文章编号:1671-4598(2023)04-0213-06 DOI:10.16526/j.cnki.11—4762/tp.2023.04.033 中图分类号:TN99; TP391 文献标识码:A

结合 Bert 与 Bi-LSTM 的英文文本分类模型

张卫娜
(西安思源学院,西安 710038)

摘要:作为自然语言处理技术中的底层任务之一,文本分类任务对于上游任务有非常重要的辅助价值;最近几年,深度学习广泛应用于 NLP 中的上下游任务的趋势,深度学习在下游任务文本分类中性能不错;但是目前的基于深层学习网络的模型在捕捉文本序列的长距离型上下文语义信息进行建模方面仍有不足,同时也没有引入语言信息来辅助分类器进行分类;针对这些问题,提出了一种新颖的结合 Bert 与 Bi-LSTM 的英文文本分类模;该模型不仅能够通过 Bert 预训练语言模型引入语言信息提升分类的准确性,还能基于 Bi-LSTM 网络去捕捉双向的上下文语义依赖信息对文本进行显示建模;具体而言,该模型主要有输入层、Bert 预训练语言模型层、Bi-LSTM 层以及分类器层搭建而成;实验结果表明,与现有的分类模型相比较,所提出的 Bert-Bi-LSTM 模型在 MR 数据集、SST-2 数据集以及 CoLA 数据集测试中达到了最高的分类准确率,分别为 86.2%、91.5%与 83.2%,大大提升了英文文本分类模型的性能。

关键词:文本分类;深度学习;上下文语义信息;Bert;Bi-LSTM

English Text Classification Model Combining Bert and Bi-LSTM

ZHANG Weina
(Xi'an Siyuan University, Xi'an 710038, China)

Abstract: As a downstream natural language processing task, text classification has very vital auxiliary value for upstream task. Deep learning is widely used in the trend on upstream and downstream tasks of NLP in recent years, deep neural networks have very good performances in text classification tasks. However, current model based on deep learning networks has the shortage of modeling the long context semantic information of the text sequence, and it also does not introduce language information to assist the classifier to classify. To solve these problems, a novel English text classification model for combining Bert with Bi-LSTM is proposed. The proposed model can not only boost the performance of classification by introducing language information into Bert pre-training language model, but also capture bi-directional context semantic dependency information based on Bi-LSTM network to model the display of text. Specifically, the model is mainly composed of input layer, Bert pre-training language model layer, Bi-LSTM layer and classifier layer. Compared with baseline models, extensive experimental results demonstrate that the proposed Bert-Bi-LSTM model achieves the highest classification accuracy in MR dataset, SST-2 dataset and CoLA dataset with 86.2%, 91.5% and 83.2% respectively, which greatly improves the performance of the English text classification model.

Keywords: text classification; deep learning; context semantic information; Bert; Bi-LSTM

0 引言

随着互联网络的快速普及以及智能手机的风靡,英语世界中各种各样的社交媒体上出现了海量的文本信息,例如日更的新闻文本、各大购物平台上的商品评价文本、影视书籍平台的评论文本、手机短信、各种垃圾邮件等[1-2]。其中,短文本的数量在节奏明快的互联网信息时代最为庞大,可谓是数不胜数,这给我们提出了一个紧迫的需求和研究课题,那就是如何从这些短文本中及时且准确地提取出富有价值的信息,协助用户快速高效地筛选出有用的信息,满足各种需求,譬如情感分类[3]、垃圾邮件与色情信息过滤及新闻分类等[4]。

英文短文本主要的特征是长度稍短,其中介于 100~200 个单词长度的短文本最为常见。短文本分类是自然语言处理中一个非常经典且关注点极高的低层级任务之一,旨在为给定的文本分配预定义好的类别,其被广泛应用在问答系统、对话系统、情感分析系统以及其他系统等等。近年来,各种各样的神经网络模型被尝试应用到文本分类任务当中,例如 CNN 卷积神经网络模型[5-7]、自注意力模型以及生成对抗式网络[8]、RNN 循环神经网络模型[9-11]模型。相比于支持向量机 SVM 模型[12]等传统的统计学习方法,基于深度学习的模型性能更加优越,提供了更好的分类结果。预训练语言模型 Bert (bidirectional encoder representations from transformers) 是基于大规模语料库利用多任务预训练技术的一种自注意力模型[13]。与基于 CNN/RNN 模型和传统模型相比,它通常在许多任务中取得优异的性能,从命名体识别任务、文本分类到阅读理解任务等。

收稿日期:2022-08-27; **修回日期**:2022-09-27。
基金项目:国家自然科学基金项目(61502290)。
作者简介:张卫娜(1981-),女,陕西西安人,大学本科,讲师,主要从事英语教学法及信息化教育方向的研究。
引用格式:张卫娜.结合 Bert 与 Bi-LSTM 的英文文本分类模型[J].计算机测量与控制,2023,31(4):213-218,251.

