Software Engineering and Applications 软件工程与应用, 2023, 12(2), 230-241
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.122023

Hans汉斯

BAG：基于注意力机制融合Bert和GCN的文本分类模型

李　想[1,2]，马致远[1*]，汪　伟[1,2]，韩士洋[1,3]

[1]上海理工大学机器智能研究院，上海
[2]上海理工大学光电信息与计算机工程学院，上海
[3]上海理工大学机械工程学院，上海

收稿日期：2023年2月17日；录用日期：2023年4月3日；发布日期：2023年4月14日

摘　要

通过图的方式来建模文本分类任务是近年来研究的热点。现有基于图神经网络的方法虽然取得了一定的性能提升，但缺乏有效利用预训练语言模型获得的文本语义和图结构语义，且建图规模相对较大，由此带来的训练开销导致相关方法难以在低算力平台上使用。针对这些问题，在通过图神经网络构建传导式文本分类模型的过程中，利用注意力机制来融合异构图中的结构语义和预训练语言模型提供的字符级语义，在保留部分模型参数进行训练的基础上，提出了一种改进的文本分类模型BAG。实验结果表明，BAG能在更低显存的机器上进行训练，且在四个数据集上的准确率比其他文本分类模型更高。在对比同样基于图的TextGCN和BertGCN时，最高时分别高出10.82%和3.14%。

关键词

深度学习，文本分类，图卷积网络，注意力机制

BAG: Text Classification Based on Attention Mechanism Combining BERT and GCN

Xiang Li[1,2], Zhiyuan Ma[1*], Wei Wang[1,2], Shiyang Han[1,3]

[1]Institute of Machine Intelligence, University of Shanghai for Science and Technology, Shanghai
[2]School of Optical-Electrical Computer Engineering, University of Shanghai for Science and Technology, Shanghai
[3]School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 17th, 2023; accepted: Apr. 3rd, 2023; published: Apr. 14th, 2023

*通讯作者。

文章引用: 李想, 马致远, 汪伟, 韩士洋. BAG: 基于注意力机制融合 Bert 和 GCN 的文本分类模型[J]. 软件工程与应用, 2023, 12(2): 230-241. DOI: 10.12677/sea.2023.122023

Abstract

In recent years, text classification based on graph is the research focus. Although existing work of text classification based on graph neural network have achieved performance improvement, they lack the text semantics and graph semantics obtained by effectively using the pretrained language model. Moreover, the scale of graph is large and the training cost caused by this leads to the difficulty of using relevant methods on low computational power platforms. For this problem, in the process of building a transductive text classification model through graph neural network, attention mechanism is used to fuse the structural semantics in heterogeneous graphs and the token-level semantics provided by the pretrained language model. On the basis of retaining some model parameters for training, an improved text classification model BAG is proposed. The experimental results show that BAG can be trained on a lower memory machine, and accuracy on four datasets is higher than other text classification models. When comparing TextGCN and BertGCN, they are 10.82% and 3.14% higher at the highest.

Keywords

Deep Learning, Text Classification, Graph Convolution Network, Attention Mechanism

1. 引言

文本分类(Text Classification)作为自然语言处理(Natural Language Processing, NLP)中一项经典且重要的任务，在垃圾邮件检测[1]、观点挖掘[2]、新闻过滤[3]和情感分类[4]等方面有广泛的应用。随着人工智能的发展，大规模的文本数据及其应用层出不穷。从海量的文本数据中筛选、提取所需要的信息，避免重复无效的信息出现，需要对文本数据进行分类。因此，如何提高文本分类的准确率，尤其是在硬件环境受限的情况下保证准确率的稳定，成为文本分类中的热点问题。

传统基于机器学习的模型利用当前文本的规则或统计信息构建特征来实现分类，但随着训练样本的增加和领域的切换，分类结果往往不够稳定。近年来，以卷积神经网络(Convolutional Neural Network, CNN) [5]和循环神经网络(Recurrent Neural Network, RNN) [6]为基础的深度学习模型被广泛应用于文本建模的研究中。与传统模型相比，深度学习模型能够将语义和语法信息通过学习形成高维嵌入表示，有利于提高模型的泛化性。

然而，上述方法对建模非连续且间隔较远单词之间的语义关系时会存在一定限制，这在很大程度上影响了模型分类的准确率。例如，“尽管这部电影有些不合常理，甚至有些难以理解，但它仍能以目前很少有电影能达到的方式引来观影者的目光”。文本表达了对电影创新性的高度赞同，但“难以理解”等负面评价的出现可能导致分类器错误地将句子分成负面类。ELMo [7]、XLNet [8]和 BERT [9]等基于 Transformer 的预训练语言模型可以有效学习文本全局语义表示，并通过无监督的方式挖掘语义信息，显著提高文本分类性能，但在处理较长的文本时会截断输入文本，导致部分信息丢失。近年来，图神经网络(Graph Neural Network, GNN) [10]因能学习丰富有效的关系结构信息，并充分考虑文档的全局信息而被用于解决上述问题[11]。以 TextGCN [12]为代表的传导式模型为整个语料库构建一张基于词

