BertGCN：传导文本分类结合 GCN 和 BERT

林育晓♠, 孟玉贤♣, 孙晓飞♣
韩庆红♣, 堃匡♠, 李继伟♠♣和吴飞♠♠浙江大学计算机科学与技术学院
♣香农人工智能
{yuxiaolinling, kunkuang, jiwei_li, wufei}@zju.edu.cn
{yuxian_meng, xiaofei_sun, qinghong_han}@shannonai.com

抽象的

在本文中，我们提出了 BertGCN，这是一种将大规模预训练和传导学习相结合用于文本分类的模型。Bert-GCN 在数据集上构建异构图，并使用 BERT 表示将文档表示为节点。通过在 Bert-GCN 中联合训练 BERT 和 GCN 模块，所提出的模型能够充分利用两方面的优势：大规模预训练利用大量原始数据，传导学习通过图卷积传播标签影响，联合学习训练数据和未标记测试数据的表示。实验表明，BertGCN 在各种文本分类数据集上都实现了 SOTA 性能。[1] [2]

1 介绍

文本分类是自然语言处理 (NLP) 中的一项核心任务，已用于许多实际应用中，例如垃圾邮件检测（王， 2010) 和意见挖掘（Bakshi 等人，2016). 传导学习（瓦普尼克，1998) 是一种特殊的文本分类方法，它在训练过程中同时使用标记和未标记的示例。图神经网络 (GNN) 是一种有效的传导学习方法（姚等人，2019；刘等人，2020在这些工作中，构建了一个图来模拟文档之间的关系。图中的节点表示文本单元，例如单词和文档，而边则基于节点之间的语义相似性构建。然后将 GNN 应用于图以执行节点分类。GNN 和传导学习的优点如下：（1）对实例（训练和测试）的决策并不不仅仅依赖于自身，还依赖于邻居。这使得模型对数据异常值具有更强的免疫力；（2）在训练时，由于模型通过图边缘在训练和测试实例中传播监督标签的影响，未标记的数据也有助于表征学习的过程，从而提高性能。

大规模预训练最近已证明其在各种 NLP 任务中的有效性（Devlin 等人，2018；刘等人，2019)。在无监督的情况下，大规模预训练模型在大规模无标记语料库上进行训练，能够大规模学习语言中隐含但丰富的文本语义。直观地看，大规模预训练模型有潜力促进传导学习。然而，现有的传导文本分类模型（姚等人，2019；刘等人，2020) 没有考虑大规模预训练，其有效性仍不清楚。

在本文中，我们提出了 BertGCN，这是一种结合大规模预训练和传导学习优势的文本分类模型。BertGCN 为语料库构建一个异构图，节点为单词或文档，节点嵌入使用预训练的 BERT 表示初始化，并使用图卷积网络 (GCN) 进行分类。通过联合训练 BERT 和 GCN 模块，所提出的模型能够充分利用两全其美的优势：大规模预训练利用大量原始数据，传导学习通过图边缘传播标签影响，联合学习训练数据和未标记测试数据的表示。所提出的 BertGCN 模型成功地结合了大规模预训练和图网络的功能，并在广泛的文本分类上取得了新的最佳性能

[1] 代码 可用的 在 https://github.com/ZeroRin/BertGCN。
[2] 被 ACL2021 的调查结果接受。

数据集。

2 相关工作

图神经网络 (GNN) 是一种联结主义模型，它通过连接节点的边传递消息来捕获图节点之间的依赖关系和关系（Scarselli 等人，2008； 汉密尔顿等人，2017；徐等人，2018)。 GNN 实际上分为（吴等人，2020）：图卷积网络（基普夫和威灵， 2016 年；吴等人，2019)、图注意力网络（Veličković 等人，2017；张等人，2018 年)、图自动编码器（曹等人，2016；基普夫和威灵， 2016b)、图生成网络（德曹和基普夫，2018；李等人，2018b) 和图形时空网络（李等人，2017；余等人，2017GNN 是利用不同对象之间关系的有力工具，已应用于交通预测等各个领域（余等人，2018；张等人，2018 年) 和建议 (张等人，2020；Monti 等人，2017在 NLP 领域，GNN 在关系提取（张等人， 2018b)、语义角色标注（马尔凯吉亚尼和蒂托夫，2017 ）、数据到文本生成（Marcheggiani 和 Perez-Beltrachini，2018)、机器翻译（Bastings 等人，2017) 和问答 (宋等人，2018；曹德等人，2018）。

