文章编号:1000-5862(2021)03-0299-06

# 基于知识增强的 ERBERT-GRU 中文图书分类方法研究

刘 磊[1],许 婕[2],周 勇[1*]
(1. 江西师范大学计算机信息工程学院,江西 南昌 330022;2. 江西师范大学图书馆,江西 南昌 330022)

**摘要:** 图书的自动分类是图书管理和图书推荐算法中的基础工作,也是难点之一,而且目前针对中文分类算法主要集中在短文本领域中,鲜有对图书等长文本分类的研究. 该文对深度学习分类算法进行了深入细致的研究,并对 BERT 预训练模型及其变体进行相应的改进. 利用复杂层级网络叠加双向 Transformer 编码器来提取隐藏在文本中的细粒度信息. 在预训练过程中,增加实体级别的遮罩,获得对传统 BERT 模型的改进,提高了模型对中文语义理解的能力. 通过添加外部知识提升了该模型的鲁棒性.

**关键词:** 图书分类; ERBERT-GRU 模型; 神经网络; 深度学习; 知识嵌入
**中图分类号:** TP 311 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2021.03.12

## 0 引言

随着人工智能和大数据等信息技术的兴起,人们开始步入一个信息爆炸的时代,信息爆炸使用户在如何高效快速地选择和过滤出自己所需的信息上成为一个难题. 图书作为信息流的主要载体之一,也在急剧增长. 为了提高图书检索的效率,需要对图书进行细化和分类,同时也为下游任务(如推荐等)做好铺垫. 传统的图书分类主要依靠图书管理员的自身职业素养和使用中国图书馆分类法来进行,不仅成本高、效率低,而且处理周期也较长. 随着基于深度学习的文本分类研究的发展,大量的科研人员开始投入到图书的自动分类研究中. 但是,目前这些研究主要聚焦于英文图书分类,随着中文图书快速增长,如何针对中文图书的特征构建一个相对精确的中文图书自动分类模型就显得格外重要.

近年来,深度学习在文本的特征表示已经取得巨大成功,在该背景下,本文在深入研究传统模型的基础上,提出一种神经网络模型(Enrich BERT-GRU with Knowledge,ERBERT-GRU),用于解决传统的循环神经网络在文本分类中过于关注全局信息和卷积神经网络过于关注局部信息,从而忽略了词在不同的上下文中存在不同语义的问题,通过改进 BERT[1] 预训练模型,以及叠加更加复杂的层级模型,用于提升 BERT 对中文语义的理解能力,通过添加外部知识,在一定程度上避免了模型在图书分类过程中因训练样本不足或干扰文本而导致的分类准确率的下降,也有效地缓解了过度依赖单一数据源来进行分类的问题,提升模型的可扩展性和鲁棒性,从而为中文图书分类的实际应用提供参考.

## 1 相关研究

随着深度学习和神经网络研究的不断深入,以词向量为主要形式的文本特征表示方法开始出现,文本分类效果相较于以往有了大幅提升,如 Y. Kim[2] 提出将卷积神经网络应用于文本分类任务,Liu Pengfei[3] 通过堆叠 2 层 LSTN[4] 和 GRU[5] 模型,解决了 LSTM 神经网络的长距离依赖难题,这些模型的成功应用主要归功于文本特征提取技术的进步.

传统的词向量模型是立足于统计学思想,如 Naive Bayes[6]、支持向量机[7]、$k$-近邻[8] 等统计方法. 基于深度学习的词向量模型 word2vec[9],其核心思想是建立语言模型[10],即词的向量表示由词的上下文确定,但它在学习的过程中忽略了词在全局上

收稿日期:2020-09-16
基金项目:江西省教育厅科学技术研究(KJLD14021)和江西省教育厅省重点教改课题(JXJG1821)资助项目.
通信作者:周 勇(1971—),男,江西南昌人,副研究员,主要从事数据库、数据挖掘和人工智能方面的研究. E-mail: zhou_yong@126.com