尽管深度学习模型在文本分类任务中的性能出色，但是这些模型中的大多数通常没有做到很好地捕捉长距离的语义信息问题[14]。针对这点不足，很多研究者也提出了相应的解决方案。Yang 等人的 HAN 模型设计了一种分层注意机制将文本分为句子和单词两个层次，并使用双向 RNN 作为编码器[15]。Lei 等人提出的 MEAN[16]模型试图通过注意力机制将 3 种情感语言知识整合到深度神经网络中，从而缓解这一问题。虽然基于自注意力机制的模型在一定程度上缓解了这个问题，但依旧没有彻底解决。究其原因，很多模型是仅仅从句子或者文档直接对句子的表示进行建模的，并没有显式地将语言知识考虑进去，导致上下文的语义联系不强。例如句子："Though this movie is a little bit repetitive and vague, it really engages our senses through the way few current movie do"，可以发现句子中既包含了负面情感和正面情感。虽然句子的前半句对电影是持否定态度，然而联合句子的上下文可以发现"through the way few current movie do"是强烈的表达了对电影的肯定，如果不考虑语言知识的话，最终的分类器给出的预测分类结果可能是错误的，将此句子判定为负类。这恰恰是自注意力模型不能够达到的效果。

为了更好地捕捉长距离上下文信息以及语义依赖，同时有效地将语言知识引入到文本分类任务中，本文基于 Bert 预训练语言模型与 Bi-LSTM 模型提出了一种新颖的英文文本分类模型。首先，我们先使用 Bert 模型得到的嵌入来表示输入，然后将该语义嵌入向量当做 Bi-LSTM 模型的输入来做进一步的特征表示和提取。具体而言，Bert 模型能够很好地对语言知识进行建模与提取，提升网络的特征表达能力。与此同时，Bi-LSTM 模型能够对输入的前后两个方向的长距离上下文语义特征进行建模，捕捉文本中单词间的语义依赖关系。在多个数据集上的实验结果表明，本文的 Bert-Bi-LSTM 模型能够高效且有效地对句子中的长距离上下文依赖语义信息进行建模，同时结合语言知识，有效提升了文本分类的性能。

1 基于深度学习的文本分类工作

1.1 深层神经网络

近来几年中，深度神经网络在自然语言处理中放异彩，在不同任务中均取得了良好的效果。其中，CNN 和 RNN，包括长短期记忆 LSTM 型网络（long-short term memory）[17]以及门控递归单元 GRU（gate recurrent unit）[18]网络非常适合与文本中单词和序列的处理和特征提取。文献[5]和[7]均利用 CNN 来构建文本分类模型。首先，它们使用 word2vec 处理输入文本，然后将语义表示向量输入给 CNN 以提取特征，最后使用 softmax 函数来判断文本的类别。基于 CNN 的方法不仅仅能够实现自动提取特征，同时长于处理高维度的文本数据。然而，CNN 处理当前时间步的输入，不能够很好地对前一个时间步输入和后一个时间步的输入同时进行建模。针对此，Zhu 等人于 2015 年设计了一种基于 LSTM 网络的分类模型，该模型将使用词语序列技术来对评论文本进行词序列建模而完成特征抽取和分类，因为 LSTM 网络具有捕捉文本句子中的长距离型上下文语义关系[19]。C-LSTM 首先使用 CNN 捕获文本的局部信息，然后使用 LSTM 网络对卷积核的每个输出进行编码以捕获全局信息[20]。