神经网络的流行激发了各种各样的研究，开发用于文本分类的神经模型。不同的神经模型架构（金，2014；周等人，2015；Radford 等人，2018；柴等人，2020）已经证明了它们对基于传统统计特征的方法的有效性（瓦拉赫，2006 其他研究则利用标签嵌入，并将它们与输入文本一起进行联合训练（王等人，2018；帕帕斯和亨德森，2019最近，大规模预训练模型所取得的成功激发了人们对采用大规模预训练框架的极大兴趣（Devlin 等人，2018) 转化为文本分类 (赖默斯和古列维奇， 2019), 从而在 fewshot 上取得了显著的进步 (穆克吉和阿瓦达拉，2020）和零射击（叶等人，2020） 学习。

我们的工作受到使用图神经网络进行文本分类工作的启发（姚等人， 2019；黄等人，2019；张、张，2020）但与这些作品不同的是，我们关注的是将大规模预训练模型与 GNN 结合起来，并表明 GNN 可以从大规模预训练中获益匪浅。现有的结合 BERT 和 GNN 的工作使用图来模拟单个文档样本中 token 之间的关系（卢等人，2020；他等人，2020b)，属于归纳学习的范畴。与这些工作不同的是，我们使用图来建模来自整个语料库的不同样本之间的关系，以利用标记和未标记文档之间的相似性，并使用 GNN 来学习它们之间的关系。

3 方法

3.1 BertGCN

在提出的 BertGCN 模型中，我们使用 BERT 样式的模型（例如 BERT、RoBERTa）初始化文本图中文档节点的表示。这些表示用作 GCN 的输入。然后，将使用 GCN 根据图结构迭代更新文档表示，其输出将被视为文档节点的最终表示，并发送到 softmax 分类器进行预测。通过这种方式，我们能够利用预训练模型和图模型的互补优势。

具体来说，我们按照 TextGCN 构建一个包含单词节点和文档节点的异构图（姚等人，2019)。我们分别基于词频-逆文档频率（TF-IDF）和正逐点互信息（PPMI）来定义词-文档边和词-词边。两个节点之间的边的权重*我*和*杰*定义为：

$$A_{我，} = \begin{cases} \text{PPMI（}我\text{，），} & 我\text{，是文字和}我6\text{=}吉\text{是} \\ \text{TF-IDF(}我\text{，），} & \text{文件，}杰\text{是单词 }我\text{=}杰 \\ 1, & \\ 0, & \text{否则} \end{cases} \tag{1}$$

在 TextGCN 中，一个单位矩阵$X=我_{n_{文档}+n_{单词}}$用作初始节点特征，其中$n_{文档}$是文档节点的数量，$n_{单词}$是单词节点的数量（包括训练和测试）。在 BertGCN 中，我们使用 BERT 风格的模型来获取文档嵌入，并将其视为文档节点的输入表示。文档节点嵌入表示为$X_{文档} \in \mathrm{R}^{n_{文档} \times d}$，在哪里$d$是嵌入维数。总体而言，

初始节点特征矩阵由下式给出：

$$X= \begin{pmatrix} X_{文档} \\ 0 \end{pmatrix}_{(n_{文档}+n_{单词}) \times d} \tag{2}$$

我们喂养X进入 GCN 模型（基普夫和威灵，2016 年）在训练和测试样本之间迭代传播消息。具体来说，输出特征矩阵$我$第 GCN 层$大号$（$我$）是
计算为

$$大号^{(我)}=\rho(埪^{(我-1)}西^{(我)}) \quad (3)$$

在哪里ρ是激活函数，A是标准化的邻接矩阵，$西^{(我)}\in \mathbb{R}^{d_{我-1}\times d_{我}}$是该层的权重矩阵。$大号^{(0)}=X$是模型的输入特征矩阵。GCN 的输出被视为文档的最终表示，然后输入到 softmax 层进行分类。