的信息;而 J. Pennington 等[11]提出的 Glove 通过使用词的共现矩阵,在保留词的局部信息情况下也考虑了词的全局信息. 但是,word2vec 和 Glove 存在一个共同的问题,那就是它们都忽视了词在特定的上下文中有不同的语义信息. M. E. Peters 等[12]提出的 ElMo 使用一个双向的 LSTM 网络来得到词的上下文表示,在获得预训练词向量的基础上,根据词的特定上下文信息对向量表示进行动态调整. 但是 ELMo 在处理模型的损失时,只是对 2 个模型的损失进行简单的叠加. 与 ELMo 采用 LSTM 模型不同,生成预训练词向量模型 GPT 是采用 Transformer 编码器来进行编码,但是其只是采用了单向的 Transformer 编码器,考虑了词的前半部分信息,而丢失了后半部分信息; A. Vaswani 等[13]针对该问题提出了 BERT 模型,BERT 模型与上述模型不同之处在于其利用双向的 Tranformer 作为编码器,充分考虑词的前后方向上的信息,这也是本文选择 BERT 用于图书分类编码器的原因.

尽管 BERT 模型、卷积神经网络和循环神经网络在文本分类上已经取得了一定进展. 但是,BERT 模型主要关注在词粒度的完形填空学习上,并没有充分利用数据中的词法以及语义信息;卷积神经网络通过卷积操作可以提取文本的局部特征,而循环神经网络则通过神经元能够获取文本的全局特征. 但是,卷积神经网络忽略了文本的上下文语义信息,循环神经网络则对局部语义信息不敏感 .

对于图书分类这种具有层级关系的任务来说,上述方法的精确度和召回率将随着分类类别的增加而显著下降,而对于 BERT 模型来说,在数据集训练不充分和存在添加干扰文本的情况下,其表现会非常差,本文针对这些不足提出一种 ERBERT-GRU 模型用于图书分类任务,通过大量实验验证了该模型的有效性.

## 2 数据集和任务

由于当前并没有公开的中文图书数据集,故利用爬虫程序爬取豆瓣读书数据作为数据源,共爬取了约 12 万本图书,去除一些无用数据(如 ISBN 为空、图书摘要为英文等),总计 117 865 本书. 采用 $k$-折交叉验证法把数据划分为训练集、验证集和测试集,其中 $k=10$,训练集和验证集占 80%,测试集占 20%. 每本图书包含以下信息:(i)图书标题;(ii)作者姓名(一本书可能对应多个作者,平均每本书的作者人数为 1.14,作者最多的一本书有 10 人,74% 的书只有 1 个作者);(iii)图书的摘要信息(简要介绍了图书的内容);(iv)作者简介信息(简要介绍作者个人风格和作品信息);(v)出版社信息(该书的出版机构);(vi)ISBN(图书的唯一标识符);(vii)图书的出版日期.

每本书由 1 级标签和 2 级标签共同组成,在豆瓣读书中共有 6 个 1 级标签. 6 个 1 级标签分别为文学、流行、文化、生活、经管、科技. 1 级标签下又存在若干个更加具体的 2 级标签(约为 120 个). 鉴于 2 级标签类型过多,故决定把实验分为 2 个部分进行,对 2 个标签分别进行实验,但是从本质上来说,它们都属于多标签分类的范畴.

## 3 研究方法

### 3.1 图书元数据

除了数据集的信息以外,在数据处理过程中提取了每本图书的元数据,元数据包含如下信息:(i)作者人数;(ii)作者是否为从事学术和科研的人员(若是,则状态为 1,否则为 0);(iii)图书标题词的长度;(iv)图书摘要词的长度;(v)作者简介词的长度;(vi)图书摘要词的最大长度;(vii)作者简介中词的最大长度;(viii)图书摘要的平均长度;(ix)作者简介的平均长度;(x)图书摘要长度的中位数;(xi)作者简介长度的中位数;(xii)出版日期距离当前的年数.

在实验过程中发现,标题的平均长度为 3.68,而漫画类型的标题平均长度为 4.33,悬疑类的标题长度为 3.22. 图书摘要的平均长度为 143.56,诗歌类型的摘要长度为 151.37,而武侠类型的摘要长度为 129.70. 通过添加这些细粒度的信息将会使得类器对细粒度的特征更加敏感,从而提高分类的效果.

### 3.2 作者嵌入

虽然不应武断地按作者来判断一本图书的风格和类型,但是与作者相关的信息(比如出生年月、国籍等特征信息)可以用来辅助支持图书分类,而且大部分作者会坚持自己的写作方向,而这些特征和规律有助于图书分类.