频的异构图，在此基础上进一步通过 BERT 提供文本节点的语义信息，初步解决了预训练模型和图网络难以结合的问题。

而 BertGCN [13]在训练过程中仅通过提取<cls>标签来提供句子级别的语义嵌入表示，缺失了句子中其他语义信息。针对此问题，提出 BERT-Attention-GCN (BAG)模型，通过在 GCN [14]和 BERT 的特征输入之间增加注意力机制[15]来补全句子中除<cls>标签之外的其他 token 级别的语义信息，使全局信息和局部信息相互引导，防止训练过程中重要语义信息的丢失。此外，本文还针对 BERT 中的 Transformer 模块在表征语义的层次上进行筛选，使用跟文本分类任务关系更密切的上层语义来实现信息融合，使模型轻量化的同时分类性能不会因此降低，进而使模型在计算机算力有限的情况下仍能进行训练。后续验证实验表明，在这种情况下，BAG 模型的分类性能明显优于同类基于图网络的算法(TextGCN, BertGCN)。

本文的主要贡献与创新点概括如下：1) 提出了 BERT-Attention-GCN (BAG)模型，将 GCN 和 BERT 进行结合，在两者的特征输入之间增加注意力机制层，充分融合全局信息和局部信息，加强邻域内多特征之间的信息交互；2) 针对 BERT 中的 Transformer 模块在表征语义的层次上提出改进，在不降低分类性能的前提下使模型轻量化；3) 将本文模型在 R8，R52，Ohsumed 和 MR 数据集上进行实验，分类效果远高于 Graph-CNN 和 TextGCN 等深度学习模型，且在低等级算力和低参数设置下，在多个数据集上的结果优于 BertGCN 模型。

2. 相关工作

2.1. 文本分类概述

文本分类是从载有信息的原始文本中提取特征，并根据所提特征预测文本数据的类别主题的过程。传统基于机器学习的方法主要利用预先设定好的规则或统计模型来表述类别。例如 k 近邻(k-Nearest Neighbor, KNN) [16]、支持向量机(Support Vector Machine, SVM) [17]和朴素贝叶斯(Naive Bayes) [18]等。

CNN、RNN 和长短期记忆网络(Long Short-Term Memory, LSTM) [19]等通过构建深度神经网络，摒弃人工构建复杂而低效的特征工程等步骤，提高了模型的稳定性和鲁棒性[20]。Kim 等人[21]提出 TextCNN，将 CNN 应用在句子级的文本分类中，用卷积操作对输入文本进行特征提取，最终输出概率分布，分类效果显著。然而，CNN 只能有效地提取输入文本的局部特征，难以获取一段文本的上下文语义信息。Wang 等人[22]将 CNN 和 RNN 结合，Yang 等人[23]在此基础上引入注意力机制，虽然在某些方面取得一定成效，但上述研究的本质缺陷依然难以改变。

Peters 等人[7]提出的 ELMo、Brown 等人[24]提出的 GPT-3 和 Devlin 等人[9]提出的 BERT 等预训练语言模型可以在不同语境下有效学习到文本的全局语义表示，充分应用于文本分类等下游任务，但上述模型在处理较长文本时会丢失重要信息，难以有效提取长文本的句法结构信息。Hu 等人[25]将提示学习(Prompt Learning)应用在文本分类中，将下游任务转换成完形填空式的任务，充分应用预训练知识，在少样本学习(Few-Shot Learning)和零样本学习(Zero-Shot Learning)上取得优异的性能。Su 等人[26]利用对比学习(Contrast Learning) [27]增强 KNN 机制，模型在多标签文本分类任务上提高了性能。

近年来，引入图结构的方法开始在文本分类任务中流行。基于图网络的模型可以通过对文本图节点编码来提取文本的句法结构特征。Hao 等人[28]提出 Graph-CNN，该模型将文本转换为词图的形式，然后对词图进行卷积，使 CNN 学习不同层次的语义信息。Kipf 等人[14]提出一种充分利用邻域信息的图卷积网络 GCN，Yao 等人[12]将 GCN 应用在文本分类任务中提出 TextGCN，为整个语料库构建一张基于词频的异构图(Graph)，以单词和文档为节点，捕获高阶邻域节点信息，由此将文本分类问题转化为节点分类问题，该方法能够在有标签的文档数量比例较小的情况下得到较高的分类效果。由于只考虑全局信息的图网络模型无法对字符(token)级别的局部语义信息进行捕捉，受 GCN 和 BERT 的启发，Lin 等人[13]

提出了 BertGCN，通过联合训练 BertGCN 中的 GCN 模块和 BERT 模块，在长文本分类数据集上取得了领先(State of The Art, SOTA)性能。