1.2 预训练语言模型

近年来，类似于计算机视觉领域中的研究工作，预训练模型在 NLP 的多个任务中均取得了非常好的结果。预训练模型通常通过利用大量未标记语料库来学习通用的语言表示，并在针对不同任务的预训练模块之后接续额外的任务特定层。例如，ELMO（embeddings from language models）模型[21]致力于从语言模型中提取上下文敏感特征，从而刷新了几个主要 NLP 基准任务的 SOTA 性能，包括问答任务[22]、情感分析[23]和命名实体识别[24]。如图 1 所示，GPT（generative pre-training）[25]是一种基于微调方法的单向预训练语言模型。在预训练阶段，GPT 使用大规模连续语料来训练多层 Transformer 编码器。而 BERT 模型是基于多层双向变换器，并在大规模语料库上进行掩码语言模型预训练任务训练和下一个句子预测预训练任务。Bert 模型通过多头自注意力机制大大增强了对输入数据的不同部分的上下文语义表示的关注。

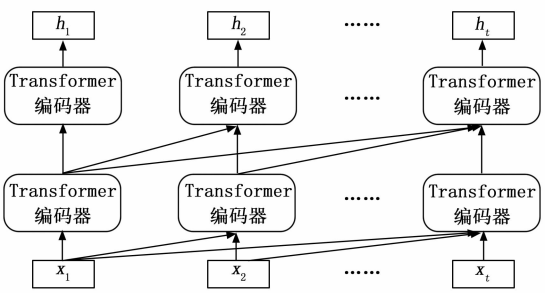

图 1 GPT 模型

2 预训练技术和双向长短期记忆网络

本节主要介绍文中所涉及到的相关模型和技术，分别是 Bert 预训练语言模型与 Bi-LSTM 网络。

2.1 Bert 预训练语言模型

图 2 中所展示的模型是 Transformer 模型的编码器部分，而这个编码器恰恰是构成 Bert 模型的基本单位。从图 2 中可以看出，编码器主要是包含了一层多头注意力网络、两层加法与归一化网络、一层前馈型网络，以及跳跃连接机制。

图 1 所示的 GPT 是一种基于预训练技术的单向预训练语言模型，其基本组成器件是 Transformer 模型的解码器部分，同时只是利用了单一方向的解码器。

如图 3 所示，Bert 模型则是利用了 Transformer 模型的编码器部分建立了双向的编码器模型。不同于 GPT 模型，Bert 模型没有使用 Transformer 模型的解码器，主要原因是解码器接触到的都是不完整的句子，而编码器则可以看到语料库中完整的句子，有助于捕捉完整的语言信息从而提

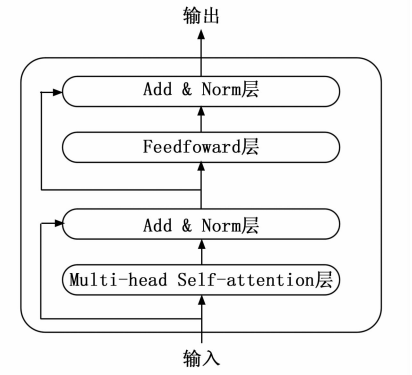

图 2　Transformer 的编码器

升文本分类的性能。

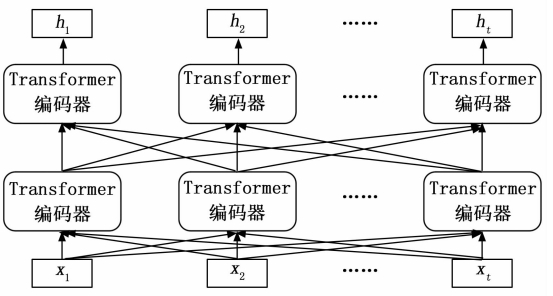

图 3　BERT 模型

2.2　Bi-LSTM 网络模型

如图 14 所示，针对标准 RNN 网络在训练阶段中一直存在梯度消失抑或爆炸的现象，HochReiter 与 Schmidhuber 两名学者于 20 世纪 90 年代提出了 LSTM 网络。

