第 40 卷 第 1 期
2025 年 2 月

天津科技大学学报
Journal of Tianjin University of Science and Technology

Vol. 40 No. 1
Feb. 2025

DOI:10.13364/j.issn.1672-6510.20230195
网络首发日期：2024–06–25；网络首发地址：http://link.cnki.net/urlid/12.1355.N.20240624.1440.001

融合图嵌入和 BERT 嵌入的文本分类模型

常慧霞，李孝忠
（天津科技大学人工智能学院，天津 300457）

摘 要：文本分类作为自然语言领域中的重要任务之一，广泛应用于问答系统、推荐系统以及情感分析等相关任务中。为了提取文本数据中的复杂语义特征信息并捕获全局的图信息，提出一种融合图嵌入和 BERT（bidirectional encoder representation from Transformers）嵌入的文本分类模型。该模型引入双级注意力机制考虑不同类型节点的重要性以及同一类型不同相邻节点的重要性，同时采用 BERT 预训练模型获得包含上下文信息的嵌入并解决一词多义的问题。该模型把所有单词和文本均视为节点，为整个语料库构建一张异构图，将文本分类问题转化为节点分类问题。将双级注意力机制与图卷积神经网络进行融合，双级注意力机制包含类型级注意力和节点级注意力。类型级注意力机制捕获不同类型的节点对某一节点的重要性，节点级注意力机制可以捕获相同类型的相邻节点对某一节点的重要性。将 BERT 模型获得的文本中局部语义信息与经图卷积神经网络得到的具有全局信息的图嵌入表示相结合，得到最后的文本嵌入表示，并完成文本分类。在 4 个广泛使用的公开数据集上与 7 个基线模型进行对比实验，结果表明本文模型提高了文本分类的准确性。

关键词：文本分类；图卷积神经网络；注意力机制；BERT
中图分类号：TP391 **文献标志码**：A **文章编号**：1672-6510（2025）01-0072-09

Text Classification Model Based on BERT and Dual-Level Attention Mechanism

CHANG Huixia，LI Xiaozhong
（College of Artificial Intelligence，Tianjin University of Science and Technology，Tianjin 300457，China）

Abstract：Text classification is one of the crucial tasks in the field of natural language and is widely used in related tasks such as question answering system，recommendation system，and sentiment analysis. In order to extract complex semantic feature information in text data and capture global graph information，a text classification model based on bidirectional encoder representation from Transformers（BERT）and a dual-level attention mechanism is proposed in this article. This model introduces a two-level attention mechanism to consider the importance of different types of nodes and the importance of different neighboring nodes of the same type. At the same time，a BERT pre-training model is used to obtain embeddings containing contextual information and solve the problem of polysemy. This method treats all words and texts as nodes，builds a heterogeneous graph for the entire corpus，and transforms the text classification problem into a node classification problem. The dual-level attention mechanism is then integrated with the graph convolutional neural network. The dual-level attention mechanism includes type-level attention and node-level attention. The type-level attention mechanism captures the importance of different types of nodes to a certain node. The node-level attention mechanism can capture the importance of neighbor nodes of the same type to a certain node. Then，the local semantic information in the text obtained by the BERT model is combined with the graph embedding representation with global information obtained by the graph convolutional neural network to obtain the final text embedding representation and complete the text classification. Comparative experiments were conducted with seven baseline models on four widely used public data sets. The results showed that the proposed

收稿日期：2023–10–17；**修回日期**：2024–03–06
作者简介：常慧霞（1999—），女，河南安阳人，硕士研究生；**通信作者**：李孝忠，教授，lixz@tust.edu.cn

model improved the accuracy of text classification.

Key words：text classification；graph convolutional neural network；attention mechanism；BERT

引文格式：

常慧霞，李孝忠. 融合图嵌入和 BERT 嵌入的文本分类模型[J]. 天津科技大学学报，2025，40(1)：72-80.

CHANG H X，LI X Z. Text Classification model based on BERT and dual-level attention mechanism[J]. Journal of Tianjin university of science and technology，2025，40(1)：72-80.

随着互联网的迅速发展，文本形式的数据信息大幅度增加。面对海量的文本数据，如何有效地进行管理和应用，并获取潜在有价值的信息是一项较为重要的任务。文本分类不仅是自然语言处理(natural language processing，NLP)[1]领域重要的分支之一，而且是许多文本相关任务的基础。文本分类任务是指对给定的一段文本内容进行特征提取分析，并将文本分配到预定义的类别或标签中，使具有相似主题或含义的文本被归类在一起。因此，高效的文本分类模型可以最大程度地提高信息检索和数据挖掘的效率。

