

融合 BERT 和自注意力机制的张量图卷积网络 文本分类^①



史文艺, 朱欣娟

(西安工程大学 计算机科学学院, 西安 710600)

通信作者: 朱欣娟, E-mail: zhuxinjuan@xpu.edu.cn

摘 要: TensorGCN 模型是图神经网络应用在文本分类领域的 SOTA 模型之一. 然而在处理文本语义信息方面, 该模型使用的 LSTM 难以完全地提取短文本语义特征, 且对复杂的语义处理效果不佳; 同时, 由于长文本中包含的语义及句法特征较多, 在进行图间异构信息共享时特征共享不完全, 影响文本分类的准确性. 针对这两个问题, 对 TensorGCN 模型进行改进, 提出融合 BERT 和自注意力机制的张量图卷积网络 (BTSGCN) 文本分类方法. 首先, 使用 BERT 代替 TensorGCN 架构中的 LSTM 模块进行语义特征提取, 通过考虑给定单词两侧的周围单词来捕获单词之间的依赖关系, 更准确地提取短文本语义特征; 然后, 在图间传播时加入自注意力机制, 帮助模型更好地捕捉不同图之间的特征, 完成特征融合. 在 MR、R8、R52 和 20NG 这 4 个数据集上的实验结果表明 BTSGCN 相比于其他对比方法的分类准确度更高.

关键词: 文本分类; 图神经网络; BTSGCN; BERT; 自注意力机制

引用格式: 史文艺, 朱欣娟. 融合 BERT 和自注意力机制的张量图卷积网络文本分类. 计算机系统应用, 2025, 34(3): 152-160. <http://www.c-s-a.org.cn/1003-3254/9831.html>

Text Classification with Tensor Graph Convolutional Network Fusing BERT and Self-attention Mechanism

SHI Wen-Yi, ZHU Xin-Juan

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract: TensorGCN model is one of the state-of-the-art (SOTA) models applied by graph neural networks in the field of text classification. However, in terms of processing text semantic information, the long short-term memory (LSTM) used by the model has difficulty in completely extracting the semantic features of short text and performs poorly in handling complex semantic information. At the same time, due to the large number of semantic and syntactic features contained in long texts, feature sharing is incomplete when heterogeneous information is shared among graphs, which affects the accuracy of text classification. To solve these two problems, the TensorGCN model is improved, and a text classification method based on the tensor graph convolutional network fusing BERT and the self-attention mechanism (BTSGCN) is proposed. Firstly, BERT is used to replace the LSTM module in the TensorGCN architecture for semantic feature extraction. It captures the dependencies between words by considering the surrounding words on both sides of a given word, thus extracting the semantic features of short texts more accurately. Then, the self-attention mechanism is added during the propagation among graphs to help the model better capture the features among different graphs and complete the feature fusion. Experimental results on MR, R8, R52, and 20NG datasets show that BTSGCN has higher

① 基金项目: 陕西省重点研发计划 (2024GX-YBXM-548)

收稿时间: 2024-09-06; 修改时间: 2024-10-10; 采用时间: 2024-11-19; csa 在线出版时间: 2025-01-21

CNKI 网络首发时间: 2025-01-22

classification accuracy than other comparison methods.

Key words: text classification; graph neural network (GNN); tensor graph convolutional network fusing BERT and the self-attention mechanism (BTSGCN); BERT; self-attention mechanism

文本分类问题作为自然语言处理领域经典的问题之一,是指将载有特定信息的文本(1个句子或者1个文档)映射到预先给定的某一类别或者某几类主题中^[1]。目前,在许多任务中都可以应用文本分类,比如情感分析^[2,3]、新闻分类^[4,5]、疾病诊断^[6,7]和垃圾邮件检测^[8,9]等。

现有的文本分类方法可以分为两大类。第1类是基于深度学习的方法,该方法主要由数据预处理、特征提取和分类处理几部分组成。这类方法涌现出众多模型,例如由Kim等人^[10]提出的TextCNN模型,首次利用卷积神经网络捕获文本信息中的特征,但是由于卷积核感受野的尺度不可变,因此该模型不能处理长序列输入的单词依赖关系。Liu等人^[11]在TextCNN模型的基础上提出TextRNN模型,通过循环神经网络代替卷积神经网络来提取文本信息中的特征,使模型更好地获取长文本序列中各个单词的依赖关系,一定程度上解决了TextCNN模型的缺陷。但是该模型在长序列非连续单词的交互方面还存在一些问题。针对这个问题,Hochreiter等人^[12]提出在模型中融合LSTM,并由Zhang等人^[13]将该网络应用于文本分类模型中,有效地解决了长序列非连续单词之间的交互问题。但是融合LSTM导致模型的计算复杂度过高,并且在处理长序列文本时,LSTM容易受到梯度消失的影响,导致训练过程不稳定。对于LSTM的缺陷,Cho等人^[14]提出GRU网络,与LSTM的三门控机制相比,GRU采用输入门、更新门的双门控机制。既成功解决了LSTM导致的计算复杂度过高、梯度消失等问题,同时还保留了LSTM的优点。但是,GRU网络还是没有解决无法捕捉文本信息局部深层次特征的问题。Yang等人^[15]在文本分类模型中加入注意力机制,为文本语料库中的序列分配不同权重来协助模型更有效地处理文本信息。

在大多数基于深度学习的文本分类方法中,文本的全局内容和局部细节通常作为两个分开的单元进行处理,这种做法容易导致两大问题。一是特征提取可能不全面,导致模型无法有效识别文本中的关键信息;二