阳离子：

$$是_{\text{地理信息网络}}=\text{softmax}(G(X,A)) \quad (4)$$

在哪里G表示 GCN 模型。我们使用标记文档节点上的交叉熵损失来联合优化 BERT 和 GCN 的参数。

3.2 插值 BERT 和 GCN 预测

实际上，我们发现，使用直接对 BERT 嵌入进行操作的辅助分类器来优化 BertGCN 可以实现更快的收敛速度和更好的性能。具体来说，我们通过直接输入文档嵌入来构建辅助分类器（表示为X）到具有softmax激活的密集层：

$$是_{\text{BERT}}=\text{softmax}(王晓玲) \quad (5)$$

最终的训练目标是 BertGCN 的预测和 BERT 的预测的线性插值，如下所示：

$$是=\lambda_{\text{地理信息网络}}+(1-\lambda)是_{\text{BERT}} \quad (6)$$

在哪里λ控制两个目标之间的权衡。λ=1意味着我们使用完整的 BertGCN 模型，并且λ=0意味着我们只使用 BERT 模块。当$\lambda\in$（0，1）、我们能够平衡两个模型的预测，并且 BertGCN 模型可以得到更好的优化。

插值取得更好性能的解释如下：$是_{\text{BERT}}$直接对 GCN 的输入进行操作，确保 GCN 的输入得到调节并朝着目标进行优化。这有助于多层 GCN 模型克服梯度消失或过度平滑等固有缺陷（李等人，2018 年），从而带来更佳的表现。

3.3 使用Memory Bank进行优化

原始 GCN 模型采用全批量梯度下降法进行训练，这对于提出的 BertGCN 模型来说是难以解决的，因为全批量方法不能应用于 BERT

由于内存限制。受到对比学习中将词典大小与小批量大小分离的技术的启发（吴等人，2018；他等人，2020 年），我们引入了一个存储所有文档嵌入的存储库，以将训练批次大小与图中的节点总数分离。

具体来说，在训练过程中，我们维护一个记忆库$米$跟踪所有文档的输入特征
文档节点。在每个 epoch 开始时，我们首先使用$当前的$BERT 模块并存储在$米$每次迭代期间，我们从标记和未标记的文档节点中抽取一个小批量，索引集$乙=\{b_0, b_1 \ldots b_n\}$，在哪里$n$是小批量大小。然后我们计算它们的文档嵌入$米_乙$还可以使用$当前的$BERT模块并更新相应的内存$米$。[3]

接下来，我们使用更新后的$米$作为输入来导出 GCN 输出并计算当前小批量的损失。对于反向传播，$米$被视为常数，除了记录$乙$。

有了记忆库，我们能够高效地训练包含 BERT 模块的 BertGCN 模型。然而，在训练过程中，记忆库中的嵌入是使用 BERT 模块在一个时期的不同步骤计算的，因此不一致。为了解决这个问题，我们为 BERT 模块设置了一个较小的学习率，以提高存储嵌入的一致性。学习率低时，训练需要更多时间。为了加快训练速度，我们在训练开始前对目标数据集上的 BERT 模型进行了微调，并使用它来初始化BertGCN中的BERT参数。

4 实验

4.1 实验设置

我们对五个广泛使用的文本分类基准进行了实验：20 个新闻组（20NG）[4], R8

[3]请注意，用于计算的 BERT 模块$米_乙$是在最后一次迭代中完成训练的，这与用于计算初始$米$。

[4]http://qwone.com/~jason/20Newsgroups/

模型	20NG	R8	R52	奥苏梅德	先生
文本GCN	86.3	97.1	93.6	68.4	76.7
新加坡国立大学	88.5	97.2	94.0	68.5	75.9
BERT	85.3	97.8	96.4	70.5	85.7
罗伯塔	83.8	97.8	96.2	70.7	89.4
伯特·GCN	89.3	98.1	96.6	72.8	86.0
罗伯特·塔格恩	89.5	98.2	96.1	72.8	89.7
伯特盖特	87.4	97.8	96.5	71.2	86.5
罗伯特·盖特	86.5	98.0	96.1	71.2	89.2