现有的文本分类方法主要分为两类，即基于传统机器学习的文本分类方法和基于深度学习的文本分类方法。相较于传统的机器学习方法而言，基于深度学习的文本分类方法可以利用其自身的网络结构自动学习数据的特征表示，不需要繁琐冗长的特征工程。基于深度学习的方法中，卷积神经网络(convolutional neural networks，CNN)[2]、循环神经网络(recurrent neural network，RNN)[3]、长短期记忆网络(long short-term memory，LSTM)[4]、胶囊网络[5]、图神经网络(graph neural network，GNN)[6]等模型相继应用于文本分类任务中。Kim[7]在 CNN 的基础上提出 Text-CNN 模型，通过引入多个不同尺寸的卷积核，从不同的上下文窗口中提取特征，以更全面地捕捉句子中的信息，从而解决了句子级别的文本分类任务。同时运用池化操作进行特征降维，保留最重要的特征信息并减少参数数量，从而提高模型的效率和泛化能力。Zhang 等[8]通过使用字符级别的卷积神经网络结构，也取得了同样显著的分类效果。Liu 等[9]的 Text-RNN 模型是在 RNN 的基础上提出的，该模型实现了循环神经网络在文本分类任务上的应用且具有短期记忆功能，能够处理带时序关系的任务。Bahdanau 等[10]提出的注意力机制能够自动捕捉输入文本中不同部分的重要程度。HAN(Heteroge-neous graph attention network)模型[11]在 Text-RNN 模型的基础上引入注意力机制，并将注意力机制从同质图扩展到节点和边有不同类型的异构图。Hochreiter 等[4]在长短期记忆网络中引入门控机制，通过控制信息的流动和记忆，能够更好地捕捉和处理时间序列数据中的长期依赖关系。

尽管神经网络模型已经能够较好地捕获到句法和语义信息，并取得了不错的分类效果。但是，传统的神经网络仅适用于处理欧几里得数据，对于文本等非欧几里得数据而言，首先需要将其转化为二维矩阵向量的形式，然后通过神经网络进行训练。这种转化过程可能会限制神经网络的表达能力，尤其是涉及包含复杂语法结构的文本。

近年来，由于 GNN 能够捕捉数据中的拓扑信息并适合处理不规则的图结构数据，在自然语言领域取得了较好的效果。图卷积神经网络(graph convolution neural networks，GCN)[12]作为图神经网络的一种，其出现在文本分类领域时引起了广泛的关注。GCN 融合了深度学习算法和图算法，以图的形式表示文本信息，通过不同类型节点之间的连接，将文本分类问题转化为节点分类问题或图分类问题。首次将图卷积神经网络应用到文本分类任务中的是 Text-GCN 模型[13]，其将整个语料库中的单词和文本都视为节点，并构建一个无向加权异构图。该模型不仅可以学习词嵌入，还可以学习文本嵌入。此外，TensorGCN 模型[14]构建了 3 种异构图，并引入 LSTM 和单词之间的句法依赖，采用图间和图内两种信息传播方式协调不同图之间的异构信息。Huang 等[15]提出的 Text-level GCN 模型为每个输入文本构建了一个具有全局参数共享的有向图。通过消息传递机制[16]可以更新节点表示和边权，解决了内存高消耗的问题，并对新文本有较好的泛化能力。为了获取上下文信息，Gao 等[17]提出了图池化层和结合图卷积以及常规一维卷积的混合卷积层。该模型不仅可以提取文本的语序信息，还能快速增大感受野并自动计算特征。Li 等[18]提出了一种基于归纳学习的文本分类方法，利用异构信息网络上的表示学习和外生知识，将非结构化文本表示为结构化的异构信息网络，扩展了文本特征的粒度，充分利用了外生的结构信息和显性的语义信息，

增强了文本信息的可解释性。Liu 等[19]提出基于 Transformer 和 GCN 的文本分类模型，将输入文本经过词嵌入和位置编码，通过 Transformer 编码器进行特征提取，然后 GCN 利用图卷积操作对隐藏层特征进行进一步编码。该模型可以同时利用全局信息和局部信息，提高对文本语境和结构的建模能力。

在上述模型中，Text-level GCN 模型[15]仅构建了文本级别的图，而没有考虑文档节点，此外 Text-GCN 模型[13]和 TensorGCN 模型[14]忽略了不同相邻节点对当前节点的重要性，在聚合相邻信息时不能有效提取关键信息。针对以上问题，本文提出融合图嵌入和 BERT（bidirectional encoder representation from Transformers）嵌入的文本分类模型。该模型主要在 Text-GCN 模型的基础上引入双级注意力机制[20]考虑不同类型节点对某一节点的重要性以及相同类型相邻节点对当前节点的重要性，同时利用 BERT 预训练模型获得包含上下文信息的嵌入并解决一词多义的问题。

1　异构文本图建模

为了能够准确地建立全局单词共现关系，并且便于进行图卷积操作，本文模型构建了一个包含单词节点和文档节点的异构图。异构图包含两类节点、三种类型的边。下面将介绍不同类型节点所构成的边及其权重的计算方法。

异构图表示为 $G=(V,E)$，其中 V 是节点集，E 是边集。异构图|V|中的节点数是文档数量加上语料库中唯一单词的数量。节点集 $V=\{X,Y\}$，其中 X 表示单词节点集，Y 表示文档节点集。边集 $E=\{(x_i,x_j),(x_i,y_j),(y_i,y_j)\}$，其中 (x_i,x_j) 表示单词节点与单词节点构成的边，(x_i,y_j) 表示单词节点和文档节点构成的边，(y_i,y_j) 表示文档节点和文档节点构成的边。