是全局内容模糊表达,使得模型难以理解文本的整体语境和语义关系。这些问题可能会影响模型在文本分类方面的表现力。

第2类是基于图神经网络(graph neural network, GNN)的方法。GNN处理过程分为两部分:文本构建图和基于图的学习。文本构建图是提取文本信息中包含的语义、句法或顺序上下文信息,根据单词与单词或单词与句子之间的依赖关系构建文本图。基于图的学习是使用GNN对构建的文本图进行处理,学习节点和边的特征表示。基于GNN的方法也涌现出众多模型,例如,Yao等人^[16]首次提出的TextGCN模型,通过构建一个包含单词和文档节点的异构图来整体处理输入序列,利用图卷积网络学习图节点的特征表示。但是该模型对内存的消耗非常严重。之后,Huang等人^[17]以TextGCN模型为基础提出Text-level GCN模型,为语料库中的每个文本序列单独构建图,并在图之间共享全局参数,在一定程度上降低了内存的消耗。Wu等人^[18]提出了SGC方法,融合每个图节点及其邻近节点的特征,减少了图卷积网络需要处理的参数数量,从而降低了系统的计算复杂度。Zhang等人^[19]提出TextING模型,对输入的每个文本序列构建一张图,结合GRU网络学习单词的特征表示。但是由于模型主要处理局部信息,所以在提取全局信息特征方面效果不佳。Wang等人^[20]提出的Induct-GCN模型利用全局信息分配文本标签进行文本分类。Zhu等人^[21]提出SSGC模型,调节低通和高通滤波器以便获得图节点局部信息和全局上下文的特征表示。Linmei等人^[22]提出了HGAT模型,通过异构图建模输入文本来解决语义特征矩阵稀疏的问题。Liu等人^[23]提出的TensorGCN模型对输入的文本序列分别提取语义、句法与顺序上下文信息特征用于构建3张异构图,同时结合两种传播方式:图内传播和图间传播来协调节点信息及不同图的异构信息。融合3种图的异构特征信息进行文本分类,使得分类准确率比以往的模型有较大的提升,但是在处理文本语义信息方面,对输入的文本序列为短文本的情况,LSTM在进行语义特征提取时的迭代次数太少而导致

提取到的语义特征不完全;同时,对于长文本序列,由于包含的语义及句法特征过多,在进行图间传播共享异构信息时特征共享不完全,从而影响最终文本分类的准确性。

与基于深度学习的方法相比,基于图神经网络的方法在处理文本信息时,通过对完整的文本序列构建一张图,同时处理全局内容和局部细节,可以提取更全面的文本特征,从而在模型效果方面,基于图神经网络的方法普遍优于基于深度学习的方法。

本文在原始 TensorGCN 模型框架基础之上进行改进,提出融合 BERT 和自注意力机制的张量图卷积网络文本分类方法 (text classification with tensor graph convolutional network fusing BERT and self-attention mechanism, BTSGCN), 本文的主要贡献如下。

(1) 使用 BERT 代替原 TensorGCN 网络架构中的 LSTM 模块进行语义特征提取, 通过考虑给定单词两侧的周围单词来捕获单词之间的依赖关系, 改善短文本特征提取不完全及复杂语义处理效果不佳的问题。

(2) 在图间传播时加入自注意力机制, 在长文本特征较多的情况下更好地融合异构图之间的特征, 改善特征融合不完全的问题。

(3) 提出 BTSGCN 模型, 经实验验证, 本文模型的

文本分类准确率与 TensorGCN 相比, 在 4 个数据集 MR、R8、R52 和 20NG 上分别提升了 2.98%、0.35%、0.85% 和 2.32%; 与表现最好的基线模型相比, 在 4 个数据集 MR、R8、R52 和 20NG 上分别提升了 1.74%、0.2%、0.66% 和 1.23%。

1 BTSGCN 模型

BTSGCN 模型的主框架如图 1 所示 (实线框为相比原始 TensorGCN 模型改进之处), 由构图模块、图卷积网络模块及池化模块组成。首先, 对输入的文本序列进行数据预处理, 分别使用 BERT、Stanford CoreNLP 解析器、滑动窗口算法和点互信息技术 (point-wise mutual information, PMI) 提取语义、句法和顺序上下文信息特征, 以单词作为节点、单词与单词的关系作为边构建文本图来获取图张量信息。其次, 使用两种传播方式: 图内传播和图间传播来进行图张量信息的学习, 利用图内传播将每一个节点及其邻近节点的信息聚合在一起, 利用图间传播来共享不同图之间的特征信息, 同时在图间传播过程中使用自注意力机制来辅助不同特征融合得更完全。最后, 将融合完成的图特征输入到池化层中进行平均池化操作来获得图节点的最终表示, 从而完成文本分类任务。

