基于 BERT 的长文本分类方法

刘　博，蒲亦非
(四川大学计算机学院，成都 610065)

摘　要：由于预训练模型输入分词数量限制，基于 BERT 的长文本分类任务效果与长文本分割后的文本段集合的处理及特征融合密切相关，现有的长文本分类研究在融合文本段特征时更关注文本段之间原始的顺序关系，而本文提出了一种基于 BERT 和集合神经网络的长文本分类模型. 该方法以 BERT 为基础，可处理从同一文本样本分割得到的任意数量文本段，经过 BERT 后得到文本段特征，再将所有文本段特征输入到具有置换不变性的集合神经网络层中，提取出集合级别特征来优化长文本的特征表达. 通过在三个数据上的实验分析，论文在平均分词长度较长的数据集上取得了 90.82%的准确率，高出目前最优方法 4.37%.

关键词：文本分类；BERT；集合神经网络；长文本
中图分类号：TP391.4　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2023.022003

BERT-based approach for long document classification

LIU Bo, *PU Yi-Fei*
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: Concerning the input limitation of pre-training model, a long document needs to be spit into a set of text segments. The performance of long document classification is closely related to the further processing of the segment set and feature fusion. Existing document classification models keep more attention on the sequential of segments in the text segment set. However, the athors consider that the sequential order of segments have a mild influence on drawing the feature of a long document. The authors propose a BERT based long document classification model, which utilizes deep sets to obtain the collection-level feature from the segment set. In the model, the authors obtain a set of text segment features after BERT, and this proposed network which is immune to permutation learns the identical feature of the set to represent the long document feature. The accuracy of our model on the 20 newsgroups dataset achieved 90.82%, which outperforms the state-of-the-art method by 4.37%.

Keywords: Document classification; BERT; Deep Sets; Long Document

1　引　言

文本分类在自然语言处理中是一个经典且重要的任务，要求模型能够从各种文本信息中学习到句子或篇章级特征信息，从而对文本语义、情感和意图等进行识别分类[1]. 网络中存在海量的各类型文本数据，包括新闻、论文、网页、演讲、对话等，对文本进行快速识别和分类整理需要大量时间消耗，

收稿日期：2022-05-12
基金项目：国家自然科学基金面上项目(62171303)；中国兵器装备集团(成都)火控技术中心项目(非密)(HK20-03)；国家重点研发项目(2018YFC0830300)
作者简介：刘博 (1998－)，女，四川泸州人，硕士研究生，主要研究领域为自然语言处理. E-mail：liu_bo@stu.scu.edu.cn
通讯作者：蒲亦非. E-mail：puyifei@scu.edu.cn

因此,借助计算机技术来完成文本快速分类成为了重要的研究方向.

近年来,深度学习的文本分类模型大量涌现,比如基于卷积神经网络的 S2Net 模型[2],循环神经网络(RNN)的 TextCNN 模型[3],基于长短时记忆(LSTM)的 Bi-LSTM 模型[4]和 LT-LSTM 模型[5],基于胶囊网络的文本分类模型[6-8]等. 上述模型在实现文本分类任务时,对文本长度都比较敏感,往往在处理较短文本时才能取得比较好的效果. RNN 和 LSTM 等模型着重考虑了句子的词序信息,但网络结构决定了它们对内存资源消耗极高. 不仅如此,RNN 模型还存在梯度消失的问题. 2018 年谷歌提出了基于双向 Transformer[9]的 BERT (Bidirectional Encoder Representations from Transformers)[10]预训练模型,由于其优秀的基于词向量的句级特征提取能力,在大量非生成式自然语言处理任务上都有突破性的贡献. BERT 模型很好地克服了 RNN 和 LSTM 模型无法并行计算的缺点,并且由于其内部 Transformer 自注意力模块的特性,在长距离依赖问题上表现出更好的"记忆"能力. 但是 BERT 模型相对复杂,本身参数规模较大,模型对输入句子中包含的分词个数进行了限制,想要训练超过输入限制词数的长文本时只能截取其中一部分内容作为输入. 因此相比短句子分类问题,BERT 模型在长文本的分类问题上表现较差.