1.1　单词节点与单词节点

为了充分利用全局词共现信息，本文模型采用了一个固定大小的滑动窗口在语料库中的所有文档上进行词共现信息的统计[21]。这种统计方式使用了 PMI（point-wise mutual information）计算两个单词节点之间的权重。初步实验结果表明，采用 PMI 相比于单纯的词共现计数取得了更好的效果。定义单词节点 x_i 和单词节点 x_j 之间的 PMI 值计算公式为

$$\mathrm{PMI}(x_i,x_j)=\lg\frac{p(x_i,x_j)}{p(x_i)p(x_j)} \tag{1}$$

$$p(x_i,x_j)=\frac{W^{\#}(x_i,x_j)}{W^{\#}} \tag{2}$$

$$p(x_i)=\frac{W^{\#}(x_i)}{W^{\#}} \tag{3}$$

其中：$p(x_i,x_j)$ 表示单词 x_i 和单词 x_j 在整个语料库中出现的联合概率，$p(x_i)$ 和 $p(x_j)$ 分别表示单词 x_i 和单词 x_j 出现的边缘概率，$W^{\#}(x_i)$ 是指语料库中包含单词 x_i 的滑动窗口数量，$W^{\#}(x_i,x_j)$ 是指语料库中同时包含单词 x_i 和单词 x_j 的滑动窗口数量，$W^{\#}$ 是语料库中滑动窗口的总数。

当 PMI 值为正时，说明语料库中的这两个词之间具有较高的语义相关性；当 PMI 值为负时，则表示这两个单词之间的语义关联较低或者不存在关联。因此，本文只对 PMI 值为正的两个词之间添加边。两个单词之间的权重计算公式为

$$A_{i,j}=\begin{cases}\mathrm{PMI}(x_i,x_j), \text{当 } \mathrm{PMI}(x_i,x_j)>0\\ 0, \text{其他}\end{cases} \tag{4}$$

1.2　单词节点与文档节点

根据该单词在文档中的词频–逆文档频率（TF-IDF，用符号 $f_{\text{TF-IDF}}$ 表示）可以计算文档节点和单词节点之间的边的权重。TF-IDF 综合考虑了单词在文档中的出现频率（词频，f_{TF}）和整个语料库中包含该单词的文档数量（逆文档频率，f_{IDF}）。逆文档频率是通过将文档总数除以包含该单词的文档数量，然后取以 10 为底的对数得出。TF-IDF 能够较全面地衡量 1 个单词对当前文档的重要性，是词频和逆文档频率的乘积，TF-IDF 值越大，说明该词对文档的重要程度越高。

$$f_{\text{TF-IDF}(x_i,y_j)}=f_{\text{TF}(x_i,y_j)}\cdot f_{\text{IDF}(x_i)} \tag{5}$$

$$f_{\text{TF}(x_i,y_j)}=\frac{c(x_i)}{N} \tag{6}$$

$$f_{\text{IDF}(x_i)}=\lg\frac{N+1}{N(x_i)+1} \tag{7}$$

其中：$f_{\text{TF}(x_i,y_j)}$ 表示单词 x_i 在文档 y_j 中的词频，$f_{\text{IDF}(x_i)}$ 表示单词 x_i 在语料库中的逆文档频率，N 表示语料库中拥有的文档数，$N(x_i)$ 表示语料库中包含单词 x_i 的文档数。

一般来说，词频越高则说明该词在文档中的重要性越高，但是仅仅用词频考虑某个单词对文档的重要

性是远远不够的。逆文档频率能够惩罚一些词频高的单词，如果该单词在语料库中的多个文档中出现，则该单词的重要性会被削减，即 IDF 与词对文档的重要性成正比。由于单词 x_i 出现在语料库中的文档总数可能为 0，即 $N(x_i)=0$，因此将它做了一些平滑处理。

1.3　文档节点与文档节点

使用 one-hot 编码作为文档节点的初始化。对于两个文档节点所构成的边，它们的权重计算方式是对于任意两个文档节点 y_i, y_j，若 $i=j$，则它们之间的权重 $A_{i,j}=1$，否则 $A_{i,j}=0$，权重计算公式为

$$A_{i,j}=\begin{cases}1, & \text{当 } y_i=y_j\\ 0, & \text{其他}\end{cases} \tag{8}$$

本文将节点之间构成的边的权重转化为邻接矩阵的形式，其邻接矩阵表示为

$$A_{i,j}=\begin{cases}\mathrm{PMI}(x_i,x_j) & x_i,x_j\in X,\text{当 } \mathrm{PMI}(x_i,x_j)>0\\ f_{\text{TF-IDF}(x_i,y_j)} & x_i\in \mathrm{X}, y_j\in \mathrm{Y}\\ 1 & y_i,y_j\in Y,\text{当 } y_i=y_j\\ 0 & \text{其他}\end{cases} \tag{9}$$

通过上述对异构图构建过程的描述可知，异构图由两类节点与三类边构成，其示意图如图 1 所示。

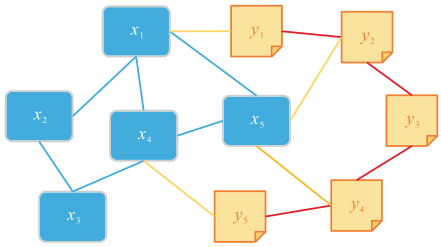


图 1　异构图示例

Fig. 1　Example of heterogeneous graph

2　基于双级注意力机制的图卷积神经网络

2.1　图嵌入

在构建完异构图之后，将图输入图卷积神经网络。采用半监督分类的图卷积神经网络作为基础网络模型[6]，GCN 的层级传播公式为

$$\hat{\boldsymbol{H}}^{L}=f\left(\tilde{\boldsymbol{A}}\boldsymbol{H}^{(L-1)}W_{L-1}\right) \tag{10}$$

$$\boldsymbol{D}_{ii}=\sum_j \boldsymbol{A}_{ij} \tag{11}$$

$$\tilde{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} \tag{12}$$

其中：$\boldsymbol{D}$ 为节点的度矩阵，$\tilde{\boldsymbol{A}}$ 为标准化的邻接对称矩阵，$\sum$为求和符号，$f(\cdot)$为激活函数，W_{L-1}为 GCN 的网络参数，$\boldsymbol{H}^{(l)}\in\boldsymbol{R}^{|v|\times q}$ 表示第 1 层节点的隐藏表示。首层可以表示为

$$\boldsymbol{H}^{(0)}=\tilde{\boldsymbol{A}}\boldsymbol{X}W_0 \tag{13}$$

为考虑不同信息类型的异构性，使 GCN 适应包含不同类型节点 $T=\{\tau_1,\tau_2\}$，将不同类型节点的特征空间连接在一起构造一个新的大特征空间，则更新传播公式为