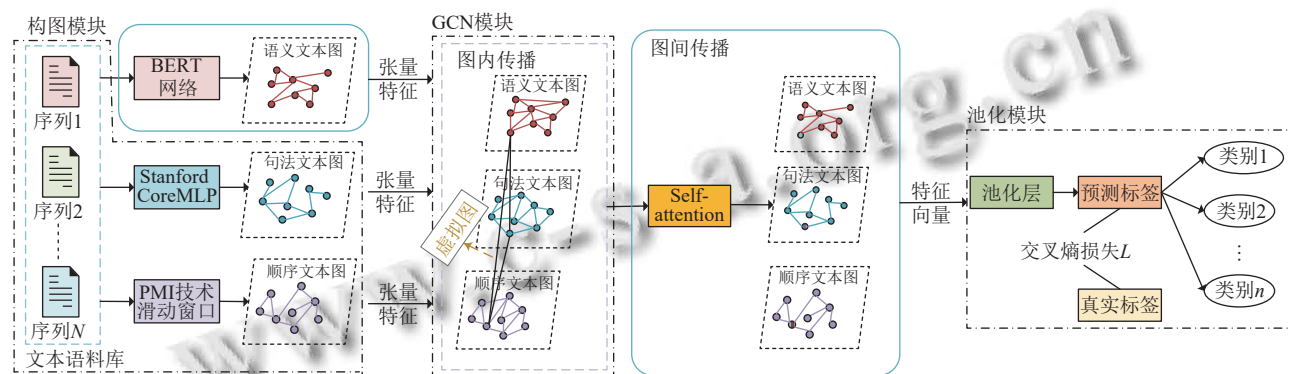


图1 BTSGCN 模型主框架图

1.1 构图模块

文本信息中包含的语义特征可以使模型更好地理解文本序列的主题和内容, 句法特征可以使模型更好地理解文本序列的结构和逻辑关系, 顺序特征可以使模型更好地捕捉文本序列的逻辑顺序。因此使用 3 种技术分别提取出输入的文本序列中包含的语义、句法以及顺序上下文信息特征, 以单词作为节点、单词与单词的关系作为边分别构建语义文本图、句法文本图

和顺序文本图, 并构建图张量。

1.1.1 语义文本图

由于 LSTM 的输入通常需要经过词向量嵌入方法进行处理, 然后根据时间单元的顺序变化来处理输入序列, 并在前向传播过程中依赖上一个时间步的隐藏态更新权重信息, 导致训练及推理过程时间较长, 效率较低且对梯度消失问题较敏感。而 BERT^[24] 的输入需要经过基于子词的分词器进行处理, 添加特殊标记 (如

[CLS][SEP]), 然后通过双向自注意力机制同时处理输入序列的双向上下文信息进行特征处理, 在训练及推理过程中可以并行处理整个输入序列的任何位置, 利用单词的双向上下文信息来捕获单词之间的依赖关系来学习全局语义表示, 更全面地提取短文本和复杂文本的语义特征, 同时训练及推理整体计算效率较高. 所以使用 BERT 代替 LSTM 提取文本序列中的语义特征.

对于输入的文本序列, 首先去除文本信息中的停用词, 利用 BERT 获取剩余单词的语义特征, 计算特征之间的余弦相似度, 如果余弦相似度值大于设定的阈值, 则表明两个单词之间具有语义关系. 然后统计文本序列中具有语义关系的单词组的次数, 用该值除以该单词组在整个文本序列同一句子中存在的次数得到这两个单词之间的边权值. 可以用式 (1) 表示:

$$E_{\text{sem}}(i, j) = \frac{N_{\text{sem}}(i, j)}{N_{\text{tot}}(i, j)} \quad (1)$$

其中, $E_{\text{sem}}(i, j)$ 表示两个单词之间具有语义关系的边权值, $N_{\text{sem}}(i, j)$ 表示 i, j 单词在整个文本序列中具有语义关系的次数, $N_{\text{tot}}(i, j)$ 表示 i, j 单词在整个文本序列同一句子中存在的次数.

最后, 以单词作为节点, 计算出的单词之间具有语义关系的边权值作为边进行语义文本图的构建.

1.1.2 句法文本图

首先进行句法关系分析, 即分析输入的文本序列中每一个句子的句法成分及各个成分之间的组合关系. 使用 Stanford CoreNLP 解析器分析文本序列, 捕捉单词之间的依赖关系. 然后统计在整个文本序列中具有句法依赖关系的单词组的次数, 用该值除以该单词组在整个文本序列同一句子中存在的次数得到这两个单词之间的边权值. 可以用式 (2) 表示:

$$E_{\text{syn}}(i, j) = \frac{N_{\text{syn}}(i, j)}{N_{\text{tot}}(i, j)} \quad (2)$$

其中, $E_{\text{syn}}(i, j)$ 表示两个单词之间具有句法依赖关系的边权值, $N_{\text{syn}}(i, j)$ 表示 i, j 两个单词在整个文本序列中具有句法依赖关系的次数, $N_{\text{tot}}(i, j)$ 表示 i, j 两个单词在整个文本序列同一句子中存在的次数.

最后, 以单词作为节点, 计算得到的单词之间具有句法依赖关系的边权值作为边进行句法文本图的构建.

1.1.3 顺序文本图

顺序上下文信息描述了单词之间局部共现的语言

属性. 首先使用 PMI 技术^[25]来描述输入文本序列的顺序上下文信息, 使用滑动窗口算法来捕获其中的共现关系并统计单词之间局部共现的次数. 每组单词之间局部共现的边权值可以用式 (3) 表示:

$$E_{\text{seq}}(i, j) = \lg \frac{N_{\text{co-occur}}(i, j) \times N_{\text{win}}}{N_{\text{occur}}(i) \times N_{\text{occur}}(j)} \quad (3)$$

其中, $E_{\text{seq}}(i, j)$ 表示 i, j 单词之间具有局部共现关系的边权值, $N_{\text{co-occur}}(i, j)$ 表示 i, j 单词在整个文本序列中同时出现于一个滑动窗口中的次数, N_{win} 表示在整个文本序列中出现的滑动窗口总数, $N_{\text{occur}}(i)$ 表示 i 单词在整个文本序列的滑动窗口中出现的次数.

最后, 以单词作为节点, 计算得到的单词之间具有局部共现关系的边权值作为边进行顺序文本图的构建.