表 1：不同模型在传导式文本分类数据集上的结果。我们运行所有模型 10 次并报告平均测试准确率。

和 R52[5]、奥苏梅德[6]和电影评论（MR）[7]。

我们将 BertGCN 与当前最先进的预训练和 GCN 模型进行比较：TextGCN（姚等人， 2019), SGC (吴等人，2019）、BERT（Devlin 等人，2018）和 RoBERTa（刘等人，2019）。数据集和基线的详细信息留在补充材料中。

我们遵循 TextGCN 中的协议来预处理数据。对于 BERT 和 RoBERTa，我们使用 [CLS] 标记的输出特征作为文档嵌入，然后使用前馈层来得出最终预测。我们使用 BERT根据以及两层 GCN 来实现 BertGCN。我们将 GCN 模块的学习率初始化为 1e-3，将微调后的 BERT 模块的学习率初始化为 1e-5。我们还使用 RoBERTa 和 GAT 实现了我们的模型（Veličković 等人， 2017）。GAT 变体与 GCN 变体在相同的图上进行训练，但通过注意力机制学习边权重，而不是使用预定义的权重矩阵。

4.2 主要结果

桌子1展示了每个模型的测试准确率。我们可以看到，BertGCN 和 RoBERTaGCN 在所有数据集上表现最佳。仅使用 BERT 和 RoBERTa 通常会比 20NG 以外的 GCN 变体表现更好，这要归功于大规模预训练带来的巨大优势。与 BERT 和 RoBERTa 相比，BertGCN 和 RoBERTaGCN 在 20NG 和 Ohsumed 数据集上的性能提升非常显著。这是因为 20NG 和 Ohsumed 中的平均长度比其他数据集中的长度要长得多：该图是使用 word-document 统计量构建的，

[5]https://www.cs.umb.edu/~smimarog/textmining/datasets/

[6]http://disi.unitn.it/moschitti/corpora.htm

[7]http://www.cs.cornell.edu/people/pabo/movie-review-data/

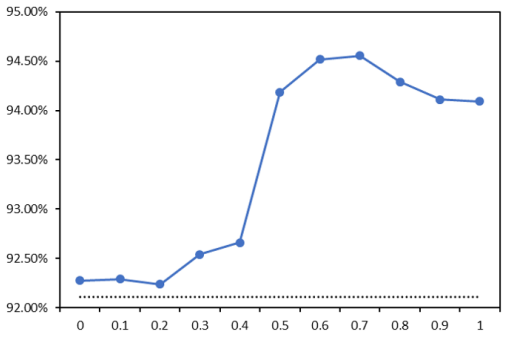

图 1：RoBERTaGCN 在不同训练集下的准确率λ在20NG开发集上。虚线表示相应的RoBERTa基线。[8]

战略	*两者皆有*	*无微调*	*无小 lr。*	*两者皆无*
准确性	94.7	93.8	10.3[9]	10.3[9]

表 2：不同策略在 20NG 开发集上的准确率。“finetune”表示我们使用微调后的 RoBERTa 作为初始化，“small lr.”表示我们对 RoBERTa 模块使用较小的学习率。

tics，这意味着长文本可能产生更多通过中间词节点中转的文档连接，这可能有利于消息通过图传递，从而与 GCN 结合时获得更好的性能。这也可以解释为什么 GCN 模型在 20NG 上的表现优于 BERT 模型。对于 R52 和 MR 等文档较短的数据集，图结构的力量有限，因此相对于 20NG 的性能提升较小。BertGAT 和 RoBERTaGAT 也可以从图结构中受益，但由于缺乏边权重信息，它们的性能不如 GCN 变体。

4.3 的影响λ

λ控制训练 BertGCN 和 BERT 之间的权衡。λ对于不同的任务可能会有所不同。图1展示了使用不同方法的 RoBERTaGCN 的准确率λ. 在 20NG 上，准确率始终较高，并且λ值。这可以通过基于图的方法在 20NG 上的高性能来解释。当λ=0。7、比仅使用 GCN 预测效果略好（λ=1).