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\sum_{\tau\in T}\tilde{\boldsymbol{A}}_{\tau}\cdot\boldsymbol{H}_{\tau}^{(l)}\cdot\boldsymbol{W}_{\tau}^{(l)}\right) \tag{14}$$

其中：$\tilde{\boldsymbol{A}}_{\tau}\in\boldsymbol{R}^{|v|\times|v_\tau|}$ 是 $\tilde{\boldsymbol{A}}$ 的子矩阵，其行表示所有节点，列表示类型为 τ 的相邻节点。$\boldsymbol{H}^{(l+1)}$ 表示通过使用不同的变换矩阵 $\boldsymbol{W}_{\tau}^{(l)}$ 聚合具有不同类型 τ 的相邻节点 $\boldsymbol{H}_{\tau}^{(l)}$ 的特征信息。

2.2　双级注意力机制

对于图中的任一节点而言，不同类型的相邻节点对该节点特征的影响也会有所不同，并且同一类型的节点也会携带不同的信息，因此即使类型相同的不同相邻节点在为该节点传递特征时应该具有不同的权重。在文本异构图中，文档节点会有文档节点和单词节点作为它的相邻节点，在传播过程中会从它的相邻节点处聚合特征，而单词传递的信息应该在重要性上有所不同，同一类型下的节点如单词节点也应该具备不同的权重。

针对这一问题，本文模型引入 Hu 等[20]提出的双级注意力机制。在节点聚合相邻节点特征时，针对相邻节点类型不同以及同一类型下的节点不同提出双级注意力机制以分配权重。双级注意力包括类型级注意力与节点级注意力。节点级注意力的计算需要建立在类型级注意力的基础上。

2.2.1　类型级注意力

给定一个节点 $\boldsymbol{v}$，其相邻节点为 $\boldsymbol{v}'$。首先将类型 τ 的嵌入表示为 $\boldsymbol{h}_{\tau}=\sum_{v'}\tilde{\boldsymbol{A}}_{vv'}\boldsymbol{h}_{v'}$，其表示为相邻节点特征 $\boldsymbol{h}_{v'}$ 的总和。根据当前节点的嵌入表示 $\boldsymbol{h}_v$ 和类型嵌入表示 $\boldsymbol{h}_{\tau}$ 计算类型级注意力得分为

$$a_{\tau}=\sigma\left(\boldsymbol{\mu}_{\tau}^{\mathrm{T}}\cdot[\boldsymbol{h}_v\,\|\,\boldsymbol{h}_{\tau}]\right) \tag{15}$$

其中：$\boldsymbol{\mu}_{\tau}$ 表示类型 τ 的注意力向量，$\|$表示拼接，$\sigma(\cdot)$表示激活函数。

随后使用 Softmax 函数标准化所有类型注意力分数，获得类型级注意力权重，为

$$\alpha_{\tau}=\frac{\exp(a_{\tau})}{\sum_{\tau'\in T}\exp(a_{\tau'})} \tag{16}$$

2.2.2　节点级注意力

通过节点级注意力可以捕捉不同相邻节点的重要性，并且减少噪声节点的权重。给定类型为 τ 的特定节点 v 及其类型 τ' 的相邻节点 v'，根据节点嵌入 $\boldsymbol{h}_v$ 和 $\boldsymbol{h}_{v'}$ 以及类型级注意力权重 $\alpha_{\tau'}$，计算节点 v' 的节点级注意力分数为

$$b_{vv'}=\sigma\left(\boldsymbol{v}^{\mathrm{T}}\cdot\alpha_{\tau'}\left[\boldsymbol{h}_v\,\|\,\boldsymbol{h}_{v'}\right]\right) \tag{17}$$

其中：$\boldsymbol{v}$ 为注意力向量。使用 Softmax 函数对节点注意力分数进行归一化处理，即

$$\beta_{vv'}=\frac{\exp\left(b_{vv'}\right)}{\sum_{i\in N_v}\exp\left(b_{vi}\right)} \tag{18}$$

最后，将包含类型级和节点级注意力的双级注意力机制融合到 GCN 中，传播规则为

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\sum_{\tau\in T}\boldsymbol{B}_\tau\cdot\boldsymbol{H}_\tau^{(l)}\cdot\boldsymbol{W}_\tau^{(l)}\right) \tag{19}$$

其中：$\boldsymbol{B}_\tau$ 表示注意力矩阵，其第 v 行第 v' 列为 $\beta_{vv'}$。

经过上述融合了双级注意力机制的图卷积神经网络之后，可以得到所有节点新的表示向量。

3　融合图嵌入和 BERT 嵌入的文本表示及文本分类

经过图卷积神经网络训练之后，文档节点和单词节点都可以通过聚合相邻节点的特征更新其特征，同时融合双级注意力机制赋予不同的权重，最后得到包含全局信息的嵌入表示。但是，在构建异构图的同时，会随之丢失文本中局部的上下文信息，这些局部信息在分类效果中起着重要作用。于是，本文模型引入 BERT 模型[22]，BERT 是一种基于 Transformer 架构的预训练语言模型，其优势在于能够双向预测，即可以通过考虑上下文信息以及目标语言的特征获得更好的文本表示。