同时,除文本分类问题以外,自然语言处理的很多方向都有处理长文本的需求和困难[11]. 大量预处理模型和基于预处理模型改进的长文本分类模型也开始尝试克服长文本处理的难题[12-15]. 受到 BERT 预处理模型的启发,卡内基梅隆大学联合谷歌发布了 XLNET 预处理模型[16],XLNET 模型吸取 BERT 自编码语言模型双向信息提取的优点,并改进了需要设置 mask 从损坏的输入中重建数据的缺点,同时将 Transfoermer-XL 的分割循环机制(Segment Recurrence Mechanism)和相对编码范式(Relative Encoding)整合到模型中,从而解决超长序列的依赖问题,在处理长文本场景下有天然的优势. 虽然 XLNET 模型在处理长文本甚至超长文本问题上,都有比较好的能力,但模型复杂度也远超于 BERT 模型,模型训练的硬件设备资源要求非常高,训练难度较高. 因此本文选择基于 BERT-base 预训练模型来完成长文本分类任务.

目前主要的基于 BERT 的文本分类算法在处理数据集中长文本样本时,都将 BERT 模型无法处理的长文本分割成多个文本段后再分别输入 BERT,将经过 BERT 提取到的所有特征以一定方式融合后投入下游的分类任务网络中. 这些长文本处理模型往往在下游网络中使用 RNN、LSTM 和 Transformer 等结构,分析提取文本段顺序中隐含的特征.

由于文字的特殊性,不同文本之间句子结构和段落间关联关系都不尽相同,例如有的文本中总结段落在文本开头,有的文本总结段落在文本结尾. 因此同一类别中,文本样本中段落之间的顺序关系对应的深层逻辑关系并不相同. 因此本文认为可以让模型学习文本段之间的深层逻辑特征. 下游网络输入应该忽视文本段顺序关系并能接收不同数量的文本段特征输入,使模型学习文本段之间的真实逻辑特征. 并且大部分长文本的主要信息都凝聚在局部的句子或段落,由文献[1,10]的实验结果可见,使用不超过 BERT 输入限制的局部文本段学习到的特征信息在文本分类任务中已经可以得到较好的结果.

受到集合神经网络模型[17]的启发,为学习文本段之间的真实逻辑特征、提取篇章级特征,我们认为可以将长文本的文本段特征集合,用具有置换不变性输入的集合神经网络来进行处理.

综上所述,使用预训练模型完成长文本分类任务时,对分割后文本段的特征信息融合处理尤为重要,以往的算法为了达到较高的准确度牺牲了模型的复杂度,太过简单的下游网络又难以学习到较好的分类特征. 过去基于 LSTM 和 RNN 等模型的下游处理网络更关注与文本段之间的真实顺序关系特征,但由于行文风格等差异,长文本之间顺序关系特征可能不尽相同. 同时我们注意到人脑是可以分析乱序文本段落之间的逻辑关系的,并且人脑在对文章进行分类时,段落之间的时序信息并不会带来特别大的影响. 因此我们猜测,长文本分类任务中,不需要着重对文本段时序信息进行提取. 本文提出了一种基于 BERT 的新颖且简单的长文本分类算法,利用集合神经网络层来融合文本段集合的多尺度特征实现分类任务. 本文的主要贡献如下:(1) 提出了一种简单高效的基于集合神经网络聚合多尺度特征信息进行长文本分类的方法;(2) 网络可以处理同一文本样本分割得到的任意多数量的文本段;(3) 在多个数据集中效果达到了 SOTA(State-Of-The-Art)水平,本文模型在参数量上

有优势.

