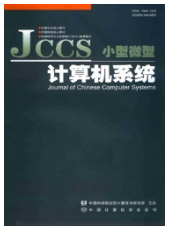

小型微型计算机系统
Journal of Chinese Computer Systems
ISSN 1000-1220,CN 21-1106/TP

《小型微型计算机系统》网络首发论文

题目：融合 BERT 词嵌入和注意力机制的中文文本分类
作者：孙红，陈强越
收稿日期：2020-10-12
网络首发日期：2021-01-06
引用格式：孙红，陈强越．融合 BERT 词嵌入和注意力机制的中文文本分类．小型微型计算机系统. https://kns.cnki.net/kcms/detail/21.1106.TP.20210106.1146.008.html

小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems

融合 BERT 词嵌入和注意力机制的中文文本分类

孙 红，陈强越

（上海理工大学 光电信息与计算机工程学院，上海 200093）

E-mail ： yueyuhai430@163.com

摘 要：文本分类是自然语言处理的一个重要领域。近年来，深度学习的方法被广泛应用于文本分类任务中。在处理大规模的数据时,为了兼顾分类的精度和处理效率,本文使用 BERT 训练词向量作为嵌入层,进一步优化输入语句的词向量，然后用双层的 GRU 网络作为主体网络，充分提取文本的上下文特征，最后使用注意力机制，将目标语句重点突出，进行文本分类。实验证明，BERT 作为嵌入层输入时，有效优化了词向量。同时，文本提出的 BBGA 模型具有高效的处理能力，在处理 THUCNews 数据集时，达到了 94.34%的精确度，比 TextCNN 高出 5.20%，比 BERT_RNN 高出 1.01%。

关键词：文本分类；自然语言处理；BERT；深度学习

中图分类号：TP391 **文献标识码：**A

Chinese text classification based on BERT and attention

SUN Hong，CHEN Qiang-yue

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Text classification is a classic problem in natural language processing. In recent years, deep learning models have drawn a lot of attention in such fields. When it comes to large amounts of data, it is important that we take both accuracy and efficiency into account, so we use the BERT pre-trained language model to obtain word vector, and which is later the embedding layer used to further optimize the word vector of the input sentence. Then the Bidirectional GRU network, which is proved to be efficient in processing text, is utilized to extract the contextual feature of the text. Finally, the attention mechanism is applied to highlight the target sentence and perform text classification. Experiments show that when the BERT pre-trained language model is used as the input of the embedding layer, it effectively optimizes the word vector. At the same time, the method used in the text has reached an accuracy of 94.34% when processing the THUCNews data set, which is 5.19% higher than TextCNN and 1.01% higher than BERT_RNN.

Key words: text classification; natural language processing; BERT; deep learning

1 引言

随着信息技术的不断发展，互联网已经成为信息传递的主要媒介。如今，每天都会产生难以估量的文本信息，对这些信息进行分类，既方便运营商的管理，又使得普通大众能够有选择的阅读自己感兴趣的内容。如何对这些文本进行快速高效的分类，是文本分类研究的热点问题。

由于文本数量庞大，人工进行文本分类显然不可取。随着机器学习和深度学习的发展与应用，相关方法被越来越多的应用到文本分类中。

Kalchbrenner[1]提出了动态卷积神经网络模型(DCNN)，使用 k-max 池化，并与卷积层交替的结构。Kim[2]提出了一个基于卷积神经网络(CNN)的文本分类模型，该模型通过 word2vec 获得输入的词向量，并将此作为卷积神经网络的输入。Johnson[3]等提出了一种新型的 CNN 结构 DPCNN，可以有效提取文本中的远程关系特征，同时避免了复杂度的堆砌。Hochreiter[4]等基于循环神经网络[5](RNN)，提出了长短时记忆网络(LSTM)，弥补了传统循环神经网络梯度消失或梯度爆炸的问题。Chung[6]等在此基础上，提出 GRU 模型，提高了训练的效率。Joulin[7]等人提出了一种简单高效的文本分类器 FastText, FastText 将文本视为词袋，并使用 n-gram 作为附加功能来捕获词序信息。Devlin[8]等人提出了双向编码模型 BERT(Bidirectional Encoder Represent- ation from Transformers)，在大量语料训练基础上，同时考虑了词语在不同上下文的特殊表达，形成动态词向量的输出。杨[9]等人提出多特征融合的模型，从不同层面对文本特征进行提取。

收稿日期：2020-10-12 收修改稿日期：2020-12-02 基金项目：国家自然科学基金项目 (61472256, 61170277, 61703277) 资助; 上海市教委科研创新重点项目 (12zz137) 资助; 沪江基金项目 (C14002) 资助。作者简介：**孙 红，**女，1964 年生，博士，副教授，CCF 会员,研究方向为大数据与云计算、控制科学与工程、模式识别与智能系统；**陈强越，**男，1997 年生，硕士研究生，研究方向为自然语言处理.