BERT 模型包含 3 层，即输入层、编码层以及输出层。文本以序列的形式传入输入层。假设一个含有 n 个单词 $\{\boldsymbol{W}_0,\boldsymbol{W}_1,\cdots,\boldsymbol{W}_n\}$ 的文本序列，经过 BERT 嵌入层可以得到包含局部信息的嵌入表示 $\{\boldsymbol{W}_0',\boldsymbol{W}_1',\cdots,\boldsymbol{W}_n'\}$，$\{\boldsymbol{G}_0,\boldsymbol{G}_1,\cdots,\boldsymbol{G}_n\}$ 为语料库中所有单词经过融合双级注意力机制的图卷积神经网络模型后得到的图嵌入表示[23]。为综合两者的优势，将两种嵌入表示进行拼接作为最终的词向量表示。

$$\boldsymbol{h}_i=\boldsymbol{G}_i\,\|\,\boldsymbol{W}_i' \tag{20}$$

其中：‖表示向量的拼接操作，$\boldsymbol{h}_i$ 为同时具有全局信息与局部信息的嵌入表示。

通过图卷积神经网络得到异构图中每一个节点的嵌入表示，将文档节点的图嵌入表示称为潜在文本向量 $\boldsymbol{\mu}_t$。为体现不同单词的重要性，再次引入注意力机制得到新的文本向量，即

$$\boldsymbol{u}_i=\tanh\left(\boldsymbol{W}_t\boldsymbol{h}_i+\boldsymbol{b}_t\right) \tag{21}$$

$$\chi_i=\frac{\exp\left(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{\mu}_t\right)}{\sum_i\exp\left(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{\mu}_t\right)} \tag{22}$$

$$\boldsymbol{V}_t=\sum_t\chi_i\boldsymbol{h}_i \tag{23}$$

其中：$\boldsymbol{W}_t$ 为词的嵌入矩阵表示，χ_i 为第 i 个单词对于文本语义表示的权重，$\boldsymbol{V}_t$ 为语料库中各单词与其求得的权重线性加权后得到的文本向量。

得到最终文本的嵌入表示 $\boldsymbol{V}_t$ 后，通过 Softmax 分类器进行文本分类。分类器使用交叉熵函数作为模型的损失函数，即

$$L=\mathrm{CrossEntropy}(\boldsymbol{x},\boldsymbol{x}') \tag{24}$$

其中：$\boldsymbol{x}$ 表示文本的真实标签，$\boldsymbol{x}'$ 表示模型预测的类别。

4　实　验

4.1　数据集

通过实验在 4 个公开数据集上对比本文模型和基线模型的有效性，采用正确率作为评价模型的性能指标。数据集分别为 R8、R52、Ohsumed 和 MR。使用与文献[24]相同的方法将数据集分为训练集和测试集，并在训练集中随机选择了 10% 的文本作为验证集。表 1 为数据集详细信息。

表 1　数据集详细信息

Tab. 1　Details of datasets

数据集	文本数	训练量	测试量	节点数	类别	平均长度
R8	7 674	5 485	2 189	15 326	8	66
R52	9 100	6 532	2 568	17 992	52	70
Ohsumed	7 400	3 357	4 043	14 157	23	136
MR	10 662	7 108	3 554	18 764	2	21

4.2　基线模型

为了验证本文模型的有效性，选择以下常见的模型作为基线模型进行比较。

Text-CNN 模型[7]：该模型利用不同的卷积核尺寸并行提取文本的信息，然后通过最大池化突出最重要的关键词实现文本分类。

LSTM 模型[4]：该模型通过引入门控机制可以更

好地处理和捕捉序列数据中的长期依赖关系，并且使用最后一个隐藏状态作为整个文本的表示。

FastText 模型[25]：由 Facebook 提出的一种快速高效的文本分类算法，能够处理大规模文本数据。

LEAM 模型[26]：该模型结合了标签增强和注意力机制且用于文本分类的深度学习模型，通过标签增强技术将标签信息嵌入文本表示中，使其能更好地捕捉标签相关的特征。

Text-GCN 模型[13]：首次将图卷积神经网络应用于文本分类任务，为整个语料库构建一张图，同时学习词嵌入和文本嵌入，并利用多层图卷积操作融合上下文信息。

Tensor-GCN 模型[14]：该模型是一种使用张量分析的图卷积网络模型，其结合了图卷积神经网络和高阶张量分解的思想，通过对张量进行低秩分解，能够有效捕捉图中的高阶关系。

Text-level GCN 模型[15]：该模型为每个输入文本创建 1 个具有全局参数共享的异构图，从而消除了单个输入文本与整个语料库之间的依赖负担，并且支持在线测试。

4.3　实验参数设置

在构建异构图的过程中，卷积层的嵌入大小设置为 200，并将滑动窗口的大小设置为 20。使用 Adam 训练器对模型参数进行优化，学习率初始化为 0.001。在训练过程中使用 Dropout 防止过拟合，Dropout 设为 0.55，L2 正则化设置为 0.001。每轮训练最大次数设为 200，当连续 10 次迭代验证损失没有减少，则停止训练。对于基线模型，使用原论文和复现中的默认参数。对于使用预训练单词嵌入的基线模型，使用 300 维的 Glove 词向量初始化单词嵌入。