图 1 是作者信息的矢量表示,从而通过图的共现矩阵可以计算出 2 个作者之间的距离,该距离表示作者之间的相似度,也可以解释为相应作者之间写作风格的相似性.

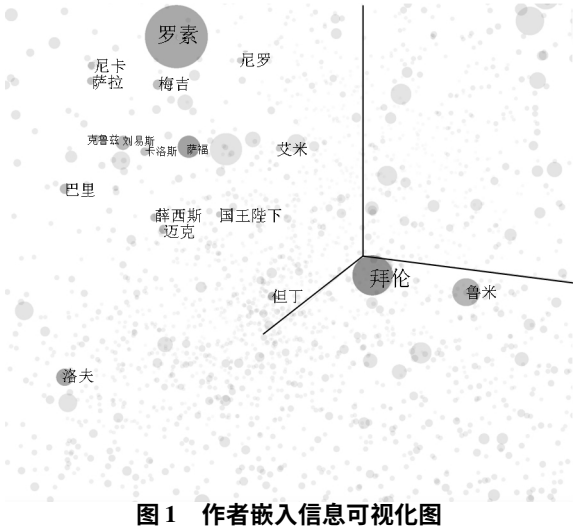


图 1　作者嵌入信息可视化图

### 3.3　ERBERT-GRU 模

ERBERT-GRU 模型的结构图如图 2 所示,由词向量嵌入层、多连接双向的 BERT 层、更新门、遗忘门、输出层等组成.

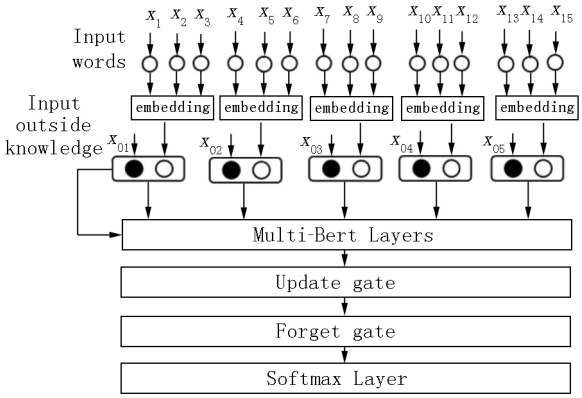


图 2　ERBERT-GRU 模型

3.3.1　BERT 网络　相关研究表明,BERT 模型(见图 3)通过双向的基于注意力机制的 Transformer 在词向量嵌入上取得较好效果. 为了更好地学习中文词的向量表示,本文在 BERT 训练中增加了实体级别的遮罩,以提升模型通过词的上下文来学习词的语义特征的能力. 其结构本质上是 Encoder-Decoder 结构,内部核心功能为 Self-Attention 操作,其计算过程如图 4 所示.

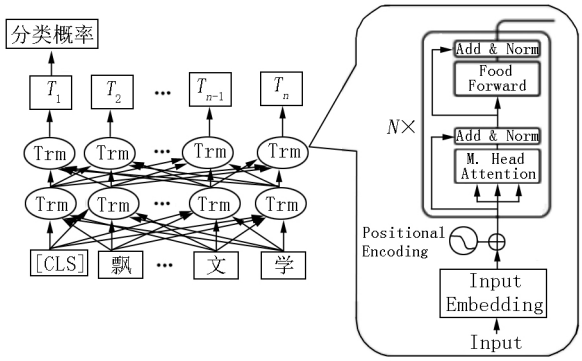


图 3　BERT 网络内部结构图

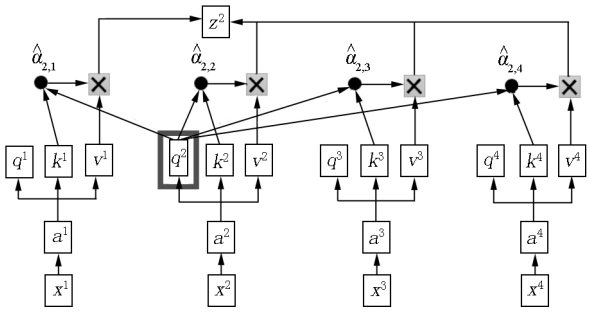

图 4　self-attention 工作流程

首先,根据 $a^i = Wx^i$ 把输入单词乘以矩阵转化为嵌入向量,根据词嵌入向量可得 $q^i = W^q a^i$,$k^i = W^k a^i$,$v^i = W^v a^i$,然后计算 Scaled Dot-Product Attention:

$$a_{1,i} = q^1 g k^i / \sqrt{d}.$$

经过 softmax 函数计算得

$$y_t = \sigma(W^{(r)} x_t + U^{(r)} h_{t-1}). \quad (1)$$

对式(1)点乘 Value 值 $v$,相加之后得到最终输出结果为

$$z^1 = \sum_i \hat{a}_{1,i} v^i.$$

3.3.2　GRU 神经网络层　循环神经网络[14]具有一定的记忆能力,当序列的距离较大时,则无记忆能力. 而 GRU 能够解决长距离依赖问题,且提供更快的执行速度以及更少的神经元数量,其结构图如图 5 所示.

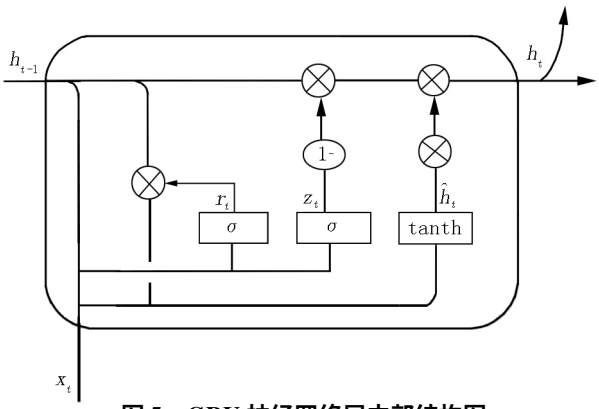


图 5　GRU 神经网络层内部结构图

在 $t$ 时刻,通过

$$z_t = \sigma(W^{(z)} x_t + U^{(z)} h_{t-1})$$

得到更新门的输出,其中 $x_t$ 为 $t$ 时刻的输入向量,$h_{t-1}$ 保存 $t-1$ 时刻信息.

$$r_t = \sigma(W^{(r)} x_t + U^{(r)} h_{t-1}). \quad (2)$$

信息过滤主要由重置门完成,其与重置门的更新式(2)类似,经过一个线性变换后相加,最后经过 Sigmoid 激活函数后输出. 如 $\hat{h}_t = \tanh(Wx_t + r_t e U_{t-1})$ 所示,在重置门中,重置门通过上一时刻的

信息来决定当前时刻的记忆内容. 输入 $x_t$ 与上一步的信息 $h_{t-1}$ 经过一个线性变换,即分别右乘矩阵 $W$ 和 $U$.

最后,当前时刻 $t$ 的信息 $h_t$ 将其记忆的信息传递给下一个单元,并由重置门完成控制记忆数据的删除与输入数据的更新,公式为

$$h_t = z_t e h_{t-1} + (1 - z_t)\, e\hat{h}_t.$$

3.3.3 层级连接的 ERBERT-GRU 网络 如图 6 所示,为了应对图书分类过程中的层级关系而导致的分类性能下降问题,设计了一种具有层次关系的网络模型,输入数据经过词向量嵌入,作为多路 BERT 网络输入,数据经过多路 BERT 网络得到图书的 1 级标签,而 GRU 网络在获知上级网络的输出后,将选取上级网络输出的 1 级标签下的所有数据进行训练.

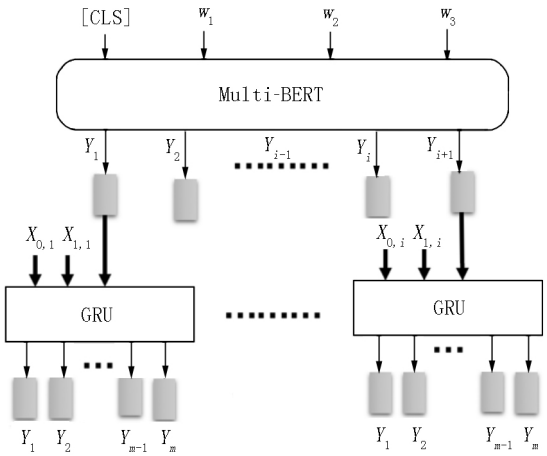


图 6 BERT-GRU 层级网络

# 4 实验

## 4.1 实验环境设置

本文实验环境为 Python 3.7.6,深度学习框架为 Pytorch 1.3.0.