上述方法对文本分类的贡献大多体现在优化词向量或者优化特征表达其中一方面，兼顾两者的方法大多结构复杂，分类耗时长。为了对文本进行分类，我们既要考虑合适的文本向量表达，又要精准地提取出文本的主要特征。为此，本文提出了一种用 BERT 训练词向量，用双向 GRU 网络进行高效的特征提取，同时融合注意力机制作为辅助特征嵌入的文本分类模型 BBGA(BERT based Bidirectional GRU with Attention)。实验证明，该方法能够相对快速并且精确地进行较大规模的文本分类任务。

2 相关工作

2.1 文本分类简述

文本分类是自然语言处理的经典问题之一，其主要目的是为目标语句分配标签。随着互联网的发展，文本的规模呈指数级上涨，自动文本分类逐步成为主流方法。自动文本分类方法可分为三类[10]：基于规则、基于机器学习和深度学习以及混合方法。基于规则的方法使用预先定义的各种规则来进行文本分类，例如“体育”这一类别会把所有包含“足球”、“篮球”或“排球”之类词语的文本纳入其中。基于规则的方法需要对待分类文本所属的领域有着深入的了解，这就抬高了这种方法的门槛。

近年来，机器学习，尤其是深度学习相关的方法开始在文本分类中流行起来。深度学习模型在学习文本特征时，能够发现一些难以定义的隐藏规则或模式。这类方法通常包含两个主要的步骤：一是构造合适的词向量来表示任务中输入的文本；二是选择合适的模型来训练提取文本特征，并通过这些特征进行文本分类。

2.2 词向量

词向量，顾名思义，就是将输入的单词或者词语用向量来表示。One-hot 向量是最简单的词向量表示方法，这种向量的维度和词库的大小相等，通过在不同位置设定值 1，其余位置设定 0 来达到唯一表示的目的。这种方法虽然简单，但是会导致维度灾难，而且无法表示一词多义，也无法表示出词与词之间的联系。

Word2vec 通过词的上下文来得到目标词的向量表达，主要方法包括 CBOW 和 Skip-gram，前者通过周围的词来预测中间词，后者则通过中间词来预测周围的词。

然而 word2vec 的表达与窗口大小密切相关，因为它只考虑窗口内的局部联系。为此，glove 利用共现矩阵，把局部信息和整体内容都加以考虑。

尽管 word2vec 和 glove 在某种程度上提升了词向量的表达效果，但是它们都无法表示一词多义，这两个模型的词在不同语境中得到的向量是相同的。为了优化这一点，ElMo 采用双向模型来预测单词。在正向模型中，使用前 1-k 个词去预测第 k 个词，在反向模型中，使用 k 后面的单词来预测第 k 个词。

BERT 通过海量语料的训练，得到了一组适用性十分广泛的词向量，同时还能在具体任务中动态优化词向量，大幅提升了相关 NLP 任务的实验效果。

2.3 深度文本分类模型

使用深度学习方法进行文本分类已经是现在的主流方法，后来的深度学习模型大多是对一些主流模型(例如 RNN 和 CNN)的优化，从而使其更适用于某一类的任务。Tai[11]等人发现传统的链式 LSTM 在 NLP 任务上的效果有限，提出了 Tree-LSTM 模型，从而学习到更加丰富的语义表示。Bieng[12]等人提出了 TopicRNN 模型，把潜在主题模型与 RNN 相结合，使用 RNN 获取局部的联系，同时使用主题模型获取全局的联系。Prusa[13]为了减少字符级文本学习的耗时，使用 CNN 进行文本的编码，可以在原始文本中更多的保留信息。Conneau[14]使用深层的 CNN(VDCNN)进行文本分类，随着层数的不断增加，分类的效果会更好。与 VDCNN 类似，现有模型大多通过复杂度的增加来提高性能。