1.1.4 构建图张量

图张量由共享相同节点的多个图组成. 为了更好地将不同图转化为图张量, 对其进行一个形式化的定义: 假设 g 是一个图张量, 则 $g = (G_1, G_2, G_3, \dots, G_n)$, 其中 $G_i = (V_i, E_i, A_i)$, $V_i = V_j (i, j = 1, 2, \dots, n)$ 且 $A_i \neq A_j (i \neq j)$. 其中, G_i 是图张量 g 的第 i 张图, V_i, V_j 是第 i, j 张图的节点集合, E_i 是第 i 张图的边集合, A_i, A_j 是第 i, j 张图的邻接矩阵. 同时, 将图邻接矩阵也转化为图邻接张量, 即图邻接张量 $A = (A_1, A_2, \dots, A_n)$.

在张量图中所有图只有边集不同, 所以合并图可以通过融合不同图的图邻接张量完成. 由于异构图中的边权重不相同, 因此在合并图时需要对不同的边权重进行处理. 图邻接矩阵的融合可以用式 (4) 表示:

$$A_{\text{merge}} = \sum_{i=1}^n W_{\text{edge}}^i \odot A_i \quad (4)$$

其中, W_{edge}^i 表示与邻接矩阵大小相同的权值矩阵.

1.2 GCN 模块

图卷积网络通过利用图传播规则来进行图结构和节点特征的学习^[26], 一种比较常见的分层图传播规则如式 (5)–式 (8):

$$H^{(l+1)} = \sigma(\hat{A}H^{(l)}W^{(l)}) (l = 1, 2, \dots, n) \quad (5)$$

$$\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} \quad (6)$$

$$\tilde{A} = A + I \quad (7)$$

$$\tilde{D}(i, i) = \sum_j \tilde{A}(i, j) (j = 1, 2, \dots, n) \quad (8)$$

其中, $H^{(l)}$ 是第 l 层的特征矩阵, l 是 GCN 的层数, $\hat{A}$ 是由多个归一化对称邻接矩阵 $\tilde{A}$ 进行自连接组成的归一化对称邻接张量, $\tilde{D}$ 是对角节点度矩阵, $W^{(l)}$ 是可调整

的参数矩阵, σ 是非线性激活函数. 通常在最后一层时, σ 被设置为 *Softmax* 函数, 并且每个节点的数量等于标签的数量. 当 $l=0$ 时, $H^{(0)}$ 是初始特征矩阵, 其中每 1 行代表 1 个节点的初始输入特征.

1.2.1 图内传播

图内传播是将图中各个节点与临近节点的特征信息进行结合, 从而输出一个特征张量. 对于给定的图邻接张量 $A = (A_1, A_2, \dots, A_n)$, 第 l 层第 i 个图的特征通过图内传播的处理过程可以用式 (9) 和式 (10) 表示:

$$H_{\text{intra}}^{(l)}(i, :, :) = \sigma(\hat{A}(i, :, :)H^{(l)}(i, :, :)W_{\text{intra}}^{(l,i)}) \quad (9)$$

$$\hat{A}(i, :, :) = \tilde{D}(i, :, :)^{-\frac{1}{2}} \tilde{A}(i, :, :) \tilde{D}(i, :, :)^{-\frac{1}{2}} \quad (10)$$

其中, $\hat{A}(i, :, :)$ 是由多个归一化对称邻接矩阵 $\tilde{A}(i, :, :)$ 进行自连接组成的对称邻接张量, $\tilde{D}(i, :, :)$ 是由 n 个对角节点度矩阵组成的节点度张量, $W_{\text{intra}}^{(l,i)}$ 是第 i 个图在第 l 层上的可训练权重矩阵, σ 是非线性激活函数.

1.2.2 图间传播

图间传播是交换并融合张量中多个图之间的异构特征信息. 由于长文本序列包含的特征信息较多, 可能会难以完全捕捉特征间的复杂关系或者遗漏某些有用的特征信息, 导致特征融合不完全, 影响文本分类的准确性. 而自注意力机制可以允许每个特征与所有特征进行交互, 捕捉不同特征之间的细微关系, 同时通过计算注意力分数, 为不同特征分配不同权重, 再进行加权平均来减少融合过程中的特征遗漏, 使特征融合更完全.

张量图中异构图的节点是相同的, 因此可以连接对应异构图上的节点来构建一系列的虚拟图. 由于有 n 个节点, 所以可以构建 n 张虚拟图. 利用自注意力机制在每张虚拟图上计算出注意力分数, 使用得到的注意力分数对原始虚拟图的特征进行加权, 将加权后的特征与原始特征结合, 对融合后的特征进行平均求和得到最终的嵌入表示, 从而得到一张新的图邻接张量. 在虚拟图上图间传播学习的处理过程可用式 (10) 表示:

$$H^{(l+1)}(:, j, :) = \sigma(A^+(:, :, J)H_{\text{intra}}^{(l)}(:, j, :)W_{\text{inter}}^{(l,j)}) \quad (11)$$

其中, $H^{(l+1)}$ 是第 l 层图间传播的输出特征张量, $W_{\text{inter}}^{(l,j)}$ 是第 l 层上可调整的参数矩阵, $A^+(:, :, J)$ 是第 j 个虚拟图的邻接矩阵. 虚拟图中所有节点彼此连接, 并且将边权重设为 1.