4.4 联合训练中策略的影响

为了克服记忆库中嵌入的不一致性，我们为 BERT 模块设置了较小的学习率，并使用微调后的 BERT 模型进行初始化。我们评估了这两种策略的效果。表2展示了 RoBERTaGCN 在 20NG 上采用和不使用上述策略的结果。如果 RoBERTa 和 GCN 的学习率相同，无论是否使用经过微调的 RoBERTa，由于记忆库不一致，模型都无法训练。当我们为 RoBERTa 模块设置较小的学习率时，模型可以成功训练，而额外使用经过微调的 RoBERTa 可获得最佳性能。

5 结论和未来工作

在本研究中，我们提出了 BertGCN，它充分利用了大规模预训练模型和文本分类传导学习的优势。我们使用一个存储库来高效地训练 BertGCN，该存储库存储所有文档嵌入并根据采样的小批量更新其中的一部分。BertGCN 的框架可以构建在任何文档编码器和任何图形模型之上。实验证明了所提出的 BertGCN 模型的强大功能。然而，在这项研究中，我们仅使用文档统计数据来构建图形，与能够自动构建节点间边的模型相比，这可能不是最优的。我们将在未来的工作中解决这个问题。

致谢

该工作得到了国家重点研发计划（2020AAA0105200）和北京人工智能研究院（BAAI）的支持。

参考

Rushlene Kaur Bakshi、Navneet Kaur、Ravneet Kaur 和 Gurpreet Kaur。 2016.观点挖掘和情绪分析。在*2016年第三届可持续全球发展计算国际会议（INDIACom）*，第 452-455 页。IEEE。

Jasmijn Bastings、Ivan Titov、Wilker Aziz、Diego Marcheggiani 和 Khalil Sima'an。2017 年。用于语法感知神经机的图卷积编码器翻译。在*2017 年自然语言处理实证方法会议论文集*，第 1957-1967 页，丹麦哥本哈根。计算语言学协会。

Shaosheng Cao、Wei Lu 和 Qiongkai Xu。2016 年。用于学习图形表示的深度神经网络。 *AAAI 人工智能会议论文集*，第 30 卷。

Duo Chai、Wei Wu、Qinghong Han、Fei Wu 和 Jiwei Li。2020 年。基于描述的文本分类与强化学习。在*国际机器学习会议*，第 1371-1382 页。PMLR。

Nicola De Cao、Wilker Aziz 和 Ivan Titov。2018 年。通过图卷积网络跨文档推理来回答问题。*arXiv 预印本 arXiv:1808.09920*。

Nicola De Cao 和 Thomas Kipf。2018 年。Molgan：小分子图的隐式生成模型。 *arXiv 预印本 arXiv:1805.11973*。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018 年。Bert：用于语言理解的深度双向转换器的预训练。 *arXiv 预印本 arXiv:1810.04805*。

Will Hamilton、Zhitao Ying 和 Jure Leskovec。2017 年。大型图上的归纳表示学习。 *神经信息处理系统的进展*，第 1024-1034 页。

Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。2020a。无监督视觉表征学习的动量对比。在*IEEE/CVF 计算机视觉和模式识别会议论文集*，第 9729-9738 页。

何奇、王涵和张悦。2020b. 通过句法增强自然语言推理的泛化能力。*2020 年自然语言处理实证方法会议论文集：研究结果*，第 4973-4978 页。

黄连哲、马德宏、李苏建、张晓东、王厚峰。 2019. 用于文本分类的文本级图神经网络。*arXiv 预印本 arXiv:1910.02356*。

金允。2014 年。 卷积神经网络用于句子分类。 *arXiv:1408.5882* *arXiv 预印本*。

Thomas N Kipf 和 Max Welling。2016a。使用图卷积网络进行半监督分类。*arXiv 预印本 arXiv:1609.02907*。

Thomas N Kipf 和 Max Welling。2016b。国 变家图自动编码器。 *论文集 预印本 arXiv:1611.07308*。

Qimai Li、Zhichao Han 和 Xiao-Ming Wu。2018a。深入了解用于半监督学习的图卷积网络。*AAAI 人工智能会议论文集*，第 32 卷。

[8]20NG 的原始训练/测试划分是基于发布日期的，但开发集是从原始训练集中随机抽取的。因此测试集的准确率远低于开发集。

[9]没有小 lr 的实验无法收敛。

Yaguang Li、Rose Yu、Cyrus Shahabi 和 Yan Liu。2017 年。扩散卷积循环神经网络：数据驱动的交通预测。*arXiv 预印本 arXiv:1707.01926*。