2　相关工作

长文本分类算法主要分为基于传统机器学习的文本分类和基于深度学习的文本分类. 基于深度学习的长文本分类算法是近年来的研究热点，本节将分析领域内主要的几种深度学习模型.

2.1　基于 LSTM 的长文本分类模型

LSTM 模型能够有效对序列信息进行建模，并很好地解决长文本的远距离依赖问题. 大量基于 LSTM 模型的变体在长文本分类任务中被提出. 为了对树结构对语义进行建模，树形结构的 LSTM 模型被提出，不同于标准 LSTM 模型中节点只能从前一个节点获取依赖信息，树形 LSTM 可以选择性地从子节点中获取信息. Tai 等[18]提出了两种 Tree-LSTM 模型，其中依存句法分析树适用于子节点个数不定或子节点乱序的树结构，成分句法分析书仅适用于在子单元之间有序的情况下，且单元的子单元个数最多为 *N*. Chen 等[19]提出了能够处理序列级别句子的 LSTM 模型，引入符号间关系的注意力机制，解决输入的结构化问题. 除此以外，为了对非连续性短语结构进行建模，Peng 等[20]首次将图卷积神经网络引入长文本分类任务中.

2.2　基于 BERT 的长文本分类模型

BERT 是由 Transformer 块组成的双向自编码预训练语言模型，相较于过去的 RNN、LSTM 语言处理模型，BERT 不仅能够同时利用上下文信息进行并发训练，还能解决远距离依赖问题. BERT 预训练模型提取文本特征示意图如图 1，从目前相关研究中[1,12,21]我们可以发现，在进行下游任务之前，使用预训练好的 BERT 模型提取到的句子特征进行简单的分类任务时就非常有效，即使只用一层全连接层实现分类，也在大部分数据集上取得了较好的结果.

但由于 BERT 模型对输入分词数量的限制，在处理长文本分类问题时，必须对文本提前进行分割，再对分割后的文本进行处理. 如何分割文本，以及分割后经过 BERT 得到的所有文本段特征如何融合都是需要认真考虑的问题. 目前基于 BERT 的长文本分类算法对文本的处理方式主要可以被分为以下两类. 一类是直接将文本循环截断成文本段，所有文本段通过 BERT 提取到其文本特征，再利用循环神经网络或长短时记忆网络等结构作为下游网络，实现对文本特征按序结合生成新的总特征. 另一类是从文中切割出单个且等于 BERT 输入限制的文本段代表整篇文章，只对这一个文本段进行特征提取、分析.

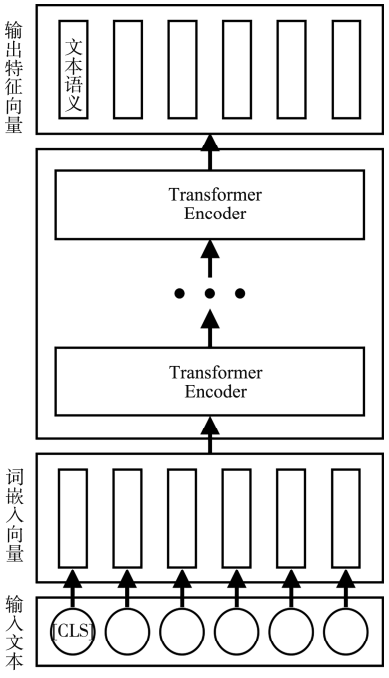

图 1　BERT 结构图
Fig. 1　Structure of BERT

2018 年，相关研究者提出的 BERT 模型在文本分类等许多自然语言处理任务上展现出巨大潜力. 在此之后基于 BERT 的模型也被大量提出，其中 Adhikari 等[10]提出的 DocBERT 模型通过微调 BERT 以及知识蒸馏实现长文本分类任务. 其他一些基于微调 BERT 模型的长文本分类算法，通过将文本序列输入预训练好的 BERT 模型，输出序列级别的文本特征，再在下游任务中进行分类. 由于 BERT 模型限制了输入文本长度，在长文本分类处理上有限制，Pappagari 等[22]提出了将长文本又重复地分割，经过 BERT 模型后，下游任务中采用 Transformer 来实现分类，Khandve 等[23]采用 LSTM 在 BERT 下游对文本段信息进行融合，这两种层级训练方法高效地将 BERT 扩展到了长文本分类任务上，并取得了比较好的效果. 但下游参数量较多，难以对模型进行加速，同时将文本段原始顺序特征作为主要特征进行提取.