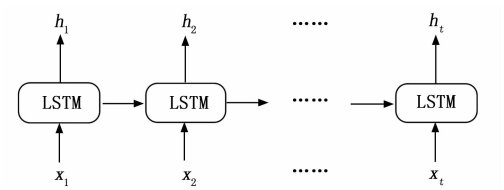

图 4　LSTM 网络

LSTM 网络的每个单元中主要包含了 3 个门，分别是输入门 i 、遗忘门 f 和输出门 o 。遗忘门主要是控制遗忘上一个时间步传来的信息类型和数量；输入门控制当前时间步内主要接收哪些信息来更新门状态；而输出门控制了当前时间步的信息流。在 LSTM 单元在第 t 个时间步的计算公式如下：

$$X = [h_{t-1}, \boldsymbol{x}_t] \tag{1}$$

$$f_t = \text{sigmoid}(W_f X + \boldsymbol{b}_f) \tag{2}$$

$$i_t = \text{sigmoid}(W_i X + \boldsymbol{b}_i) \tag{3}$$

$$c_t = f_t * c_{t-1} + i_t * \tanh(W_c X + \boldsymbol{b}_c) \tag{4}$$

$$o_t = \text{sigmoid}(W_o X + \boldsymbol{b}_o) \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

其中：$*$ 表示矩阵对应元素的乘法；$\boldsymbol{x}_t \in R^n$ 为输入向量，而 $W \in R^{m*n}$ 是各个门的参数，$\boldsymbol{b} \in R^m$ 是偏置向量。而上标 n 与 m 分别代表了单词向量的维度数目与语料库中的词汇表尺寸。此外，$[\cdot]$ 则表示拼接操作。

尽管 LSTM 网络基于门控机制和记忆单元提取了长距离的依赖信息，但是它忽略了从后往前这个方向的上下文信息，而这对于文本分类任务来说至关重要。自然而言地，我们想到了利用从前往后与从后往前两个方向的 LSTM 网络，即 Bi-LSTM，这样便能够实现双向的上下文依赖信息的捕捉。

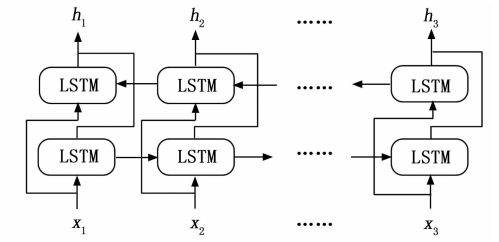

图 5　Bi-LSTM 网络

图 5 展示了 Bi-LSTM 网络的基本结构。给定上一个时间步的隐藏状态 h_{t-1} 和当前时间步的输入 x_t ，两个方向的 LSTM 的输出可通过以下公式得到：

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}}) \tag{7}$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t-1}}) \tag{8}$$

Bi-LSTM 的输出是通过拼接当前时间步下从前往后和从后往前双方向的隐藏向量而得到的，即 $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$ 。

3　结合 Bert 与 Bi-LSTM 分类模型搭建

3.1　Bert 层

3.1.1　网络的输入

Bert 双向编码器预训练语言模型，其一般是接受一个个由于多个单词组成的英文句子作为输入．每个句子在输入之前需要做一些预处理，即在句首加上表示句子开始的［CLS］标识以及在句末插入表示句子结束的［SEP］标识。随后，输入的英文句子经过预处理转换为 3 种输入向量嵌入，分别是单词向量、段向量以及位置向量嵌入等，而最终的输入是将以上 3 种向量嵌入相加而来。

其中，由于预训练的过程中需要确定两个不同句子的顺序先后，因此段向量嵌入主要是通过［CLS］标识将输入文本的句子两两之间进行相连，从而达到区分的目的。位置向量嵌入主要用于区分文本序列中处于不同位置的单词各自间的语义信息的差别。

给定 x 是由 k 个单词所组成的一个输入的英文句子，该英文句子可被形式化表示为 $x = x_1 x_2 \cdots x_i \cdots x_k$ ，其中 $x_i (1 \leqslant i \leqslant k)$ 是第 i 个英文单词。

3.1.2　掩码语言模型预训练任务

掩码语言模型（MLM，masked language model）预训练任务主要目的在于让模型学习到句子的上下文特征，该任务随机地将文本中的单词用特殊标识［MASK］进行掩盖，然后在输入给模型进行预测被掩盖掉的单词。具体而言，我们需要对语料库中 15％的单词进行掩盖，然后通过

softmax 对掩盖的单词位置所输出的最终隐藏输出向量进行单词预测。掩码操作的例子如下：

原本的句子：After watching the movie，I think it is better than the one I saw last week.

掩盖后的句子：After watching the movie，I think it is [MASK] than the one I saw last week.