2.2. 图卷积网络(GCN)概述

图卷积网络(GCN)可以从 graph 中提取特征并生成对应的表示向量，主要用于处理具有广义拓扑图结构的数据。GCN 将文本中的词以图节点的形式进行组合，通过在单词节点上添加更多的关系，将文本中的词以结构图的形式进行连接，并对所建图中相邻节点进行卷积计算，在一定程度上整合了特定领域文本的全局上下文信息。GCN 的每次计算，将每个节点和它的相邻节点的信息通过卷积操作聚集起来，并根据节点的邻域性质，推导节点的嵌入向量。Kipf 等人将时域的谱卷积运算转化为频域的矩阵乘法运算。

$$g * x \approx \theta\left(D^{-1/2} A D^{-1/2}\right) x \tag{1}$$

其中 g 是卷积核；x 是输入信号；*是卷积运算；A 是邻接矩阵；D 是 A 的对角化度矩阵，$D_{ij} = \sum_j A_{ij}$；$D^{-1/2} A D^{-1/2}$ 是邻接矩阵 A 的标准化形式，用于防止多层网络优化时梯度消失或爆炸。公式(1)表示单层 GCN 的卷积计算。

然而无论是建图还是特征提取，都要考虑计算和内存消耗。一方面，TextGCN 建立的图结构中边权重的表达会受到限制，影响文本分类性能，另一方面，只考虑全局词信息的 GCN 可能无法捕捉到局部信息(如文档中单词的顺序)，而这些信息对于理解语义更好地进行文本分类也是非常重要的。BertGCN 考虑利用预训练模型给图网络模型补充文本语义特征，但其对句子中 token 级别的语义信息利用并不充分，为了将文本的语义特征表示真正有效融入到图卷积网络中，本文提出在 GCN 和 BERT 的特征输入之间增加注意力机制层，充分利用全局信息和局部信息，增强文本之间的相关性，提升文本分类效果。

3. 方法

BAG 模型如图 1 所示。首先在语料库上构建异构图，文本图中的文档节点表示用 BERT 初始化，图中各节点之间的边权重基于术语频率逆文档频率(TF-IDF)和点互信息(PMI)定义[12]，被用作 GCN 的输入。GCN 通过卷积操作聚合节点表示，迭代更新，得到符合文本图特征的文本特征表示 f_{GCN}，将其与利用 BERT 预训练模型获得的包含文本信息的特征表示 f_{Bert} 联合训练，充分融合二者对数据不同层次的处理能力，引入注意力机制补全 token 级别的语义信息，将 GCN 输出的全局图特征和 BERT 输出的局部单词节点特征进行融合，将输出作为文本节点的最终表示通过线性层和分类器进行文本标签分类预测，分别得到输入文本的最终概率表达 y_{BAG} 和 y_{GCN}，最后将两部分的预测相结合，通过调节 λ 系数使模型参数达到最优解。

3.1. 图卷积网络模块

首先构建文本图 Graph = (V, E)，其中 V、E 分别是图中节点(node)、边(edge)的集合，表示节点的数量，GCN 的单层卷积如公式(2)所示：

$$L = \sigma\left(D^{-1/2} A D^{-1/2} X W\right) = \sigma\left(\tilde{A} X W\right) \tag{2}$$

其中 $\tilde{A} = D^{-1/2} A D^{-1/2}$，$\tilde{A} \in \mathbb{R}^{V \times V}$ 是 A 的归一化形式，$\tilde{A}$ 是具有自循环(Self-Loop)的邻接矩阵(对角线元素均为 1)，它反映图中节点的互连关系，$\tilde{A}$ 可以被分解为 $A + I_N$，I_N 是单位矩阵；σ 是激活函数；$W_0 \in \mathbb{R}^{m \times k}$ 是参数权重矩阵；X 是输入节点特征，设 $X \in \mathbb{R}^{n \times m}$ 为包含所有 n 个节点及其特征的矩阵，其中 m 为特征向量的维数；每一行 $x_v \in \mathbb{R}^m$ 为 v 的特征向量；$L \in \mathbb{R}^{n \times k}$ 是 k 维节点特征矩阵。