通过对已有方法的深入学习，也为了达到快速准确地进行文本分类的目标，本文提出了 BBGA 模型。该模型综合考虑了各种词向量表达的优劣，选择使用 BERT 模型训练输入文本的词向量，得到新的词向量之后，将其作为新的输入送到 GRU 网络进行文本特征的提取，同时考虑到文本上下文的联系，将 GRU 扩展为双向网络。最后引入注意力机制，使分类过程中各文本的权重分配更加合理。

3 系统框架

BBGA 模型结构如图 1 所示，该模型的整体运作流程是：首先输入文本数据，利用 BERT 预训练模型，获得包含文本总体信息的动态词向量，接着将新的词向量输入到双层的 GRU 网络进行特征提取，捕捉文本的特征信息，最后引入注意力机制，得到输入文本的最终概率表达，从而达到文本分类的目的。

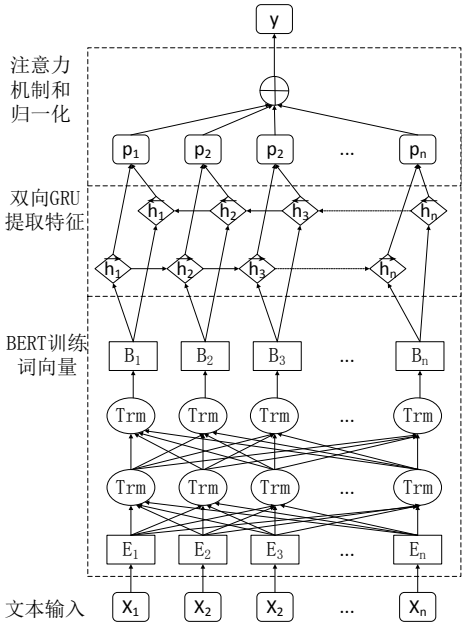

图 1 BBGA 模型结构图

Fig.1 Model structure of BBGA

3.1 BERT 词嵌入

BERT 作为一种预训练语言模型，是在海量语料中训练得到的，既可以直接进行文本分类任务的训练，也可以将其作为词向量嵌入层输入到其他训练模型，本文选择后者。BERT 模型的结构图如图 2 所示：

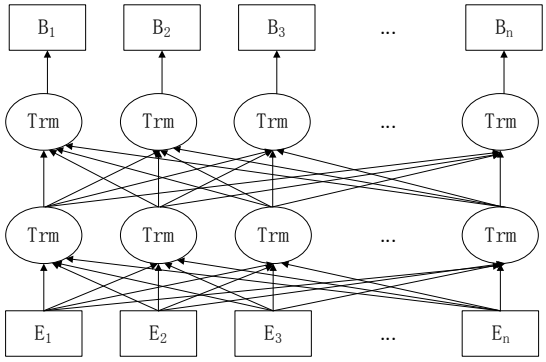

图 2 BERT 结构图

Fig.2 Model of BERT

输入的向量不仅包含当前文本的词向量，还有表示词在文本中位置的位置向量，以及词所在句子的分段向量。三个向量求和之后，分别加上 CLS 和 SEP 作为一个文本开头与结尾的标志。

3.2 循环神经网络与 BiGRU

为了综合考虑文本上下文之间的联系，开始出现一些尝试使用 RNN 处理文本数据的方法。基于 RNN 的模型将文本处理成一个较长的序列，但是，在更新网络参数的反向传播过程中，容易出现梯度弥散的问题。为了克服 RNN 存在的缺点，许多基于 RNN 的变体模型被提出。其中长短时记忆模型(LSTM)是 RNN 众多变体中最为流行的一个，它由输入门、记忆单元、遗忘门和输出门四个主要部分组成，通过对输入向量的“记忆”与“遗忘”，保留了文本中的重要特征，剔除了相对无用的内容。

但是，随着文本数量的增多，由于参数多、各个门之间的计算相对复杂，LSTM 进行网络训练耗时长的问题逐渐暴露。同时，LSTM 还会产生过拟合的现象。为了解决 LSTM 存在的弊端，一种更为简单的基于 RNN 的神经网络模型 GRU 被提出，其模型结构如图 3 所示：