如果直接对语料库中 15%的单词进行掩盖，而如前所述输入向量中并没有包含 [MASK] 标识，因此需要想办法解决这个小问题。针对要使用 [MASK] 标识随机掩盖的单词，具体的策略是：1）其中的 80% 直接使用 [MASK] 去替代；2）用任意的单词替代原有单词的比例为 10%；3）而余下另外的 10%保持原样。举例如下：

1）After watching the movie，I think it is [MASK] than the one I saw last week.

2）After watching the movie，I think it is no than the one I saw last week.

3）After watching the movie，I think it is better than the one I saw last week.

3.1.3 下一个句子预测预训练任务

为了提升模型理解两个句子之间的语义联系，强化模型的长距离上下文语义信息的捕捉能力，需要在下一个句子预测类型的预训练任务里面，对给定文本中的两个句子的相邻关系。具体而言，我们需要从文本语料库中任意挑选两个句子 A 和 B 构成一个训练样本，按照 B 属于句子 A 下一句的语言逻辑关系的样本的数学概率为 50%（其标识为 IsNext），而从数据集中随机抽选句子 B 的数学概率为 50%（其标识为 NotNext）。

3.1.4 Bert 模型的输出

如图 6 所示，BERT 预训练语言模型的输出为隐藏状态向量或隐藏状态向量的时间步长序列，其数学表示如下：

$$\boldsymbol{h}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_i,\cdots,\boldsymbol{h}_k] \tag{9}$$

其中：k 为单词向量的维度（一般而言 $k=768$），i 的取值范围为 $1\leqslant i\leqslant k$。需要指出的是，英文文本的最大长度取值为 150，这么做是为了降低模型的计算复杂度和推理速度。

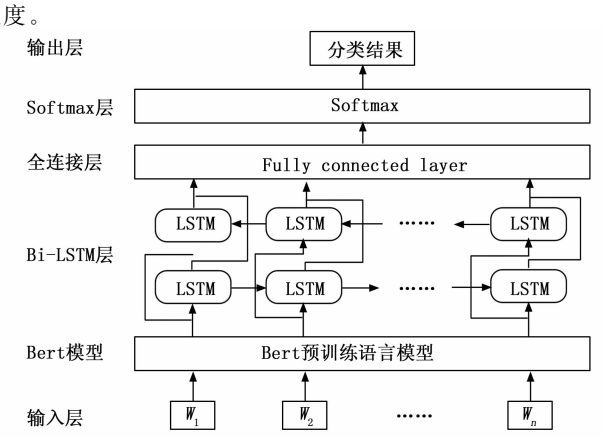

图 6 模型的整体结构

3.2 Bi-LSTM 层

如前所述，单方向的 LSTM 网络只能够捕捉一个方向的序列信息，不能做到同时对从左向右和从右向左的两个方向的上下文语义信息进行建模吗。因此本文结合从前往后和从后往前双方向的 LSTM 网络，即 Bi-LSTM 网络，来捕捉前后两个方向的全局语义信息。

如图 6 所示，本文利用 Bi-LSTM 网络作为深层语义特征抽取层，对 Bert 预训练语言模型输出的隐藏状态的时间步长序列向量 $\boldsymbol{h}$ 进一步对文本进行全局语义特征抽取。

具体而言，Bi-LSTM 网络层的真实输入为上一个时间步的隐层向量 h_{t-1} 和当前时间步输入 x_t 的拼接。需要注意的是，在第 0 个时间步中，上一个时间步的隐层向量和当前时间步的输入是相同的。在第 t 个时间步，Bi-LSTM 网络的输入为 $[\boldsymbol{h}_{t-1},x_t]$，那么从前往后方向的 LSTM 网络的隐藏状态输出向量为：

$$\overrightarrow{\boldsymbol{h}_t}=\sigma(\boldsymbol{W}_t[\overrightarrow{\boldsymbol{h}}_{t-1},x_t]+\boldsymbol{b}) \tag{10}$$

而从后往前的 LSTM 网络的隐藏状态的输出向量为：

$$\overleftarrow{\boldsymbol{h}_t}=\sigma(\boldsymbol{W}_t[\overleftarrow{\boldsymbol{h}}_{t-1},x_t]+\boldsymbol{b}) \tag{11}$$

其中：$\boldsymbol{W}_t$ 代表的是 LSTM 网络的权重向量矩阵，$\boldsymbol{b}$ 代表了网络中各层的偏置向量，$\overrightarrow{\boldsymbol{h}}_{t-1}$ 与 $\overleftarrow{\boldsymbol{h}}_{t-1}$ 分别表示前后两个序列方向的上一个时间步 $t-1$ 的隐藏状态向量输出。