本文参考 TextGCN 的建图方式，将每个单词或文档都用 one-hot 编码向量化作为输入，如图 2 所示。

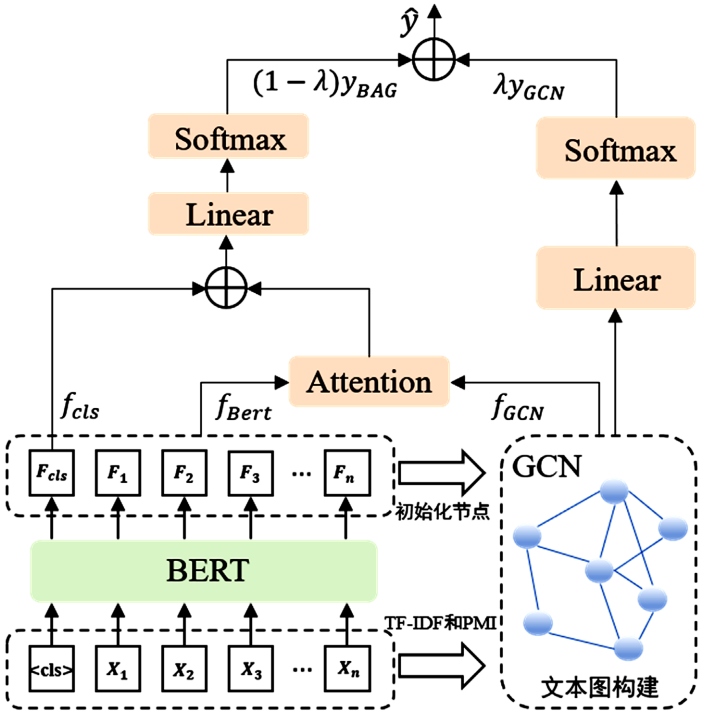

Figure 1. Model structure of BAG
图 1. BAG 模型结构图

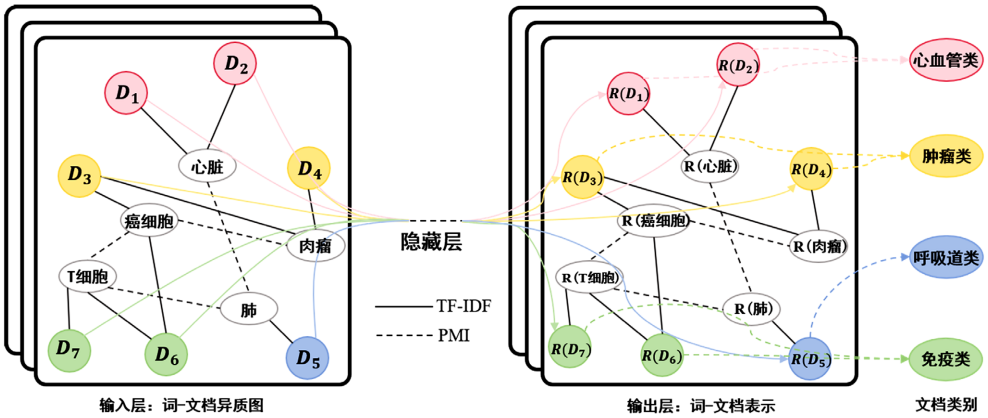

Figure 2. Schematic of graph (example taken from Ohsumed corpus)
图 2. 文本图构建原理图(以 Ohsumed 数据集为例)

单词节点单独表示，文档节点用 D_i 表示，不同颜色表示不同文档类别，单词节点与文档节点之间的边根据单词在文档中的出现来构建，用黑色实线表示，其权值是单词在文档中的 *TF-IDF* 值；单词与单词之间的边根据单词在整个语料库中的共现来构建，图中用黑色虚线表示，使用点对点互信息(*PMI*)来计算两个词节点之间的权重。定义节点 i 与节点 j 之间的边的权值如公式(3)所示：

$$A_{ij}=\begin{cases} PMI(i,j) & i,j \text{ are words, } PMI(i,j)>0 \\ TF\text{-}IDF_{ij} & i \text{ is document, } j \text{ is word} \\ 1 & i=j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

其中 $PMI(i,j)=\log\frac{p(i,j)}{p(i)p(j)}$，用来衡量两个事物之间的相关性；*TF-IDF* 为将词频(*TF*)和逆文档频率(*IDF*)相乘的结果，某个词在文档中的 *TF-IDF* 值越大，表示这个词在这篇文档的重要性就越高。

图卷积网络可以捕获单词和文档之间以及单词和单词之间的关系信息，在训练过程中，每个节点的标签信息可以通过它们的相邻节点传递给其他的单词和文档。将 BERT 初始化后得到的文档节点的语义特征表示输入到一个图卷积网络中，模型能在该网络中凭借已标注节点的信息推理未标注节点的特征，得到符合文本图特征的文本特征表示 f_{GCN}。具体而言，第 i 个 GCN 层的输出特征矩阵如公式(4)所示，第二层节点(单词或文档)嵌入具有与标签集相同的大小。使用 *Softmax* 作为分类器，形式化如公式(5)：

$$L^{(i)}=ReLU\left(\tilde{A}L^{(i-1)}W^{(i)}\right) \tag{4}$$

$$y_{GCN}=Softmax\left(g\left(X,A\right)\right) \tag{5}$$

其中，模型的第一层使用 *ReLU* 作为激活函数，输出视为文档的最终表示形式，输入到第二层使用 *Softmax* 函数进行分类优化，g 代表 GCN 层。

3.2. BERT 模块

BERT 作为一种预训练语言模型，可以作为词向量嵌入层输入到其他训练模型中，具有在大规模无标注语料库上预训练的能力。本文使用 BERT 模型对输入文本序列进行处理，进一步获取文本的字符、单词和上下文信息，生成文档嵌入，并将其作为文档节点的输入表示。使用 BERT 初始化文档节点，节点嵌入用 $X_{doc}\in\mathbb{R}^{n_{doc}\times d}$ 表示，其中 d 为特征向量的维度。把所有的单词节点 X_{word} 初始化为 0，得到文本输入如公式(6)所示：

$$X=\begin{bmatrix}X_{doc}\\0\end{bmatrix}_{(n_{doc}+n_{word})\times d} \tag{6}$$

文本输入经过 BERT 最终输出为包含文本信息的特征表示 f_{Bert}，如公式(7)所示：

$$f_{Bert}=\left[F_1,F_2,\cdots,F_n\right]=Bert\left[X_1,X_2,\cdots,X_n\right] \tag{7}$$