1.3 池化模块

在完成最后一层虚拟图的图间传播后, 对得到的 n 个图的特征向量做平均池化处理, 来获取每个图节点的全局特征表示. 训练阶段, 从这些节点中提取出与训练数据集对应的特征, 并将其与真实标签的 one-hot 向量进行比较, 使用交叉熵损失函数计算损失, 之后利用反向传播算法来计算梯度并修改模型参数; 测试阶段, 提取出与测试数据集对应的特征, 计算模型损失和分类准确率以评估模型性能. 损失函数表达式为:

$$Z = WH + b \quad (12)$$

$$\hat{y} = \text{Softmax}(Z) \quad (13)$$

$$L = -\sum y \lg(\hat{y}) \quad (14)$$

其中, W 是训练参数, H 是输出特征表示, b 是偏置项, Z 是全局特征表示, $\hat{y}$ 是预测样本属于某一类别的概率, y 是真实标签.

2 实验

为了更好地评估融合 BERT 和自注意力机制的张量图卷积网络文本分类方法的结果, 在 4 个公开的数据集上与 6 种不同的基线模型进行对比试验, 并分析实验结果.

2.1 数据集介绍

模型训练采用 4 个被广泛应用于文本分类领域的公开数据集: MR、R8、R52 以及 20NG. MR 是一个用于正负情感二分类任务的电影评论数据集, 包含 10 062 条数据, 数据集中每 1 行代表 1 条文本数据; R8 和 R52 是路透 21 578 数据集的两个子集, R8 数据集分为 8 类, 包含 7 674 条数据, R52 数据集分为 52 类, 包含 9 100 条数据, 数据集中每 1 行也代表 1 条文本数据; 20NG 是常用于文本分类领域的标准数据集之一, 包含 18 846 条数据, 分为 20 个不同的主题. 根据数据集中文本序列包含的平均单词数将数据集分为 3 类: 短文本数据集, 正常文本数据集以及长文本数据集. 短文本数据集的平均单词数在 0–45 之间 (MR), 正常文本数据集的平均单词数在 46–90 之间 (R8、R52), 平均单词数超过 91 为长文本数据集 (20NG). 4 个数据集的详细信息如表 1 所示.

2.2 实验环境和参数设置

实验的运行环境为: CPU 为 Intel Core i7-12650H,

GPU 为 Nvidia GeForce RTX4060. 开发环境为 Windows 11 操作系统, TensorFlow 1.15 版本.

表1 数据集详细信息

数据集	文本数	训练集	测试集	单词数	类别	平均单词数
MR	10062	7108	3554	18764	2	20.39
R8	7674	5485	2189	7688	8	65.72
R52	9100	6532	2568	8892	52	69.82
20NG	18846	11314	7532	42757	20	221.26

在构建语义文本图时, 初始词嵌入使用 Glove 预训练, 维数设为 300; 在构建顺序文本图时, 滑动窗口的大小设定为 20; 采用的两层张量图卷积层, 第 1 层嵌入节点的维数设定为 200, 第 2 层嵌入节点的维数设定为标签的个数, 隐藏层维数设定为 200, dropout 设定为 0.9, L2 正则衰减权重设为 0.000 001. 使用 Adam 作为实验的优化器, epoch 设定为 1000, 学习率设定为 0.0002.

2.3 评估指标

采用文本分类准确率 (accuracy) 对实验结果进行评估, 计算公式为:

$$accuracy = \frac{\text{分类正确的文本数}}{\text{总文本数}} \quad (15)$$

2.4 对比基线模型

采用的 6 个对比基线模型为近年在文本分类领域比较有代表性的模型, 可以分为 3 类, 第 1 类是基于深度学习的模型: CNN-rand; 第 2 类是基于图神经网络的模型: Graph-CNN、HINT、TextGCN 和 TensorGCN; 第 3 类是基于门控图神经网络的模型: TextFCG.

(1) CNN-rand^[27]: 利用卷积层及池化层对随机初始化的词向量进行特征提取, 然后通过全连接层对提取到的特征向量进行处理, 完成文本分类.

(2) HINT^[28]: 将文本信息解析为树状结构并构建为一张关系依赖树状图, 从而学习文本信息中包含的特征信息.

(3) Graph-CNN^[29]: 对输入的文本序列构建词图, 利用卷积神经网络对词图进行处理得到文本语义特征.

(4) TextGCN^[16]: 通过构建包含单词和文档节点的异构图来整体处理输入序列, 利用图卷积网络学习图节点的特征表示.

(5) TensorGCN^[23]: 对输入的文本序列分别构建语义、句法与顺序文本图, 使用图内传播和图间传播来协调节点信息及不同图的异构信息.

(6) TextFCG^[30]: 把输入序列中的上下文关系信息整合到单个图中, 利用图神经网络将局部单词与全局文本信息进行交互, 并增强节点的序列表示, 从而完成文本分类.

2.5 实验结果分析

2.5.1 分类准确率对比

根据对比表 2 中文本分类准确率可知, 本文提出的模型在 4 个数据集上对比基线模型都有不同程度的提升. 相比于原模型 TensorGCN, 本文模型的文本分类准确率在 4 个数据集 MR、R8、R52 和 20NG 上分别提升了 2.98%、0.35%、0.85% 和 2.32%.

分析实验结果可以发现, BTSGCN 与基于深度学习的模型 (CNN-rand) 相比, 在 4 个数据集 MR、R8、R52 和 20NG 上分别提升了 6.37%、4.24%、10.76% 和 12.58%, 提升比较明显, 这是因为将文本序列构建为图之后可以更好地捕捉单词之间的依赖关系, 同时处理文本信息的全局内容和局部细节, 提取更全面的文本特征, 从而在分类准确率上得到较大提升.