最终，在第 t 个时间步，Bi-LSTM 层的隐藏状态向量为当前时间步 t 的由后往前和由前往后双方向产生的隐层向量的拼接结果，具体为公式（12）：

$$\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_{t-1},\overleftarrow{\boldsymbol{h}}_{t-1}]$$

经过所有的时间步，我们可以得到一个包含长距离型上下文语义信息隐层向量的集合，即 $\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_i,\cdots,\boldsymbol{h}_t\}(1\leqslant i\leqslant t)$。

3.3 全连接层

在将隐藏状态向量输入到分类器层之前，本文使用了一个全连接层作为线性层将句子的高级全局上下文语义信息表征向量处理为一个实值向量，其维度数目等于文本的类别总数 n。上述全连接层的处理过程的数学化表示为：

$$g(h_t)=\boldsymbol{W}_g\boldsymbol{h}_t+\boldsymbol{b}_g$$

其中：$\boldsymbol{W}_g$ 表示全连接层的权重向量矩阵，$\boldsymbol{b}_g$ 表示全连接层的偏置向量。

3.4 分类器层

一般而言，对于文本多分类任务，我们可以使用 softmax 函数输出各个文本类别的条件概率值的，其中的最大值则是预测输入文本所对应的标签类别。Softmax 层的条件概率值计算的数学公式为：

$$\hat{y}_i=\frac{e^{g_i}}{\sum_{k=1}^{n}e^{g_k}}$$

3.5 损失函数

为了更好地配合端到端的深度模型设计和训练，本文主要是利用交叉熵损失函数来评估样本的真值标签和模型输出的预测值之间的误差：

$$L(\theta) = -\frac{1}{N}\sum_{i}\sum_{j} y_i^j \log(\hat{y}_i^j) + \lambda \|\theta\|^2$$

其中：N 表示类别总数，i 与 j 分别表示文本索引和类别索引，y_i 表示的是模型输出的预测值，而 θ 是模型的待优化的参数集合，分别是各层的权重参数和偏置参数。

4　实验结果与分析

4.1　实验设置

本文主要在英伟达 3 080 ti 型号的显卡上利用 PyTorch 框架搭建模型，而后完成训练和推理实验，选用 VSCode 作为编程软件，Python 的版本为 V3.6.8。本文主要基于 Word2Vec 词向量来建立所有的词向量，其维度数量为 300，文本最大长度为 150，至于 Bi-LSTM 网络层的节点数量设置为 16。训练过程总的 Epoch 数量是 10，数据集的批次大小设置为 32。此外，所选 Adam 优化器的学习率的初值被这设定陈恒 0.001，而 $\lambda = 0.1$。为了更好地拟合模型的同时保持训练过程的稳定，失活概率为 0.3 的 Dropout 引入模型的训练。此外，Bert 预训练语言模型的词嵌入的维度默认为 768 维。

4.2　数据集和评价指标

本文主要使用下列 3 个常见的英文数据集来训练和测试模型：

1）MR 数据集：该数据集是基于简短电影评论文本构建而成的，由 Pang 和 Lee 两人于 2005 年首次应用到自然语言处理中的情感分类任务中。其中，训练集主要包含 5 331 条正面评论的负面样本以及 3 610 条正面样本。

2）SST-2 数据集：该数据集是 MR 数据集的变体。需要注意的是，非常积极和正向的评价文本被标记为正样本，而负面积极和极度负面的评价文本则会被标记为负面样本。总体上，训练集被划分为 3 310 条负面样本以及 3 610 条正面样本。

3）CoLA 数据集：CoLA 语料库是一个二元单句分类任务所使用的数据集。CoLA 数据集是被经过人工标注的，主要是针对语法的接受性。本文使用该语料库的公开获取版本，其中包含 8 551 条训练数据和 1 043 个测试数据 5，总共有 6 744 个正面样本和 2 850 个负面样本。该数据集的文本平均长度为 7.7 个单词。由于 CoLA 的测试集没有进行标注，因此本文从训练集中划分出 5%样本作为验证集，并使用原始验证集作为测试集。