Yujia Li、Oriol Vinyals、Chris Dyer、Razvan Pascanu 和 Peter Battaglia。2018b。学习图的深度生成模型。*arXiv 预印本 arXiv:1803.03324*。

Xien Liu、Xinxin You、Xiao Zhang、Ji Wu 和 Ping Lv。2020 年。用于文本分类的张量图卷积网络。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019 年。Roberta：一种稳健优化的 bert 预训练方法。 *预印本论文集 arXiv:1907.11692*。

鲁志斌、杜攀和聂建云。2020 年。bert：使用图嵌入增强 bert 进行文本分类。在*欧洲信息检索会议*，第 369-382 页。Springer。

Diego Marcheggiani 和 Laura Perez-Beltrachini。2018 年。用于结构化数据到文本生成的深度图卷积编码器。*第 11 届国际自然语言生成会议论文集*，第 1-9 页，蒂尔堡大学，荷兰。计算语言学协会。

Diego Marcheggiani 和 Ivan Titov。2017 年。使用图卷积网络对句子进行编码以进行语义角色标注。在*2017 年自然语言处理实证方法会议论文集*，第 1506-1515 页，丹麦哥本哈根。计算语言学协会。

Federico Monti、Michael M Bronstein 和 Xavier Bresson。2017 年。使用循环多图神经网络进行几何矩阵补全。*第 31 届神经信息处理系统国际会议论文集*，第 3700-3710 页。

Subhabrata Mukherjee 和 Ahmed Hassan Awadallah。2020 年。使用少量标签进行文本分类的不确定性感知自训练。

Bo Pang 和 Lillian Lee。2005 年。《看星星：利用类别关系进行评分量表的情绪分类》。*第 43 届计算语言学协会年会论文集（ACL'05）*, 第 115–124 页，密歇根州安娜堡。计算语言学协会。

Nikolaos Pappas 和 James Henderson。2019 年。Gile：用于文本分类的通用输入标签嵌入。*计算语言学协会会刊*, 7: 139–155。

Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。2018 年。通过生成式预训练提高语言理解能力。

Nils Reimers 和 Iryna Gurevych。2019 年。Sentencebert：使用暹罗 bertnetworks 进行句子嵌入。*arXiv 预印本 arXiv:1908.10084*。

Franco Scarselli、Marco Gori、Ah Chung Tsoi、Markus Hagenbuchner 和 Gabriele Monfardini。2008.图神经网络模型。*IEEE 神经网络学报*《细胞与分子生物学杂志》, 20(1): 61–80。

Linfeng Song、Zhiguo Wang、Mo Yu、Yue Zhang、Radu Florian 和 Daniel Gildea。2018 年。使用图神经网络探索图结构段落表示以实现多跳阅读理解。*arXiv 预印本 arXiv:1809.02040*。

Jian Tang、Meng Qu 和 Qiaozhu Mei。2015 年。Pte：通过大规模异构文本网络进行预测文本嵌入。在*第 21 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*，第 1165-1174 页。

弗拉基米尔・N・瓦普尼克 (Vladimir N. Vapnik)。1998 年。*统计学习理论*. 威利-Interscience。

Petar Veličković、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。2017 年。图注意力网络。 *arXiv 预印本 arXiv:1710.10903*。

Hanna M Wallach。2006 年。主题建模：超越词袋。在*第 23 届国际机器学习会议论文集*，第 977-984 页。

Alex Hai Wang。2010 年。不要关注我：Twitter 中的垃圾邮件检测。在*2010年安全与密码国际会议（SECRYPT）*，第 1-10 页。IEEE。

王国印、李春元、王文林、张一哲、沈定汉、张新元、Ricardo Henao 和 Lawrence Carin。 2018.用于文本分类的单词和标签的联合嵌入。 *arXiv 预印本 arXiv:1805.04174*。

Felix Wu、Tianyi Zhang、Amauri Holanda de Souza Jr、Christopher Fifty、Tao Yu 和 Kilian Q Weinberger。2019 年。简化图卷积网络。 *arXiv 预印本 arXiv:1902.07153*。

Zhirong Wu、Yuanjun Xiong、Stella X Yu 和 Dahua Lin。2018 年。通过非参数实例判别进行无监督特征学习。在*IEEE 计算机视觉和模式识别会议论文集*，第 3733-3742 页。