2.3　基于集合神经网络的分类方法

Zaheer 等[17]提出了集合神经网络的编码器-

解码器结构,编码器部分先单独作用于集合中所有元素,解码器聚合编码器输出的所有元素特征,得到集合级别的特征,常使用池化结构来聚合集合特征,在集合统计量估计和集合异常检测实验中取得了比较好的结果. Zaheer 等也给出了解码器部分处理样本为元素集合的函数规则,对集合级别特征的提取函数建模,需要遵循排列不变性,使集合级别特征与输入样本集合的排列顺序无关,只与输入样本集合数量和特征值有关. 为了同时学习到元素之间的交互信息,Lee 等[24]提出了 Set Transformer 模型,在编码器部分设计了集合注意力模块(Set attention Block),在集合元素之间执行自注意力,输出元素个数相同的集合. 集合神经网络在集合统计量估计、点云分类、小样本图像分类和集合异常检测等领域都取得了比较好的结果. 因此本文也使用集合神经网络来对文本样本的所有文本段集合进行处理,利用排列不变性函数提取出集合级别特征解决长文本分类任务.

3　方　法

3.1　基于 BERT 的长文本分类算法

对文本段的截取以及对各文本段特征的处理,是长文本分类处理任务中的重点. 本文提出的算法中,BERT 下游采用置换不变性的集合神经网络提取文本段集合级别特征,利用集合级别特征完成长文本分类,并引入 ResNet 残差网络[25]解决退化问题. 目前的算法大多都是对长文本的有序文本段特征进行分析处理,但是即使是同一类别的文本,由于行文思路和写作手法的不同,文本的逻辑结构和叙事顺序并不具备一致性,文章有序文本段之间的依赖关系也并不能在分类中其决定性作用. 于是本文提出了一种基于 BERT 的长文本分类模型,如图 2 所示,我们将长文本按照 BERT 可处理的最大长度进行不重叠的分割,经过 BERT 处理后得到文本段特征集合,再对该文本段特征集合进行无序的多尺度集合级别特征提取,最后完成文本分类.

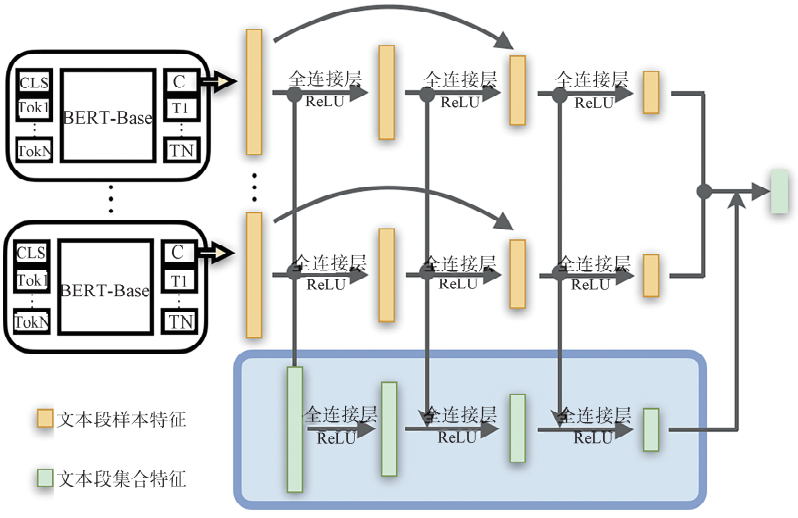

图 2　模型结构图

Fig. 2　Model structure

我们将数据集中单个长文本 X 分割为 N 个文本段,表示为 $x_i, i = 1,2,\ldots,N$,将经过 BERT 模型后得到的文本段特征表示为 $y_i, i = 1,2,\ldots,N$,于是构建出函数模型如下.