表2 不同模型在不同数据集上的文本分类准确率 (%)

基线模型	MR	R8	R52	20NG
CNN-rand	74.02±0.53	93.78±0.75	84.98±0.62	77.21±0.56
Graph-CNN	77.19±0.31	96.62±0.21	92.38±0.57	81.36±0.35
TextGCN	76.53±0.49	96.99±0.18	93.27±0.39	85.96±0.27
HINT	77.03±0.12	97.77±0.17	95.02±0.15	86.27±0.13
TensorGCN	77.41±0.16	97.68±0.08	94.89±0.23	87.47±0.09
TextFCG	78.65±0.13	97.82±0.05	95.08±0.16	88.56±0.20
BTSGCN	80.39±0.25	98.02±0.15	95.74±0.17	89.79±0.22

与基于图神经网络的 4 个对比模型: Graph-CNN、HINT、TextGCN 和 TensorGCN 相比, 2023 年提出的 TextFCG 模型通过引入自注意力机制和优化特征融合过程, 增强模型对特征间复杂关系的捕捉能力, 同时动态调整特征权重, 增强对重要特征的利用, 更有效地整合多源异构特征, 在对比基线模型里面表现最佳. 与之相比, 本文 BTSGCN 对短文本数据集 (MR) 以及长文本数据集 (20NG) 的分类准确率提升较为明显, 相比改进前的基线模型 TensorGCN 在 MR 和 20NG 数据集上分类准确率分别提升 2.98% 与 2.32%; 在正常文本数据集 R8、R52 上分类准确率提升效果较小. 这是因为在提取语义特征时, BTSGCN 使用 BERT 来处理短文本序列信息, 更好地捕获单词之间的相关性, 提取到更深层次的语义特征; 同时在处理图张量时加入自注意力机制, 帮助模型在处理长文本中包含特征

信息太多的情况时更好融合特征信息,提高模型性能。

与基于门控图神经网络的模型 (TextFCG) 相比, BTSGCN 在 4 个数据集 MR、R8、R52 和 20NG 上的分类准确率都有提升,分别提升了 1.74%、0.2%、0.66% 和 1.23%。这是因为不同于 TextFCG 只使用上下文关系信息提取特征, BTSGCN 模型同时提取上下文信息、句法信息以及语义信息特征,更深层次地挖掘文本信息中所包含的特征,并结合 BERT 以及自注意力机制帮助模型更好地提取、融合特征,提高模型的性能。

2.5.2 消融实验

为了验证 BERT 模块与自注意力机制模块的有效性,在 4 个数据集上进行两种消融实验: (1) TensorGCN 加入自注意力机制; (2) TensorGCN 使用 BERT 替换 LSTM 模块。将消融实验的准确率与完整模型的准确率进行比较,结果如表 3 所示。其中, N-BERT 表示未使用 BERT 替换 LSTM 模块,其他部分与 BTSGCN 模型保持一致; N-Attention 表示未加入自注意力机制模块,其他部分与 BTSGCN 模型保持一致。

表 3 消融实验的结果 (%)

模型	MR	R8	R52	20NG
TensorGCN	77.41±0.16	97.68±0.08	94.89±0.23	87.47±0.09
N-BERT	78.56±0.12	97.91±0.04	95.47±0.09	89.46±0.15
N-Attention	79.93±0.21	97.81±0.11	95.11±0.14	88.19±0.18
BTSGCN	80.39±0.25	98.02±0.15	95.74±0.17	89.79±0.22

根据表 3 的结果可以发现,无论是使用 BERT 替换 LSTM 还是加入自注意力机制,相比原模型准确率均有一定程度的提高,融合两个模块的准确率高于只融合一个模块的准确率,因此这两个模块都是有效的,且可以相互作用更好地提高模型的性能。

对于不同的数据集,融合不同模块之后的文本分类准确率也不同。TensorGCN 加入自注意力机制之后,模型的准确率 4 个数据集 MR、R8、R52 和 20NG 上分别提升了 1.15%、0.23%、0.58% 和 1.99%。对长文本数据集 20NG 提升的程度最大,这是因为长文本序列包含的信息太多,进行图间传播时特征共享不完全;加入自注意力机制之后,通过自注意力机制捕捉全局信息和长距离单词之间的特征依赖关系,同时分配不同的权重对不同特征进行加权平均以更好地融合特征,证明了自注意力机制模块的有效性。

TensorGCN 使用 BERT 替换 LSTM 模块之后,模型的准确率在 4 个数据集 MR、R8、R52 和 20NG 上分别提升 2.52%、0.13%、0.22% 和 0.72%。对短文本数据集

MR 提升程度最大,这是因为短文本序列包含的信息太少,导致 LSTM 在进行语义特征提取时的迭代次数太少而造成语义特征提取不全面;而 BERT 通过考虑每个单词两侧的单词来捕获单词之间的依赖关系以提取更全面的语义特征,证明了 BERT 模块的有效性。

2.5.3 可视化分析

为了更清晰地体现 BTSGCN 模型对于文本分类任务的效果,使用 t 分布-随机邻近嵌入 (t-distributed stochastic neighbor embedding, t-SNE)^[31,32]技术与线性判别分析 (linear discriminant analysis, LDA)^[33,34]分别对 BTSGCN 和 TensorGCN 在 MR 与 20NG 数据集上学习的特征信息进行聚类并将结果可视化。结果如图 2 所示,图中每个点代表 1 条文本信息,同一种颜色表示同一类别的文本信息。其中,图 2(a)、(b) 代表 BTSGCN 模型在 MR 与 20NG 数据集上的可视化结果;图 2(c)、(d) 代表 TensorGCN 模型在 MR 与 20NG 数据集上的可视化结果。