4.4　实验结果

通过在 4 个公开数据集上进行对比实验，采用评价指标为准确率，即预测正确的样本数占样本总数的百分比，结果见表 2。实验结果表明，本文模型在分类准确率上均优于其他基线模型。

本文模型与基于深度学习的模型 Text-CNN、LSTM、FastText 以及 LEAM 相比，其分类准确率显著提高，说明将图卷积神经网络应用于文本分类，其图结构有利于处理文本信息，并可以捕获高阶邻域信息，同时提取更丰富的文本结构关联信息，以达到较好的分类效果。与基于 GCN 的模型 Text-GCN、Tensor-GCN 以及 Text-level GCN 对比，本文模型在 4 个数据集上也都取得 1% 左右的提升，说明引入预训练模型 BERT 可以在文本语义上进行有效训练学习，从而获得更为丰富的语义特征。此外，本文还引入双级注意力机制，考虑了不同类型的相邻节点以及相同类型的相邻节点的重要性，使本文模型能够剔除一些噪声信息，从而提取到与节点相关的重要信息。上述实验分析进一步验证了融合图嵌入和 BERT 嵌入的文本分类模型的有效性。

表 2　不同数据集上各个模型的准确率

Tab. 2　Accuracy of various models on different datasets

模型	准确率			
	R8	R52	Ohsumed	MR
Text-CNN	0.951 7	0.875 9	0.584 4	0.777 5
LSTM	0.936 8	0.855 4	0.411 3	0.750 6
FastText	0.961 3	0.928 1	0.577 0	0.751 4
LEAM	0.933 1	0.918 4	0.585 8	0.769 5
Text-GCN	0.970 7	0.935 6	0.683 6	0.767 4
Tensor-GCN	0.980 4	0.950 5	0.701 1	0.779 1
Text-level GCN	0.978 9	0.946 0	0.694 0	0.754 7
本文模型	0.981 1	0.951 5	0.714 2	0.815 8

4.5　消融实验

为了验证本文模型引入的双级注意力机制模块和 BERT 嵌入模块对文本分类效果的有效性，通过消融实验对各个模块在不同数据集上进行分析验证，w/o 表示“不包括”，具体设置如下。

w/o D-att：表示没有双级注意力机制模块，把构建好的异构图直接通过图卷积神经网络训练得到图嵌入，然后再融合 BERT 嵌入。

w/o T-att：表示没有类型级注意力机制，模型只引入节点级注意力机制。

w/o N-att：表示没有节点级注意力机制，模型只有类型级注意力机制。

w/o Bert：表示没有 BERT 模块，模型仅引入双级注意力机制。

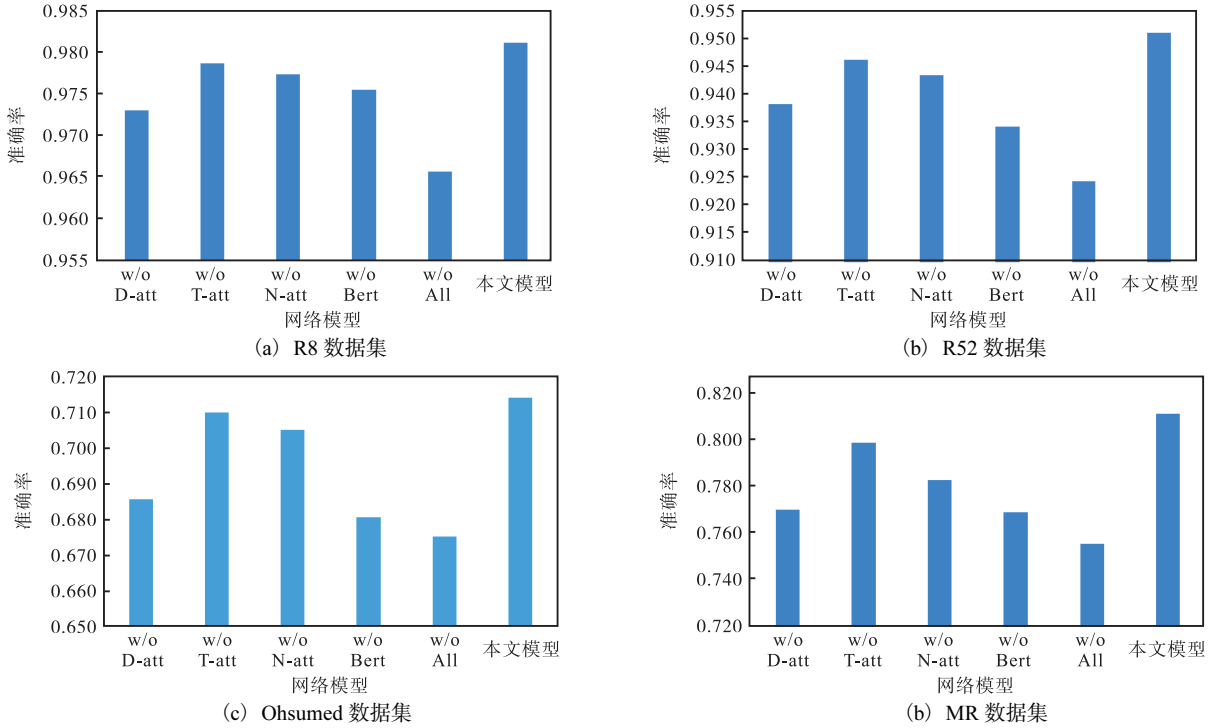
w/o All：表示同时没有 BERT 模块和双级注意力机制模块。

由图 2 可知，模型中删除任一模块都会导致准确率下降。不同模块在不同数据集中起到的作用也不一样。与原模型相比，如果删除双级注意力机制模块，就会导致模型的提取能力变差，引入较多噪声信息，增加无关特征对分类准确率的干扰，因此在每个数据集上分类效果均不太理想。如果删除 Bert 模块，其在数据集 Ohsumed 和 MR 上的准确率下降 3% ~ 4%，分类效果较差。这表明通过 BERT 嵌入得到含有局部信息的嵌入表示可以和含有全局信息的图嵌入