吴宗翰、潘诗睿、陈凤文、龙国栋、张承启和 S Yu Philip。 2020.图神经网络的全面调查。*IEEE 神经网络和学习系统学报*。

Keyulu Xu、Weihua Hu、Jure Leskovec 和 Stefanie Jegelka。2018 年。图神经网络有多强大？*arXiv 预印本 arXiv:1810.00826*。

Liang Yao、Chengsheng Mao 和 Yuan Luo。2019 年。用于文本分类的图卷积网络。 *AAAI 人工智能会议论文集*，第 33 卷，第 7370–7377 页。

叶志全、耿玉霞、陈娇燕、陈静敏、徐晓晓、郑肃航、王峰、张俊、

和陈华军。2020 年。通过强化自训练实现零样本文本分类。在 *计算语言学协会第 58 届年会论文集*, 第 3014–3024 页, 在线。计算语言学协会。

余兵、尹浩腾、朱占兴。2017 年。时空图卷积网络：一种用于交通预测的深度学习框架。*论文集预印本 arXiv:1709.04875*。

余兵、尹浩腾和朱占兴。2018 年。时空图卷积网络：用于交通预测的深度学习框架。*第 27 届国际人工智能联合会议论文集*，第 3634-3640 页。

张浩鹏和张嘉伟。2020.用于文档分类的文本图表转换器。在 *继续-2020 年自然语言处理经验方法会议（EMNLP）*，第 8322-8327 页。

Jiani Zhang、Xingjian Shi、Junyuan Xie、Hao Ma、Irwin King 和 Dit Yan Yeung。2018a。Gaan：用于大型和时空图学习的门控注意力网络。在 *2018 年第 34 届人工智能不确定性会议，UAI 2018*。

张胜宇、谭子琪、赵洲、余瑾、匡昆、蒋谭、周景仁、杨红霞、吴飞。 2020.视频字幕综合信息集成建模框架。在 *第 26 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*，第 2744-2754 页。

Yuhao Zhang、Peng Qi 和 Christopher D Manning。2018b。修剪依赖树上的图卷积改进了关系提取。*arXiv 预印本 arXiv:1809.10185*。

Chunting Zhou、Chonglin Sun、Zhiyuan Liu 和 Francis Lau。2015 年。用于文本分类的 c-lstm 神经网络。*arXiv 预印本 arXiv:1511.08630*。

数据集详细信息

- 20NG 数据集[10]包含来自 20 个不同主题的 18,846 个新闻组帖子。我们使用按日期划分的版本，根据发布日期将数据集分为 11,314 个训练样本和 7,532 个测试样本。

- R8 和 R52[11]是路透社数据集的两个子集，分别有 8 个类别和 52 个类别。R8 有 5,485 个训练文档和 2,189 个测试文档。R52 有 6,532 个训练文档和 2,568 个测试文档。

- OHSUMED 测试集[12]是来自在线医学信息数据库 MEDLINE 的一组参考文献。根据前人的研究，我们使用属于 23 个疾病类别之一的 7,400 篇文档作为分类数据集，其中 3,357 篇文档用于训练，4,043 篇文档用于测试。

- 先生 （彭和李，2005）[十三]是用于二元情绪分类的电影评论数据集。该语料库有 10,662 条评论。我们使用训练/测试拆分唐等人（2015）

乙 基线

- TextGCN（姚等人，2019): TextGCN 是一个对单词-文档异构图进行图卷积的模型，使用单位矩阵初始化节点特征。

- SGC（吴等人，2019): 简单图卷积是 GCN 的一种变体，通过消除非线性和折叠连续层之间的权重矩阵来降低 GCN 的复杂性。

- BERT（Devlin 等人，2018): BERT 是一个大规模预训练的 NLP 模型。
- RoBERTa（刘等人，2019）：一种稳健优化的 BERT 模型，通过不同的预训练方法对 BERT 进行改进。

[10] http://qwone.com/~jason/20Newsgroups/

[11] https://www.cs.umb.edu/~smimarog/textmining/datasets/

[12] http://disi.unitn.it/moschitti/corpora.htm

[十三] http://www.cs.cornell.edu/people/pabo/movie-review-data/