从图 2 可以看出,两种模型在两个数据集上的可视化结果图中都存在不同程度的点重叠部分,这是因为这些点代表的文本特征信息在转化为向量时具有相似的表示,从而在聚类时出现了重叠。然而 BTSGCN 在两个数据集上可视化结果图中的重叠现象明显少于 TensorGCN 的结果图,且不同的类别信息相较于 TensorGCN 能更加清晰地区别开来。这表明 BTSGCN 学习的文本特征信息比 TensorGCN 更全面,从而处理文本分类任务的效果更优。

3 结束语

本文在 TensorGCN 模型基础之上进行改进,提出了一种 BTSGCN 文本分类方法。该模型使用 BERT、Stanford CoreNLP 解析器以及滑动窗口算法和 PMI 技术来提取语义、句法和顺序上下文信息特征,将单词作为图节点,利用单词之间的依赖关系构建语义、句法以及顺序文本图,通过图内传播和图间传播两种传播方式进行图张量学习,并在图间传播过程中加入自注意力机制帮助模型更好地融合图间异构特征,最后通过池化层进行平均池化操作获得图节点的全局特征表示以完成分类。实验表明,本文模型比基线模型的准确率更高。未来,针对短文本数据集语义特征较少的问题,研究在 BERT 中融合知识增强技术,以进一步提升文本分类准确率。

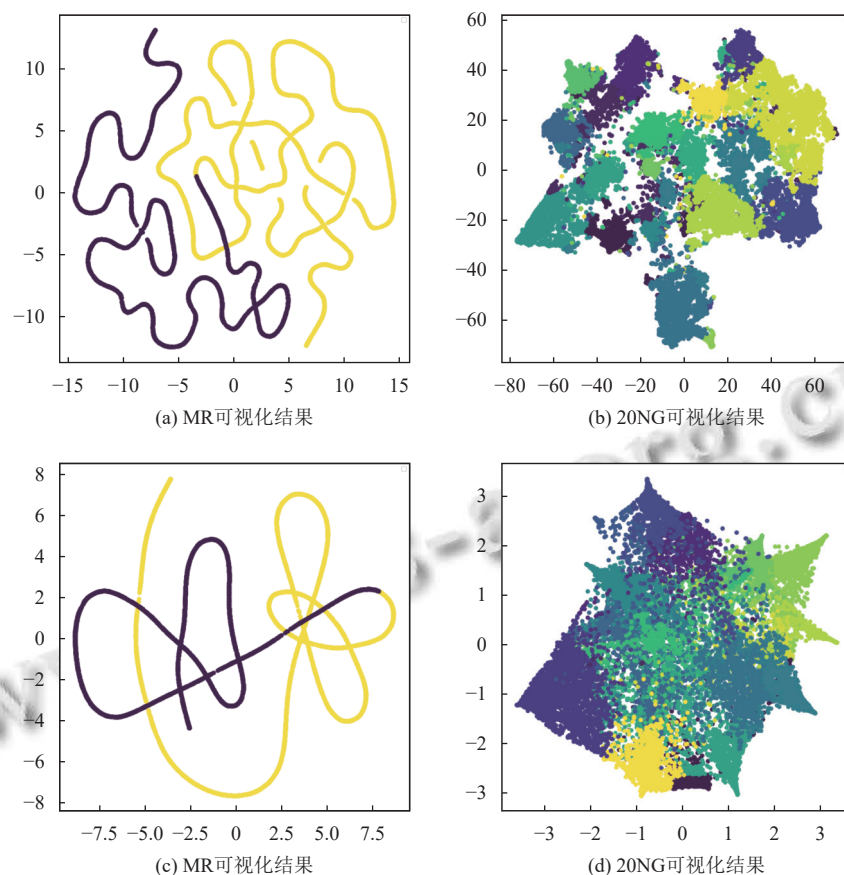


图2 可视化文本信息特征分类结果

参考文献

- 1 郑诚, 肖双. 结合语法规则和图神经网络的文本分类方法. 小型微型计算机系统, 2024, 45(11): 2594–2601.
- 2 Hu ZK, Hu JQ, Ding WF, *et al.* Review sentiment analysis based on deep learning. Proceedings of the 12th International Conference on e-Business Engineering. Beijing: IEEE, 2015. 87–94.
- 3 Atanasova P, Simonsen JG, Lioma C, *et al.* A diagnostic study of explainability techniques for text classification. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. ACL, 2020. 3256–3274.
- 4 Sun XF, Li XY, Li JW, *et al.* Text classification via large language models. Proceedings of the 2023 Association for Computational Linguistics. Singapore: ACL, 2023. 8990–9005.
- 5 Yu WB, Yin L, Zhang CJ, *et al.* Application of quantum recurrent neural network in low-resource language text classification. IEEE Transactions on Quantum Engineering, 2024, 5: 2100213.
- 6 Chai YY, Li Z, Liu JH, *et al.* Compositional generalization for multi-label text classification: A data-augmentation approach. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2024. 17727–17735.
- 7 Ding K, Li JD, Bhanushali R, *et al.* Deep anomaly detection on attributed networks. Proceedings of the 2019 SIAM International Conference on Data Mining. Tempe: SIAM, 2019. 594–602.
- 8 Papageorgiou G, Economou P, Bersimis S. A method for optimizing text preprocessing and text classification using multiple cycles of learning with an application on shipbrokers emails. Journal of Applied Statistics, 2024, 51(13): 2592–2626. [doi: [10.1080/02664763.2024.2307535](https://doi.org/10.1080/02664763.2024.2307535)]
- 9 Roy PK, Singh JP, Banerjee S. Deep learning to filter SMS Spam. Future Generation Computer Systems, 2020, 102: 524–533. [doi: [10.1016/j.future.2019.09.001](https://doi.org/10.1016/j.future.2019.09.001)]
- 10 Kim Y. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014. 1746–1751.
- 11 Liu PF, Qiu XP, Huang XJ. Recurrent neural network for text classification with multi-task learning. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: AAAI Press, 2016. 2873–2879.
- 12 Hochreiter S, Schmidhuber J. Long short-term memory.