$$y_i = B(x_i) \tag{1}$$

$$F(X) = S(R(y_1), R(y_2), \ldots, R(y_N)) \tag{2}$$

其中,B 表示 BERT 网络,输入长文本样本分割得到的文本段集合,得到对应的文本段特征集合;$R()$为残差结构的全连接网络,用于获取不同尺度特征并避免网络退化;$S()$依据置换不变性原理,利用无序文本段特征集合得到集合级别特征.

3.2　残差网络提取多尺度特征

为处理任意多的文本段特征,我们采用所有文本段特征提取网络之间参数共享的方案. 为避免下游分类网络退化,我们设计残差网络结构来提取多尺度文本段特征. 文本段特征提取网络层函数 R 为

$$y = R(y) + \boldsymbol{W} * y \tag{3}$$

其中,$\boldsymbol{W}$ 为转换特征维度的参数矩阵.

3.3　集合神经网络提取元素集合级别特征

为设计提取集合级别特征的网络结构，我们遵循置换不变性原则，使用最大池化层统计函数，从数量不定的无序文本段特征集合中提取集合级别特征 ，集合特征提取函数可表示为

$$z=S(Y)=\gamma\text{maxpool}(y_i\mid i=1,2,\ldots,N) \quad (4)$$

其中 γ 是最大归一化的系数；maxpool 最大池化函数的提出是为了缓解卷积层对位置的过度敏感性，本质是减少特征的同时并保持特征不变性. 本文使集合级别特征，maxpool 函数遵循置换不变性的原则，对文本段特征顺序不敏感，提取的集合特征具有唯一性，并且可处理任意多数量多文本段特征.

4　实验与结果

4.1　数据集

本论文使用三个文本数据集来评估我们的模型，分别是 20 NG(20 newsgroups)数据集[26]，IMDB 数据集[27]以及 R8 数据集.

其中 20 NG 数据集[26]总共包含约 19 000 篇新闻，由 20 个不同类别的新闻文本组成，包括了一些相关主题，如“comp. sys. mac. hardware”和“comp. sys. ibm. pc. hardware”，也包括一些完全无关的类别，如“sci. electronics”和“talk. politics. misc”等. 其中文本的平均长度为 266 个词，最长长度为 10 334 个词. 本文按照 6∶4 的比例将数据集分为训练集和测试集.

IMDB 数据集[27]包含 50 000 篇电影评论，分为正向和负向两种情感倾向，平均长度为 231 个词，最长长度. 我们按照 1∶1 的比例将数据集分为训练集和测试集.

R_8 数据集包含了 7 674 篇路透社的文章，分为 8 种不同的类别，平均长度为 64 个词，我们按照 7∶3的比例将数据集分为训练集和测试集.

4.2　分组实验

本文提出的长文本分类模型是在 BERT 模型的基础上，使用 BERT 模型输出的文本段特征，在下游网络中提取文本段集合级别特征并完成分类任务. 由于数据集中存在大量样本的文本长度比 BERT-base 能够处理的输入单词数量更长，于是进行无覆盖的分割，得到长文本的文本段集合，再对经过 BERT 输出的文本段特征集合进行融合，完成分类任务.

本文所有实验都使用单张 NVIDIA GeForce RTX-3090，预训练模型使用转换到 Pytorch 框架下的 BERT-base 模型，该模型输入限制为 512 个分词，包含 12 层 Transformer 编码器结构块，输出文本段特征维度为 768. 同时，在训练过程中同时对 BERT 预训练模型进行微调，并使用了交叉熵损失函数，具体参数取值见表 1.