在模型评价指标的选取上，本文主要使用准确率 Accuracy 去衡量和评估所设计的模型以及对比模型的有效性和性能。Accuracy 可通过如下公式进行计算：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

其中：TP 和 TN 分别代表 l 了输入的英文句子被正确地预估为正样本的数量和被错误地预估为正样本的数量，而 FN 和 TN 分别代表了被错误地预估为负样本的数量，以及被正确地预估为负样本的数量。

4.3　基准对比模型

为了充分地评估本文所设计的 Bert-Bi-LSTM 分类模型是否能够准确地提取到局部和全局的高级上下文语义信息，我们将该模型与多个基准模型进行了对比和评估实验。除了原本的 Bert 预训练语言模型直接应用于英文文本分类任务，本文还选取了几个相关的深度学习网络模型作为基准模型进行对比实验。关于选取的基准模型的详细信息说明如下：

1）SVM：为体现对比实验的全面性，本文特地选取了传统分类模型 SVM 作为一个基准对比模型。

2）MLP：传统模型多层感知机的隐藏层数量为 2，分别包含 512 和 100 个隐藏单元，该感知机使用了经过 TF 词频加权的词袋向量。

3）CNN-non-static：该模型的输入向量和 MLP 的保持一致，同时没有改变模型原本的训练参数。

4）LSTM：Bert 模型的隐层输出向量是单向长短期记忆网络的输入。

5）Bi-LSTM：双向 LSTM 网络的输入和单向的保持一致。

6）Bert：在本文的实验中，我们使用的是原始的预训练 Bert 模型。

为了保证对比实验是公平比较，所有的模型均是从零开始训练。

4.4　实验结果

本文在表 1 中提供了上述各个基准对比模型在 MR 数据集、SST-2 数据集与 CoLA 数据集上的准确率结果。从表 1 中列出的实验结果可以发现，基于深度学习的分类模型在 3 个数据集上的性能均一致高于传统的基于机器学习的分类模型。最为重要的是，经过实验对比可以看出，本文所提出的结合 Bert 与 Bi-LSTM 的模型的性能超越了所有的基准对比模型。特别地，本文的模型的性能同时优于单独利用那个 Bert 模型和 Bi-LSTM 网络去实现分类的情况，这个结果说明结合二者构建的分类模型性能优越，兼备二者的强大的上下文语义信息的挖掘能力和双向长距离依赖捕捉能力。

表 1　不同模型在 3 个数据集上的准确率结果

模型	MR	SST-2	CoLA
SVM	0.745	0.794	0.572
MLP	0.759	0.808	0.608
CNN- non-static	0.815	0.872	0.617
LSTM	0.804	0.859	0.612
Bi-LSTM	0.813	0.882	0.625
Bert	0.816	0.912	0.811
本文所提模型	0.862	0.915	0.832

在仅仅利用到局部语义信息的模型之中，Bert 模型超越了 SVM、MLP 与 LSTM 模型。此外，Bi-LSTM 模型的性能相比于单向的 LSTM 模型有所提升，这得益于 Bi-LSTM 同时捕捉到了前后两个方向的上下文信息。注意到，Bi-LSTM 模型在 MR 数据集上的性能是略低于 CNN- non-static 模型的，但是在结合 Bert 预训练语言模型之后，这在

此证明了 Bert 预训练模型强大的语义信息表达和提取能力对文本分类模型的性能提升很有帮助。

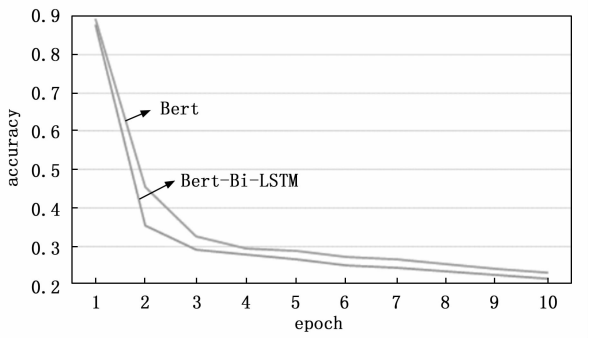

图 7 模型的训练曲线

4.5 消融实验

为了探究不同数据集中文本的不同的最大长度对模型性能的影响，本文将所提出的 Bert-Bi-LSTM 模型在 3 个数据集上均完成了消融实验。从表 2 中可以观察到，随着文本最大长度边长，本文所提的模型在 3 个数据集上的性能提升趋势基本一致，但是在长度越过 150 时，模型性能的增益变小，甚至出现了轻微下降的现象。因此为了在模型的性能与计算成本之间做出较好的权衡，本文选择了 150 作为最终的文本最大长度值。