本文选择卷积神经网络、循环神经网络、BERT 预训练模型作为基准对照模型,模型初始学习率为 $2 \times 10^{-5}$,batch-size 大小为 24,最大文本长度为 400,Epoch 为 20,Adam [15] 为优化器.

## 4.2 指标分析及优化

4.2.1 评估指标 评估指标主要为准确率、召回率、$F_1$,但由于模型是一个多分类任务,依据前人的经验和相关理论,本文选择微平均作为评价指标.

对于宏平均来说,首先计算每个混淆矩阵的准确率和召回率,然后求平均值:

$$P_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n} P_i, R_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n} R_i. \quad (3)$$

根据式(3)计算结果可以计算宏 $F_1$ 的值:

$$F_{\text{macro}} = 2P_{\text{macro}}R_{\text{macro}}/(P_{\text{macro}} + R_{\text{macro}}).$$

微平均首先计算每个混淆矩阵的平均 $T_P$(真正例)、$F_P$(假正例)、$T_N$(真负例)、$F_N$(假负例),然后求其微准确率和微召回率:

$$P_{\text{macro}} = \overline{T}_P/(\overline{T}_P + \overline{F}_P) = \sum_{i=1}^{n} T_{P_i}/(\sum_{i=1}^{n} T_{P_i} + \sum_{i=1}^{n} F_{P_i}),$$

$$R_{\text{macro}} = \overline{T}_P/(\overline{T}_P + \overline{F}_N) = \sum_{i=1}^{n} T_{P_i}/(\sum_{i=1}^{n} T_{P_i} + \sum_{i=1}^{n} F_{N_i}).$$

通过上述计算结果可求解微 $F_1$ 值为

$$F_{\text{macro}} = 2P_{\text{macro}}R_{\text{macro}}/(P_{\text{macro}} + R_{\text{macro}}).$$

4.2.2 优化器 一个合适的优化器对网络来说同样至关重要,本文选择 Adam 优化器来计算梯度,Adam 结合了 AdaGrad 和 RMRrop 的优点,同时考虑了梯度的 1 阶矩估计和 2 阶矩估计,具有实现简单、计算速度快、对资源消耗较小、能自动地调整学习率等优点而得到广泛应用,方法如下:

(i) 利用 $g_t = \nabla_\theta J(\theta_{t-1})$ 获取 $t$ 时刻的随机目标的梯度;

(ii) 通过 $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$ 更新 1 阶矩估计;

(iii) 通过 $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ 对 2 阶矩估计进行动态调整;

(iv) 通过 $m_t = m_t/(1 - \beta_1^t)$ 计算 1 阶矩估计的偏差校正;

(v) 通过 $v_t = v_t/(1 - \beta_2^t)$ 计算 2 阶矩估计的偏差校正;

(vi) 通过 $\theta_t = \theta_{t-1} - \alpha m_t/(\sqrt{v_t} + \varepsilon)$ 更新模型的参数.

## 4.3 实验结果及分析

4.3.1 整体表现 实验结果如表 1 所示. 在测试数据集的 1 级标签与 CNN 模型、RNN 模型,以及 Bert 模型等相对比,本文的模型在 1 级标签的分类上取得最好的效果. 在 1 级标签上的分类上,本文提出的模型相对基准模型也提高了 4.80%、3.71%、2.26%.

表 1　1 级标签试验结果对比表　　%

| 模型 | $F_1$ | 准确率 | 召回率 |
|---|---|---|---|
| CNN | 82.54 | 84.83 | 78.75 |
| RNN | 83.63 | 85.59 | 78.36 |
| Bert | 85.08 | 87.75 | 77.47 |
| ERBERT-GRU | 87.34 | 90.16 | 74.61 |

在 120 个 2 级标签的分类上,实验结果如表 2 所示. 本文提出的模型相较于 CNN、RNN、BERT 模型在测试集上表现更加优异,在分类指标上分别取得了 69.30%、83.34%、62.12% 的成绩,这也在一定程度上说明了像 CNN、RNN 这种网络架构并不适合对层级关系的任务进行分类,在图书分类层级并不是很深的情况下,其表现就不佳,在层级更深的任务中表现会更差,本文提出的模型有效地解决了这一问题. 对于 BERT 模型来说,在训练数据中存在干扰文本以及训练数据不足导致模型理解能力下降都会影响分类效果,而本文提出的模型可以避免上述 2 种情况的发生.