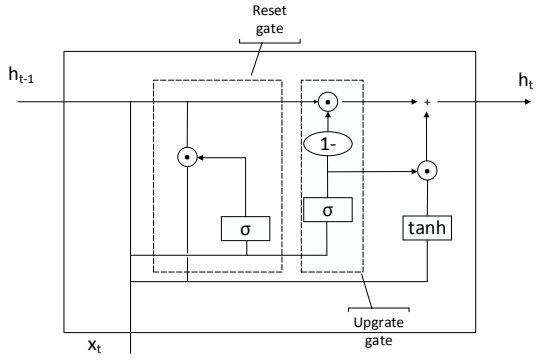

图 3 GRU 结构图

Fig.3 Model of GRU

GRU 主要由更新门和重置门组成，其中更新门用于决策上一时刻隐层状态对当前层的影响，更新门中的值越大，说明上一时刻对当前层的影响越大。更新门的计算方法如公式(1)所示:

$$r_t = \sigma(W_r x_t + W_r h_{t-1} + b_r) \tag{1}$$

重置门用于剔除上一时刻的无效信息，重置门的值越小，剔除的无效信息就越多。重置门的计算方法如公式(2)所示：

$$z_t = \sigma(W_z x_t + W_z h_{t-1} + b_z) \tag{2}$$

当前状态的计算方法如公式(3)和公式(4)所示：

$$\widetilde{h_t} = tanh(W_h x_t + W_h(r_t * h_{t-1}) + b_h) \tag{3}$$

$$h_t = z_t * \widetilde{h_t} + h_{t-1}(1 - z_t) \tag{4}$$

其中h_t, h_{t-1} 分别表示当前时刻和上一时刻的隐层状态，$\widetilde{h_t}$是候选激活状态，x_t表示当前时刻的输入，r_t和z_t分别是更新门和重置门的计算结果，W_r是权重矩阵，b为偏置量。

通过 GRU，我们既能很好的捕获文本的总体特征，又能减少计算量，达到高效训练的效果。但是，在单层的 GRU 网络中，状态的传播也是单向的，为了充分利用文本上下文的关系，我们建立了双向的 GRU 网络，双向 GRU 网络的结构模型如图 4 所示：

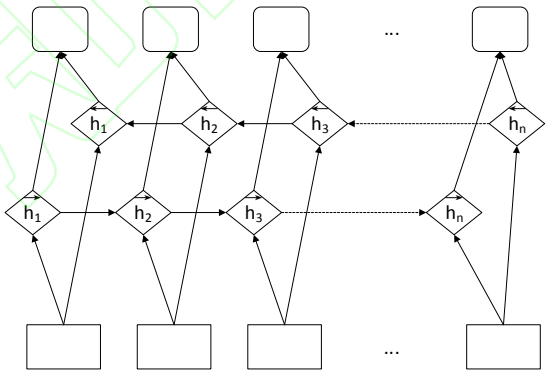

图 4 双向 GRU 结构图

Fig.4 Model of Bidirectional GRU

可以看出，当前时刻的隐层状态h_t不仅与上一时刻的正向隐层输出$\overrightarrow{h_{t-1}}$有关，还和上一时刻反向的隐层输出$\overleftarrow{h_{t-1}}$有关。双向 GRU 的更新公式如式(5)所示：

$$h_t = \overrightarrow{W_t h_t} + \overleftarrow{W_t h_t} + b_t \tag{5}$$

其中$\overrightarrow{W_t}$和$\overleftarrow{W_t}$分别表示正向和反向传播时的权重矩阵，b_t为偏置量。通过正反两个方向的训练与叠加，这样得到的每个新的向量，都包含整个文本的信息，在最后进行分类时，会得到更好的效果。

3.3 注意力机制

注意力机制是深度学习中极为重要的核心技术之一，已经被广泛地应用在模式识别和自然语言处理相关领域。注意力机制模仿了人类视觉的注意力，人类在观察事物时，总是会把注意力重点放在关键区域，从而获得视野内的重要信息，同时忽略非重要区域，提高我们视觉信息处理的效率。

对于本文的文本分类任务而言，每条文本数据的类别，通常由其中的关键词和关键语句决定，引入注意力机制的目

的在于提升分类过程中这些关键词的权重，同时降低非关键词的权重，从而获得更好的分类效果。

4 实验结果与分析

4.1 实验数据集和评价指标

本文使用的数据集是清华 NLP 组提供的 THUCNews 新文文本分类的数据集，从中抽取了 10 万条新闻数据，包含经济、房产、股票、教育、科学、社会、时政、体育、游戏和娱乐，数据分布情况如表 1 所示。同时，将数据集按照 8:1:1 划分训练集、测试集和验证集。其中，训练集八万条，测试集和验证集各 1 万条。