表 2 不同的最大句子长度对本文所提模型性能的影响

文本最大长度	MR	SST-2	CoLA
120	0.858	0.912	0.828
130	0.859	0.912	0.829
140	0.860	0.913	0.830
150	0.862	0.915	0.832
160	0.861	0.916	0.831
170	0.861	0.915	0.832

5 结束语

针对英文文本分类任务中现存的主要问题和不足，本文主要提出了一种新颖的基于 Bert-Bi-LSTM 分类模型。相比于现有的分类方法，该模型能够通过 Bert 预训练语言模型引入外部的语言知识提升分类的准确性，同时还能基于 Bi-LSTM 网络对词嵌入向量进行前后两个方向的上下文语义依赖信息进行建模。充分的对比实验结果表明，本文结合 Bert 预训练元模型和 Bi-LSTM 网络搭建的模型在 3 个公共数据集上的性能显著胜于其他基准对比模型，其能够借助外部的语言知识去提升分类的准确率。此外，本文提出的算法同时利用 Bi-LSTM 网络来捕获双向的长距离型上下文语义依赖信息，大大提高了分类模型的特征提取与表示能力。

参考文献：

[1] 贾澎涛，孙 炜. 基于深度学习的文本分类综述 [J]. 计算机与现代化，2021 (7)：29－37.

[2] 万齐斌，董方敏，孙水发，等. 基于 BiLSTM-Attention-CNN 混合神经网络的文本分类方法 [J]. 计算机应用与软件，2020，37 (9)：94－98，201.

[3] 张铭泉，周 辉，曹锦纲，等. 基于注意力机制的双 BERT 有向情感文本分类研究 [J]. 智能系统学报，2022 (8)：1－9.

[4] 杨森淇，段旭良，肖 展，等. 基于 ERNIE＋DPCNN＋BiGRU 的农业新闻文本分类 [J/OL]. 计算机应用，2022：1－9 [2022－11－23]. http://kns.cnki.net/kcms/detail/51.1307.tp.20220805.1037.006.html.

[5] 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.

[6] 祝 亮. 基于 CNN 深度学习的自媒体文本分类方法的研究 [J]. 电脑知识与技术，2021，17 (21)：97－100.

[7] CONNEAU A，SCHWENK H，BARRAULT L，et al. Very deep convolutional networks for text classification [C] //Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics：Volume 1，Long Papers. 2017.

[8] MIYATO T，DAI A M，GOODFELLOW I，et al. Adversarial training methods for semi-supervised text classification [C] // Proc. 5th Int. Conf. Learn Represent (ICLR)，2017：1－12.

[9] 卢 健，马成贤，杨腾飞，等. Text-CRNN＋attention 架构下的多类别文本信息分类 [J]. 计算机应用研究，2020，37 (6)：1693－1696，1701.

[10] 李幼军，黄佳进，王海渊，等. 基于 SAE 和 LSTM RNN 的多模态生理信号融合和情感识别研究 [J]. 通信学报，2017，38 (12)：109－120.

[11] HUANG M，QIAN Q，ZHU X，et al. Encoding syntactic knowledge in neural networks for sentiment classification [J]. ACM Transactions on Information Systems (TOIS)，2017，35 (3)：1－27.

[12] 王文韬，张士豹. 基于情感词典和 SVM 的微博网民情感分析 [J]. 现代信息科技，2021，5 (24)：24－27，31.

[13] DEVLIN J，CHANG M W，Lee K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding [C] //Proceedings of NAACL-HLT. 2019：4171－4186.

[14] 吴汉瑜，严 江，黄少滨，等. 用于文本分类的 CNN _ BiLSTM _ Attention 混合模型 [J]. 计算机科学，2020，47 (S2)：23－27，34.

[15] YANG Z，YANG D，DYER C，et al. Hierarchical attention networks for document classification [C] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2016：1480－1489.

[16] LEI Z，YANG Y，YANG M，et al. A multi-sentiment-resource enhanced attention network for sentiment classification [C] //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics，2018 (2)：758－763.

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

(下转第 251 页)