|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 论文题目 | Graph Convolutional Networks for Text Classifification | | | |
| Paper URL | [https://www.cs.virginia.edu/~ms5sw/GCN\_for\_NLP.pdf](http://arxiv.org/abs/1809.05679" \l ":~:text=Text classification is an important,e.g., sequence) to classification.) | | | |
| Project URL | [https: //github.com/yao8839836/text\_gcn](https://github.com/Weifeng-Liu/deeprl4fl-2021-icse) | | | |
| 综述/背景介绍 | 发展状况 | 原因 | 意义 | 关键词（速记词汇、信息索引词汇） |
| 文本分类的一个基本中间步骤是文本表示。传统的方法用手工制作的特征来表示文本，例如(例如，bag-of-words 和n-grams)。深度学习模型已被广泛用于学习文本表示，包括(CNN)、(RNN)、(LSTM)。忽略了携带非连续和长距离语义的语料库中的全局词共现。图神经网络是一种有效的处理方法，本文提出了一种基于图嵌入的图神经网络的文本分类方法。 | | | 文本分类、TextGCN |
| 假设 |  | | |  |
| 方法描述(含图) | 基于单词共现和文档单词关系为语料库的单文本图，然后学习语料库的文本图卷积网络(Text GCN)。我们的文本GCN用单词和文档的一个热表示初始化，然后它共同学习单词和文档的嵌入，由文档的已知类标签监督。  对于一层GCN，新的k维节点特征矩阵 L(1) ∈ Rn×k 计算为：    堆叠多个GCN层，以合并更高阶的邻域信息：    在构建了文本图之后，我们将该图输入一个简单的两层GCN： | | |  |
| 实验设计 | 1. 研究问题 2. 在有限的标记数据条件下，我们的模型能否在文本分类方面取得令人满意的结果？ 3. 我们的模型可以学习预测词和文档嵌入吗？ 4. 改变窗口大小和第一个GCN层维度大小这两个参数，其准确率的变化。 5. 标记数据大小的影响。 6. 数据集   20-Newsgroups (20NG), Ohsumed, R52 and R8 of Reuters 21578 and Movie Review (MR).   1. 评估指标   accuracy | | |  |
| 数据处理 | 输入 | 筛除特例 | 处理方式 | 关键词（速记词汇、信息索引词汇） |
| 单词和文档 | 将文本进行清理和标记为(Kim2014)。然后，我们删除了NLTK6中定义的停止单词，以及在20NG、R8、R52和Ohsumed中出现的次数小于5次的低频单词。唯一的例外是Mr，我们不会在原始文本进行清理和标记后删除单词，因为文档非常短。 |  |  |
| 结论 | 我们提出了一种新的文本分类方法，即文本图卷积网络(Text-GCN)。对于整个语料库，我们构建了一个异构的词文档图，并将文档分类转化为一个节点分类问题。利用有限的标记文档，文本GCN可以很好地捕获全局单词并发信息。一个简单的两层文本GCN在多个基准数据集上优于最先进的方法。未来的一些发展方向包括利用注意机制提高分类性能，以及开发无监督文本GCN框架，用于大规模无标记文本数据的表示学习。 | | | |
| 局限性分析 | GCN模型在GCN训练包括测试文档节点（没有标签）。因此，文本GCN不能快速生成嵌入，并对未看到的测试文档进行预测。 | | | |

2.论文总结：

**（论文名）：**用于文本分类的图卷积网络

**（题目）**：Graph Convolutional Networks for Text Classifification

**（论文URL）：**[https://www.cs.virginia.edu/~ms5sw/GCN\_for\_NLP.pdf](http://arxiv.org/abs/1809.05679" \l ":~:text=Text classification is an important,e.g., sequence) to classification.)

**（项目URL）：**[https: //github.com/yao8839836/text\_gcn](https://github.com/Weifeng-Liu/deeprl4fl-2021-icse)

**（总结）：**

我们提出了一种新的文本分类方法，即文本图卷积网络(Text-GCN)。对于整个语料库，我们构建了一个异构的词文档图，并将文档分类转化为一个节点分类问题。利用有限的标记文档，文本GCN可以很好地捕获全局单词并发信息。一个简单的两层文本GCN在多个基准数据集上优于最先进的方法。未来的一些发展方向包括利用注意机制提高分类性能，以及开发无监督文本GCN框架，用于大规模无标记文本数据的表示学习。