表 2　2 级标签实验结果对比图　　%

| 模型 | $F_1$ | 准确率 | 召回率 |
|---|---|---|---|
| CNN | 61.78 | 72.64 | 70.34 |
| RNN | 62.26 | 73.53 | 70.21 |
| Bert | 64.96 | 78.03 | 67.39 |
| ERBERT-GRU | 69.30 | 83.34 | 62.12 |

4.3.2　**最大单词输入长度对模型的影响**　为了研究在训练过程中输入单词的长度对于图书分类准确率的影响,本文进行了相关了实验分析.

实验结果如图 7 所示. 当输入单词长度为 400 时模型获得最好表现. 当单词长度较小时,截断了部分信息,但是由于本文的模型进行知识增强,从而弥补了信息缺失造成的性能下降; 当单词最大长度较大时,大量数据需通过 padding 填充进行对齐,这得益于外部知识填充抵消了这部分对模型可能造成的影响,提升了模型的鲁棒性.

4.3.3　**epoch 和 batch-size 对模型的影响**　为了研究 batch-size 和 epoch 对模型准确率的影响,本文实验在其他条件不变的情况下,改变 batch-size 和 epoch 的值,观察 ERBERT-GRU 网络在数据集上的表现,试验结果如图 8 所示.

由图 8 知,batch-size 值越大,模型训练所需时间越少,模型训练时间减少的同时,当 batch-size 值大于 24 时,模型的表现开始下降.

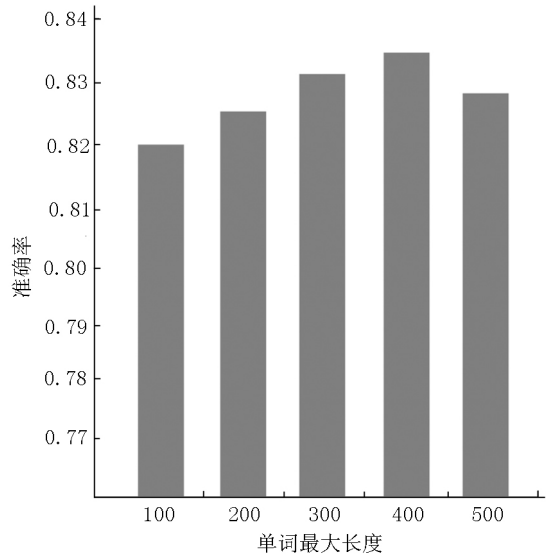


图 7　最大输入长度对模型影响

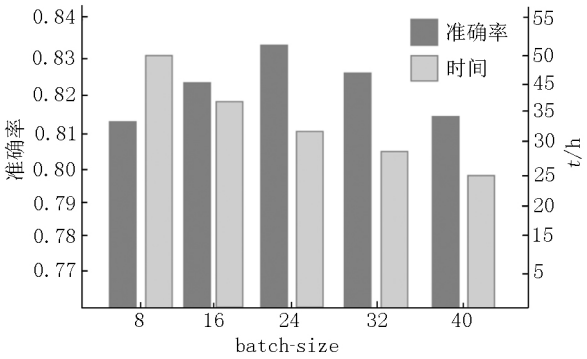


图 8　batch-size 对模型的影响

由图 9 知,epoch 越大,训练时间越长,预测结果越好,但是当值大于 20 时,模型表现开始变差,当值为 20 时取得最好结果.

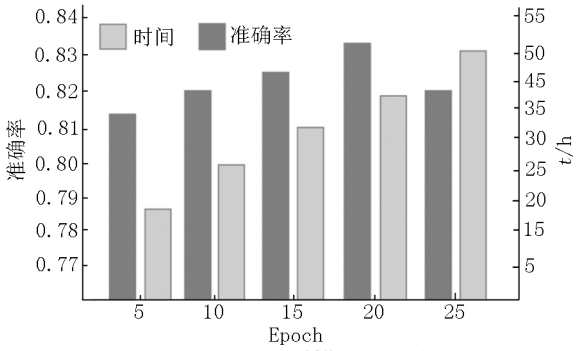


图 9　Epoch 对模型的影响

## 5　结论

在图书分类任务中,特征提取和语义表示对一个图书分类来说至关重要. 本文提出了一种新的 ERBERT-GRU 模型用于图书分类,将文本映射为向量和文本的细粒度特征作为网络输入,然后基于 Transform 模型的 BERT 来对语义关系进行建模,通

过层次连接的 ERBERT-GRU 网络,从而较好地学习了输入数据的低维和高维语义特征以及标签之间的层级关系,通过添加外部知识,弥补了在训练过程中因数据稀疏而导致的分类性能下降问题. 经过实验验证,本文提出的 ERBERT-GRU 网络在中文图书分类领域中十分有效.