有效融合，对这些过长或者过短的文本数据集的分类准确率的提高起到重要作用。针对不同级别的注意力机制，当去掉任意一种注意力机制时都会使模型的分类效果变差，并且由图 2 可以看出节点级注意力在文本分类任务中起到更为明显的作用。同时删除这两种模块后，模型的分类性能大幅降低，尤其针对 MR 数据集。因此，本文模型中使用的模块都必不可少，可以有效提高分类性能。

4.6　样例分析

测试不同迭代次数下 3 种不同的模型在数据集 Ohsumed 和 R52 上的损失和准确率，实验结果如图 3 所示，Epoch 为迭代次数。

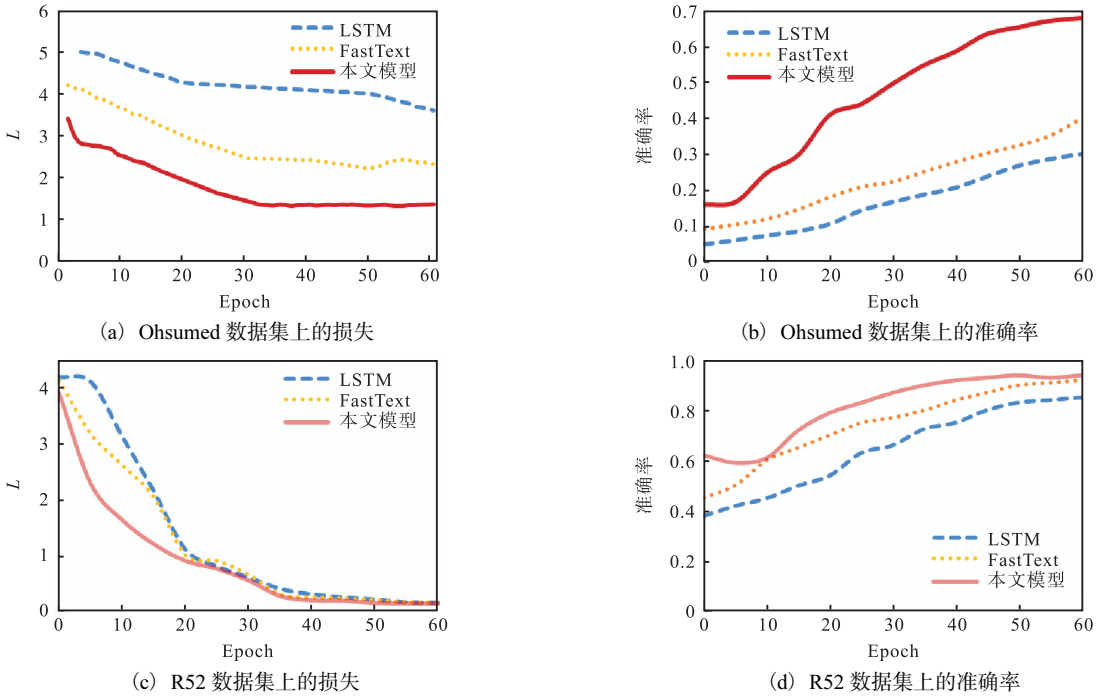


（a）R8 数据集　（b）R52 数据集

（c）Ohsumed 数据集　（b）MR 数据集

图 2　消融实验在不同数据集上的结果

Fig. 2　Results of ablation experiments on different datasets



（a）Ohsumed 数据集上的损失　（b）Ohsumed 数据集上的准确率

（c）R52 数据集上的损失　（d）R52 数据集上的准确率

图 3　不同迭代次数下 3 种模型在数据集 Ohsumed 和 R52 上的损失和准确率

Fig. 3　Loss and accuracy of three models on the Ohsumed and R52 datasets under different iteration times

根据图 3 中的曲线走势可以看出，相较于传统的神经网络模型 LSTM[4]和 FastText[25]，融合图嵌入和 BERT 嵌入的文本分类模型具有更好的收敛性和鲁棒性。由此可以看出，引入双级注意力机制并融合 BERT 嵌入有助于提升文本分类效果。

5 结　语

针对文本分类任务，本文提出了一种融合图嵌入和 BERT 嵌入的文本分类模型，将文本分类问题转化为节点分类问题。该模型首先为整个语料库构建了一个异构图，其次将双级注意力机制融入图卷积神经网络中，并把构建好的异构图传入网络模型中，经训练得到具有全局信息的图嵌入表示；然后再将文本以序列的形式通过预训练模型 BERT 训练，得到带有局部信息的词嵌入，可以有效解决一词多义的问题。随后将 BERT 嵌入表示和图嵌入表示进行融合，得到既包含全局信息又包含局部信息的词嵌入表示。通过文本潜在向量计算出每个单词与文本潜在向量的关联程度并赋予单词注意力权重，最后得到文本嵌入表示，从而实现文本分类。在 4 个广泛使用的公开数据集上的实验结果表明，相比于基线模型，本文模型取得了更好的分类效果。

在应对大规模语料库时，复杂的语义关系会使构建的异构图变得庞大，而采用全批量训练方式意味着需要将所有节点的中间状态存储在内存中，因此会带来严重的内存消耗问题，这是本文模型的不足之处。未来工作可以探索如何处理大型文本数据，减少模型复杂度，同时提升模型的可解释性，从而进一步提高文本分类准确率并减少内存消耗与计算成本。

参考文献：

[1] 车万翔. 自然语言处理[M]. 北京：电子工业出版社，2021：174–215.