BERT 模型共有 12 层的隐层向量，可以捕获丰富的语义信息，其中低层向量和中层向量分别可以学习到基础词法信息和句法特征，而高层向量可以学习到更多跟下游任务相关的语义特征。因此，本文考虑到 GCN 占用的计算和内存消耗较大，为了加快模型的收敛速度，在训练模型的过程中对 BERT 进行层数上的筛选，只保留 BERT 模型后四层中的输出向量，使模型轻量化的同时不会影响最终的分类性能。

3.3. 注意力机制模块

注意力机制可以在众多输入信息中聚焦于对当前任务更关键的信息，提高文本分类任务的效率和准确性[29]。由于每条文本的类别通常由关键词和关键语句决定，引入注意力机制有助于提升分类过程中关键部分的权重，从而获得更高的分类精度。

由于 BertGCN 用 BERT 初始化节点给 GCN 补充文本的局部语义特征时，仅通过<cls>标签来提供句子级别的语义嵌入表示，缺失了句子中大量 token 级别的语义信息，本文使用注意力机制将剩余 token 级别的语义信息进行补全，如图 3 所示。

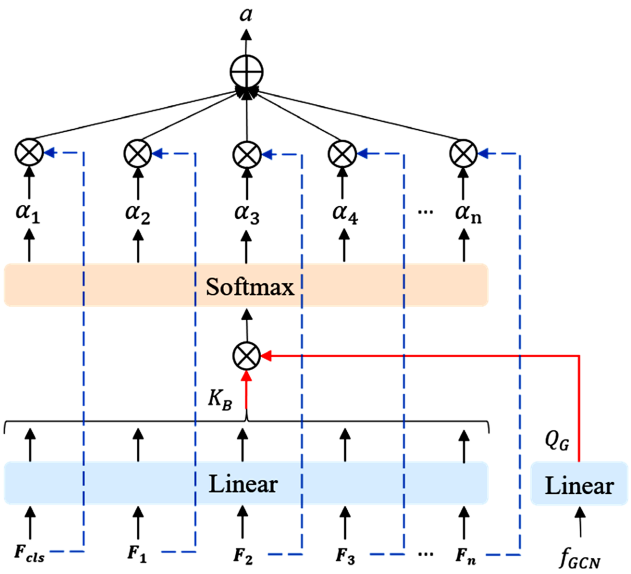

Figure 3. Attention mechanism
图 3. 注意力机制结构图

将 GCN 模块得到的符合文本图特征的文本特征表示 f_{GCN} 经过线性层后的输出作为查询向量 Q，将 BERT 模块的最终输出 f_{Bert} 作为键向量 K，由向量组成文本的查询矩阵 Q_G 和键矩阵 K_B，值矩阵 V 设置为 1，如公式(8)。其中，W^Q 和 W^K 是权重矩阵，$F=\mathbb{R}^{n\times \dim}$ 是输入序列；利用公式(9)得到第 i 个输入信息概率 α_i，其中 α_i 构成的概率向量称为注意力分布。

$$Q_G = FW^Q, K_B = FW^K, V = 1 \tag{8}$$

$$\alpha = Attention(Q,K,V) = Softmax\left(\frac{Q\cdot K^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{9}$$

将注意力分布与输入信息以加权平均的方式进行汇总，得到最终的注意力分数表示为公式(10)：

$$a = attention(F,Q) = \sum_{i=1}^{N}\alpha_i F_i \tag{10}$$

最后将 a 输入到全连接层，经过 $Softmax$ 函数得到文本标签类别的分布如公式(11)所示：

$$y_{BAG} = Softmax(a, g(X,A)) \tag{11}$$

3.4. 损失函数

模型将 BAG 模块和 GCN 模块的输出相结合，得到最终的预测概率 y，同时将两个模块进行联合优化，如公式(12)所示：

$$\hat{y} = (1-\lambda)y_{BAG} + \lambda y_{GCN} \tag{12}$$

其中，λ 是 BAG 模块和 GCN 模块的权衡系数。

模型训练时使用交叉熵损失函数，形式化如公式(13)：

$$L = -\sum_{d\in y_D}\sum_{f=1}^{F} Y_{df}\ln y_{df} \tag{13}$$

其中，y_D 为所有带标签的文档的索引集；F 是输出特征向量维度；Y 是标签指标矩阵，用于表示多标签数据的格式。

4. 实验结果与分析

4.1. 实验数据集与评价指标

本文在四个经典文本分类数据集上进行实验，验证 BAG 模型的有效性。R8 和 R52 数据集分别包含 8 种类别标签和 52 种类别标签[30]。Ohsumed 数据集[31]从医药类相关杂志中选择其中 23 种心血管疾病医学摘要的文档集合，并筛选出 7400 个具有单一文本标签的文档。MR 数据集[11]是用于二元情感分类的电影评论合集，每篇文档包含一个积极或消极的情感标签，整个数据集包含 5331 个积极评论文档和 5331 个消极评论文档。数据集相关统计信息如表 1 所列。

Table 1. Summary statistics of datasets
表 1. 数据集总体统计

数据集	文档数	单词数	节点数	类别数	平均长度
R8	7674	7688	15,362	8	65.72
R52	9100	8892	17,992	52	69.82
Ohsumed	7400	14,157	21,557	23	135.82
MR	10,662	18,764	29,426	2	20.39