表 1　参数取值

Tab. 1　Parameter settings

参数	数值
输入最大长度	512
训练轮次	10
批处理大小	12
学习率	0.01
激活函数	Leaky-ReLU

4.3　实验结果与分析

本文方法与其他方法模型参数量对比如表 2 所示. 本文提出算法是基于 BERT-base 模型，相较于其他一些基于 BERT-base 的算法，本文算法只增加了 0.2 M 的参数量.

本部分的其他算法模型准确率都来自于提出型的原论文中在同一数据集上的最好结果. 我们将本文模型与其他表现 SOTA 水平的模型进行比较，BERT 作为分类任务的基线，LongFormer 是专为长文本处理提出的模型，SGC(Simple Graph Convolution)是简单图卷积神经网络模型. 如表 3 所示，通过对比本模型与其他几种不同模型在数据集上的准确率，我们可以看出，本文提出的算法模型在三个数据集上都表现出了 SOTA 水平，其中在 20NG 数据集上，远远超过了其他几种模型的分类准确率，将准确率提升了. 除此以外，在 IMDB 数据集上，效果也略好于其他四个模型，在 R8 数据集上，与其他模型效果相当. 这些模型都着重提取了输入顺序带来的时间顺序特征，即原始长文本的文本段顺序特征. 而本文采用具有置换不变性的集合神经网络作为 BERT 下游网络，来对长文本集合级别特征进行提取，并且取得的效果优于目前主要的几种长文本分类算法. 因此在训练时，不需要着重去提取文本段之间的顺序特征，也可以在长文本分类任务中取得比较好的效果.

同时，为了验证模型的有效性，我们在 20NG 数据集上对模型的收敛性进行测验. 如图 3 损失曲线图，我们可以看出，在总共 20 轮次的迭代轮次-损失的曲线图中，模型训练迭代轮次为 10 时，模型损失已经降低到较低水平，模型基本收敛. 但是在第 13 轮时突然产生损失小幅度增加的情况，同

时在 19 轮也有一定的损失小幅度增加,我们认为这是因为数据集比较小从而产生了一定的过拟合,但是整体的损失都维持在比较合理的变化范围内.

表 2　模型参数量

Tab. 2　Comparison of the number of parameters

模型	参数量/M
BERT-base[10]	110
LongFormer[14]	102
BERT+LSTM[23]	112.5
BERT+Transformer[22]	123
Our	110.2

表 3　对比实验结果

Tab. 3　Comparison experimental results

数据集	模型	准确度
20NG	BERT-base[10]	85.78
	LongFormer[14]	86.45
	BERT+LSTM[23]	85.57
	DistilBERT[28]	85.43
	BERT+Transformer[22]	85.52
	SGC[29]	86.12
	Our	90.82
IMDB	BERT-base[10]	89.58
	LongFormer[14]	93.3
	BERT+LSTM[23]	93.63
	DistilBERT[28]	92.18
	BERT+Transformer[22]	93.97
	Our	94.54
R8	BERT-base[10]	97.62
	LongFormer[14]	97.85
	BERT+LSTM[23]	96.35
	BERT+Transformer[22]	97.13
	Our	97.25

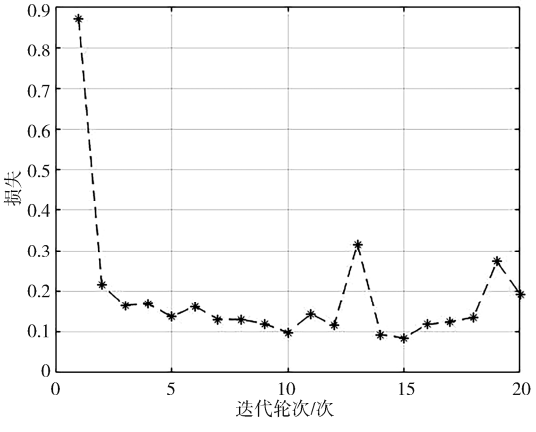

图 3　Loss 损失曲线图

Fig. 3　Loss of our models

4.4　消融实验

为了更好地分析本文提出的模型中集合特征提取模块的效果,我们设计了一组消融实验,实验采用 20NG 数据集,如表 4 所示.