## 6 参考文献


[1] Devlin J,Chang Mingwei,Lee K,et al. BERT: pre-training of deep bidirectional transformers for language understanding [J]. Computation and Language,2018,26(1):4171-4186.

[2] Kim Y. Convolutional neural networks for sentence classification [J]. Eprint Arxiv,2014,14(10):1746-1751.

[3] Liu Pengfei,Qiu Xipeng,Huang Xuanjing. Recurrent neural network for text classification with multi-task learning [EB/OL]. [2016-06-17]. https://arxiv.org/abs/1605.05101.

[4] Malhotra P,Vig L,Shroff G,et al. Long short term memory networks for anomaly detection in time series [EB/OL]. [2019-01-19]. https://www.researchgate.net/publication/304782562_Long_Short_Term_Memory_Networks_for_Anomaly_Detection_in_Time_Series.

[5] Cho K,Van M B,Gulcehre C,et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. [2019-01-19]. https://arxiv.org/abs/1406.1078v3.

[6] Chen Zhenguo,Shi Guang,Wang Xiaoju. Text classification based on naive Bayes algorithm with feature selection [J]. International Journal on Information,2012,24(10):4255-4260.

[7] Joachims T. Text categorization with support vector machines: learning with many relevant features [C]. Berlin: Springer,1998:127-142.

[8] Vries A D,Mamoulis N,Nes N,et al. Efficient KNN searchonverticallydecomposeddata [C]. Madiso: ACM Press,2002:322-333.

[9] Mikolov T. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems,2013,2(26):3111-3119.

[10] Bengio Y,Ducharme R,Vincent P,et al. A neural probabilistic language model [J]. Journal of Machine Learning Research,2003,3(2):1137-1155.

[11] Pennington J,Socher R,Manning C. Glove: global vectors for word representation [C]. Doha: Association for Computational Linguistics,2014:1532-1543.

[12] Peters M E,Neumann M,Iyyer M,et al. Deep contextualized word representations [EB/OL]. [2019-01-19]. https://www.aclweb.org/anthology/N18-1202.pdf.

[13] Vaswani A,Shazeer N,Parmar N,et al. Attention is all you need [EB/OL]. [2019-01-19]. https://arxiv.org/abs/1706.03762v5.

[14] Mikolov T,Martin Karafiát,Burget L,et al. Recurrent neural network based language model [C]. Chiba: DBLP,2015:1045-1048.

[15] Kingma D P,Ba J. Adam: a method for stochastic optimization [EB/OL]. [2019-01-19]. https://arxiv.org/abs/1412.6980v9.


# The Study on ERBERT-GRU Chinese Book Classification Method Based on Knowledge Enhancement


LIU Lei[1],XU Jie[2],ZHOU Yong[1*]

(1. College of Computer and Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China;
2. Libray,Jiangxi Normal University,Nanchang Jiangxi 330022,China)



**Abstract:** The automatic classification of books is the basic work in book management and book recommendation algorithms,and it is also one of the difficulties. At present,the Chinese classification algorithms are mainly concentrated in the field of short texts,and there are few studies on the classification of long texts such as books. The in-depth and detailed study on deep learning classification algorithms is conducted,mainly studying the BERT pre-training model and its variants and making corresponding improvements. A complex hierarchical network is used to superimpose a two-way transformer encoder to extract the fine-grained information hidden in the text. By adding an entity-level mask in the pre-training process,the traditional BERT model is improved,and the model is improved in Chinese semantic comprehension. The ability to understand semantics improves the robustness of the model by adding external knowledge.



**Key words:** book classification; ERBERT-GRU model; neural network; deep learning; knowledge embedding

(责任编辑:冉小晓)