实验采用准确率(Accuracy)作为评价指标，用所有正确分类的样本数除以总样本数进行计算。

4.2. 实验设置

本文实验基于 PyTorch1.4.0 框架，在 Ubuntu16.04 的运行环境中使用 Python3.7.9 进行编程，模型训练使用 TITAN V 12G GPU，训练模型的学习率等参数设置如表 2 所示。

Table 2. Parameters of the model
表 2. 模型参数

参数	数值
初始学习率	1e−3
微调后 BERT 模块的学习率	1e−5
Dropout 比率	0.5
GCN 层数	2
BERT 层数	4
词嵌入维度	200

实验参考 TextGCN 的方法对数据进行预处理，随机选取 10%的训练集作为验证集，分别在四个数据集上对模型进行最多 200 个 epoch 的训练，如果连续 10 个 epoch 验证损失没有下降，则停止训练。

4.3. 实验结果与分析

4.3.1. 对比实验

本文选择 CNN、LSTM、FastText、Graph-CNN、TextGCN 和 BertGCN 作为基准模型进行对比实验。由于 GCN 训练非常消耗计算资源，为了使模型能够有效训练，在将本文模型与 BertGCN 模型进行对比

时，选择将数据集的最大长度(max_length)和训练批大小(batch_size)统一降低设置，确保实验在相同的参数设置下进行，结果如表 3 所示(除 BertGCN 之外，其余对比模型的实验结果与其原有文献一致)。

Table 3. Accuracy comparison of different methods
表 3. 不同方法的准确率对比

Models	R8	R52	Ohsumed	MR
CNN	94.02	85.37	43.87	74.98
LSTM	93.68	85.54	41.13	75.06
FastText	96.13	92.81	57.70	75.14
Graph-CNN	96.99	92.75	63.86	77.22
TextGCN	97.07	93.56	68.36	76.74
BertGCN	97.90	94.12	68.37	85.48
BAG	97.90	96.03	71.51	87.56

表 3 可以看出，基于图网络的模型明显比其他不使用图结构的深度学习模型在文本分类任务上具有更加出色的分类效果。这是因为文本的上下文信息具有一定的关联性，而基于神经网络的方法如 CNN 和 LSTM，虽然能够捕获到前后单词之间的局部信息，但对于非连续单词或者单词之间距离较远的文本，往往不能捕捉其有效的全局信息。而如果引入图网络模型为文本数据建立结构图，便能通过捕捉单词之间的共现信息来为其建立语义信息交互结构，并以此实现不同节点和边之间语义信息的迭代和更新，这正是众多不基于图结构的深度学习模型所达不到的效果。

由于只考虑全局信息的图卷积网络无法捕捉到词序等局部语义信息，而这些信息对于理解文本真实含义准确进行文本分类是非常重要的。本文在图卷积网络和 BERT 之间引入注意力机制进行特征融合，从实验结果可以看出，与其他图网络模型相比，BAG 模型在 R52 数据集上较 TextGCN、BertGCN 分别提升了 2.47%、1.91%，在 Ohsumed 数据集上分别提升 3.15%、3.14%，在 MR 数据集上分别提升了 10.82%、2.08%，在 R8 数据集上较 TextGCN 提升了 0.83%，与 BertGCN 的结果持平。

4.3.2. 消融实验

如前所述，本文所提出的 BAG 模型通过引入注意力机制将输入文本的语义信息进行补全，且分别将 BAG 模块和 GCN 模块的概率预测表达相结合得到最终的分类结果，本节通过消融实验验证模型结构的合理性和有效性。在都不使用 GCN 模块的前提下，表 4 列出了模型在四个数据集上的消融实验结果，其中“w/o att”表示 BAG 模型不使用注意力机制，“w/att”表示引入注意力机制后的模型。表 4 中的粗体部分表示分类准确率在数据集上表现最优。

Table 4. Results of ablation experiment
表 4. 消融实验结果

Models	R8	R52	Ohsumed	MR
w/o att	96.85	94.70	70.37	87.45
w/ att	97.90	96.03	71.51	87.56

由表 4 可知，在仅使用 BERT 提取文本序列中的<cls>标签来提供句子级别的语义嵌入特征而不使用

注意力机制去提取句子中其他 token 级别的语义信息进行文本分类的情况下的分类效果在四个数据集上均落后于在 BERT 和 GCN 中间加入一层注意力机制的 BAG 模型的分类精度。其原因是，虽然 BertGCN 可以利用<cls>标签提供句子级别的语义信息，但忽略了句子中其他 token 级别的语义信息的使用，BAG 模型则通过注意力机制的引入充分利用 GCN 学习到剩余的语义信息，弥补剩余语义信息能力不足的缺陷，综合表中的结果可知，引入注意力机制的 BAG 模型是有效的。

4.3.3. 参数的影响

表 1 列出了四个数据集中文本的平均长度，本文选择在平均长度最大的 Ohsumed 数据集上进行补充实验，验证文本长度的设置对分类精度的影响，实验结果如表 5 所示。