[2] KALCHBRENNER N，GREFENSTETTE E，BLUNSOM P. A convolutional neural network for modelling sentences[EB/OL]. [2023–06–01]. http://www.arxiv.org/pdf/1404.2188.pdf.

[3] MIKOLOV T，KARAFIAT M，BURGETL，et al. Recurrent neural network based language model[J]. Interspeech，2010，2(3)：1045–1048.

[4] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural computation，1997，9(8)：1735–1780.

[5] ZHAO W，YE J，YANG M，et al. Investigating capsule networks with dynamic routing for text classification [EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1804.00538.

[6] KIPF T N，WELLING M. Semi-supervised classification with graph convolutional networks [EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1609.02907.

[7] KIM Y. Convolutional neural networks for sentence classification[EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1408.5882.

[8] ZHANG X，ZHAO J，LECUN Y. Character-level convolutional networks for text classification[J]. Advances in neural information processing systems，2015，28：649–657.

[9] LIU P，QIU X，HUANG X. Recurrent neural network for text classification with multi-task learning[EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1605.05101.

[10] BAHDANAU D，CHO K，BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1409.0473.

[11] YANG Z，YANG D，DYER C，et al. Hierarchical attention networks for document classification[C]//ACL. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologics. New York：ACL，2016：1480–1489.

[12] XU B，CEN K T，HUANG J J，et al. A survey on graph convolutional neural network[J]. Chinese journal of computers，2020，43(5)：755–780.

[13] YAO L，MAO C，LUO Y. Graph convolutional networks for text classification[C]//AAAI. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. California：AAAI，2019：7370–7377.

[14] LIU X，YOU X，ZHANG X，et al. Tensor graph convolutional networks for text classification[J]. Proceedings of the AAAI conference on artificial intelligence，2020，34(5)：8409–8416.

[15] HUANG L，MA D，LI S，et al. Text level graph neural network for text classification [EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1910.02356.

[16] GILMER J，SCHOENHOLZ S S，RILEY P F，et al. Neural message passing for quantum chemistry[C]//PMLR. Proceedings of the 34th International Conference

on Machine Learning. New York: PMLR, 2017: 1263–1272.

[17] GAO H, CHEN Y, JI S. Learning graph pooling and hybrid convolutional operations for text representations[C]//ACM. The World Wide Web Conference. New York: ACM, 2019: 2743–2749.

[18] LI H, YAN Y, WANG S, et al. Text classification on heterogeneous information network via enhanced GCN and knowledge[J]. Neural computing and applications, 2023, 35(20): 14911–14927.

[19] LIU B, GUAN W, YANG C, et al. Transformer and graph convolutional network for text classification[J]. International journal of computational intelligence systems, 2023, 16(1): 161.

[20] HU L, YANG T, SHI C, et al. Heterogeneous graph attention networks for semi-supervised short text classification[C]//ACL. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. New York: ACL, 2019: 4821–4830.

[21] 王婷, 朱小飞, 唐顾. 基于知识增强的图卷积神经网络的文本分类[J]. 浙江大学学报(工学版), 2022, 56(2): 322–328.

[22] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2023–06–01]. https://arxiv.org/pdf/1810.04805.pdf.

[23] 李全鑫, 庞俊, 朱峰冉. 结合 Bert 与超图卷积网络的文本分类模型[J]. 计算机工程与应用, 2023, 59(17): 107–115.

[24] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1710.10903.

[25] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1607.01759.

[26] WANG G Y, LI C Y, WANG W L, et al. Joint embedding of words and labels for text classification[EB/OL]. [2023–06–01]. https://doi.org/10.48550/arXiv.1805.04174.

责任编辑: 郎婧

(上接第 63 页)

[12] HAZARIKA D, ZIMMERMANN R, PORIA S. Misa: modality-invariant and -specific representations for multimodal sentiment analysis[C]//ACM. Proceedings of the 28th ACM international conference on multimedia. New York: ACM, 2020: 1122–1131.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2023–10–11]. http://www.aiotlab.org/teaching/intro2ai/slides/10_attention_n_bert.pdf.

[14] ZADEH A, ZELLERS R, PINCUS E, et al. Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages[J]. IEEE Intelligent systems, 2016, 31(6): 82–88.

[15] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42: 335–359.

[16] ZADEH A A B, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]//ACM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. New York: ACM, 2018: 2236–2246.

[17] DEVLIN, JACOB et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//NAACL-HLT. Proceedings of NAACL-HLT. Minneapolis: NAACL-HLT, 2019: 4171–4186.

[18] ZENG J, LIU T, ZHOU J. Tag-assisted multimodal sentiment analysis under uncertain missing modalities [C]//ACM. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 1545–1554.

责任编辑: 郎婧