表 4　消融实验

Tab. 4　Ablation experiments

输入文本表示		池化		集合特征		准确率
单个文本段	文本段集合	最大值	平均	单尺度	多尺度	
√						85.78
	√	√		√		89.02
	√	√			√	90.82
	√		√		√	90.63

表 4 中,前两列对比了不同文本样本的表示方法对准确率的影响,可以看出如果只从文本中提取单个 512 个分词的文本段来代表这个样本,效果比使用样本的文本段集合要差很多. 第三、四列比较了提取集合级别特征时,采用两种不同的置换不变的池化函数产生的不同效果. 由表 4 可见,在其他条件同等时,最大池化与平均池化效果相当. 第五、六列比较了提取集合级别特征的区别,在不同网络层提取多尺度的集合级别特征,比仅仅使用最后一层单尺度集合级别特征来进行分类任务效果更好.

由表 4 可见,没有下游网络处理的简单 BERT-base 模型在实现分类任务时,也能有一定的效果,在 BERT 模型下提取单尺度的集合级别特征用于分类,模型的准确度有一定提高,说明在长文本分类任务中,截断的文本段之间的位置关系特征对集合级别特征提取并没有那么重要,使用简单的置换不变池化函数就可以得到比较有效的集合级别特征[30-32]. 不仅如此,在加入了多尺度集合级别特征提取模块后,模型的分类准确度有了较大的提升,可以说明多尺度文本段集合级别特征信息有利于长文本分类.

5　结　论

本文提出了一种基于 BERT 的长文本分类方法,通过将长文本分割后的所有文本段输入 BERT 得到文本段特征集合,再使用置换不变的函数提取多尺度集合级别特征用于分类任务,减少了模型参数量. 本方法可以对文本样本分割得到的任意数量文本段进行特征融合. 从实验结果可见,本文模型

在长文本数据集中取得了非常好的效果，表明引入集合神经网络结构提取到的集合级别特征，极大地改善了长文本分类任务. 下一阶段，我们考虑拓展并检验本模型的分类能力，将本文模型应用到多标签分类问题上.

参考文献：

[1] Minaee S, Kalchbrenner N, Cambria E, *et al*. Deep learning-based text classification: a comprehensive review [J]. ACM Comput Surv, 2021, 54: 1.

[2] He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015.

[3] Chen Y. Convolutional neural network for sentence classification [D]. Waterloo: University of Waterloo, 2015.

[4] Liu P, Qiu X, Chen X, *et al*. Multi-timescale long short- term memory neural network for modelling sentences and documents[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015.

[5] ZhouP, Qi Z, Zheng S, *et al*. Text classification improved by integrating bidirectional lstm with two-dimensional maxpooling [C]//Proceedings of the 26th International Conference on Computational Linguistics. [S. l.]: Technical Papers, 2016.

[6] 朱海景，余谅，盛钟松，等. 基于静态路由分组胶囊网络的文本分类模型[J]. 四川大学学报：自然科学版，2021，58：062001.

[7] 高云龙，吴川，朱明. 基于改进卷积神经网络的短文本分类模型[J]. 吉林大学学报：理学版，2020，58：923.

[8] 王进，徐巍，丁一，等. 基于图嵌入和区域注意力的多标签文本分类[J]. 江苏大学学报：自然科学版，2022，43：310.

[9] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need [J]. Adv Neural Inf Process Syst, 2017, 30.

[10] Devlin J, Chang M, Lee K, *et al*. BERT: pretraining of deep bidirectional transformers for language understanding [EB/OL]. [2022-02-23]. http://arxiv.org/abs/1810.04805.