- Neural Computation, 1997, 9(8): 1735–1780. [doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)]
- 13 Zhang Y, Liu Q, Song LF. Sentence-state LSTM for text representation. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 317–327.
- 14 Cho K, Van Merriënboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014. 1724–1734.
- 15 Yang ZC, Yang DY, Dyer C, *et al.* 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: ACL, 2016. 1480–1489.
- 16 Yao L, Mao CS, Luo Y. Graph convolutional networks for text classification. Proceedings of the 33th AAAI Conference on Artificial Intelligence and the 31th Innovative Applications of Artificial Intelligence Conference and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019. 905.
- 17 Huang LZ, Ma DH, Li SJ, *et al.* Text level graph neural network for text classification. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 3444–3450.
- 18 Wu F, Souza A, Zhang TY, *et al.* Simplifying graph convolutional networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 6861–6871.
- 19 Zhang YF, Yu XL, Cui ZY, *et al.* Every document owns its structure: Inductive text classification via graph neural networks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 334–339.
- 20 Wang KZ, Han SC, Poon J. InducT-GCN: Inductive graph convolutional networks for text classification. Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Montreal: IEEE, 2022. 1243–1249.
- 21 Zhu H, Koniusz P. Simple spectral graph convolution. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
- 22 Linmei H, Yang TC, Shi C, *et al.* Heterogeneous graph attention networks for semi-supervised short text classification. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 4821–4830.
- 23 Liu XE, You XX, Zhang X, *et al.* Tensor graph convolutional networks for text classification. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 8409–8416.
- 24 Devika R, Vairavasundaram S, Mahenthara C S J, *et al.* A deep learning model based on BERT and sentence Transformer for semantic keyphrase extraction on big social data. IEEE Access, 2021, 9: 165252–165261. [doi: [10.1109/ACCESS.2021.3133651](https://doi.org/10.1109/ACCESS.2021.3133651)]
- 25 Shen DH, Zareapoor M, Yang J. Multimodal image fusion based on point-wise mutual information. Image and Vision Computing, 2021, 105: 104047. [doi: [10.1016/j.imavis.2020.104047](https://doi.org/10.1016/j.imavis.2020.104047)]
- 26 Yan MY, Chen ZD, Deng L, *et al.* Characterizing and Understanding GCNs on GPU. IEEE Computer Architecture Letters, 2020, 19(1): 22–25. [doi: [10.1109/LCA.2020.2970395](https://doi.org/10.1109/LCA.2020.2970395)]
- 27 Widiastuti NI. Convolution neural network for text mining and natural language processing. IOP Conference Series: Materials Science and Engineering, 2019, 662(5): 052010. [doi: [10.1088/1757-899X/662/5/052010](https://doi.org/10.1088/1757-899X/662/5/052010)]
- 28 Zhang C, Zhu H, Peng X, *et al.* Hierarchical information matters: Text classification via tree based graph neural network. Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, 2022. 950–959.
- 29 Peng H, Li JX, He Y, *et al.* Large-scale hierarchical text classification with recursively regularized deep graph-CNN. Proceedings of the 2018 World Wide Web Conference. Lyon, 2018. 1063–1072.
- 30 Wang YZ, Wang CX, Zhan JY, *et al.* Text FCG: Fusing contextual information via graph learning for text classification. Expert Systems with Applications, 2023, 219: 119658. [doi: [10.1016/j.eswa.2023.119658](https://doi.org/10.1016/j.eswa.2023.119658)]
- 31 王宁, 郭梓昱, 田淑珂, 等. 基于融合特征 t-SNE 降维的控制图质量异常模式识别. 系统工程理论与实践, 2024, 44(7): 2381–2393. [doi: [10.12011/SETP2023-1577](https://doi.org/10.12011/SETP2023-1577)]
- 32 Poličar PG, Stražar M, Zupan B. Embedding to reference t-SNE space addresses batch effects in single-cell classification. Machine Learning, 2023, 112(2): 721–740. [doi: [10.1007/s10994-021-06043-1](https://doi.org/10.1007/s10994-021-06043-1)]
- 33 蒋志强, 陶屹, 崔岩, 等. 文本主题特征和医疗众筹筹资绩效: 基于 LDA 模型的研究. 中国管理科学. <https://link.cnki.net/urlid/11.2835.g3.20240517.2042.005>. (2024-05-21) [2024-07-04]. [doi: [10.16381/j.cnki.issn1003-207x.2022.1311](https://doi.org/10.16381/j.cnki.issn1003-207x.2022.1311)]
- 34 Li SY, Zhang HM, Ma RJ, *et al.* Linear discriminant analysis with generalized kernel constraint for robust image classification. Pattern Recognition, 2023, 136: 109196. [doi: [10.1016/j.patcog.2022.109196](https://doi.org/10.1016/j.patcog.2022.109196)]

(校对责编: 王欣欣)