**（附图）：**

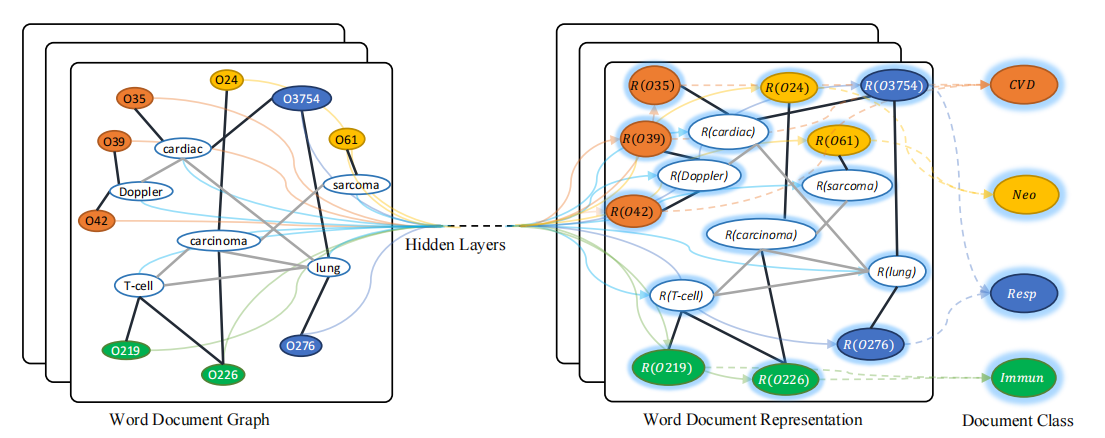


图1：TextGCN的示意图描述

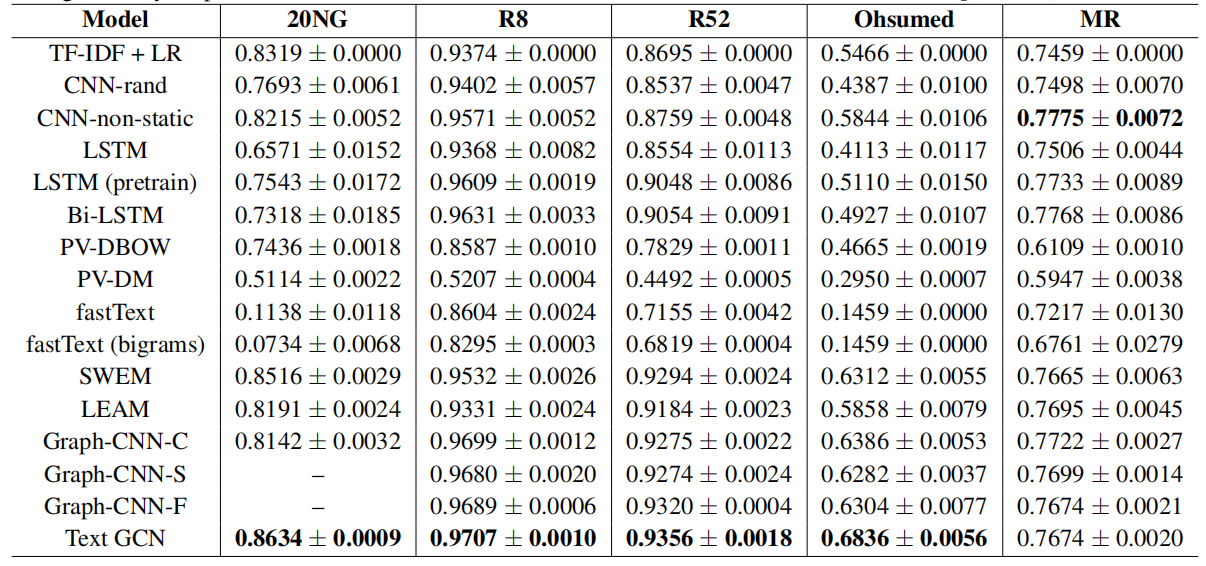


图2：对文档分类任务的测试准确性

文本GCN的性能最好，且显著优于所有的基线模型，这表明了该方法在长文本数据集上的有效性。

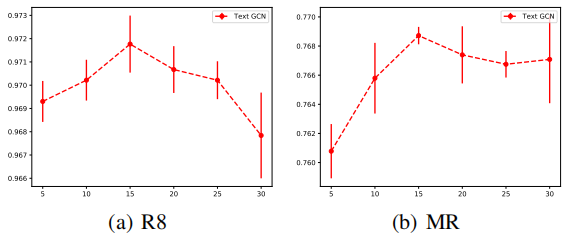


图3：在不同的滑动窗口尺寸下的测试精度

这表明，太小的窗口大小不能生成足够的全局词共线信息，而太大的窗口大小可能会增加节点之间不是非常相关的边。

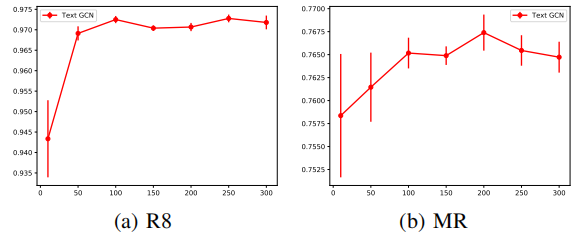


图4：使用不同尺寸的第一层嵌入物的测试精度

过低的维嵌入可能不能很好地将标签信息传播到整个图中，而高维的嵌入则不能提高分类性能，并可能花费更多的训练时间。

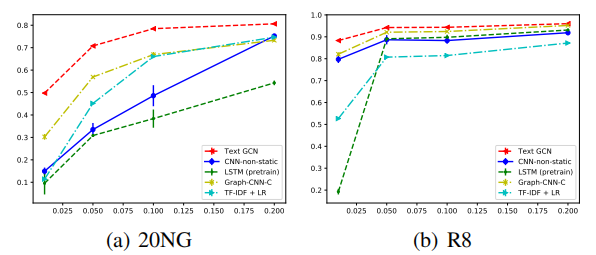


图5：使用不同比例的训练数据进行测试精度

GCN在低标签率下表现良好，说明GCN可以很好地将文档标签信息传播到整个图中，我们的词文档图保留了全局词共现信息。