Table 5. Experimental results of different max_length on dataset Ohsumed
表 5. 在 Ohsumed 数据集调整 max_length 的实验结果

batch_size	max_length	accuracy
32	8	41.91
16	16	63.74
8	32	71.51

结果表明如果将文本最大长度(max_length)参数设置为足够覆盖文本平均长度的数值，将有利于模型性能和分类结果的提升。但由于硬件设备限制，本文实验无法将四个数据集的最大长度设置为比 32 更高的数值，这在一定程度上也影响了最终的分类效果。通过平均长度最短的 MR 数据集可以验证上述结论，即使 BertGCN 在 MR 数据集上训练时将最大长度设置为 64，但 BAG 模型只要将最大长度设置为 32 就可以获得更好的分类效果，也验证了本文模型在硬件条件不足，计算机算力不够的情况下进行文本分类的有效性。虽然通过降低 batch_size 的数值可以使文本最大长度设置的更高，虽然可以使分类精度小幅提升，但同时作为补偿会使模型训练速度大幅降低。

4.3.4. λ 值的影响实验

λ 值控制着训练 BAG 和 GCN 之间的权衡。为了将模型调整为最优结构，本文在 Ohsumed 数据集上调整权衡系数进行实验，图 4 显示了不同 λ 值下 BAG 的准确率。对于不同的数据集，使模型达到最佳状态的 λ 值会有所不同。总体来看，在 Ohsumed 数据集上，λ 值越小，模型准确率越高，表示在训练的时候需要

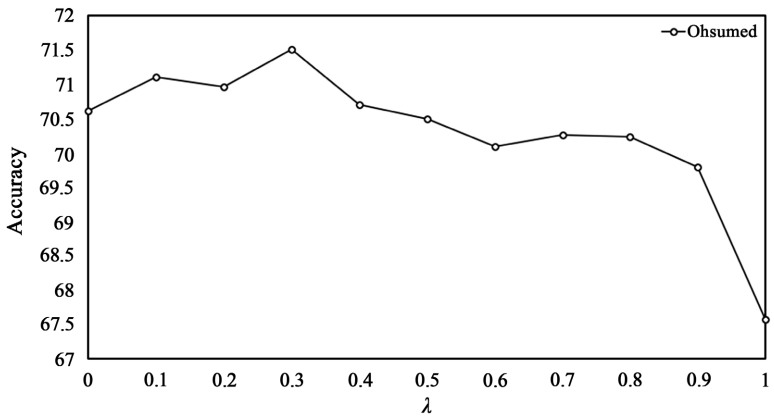

Figure 4. Experimental results of different λ on Ohsumed
图 4. 在 Ohsumed 数据集上调整 λ 的实验结果

重点关注 BAG 模型，这也可以证明融入了注意力机制的 BAG 模型是有效的，同时也证明了在训练中对 GCN 加入的文档的词频信息的学习是有效的。

5. 结论

现有的文本分类模型或无法有效捕捉一段文本中非连续或远距离单词之间的语义信息，或局限于单独利用图的全局结构特征而忽略了文本的局部结构特征。本文在 GCN 和 BERT 之间引入注意力机制，来充分融合预训练模型和图卷积网络处理数据和提取特征的能力，提出了一种改进的文本分类模型 BAG。相关实验结果表明，在四个常用文本分类数据集下，BAG 的分类性能优于 CNN、LSTM 和其他深度学习模型。同时，与 BertGCN 相比，BAG 在使用更小的批数量和文本长度等参数设置下，能获得更高的准确率。在硬件资源受限，特别是算力不足的场景下，BAG 是更好的分类模型。

本文方法未来可能的改进包括：1) 在数据层面，未来的研究将围绕如何通过添加外部知识增强模型的分类性能；2) 在模型层面，提示学习和对比学习等方法已经在很多 NLP 任务中达到理想效果，未来的研究将探索提示学习和对比学习等新兴模型在文本分类任务方面的应用。

基金项目

国家自然科学基金资助项目(92048205)；
南京大学计算机软件新技术国家重点实验室开放课题项目(KFKT2021B39)。

参考文献

[1] 闫秘. 基于 fastText 的垃圾邮件过滤算法研究[D]: [硕士学位论文]. 广州: 华南理工大学, 2020.

[2] Li, Z.H., Fan, Y.Y., Jiang, B., *et al*. (2019) A Survey on Sentiment Analysis and Opinion Mining for Social Multimedia. *Multimedia Tools and Applications*, **78**, 6939-6967. https://doi.org/10.1007/s11042-018-6445-z

[3] Kpiebaareh, M., Wu, W.P., *et al*. (2021) A Graph-Based Opinion Mining Approach for Reducing Information Loss and Overload in Product Reviews Analysis. *Proceedings of International Conference on Compute and Data Analysis*, Sanya, 2-4 February 2021, 143-148. https://doi.org/10.1145/3456529.3456561

[4] Kalaivani, K.S., Uma, S. and Kanimozhiselvi, C.S. (2020) A Review on Feature Extraction Techniques for Sentiment Classification. *Proceedings of International Conference on Computing Methodologies and Communication*, Erode, 11-13 March 2020, 679-683. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000126

[5] Song, P., Geng, C.Y. and Li, Z.J. (2019) Research on Text Classification Based on Convolutional Neural Network. *Proceedings of International Conference on Computer Network*, *Electronic and Automation*, Xi'an, 27-29 September 2019, 229-232. https://doi.org/10.1109/ICCNEA.2019.00052

[6] Li, Q., Peng, H., Li, J.X., *et al*. (2020) A Survey on Text Classification: From Shallow to Deep Learning.

