833	0.843

Nashvile Tech Meetups 的训练结果

由于 Nashvile Tech Meetups 数据集中 group 的数目较大（81），我们选取含有 10 个 group 的子集进行训练和测试，得到的结果如下表所示：

	训练集	验证集	测试集
accuracy	0.860	0.736	0.729

总结

通过对 Cora 和 Nashvile Tech Meetups 这两个图数据集进行训练，我们均得到了较高的测试正确率。同时我们也注意到，在两个数据集上的训练速度都很快，特别是在 Cora 数据集上仅用了 3.4395 秒就完成训练，这说明 GCN 的计算效率很高。结合两个问题的特点，可以说明 GCN 对于图网络模型有较好的适应能力。

References

- https://arxiv.org/abs/1609.02907
- https://arxiv.org/abs/1606.09375
- http://geometricdeeplearning.com
- https://www.experoinc.com/post/node-classification-by-graph-convolutional-network