表 1 THUCNews 实验数据集分布情况

Table 1 Distribution of THUCNews data set

数据类别	训练集	测试集	验证集
经济	8000	1000	1000
房产	8000	1000	1000
股票	8000	1000	1000
教育	8000	1000	1000
科学	8000	1000	1000
社会	8000	1000	1000
时政	8000	1000	1000
体育	8000	1000	1000
游戏	8000	1000	1000
娱乐	8000	1000	1000

本文分别采用准确率(Precision)、召回率(Recall) 和 F1-Score 作为测评指标，对 BBGA 模型的分类性能进行评价。

准确率是一个统计测量，其计算公式如(6)所示，本实验中用于对特征提取的效果验证以及 F1-Score 的计算。

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

召回率 Recall 用于计算 F1-Score，其计算公式如(7)所示：

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

F1-Score 是衡量分类器分类准确性的指标，其计算公式如(8)所示：

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

4.2 模型与参数设置

本文中的实验基于 pytorch 框架，用于训练的 GPU 是 Quadro RTX 6000，模型参数设置如表 2 所示：

表 2 模型参数

Table 2 Parameters of the model

参数	数值
词向量维度	768
BiGRU 维度	(768,2)
Dropout 值	0.2
学习率	1e-5
优化器	Adam
Attention 维度	64

实验中其他对比模型的参数设置都参考了经典的模型，例如 Kim 等人对于 CNN 的实验工作，Lai 等人对于 RNN 的实验工作。

4.3 实验结果与分析

为了验证将 BERT 作为嵌入层来训练词向量的有效性，本文首先选择 TextCNN、TextRNN、DPCNN 和 FastText 作为对比实验的模型，与 BERT_CNN 和 BERT_RNN 的实验结果相比较，试验结果如表 3 所示：

表 3 第一组实验结果

Table 3 The first group of results

方法	精确率	召回率	F1 值
TextCNN	0.8914	0.8912	0.8911
TextRNN	0.8918	0.8911	0.8909
FastText	0.9052	0.9054	0.9052
DPCNN	0.8918	0.8913	0.8912
BERT_CNN	**0.9352**	**0.9352**	**0.9352**
BERT_RNN	**0.9333**	**0.9333**	**0.9333**

可以看出，将 BERT 作为嵌入层训练词向量，可以有效的优化输入文本的向量表达，从而获得更好的训练效果。相比于 TextCNN，BERT_CNN 的 F1 值上升了 4.41%，相比于 TextRNN，BERT_RNN 的 F1 值上升了 4.24%。

接着，为了验证本文提出的 BBGA 模型在进行文本分类任务时的有效性，选择上述实验中表现相对优秀的 BERT_CNN 和 BERT_RNN 模型作比较，同时，选择了纯粹的 BERT 模型来做对比，实验结果如表 4 所示：

表 4 第二组实验结果

Table 4 The second group of results

方法	精确率	召回率	F1 值
BERT	0.9279	0.9276	0.9276
BERT_CNN	0.9352	0.9352	0.9352
BERT_RNN	0.9333	0.9333	0.9333
BBGA	**0.9434**	**0.9429**	**0.9430**

从表中的结果可以看出，相比于 BERT 词嵌入后只接一层训练网络，本文提出的 BBGA 模型，有效的优化了分类时的相对权重，提高了文本分类的精度。同时，可以看出，虽然纯粹的 BERT 在许多任务中都有不俗的表现，但是效果是有限的，将其训练的结果作为进一步的词向量，可以达到更好的效果。

接着，为了分析本文的 BBGA 模型在每个类别分类上的准确率，同时也为后续的优化工作做准备，我们挑选了具有代表性的 BERT_CNN 模型和 FastText 模型作为对比实验，实验结果如图 5 所示。