[11] Kowsari K, Jafari K, Heidarysafa M, *et al*. Text classification algorithms: a survey [J]. Information, 2019, 10: 150.

[12] Adhikari A, Ram A, Tang R, *et al*. Docbert: bert for document classification [EB/OL]. [2022-02-23]. http://arxiv.org/abs/1904.08398.

[13] Liu Y, Ott M, Goyal N, *et al*. Roberta: a robustly optimized bert pretraining approach [EB/OL]. [2022-02-23]. http://arxiv.org/abs/1907.11692.

[14] Beltagy I, Peters M E, Cohan A. Long-former: the long-document transformer [EB/OL]. [2022-02-23]. http://arxiv.org/abs/2004.05150.

[15] Ding S, Shang J, Wang S, *et al*. Ernie-doc: a retrospec-tive long-document modeling transformer [EB/OL]. [2022-02-23]. http://arxiv.org/abs/2012.15.

[16] Yang Z, Dai Z, Yang Y, *et al*. XLNet: generalized autoregressive pretraining for language understanding [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Neurlps, 2019.

[17] Zaheer M, Kottur S, Ravanbakhsh S, *et al*. Deep sets [J]. Adv Neural Inf Process Syst, 2017, 30: 3391.

[18] Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. [S. l: S. n.], 2015.

[19] Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. [S. l]: EMNLP, 2016.

[20] Peng H, Li J, He Y, *et al*. Large-scale hierarchical text classification with recursively regularized deep graph-cnn [C]// Proceedings of the 2018 World Wide Web Conference. [S. l: S. n.], 2018.

[21] Adhikari A, Ram A, Tang R, *et al*. Exploring the limits of simple learners in knowledge distillation for document clas- sification with docbert [C]//Proceedings of the 5th Workshop on Representation Learning for NLP. Seattle: ACL, 2020.

[22] Pappagari R, Zelasko P, Villalba J, *et al*. Hierarchical transformers for long document classification [C]//Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). [S. l.]: IEEE, 2019.

[23] Khandve S, Wagh V, Wani A, *et al*. Hierarchical neural network approaches for long document classi-

fication [C]//Proceedings of the 2022 14th International Conference on Machine Learning and Computing (ICMLC). Guangzhou: ICML, 2022.

[24] Lee J, Lee Y, Kim J, *et al*. Set transformer: a framework for attention-based permutation-invariant neural networks [C]//Proceedings of the International Conference on Machine Learning. Edinburgh: Opnnipress, 2019.

[25] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Juan: IEEE, 2016.

[26] Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine learning in python [J]. J Mach Learn Resh, 2011, 12: 2825.

[27] Maas A, Daly R E, Pham P T, *et al*. Learning word vectors for sentiment analysis [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. [S. l: S. n.], 2011.

[28] Sanh V, Debut L, Chaumond J, *et al*. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter [EB/OL]. [2022-02-23]. http://arxiv.org/abs/1910.01108.

[29] Wu F, Souza A, Zhang T, *et al*. Simplifying graph con-volutional networks [C]//Proceedings of the International Conference on Machine Learning. Long Beach: PMLR, 2019.

[30] Adhikari A, Ram A, Tang R, *et al*. Rethinking com-plex neural network architectures for document classification [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019.

[31] Wagh V, Khandve S, Joshi I, *et al*. Comparative study of long document classification [C]//TENCON 2021-2021 IEEE Region 10 Conference (TENCON). Auckland: IEEE, 2021.

[32] Liu L, Liu K, Cong Z, *et al*. Long length document classification by local convolutional feature aggregation [J]. Algorithms, 2018, 11: 109.

引用本文格式:

中　文: 刘博, 蒲亦非. 基于 BERT 的长文本分类方法[J]. 四川大学学报: 自然科学版, 2023, 60: 022003.

英　文: Liu B, Pu Y F. BERT-based approach for long document classification [J]. J Sichuan Univ: Nat Sci Ed, 2023, 60: 022003.