[7] Peters, M.E., Neumann, M., Iyyer, M., *et al*. (2018) Deep Contextualized Word Representations. *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 2227-2237. https://doi.org/10.18653/v1/N18-1202

[8] Yang, Z.L., Dai, Z.H., Yang, Y.M., *et al*. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 5753-5763.

[9] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* 2019, Minneapolis, 2-7 June 2019, 4171-4186.

[10] Huang, L.Z., Ma, D.H., Li, S.J., *et al*. (2019) Text Level Graph Neural Network for Text Classification. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, November 2019, 3444-3450. https://doi.org/10.18653/v1/D19-1345

[11] 邓朝阳, 仲国强, 王栋. 基于注意力门控图神经网络的文本分类[J]. 计算机科学, 2022, 49(6): 326-334.

[12] Yao, L., Mao, C.S. and Luo, Y. (2019) Graph Convolutional Networks for Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, 29-31 January 2019, 7370-7377. https://doi.org/10.1609/aaai.v33i01.33017370

[13] Lin, Y.X., Meng, Y.X., Sun, X.F., *et al*. (2021) BertGCN: Transductive Text Classification by Combining GCN and BERT. *Findings of the Association for Computational Linguistics*: *ACL-IJCNLP* 2021, August 2021, 1456-1462. https://doi.org/10.18653/v1/2021.findings-acl.126

[14] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks.

[15] Vaswani, A., Shazeer, N., Parmer, N., *et al*. (2017) Attention Is All You Need. *Proceedings of International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[16] Sun, X.F., Yang, D.Y., Li, X.Y., *et al*. (2021) Interpreting Deep Learning Models in Natural Language Processing: A Review.

[17] Deng, X.L., Li, Y.Q., Weng, J., *et al*. (2019) Feature Selection for Text Classification: A Review. *Multimedia Tools and Applications*, **78**, 3797-3816. https://doi.org/10.1007/s11042-018-6083-5

[18] Wang, Q., Xu, H.L. and Li, Y.L. (2021) Classification of News Texts Based on Bayes Algorithm. *Proceedings of International Conference on Electronic Information Technology and Computer Engineering*, Xiamen, 22-24 October 2021, 1288-1291. https://doi.org/10.1145/3501409.3501636

[19] Yu, Y., Si, X.S., Hu, C.H., *et al*. (2019) A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, **31**, 1235-1270. https://doi.org/10.1162/neco_a_01199

[20] 闫跃, 霍其润, 李天昊, 等. 融合多重注意力机制的卷积神经网络文本分类设计与实现[J]. 小型微型计算机系统, 2021, 42(2): 362-367.

[21] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Doha, 25-29 October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[22] Wang, X.Y., Jiang, W.J. and Luo, Z.Y. (2016) Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts. *Proceedings of International Conference on Computational Linguistics*, Osaka, 11-16 December 2016, 2428-2437.

[23] Yang, Z.C., Yang, D.Y., Dyer, C., *et al*. (2016) Hierarchical Attention Networks for Document Classification. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 1480-1489. https://doi.org/10.18653/v1/N16-1174

[24] Brown, T.B., Mann, B., Ryder, N., *et al*. (2020) Language Models Are Few-Shot Learners.

[25] Hu, S.D., Ding, N., Wang, H.D., *et al*. (2021) Knowledgeable Prompt-Tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. *Proceedings of the* 60*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 2225-2240. https://doi.org/10.18653/v1/2022.acl-long.158

[26] Su, X., Wang, R. and Dai, X.Y. (2022) Contrastive Learning-Enhanced nearest Neighbor Mechanism for Multi-Label Text Classification. *Proceedings of Annual Meeting of the Association for Computational Linguistics*, Dublin, 22-27 May 2022, 672-679. https://doi.org/10.18653/v1/2022.acl-short.75

[27] Gunel, B., Du, J.F., Conneau, A., *et al*. (2020) Supervised Contrastive Learning for Pre-Trained Language Model Fine-Tuning.

[28] Peng, H., Li, J.X., He, Y., *et al*. (2018) Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN. *Proceedings of World Wide Web Conference*, Lyon, 23-27 April 2018, 1063-1072. https://doi.org/10.1145/3178876.3186005

[29] 杨慧敏. 基于交互孪生网络的复合对话模型[D]: [硕士学位论文]. 南京: 南京信息工程大学, 2020.

[30] 蒋浩泉, 张儒清, 郭嘉丰, 等. 图卷积网络与自注意机制在文本分类任务上的对比分析[J]. 中文信息学报, 2021, 35(12): 84-93.

[31] Minaee, S., Kalchbrenner, N., Cambria, E., *et al*. (2021) Deep Learning-Based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, **54**, 1-40. https://doi.org/10.1145/3439726