可以看出，BBGA 模型不仅在每个具体的分类任务中的效果优于其他模型，同时每个分类任务的准确率都超过了

90%，某些特征明显的类别，例如体育和教育，精度更是达到了 98.14%和 96.48%。

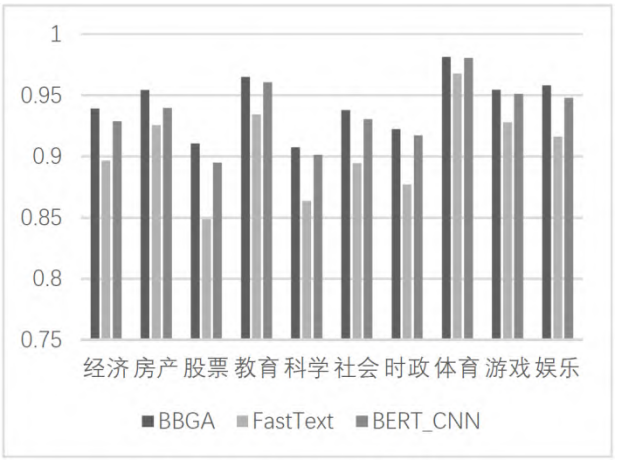

图 5 各类别分类精度对比

Fig.5 Comparison of classification accuracy of each category

最后，为了验证 BBGA 模型的分类效率，我们对比了几个深度学习模型达到收敛所需的时间，在学习率统一设置为 1e-5 的条件下，实验结果如表 5 所示：

表 5 收敛速度对比

Table 5 Comparison of convergence rates

模型	收敛耗时
BERT	8min12s
BERT_CNN	12min37s
BERT_RNN	15min19s
BBGA	**13min35s**

从表 5 中的结果可以看出，RNN 由于无法进行并行计算，收敛速度相比于 RNN 较慢，但是 BBGA 模型采用的神经网络结构相对简单，收敛速度要快于传统的 RNN 模型，同时相比于 BERT_CNN 耗时稍长一点，这是 RNN 相关模型特点决定的。尽管如此，BBGA 模型还是在保证较高精确度的同时，一定程度上减少了耗时，达到了高效准确地进行大规模文本分类的效果。

5 结论

我们正处于信息极速增长的时代，对于文本分类任务而言，准确率和效率都是极为重要的衡量指标。传统方法虽然速度快，但是准确率普遍偏低，深度学习方法通过对模型的改进提高了准确率，混合方法以耗时为代价进一步提高了准确率。为此，本文提出了 BBGA 模型，旨在保证较高分类精度的前提下，尽可能减少文本分类的耗时。

首先对比了几种常用的词向量表示方法，选择了以大量语料的训练为基础得到的动态词向量模型 BERT。接着，为了保证文本提升特征提取的质量和速度，使用 GRU 网络，并加入正反两层网络，充分捕捉文本的上下文联系。最后加入了注意力机制，用于调整分类时的权重比例。

实验表明，在 THUCNews 数据集下，BBGA 模型的性能完全优于 TextCNN、TextRNN、DPCNN 和 FastText，而相对于 BERT、BERT_CNN 和 BERT_RNN，也有着优秀的表现。

但是，本文的工作还存在优化的空间，例如进一步优化词向量的表示，优化神经网络的结构等，从而更加快速准确地进行更大规模的文本分类。

References:

[1] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics(ACL),2014:655-665.

[2]Kim Y, Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP), 2014:1746-1751.

[3] Johnson R，Zhang T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,2017:562-570.

[4] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997,9(8):1735-1780.

[5] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//AAAI,2015：2267—2273.

[6] Chung J, Gulcehre C, Cho K.H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv e-prints arXiv:1412.3555,2014.

[7] Joulin A, Grave E, Bojanowski P, et al. Fasttext.zip: Compressing text classification models[J]. arXiv e-prints arXiv:1612.03651, 2016.

[8] Devlin J, Chang Ming-wei, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv e-prints arXiv:1810.04805,2018.

[9]Yang Zhao-qiang,Shao Dang-guo,Yang, et al. Chinese short text classification model with multi-feature fusion[J]. Journal of Chinese Computer Systems, 2020, 41(7): 1421-1426.

[10] Minaee S, Kalchbrenner N, Cambria E, et al. Deep learning based text classification: a comprehensive review[J]. arXiv e-prints arXiv:2004.03705,2020.

[11]Tai K S, Socher R,Manning C D. Improved semantic representations from tree-structured long short-term memory networks[J]. arXiv e-prints arXiv:1503.00075, 2015.

[12]Dieng A B, Wang C, Gao J.et al. TopicRNN: a recurrent neural network with long-range semantic dependency[J]. arXiv e-prints arXiv:1611.01702,2016.

[13]Prusa J, Khoshgoftaar M. Designing a better data

representation for deep neural networks and text classification[C]//2016 IEEE 17th International Conference on Information Reuse and Integration, IRI 2016, 2016.

[14] Conneau A, Schwenk H, Barrault L, et al. Very deep convolutional networks for text classification[J]. arXiv e-prints arXiv:1606.01781, 2016.

附中文参考文献：

[9]杨朝强,邵党国,杨志豪,等.多特征融合的中文短文本分类模型[J].小型微型计算机系统,2020, 41(7): 1421-1426.