文章编号：2095-2163(2023)07-0001-07 中图分类号：TP391 文献标志码：A

结合双预训练语言模型的中文文本分类模型

原明君，江开忠
（上海工程技术大学 数理与统计学院，上海 201620）

摘 要：针对 Word2Vec 等模型所表示的词向量存在语义模糊从而导致的特征稀疏问题，提出一种结合自编码和广义自回归预训练语言模型的文本分类方法。首先，分别通过 BERT、XLNet 对文本进行特征表示，提取一词多义、词语位置及词间联系等语义特征；再分别通过双向长短期记忆网络（BiLSTM）充分提取上下文特征，最后分别使用自注意力机制（Self_Attention）和层归一化（Layer Normalization）实现语义增强，并将两通道文本向量进行特征融合，获取更接近原文的语义特征，提升文本分类效果。将提出的文本分类模型与多个深度学习模型在 3 个数据集上进行对比，实验结果表明，相较于基于传统的 Word2Vec 以及 BERT、XLNet 词向量表示的文本分类模型，改进模型获得更高的准确率和 F_1 值，证明了改进模型的分类有效性。

关键词：预训练语言模型；双向长短期记忆网络；自注意力机制；层归一化

Chinese text classification model based on dual pre-trained language model

YUAN Mingjun, JIANG Kaizhong
(School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 To solve the problem of sparse features caused by Word2Vec models, a text classification method based on autocoding and generalized autoregressive pretrained language model is proposed. Firstly, BERT and XLNet are used to represent features such as polysemy, word location and relationship of the text respectively. Then, context features are extracted through Bi-directional Long Short-Term Memory (BiLSTM). Finally, features through self-attention and layer normalization of the two channel are fused to obtain the features closer to the original text and improve the effect of text classification. The proposed text classification model is compared with several deep learning models on three datasets. Experimental results show that compared with traditional text classification models based on Word2Vec, BERT and XLNet word vector representation, the improved model achieves higher accuracy and F_1, which proves the validity of the improved model.

【Key words】 pre-trained language model; Bi-directional Long Short-Term Memory; self-attention mechanism; layer normalization

0 引 言

文本分类是根据文本所蕴含的信息将其映射到预先定义的带主题标签的 2 个或几个类的过程，同时也是信息检索与数据挖掘的基础。文本分类通常包括特征表示、特征提取和分类三个主要步骤。特征表示是文本分类任务的首要阶段，也是文本分类的基础。Mikolov 等学者[1]和 Pennington 等学者[2]分别提出 Word2Vec 模型、全局向量（Global Vectors, GloVe）对文本特征进行表示。袁磊[3]和宋呈祥等学者[4]使用了改进的 CHI 特征选择方法对文本进行分析，但在模型训练时，存在利用文本上下文信息范围有限等问题。

为了改进文本信息表示不准确等问题，Devlin 等学者[5]和 Yang 等学者[6]分别提出基于 Transformers 的双向编码预训练语言模型（Bidirectional Encoder Representations from Transformers, BERT）和广义自回归预训练语言模型（Generalized Autoregressive Pretraining for Language Understanding, XLNet），进一步提升了词向量分类模型的性能。

基金项目：全国统计科学研究项目（2020LY080）。

作者简介：原明君（1996-），女，硕士研究生，主要研究方向：深度学习、文本分析、自然语言处理；江开忠（1965-），男，博士，副教授，主要研究方向：知识发现、搜索算法、文本挖掘。

通讯作者：江开忠 Email：kzjpub@163.com

收稿日期：2022-08-29

1 相关工作

随着深度学习的发展,深度学习模型被用于不同场景的文本分类任务中。Adhikari 等学者[7]在 BiLSTM 模型中设置正则化等步骤,陈立潮等学者[8]通过改造传统的 BiLSTM 模型,加入对抗训练等步骤,都达到了较好的文本分类效果。杨青等学者[9]将注意力机制与 BiGRU 相结合,提出 FFA-BiAGRU 文本分类模型,进一步提高了文本分类的准确性。

孙红等学者[10]用 BERT 训练词向量,通过 BiGRU 融合注意力机制进行特征提取,有效提高了模型分类的准确率。梁淑蓉等学者[11]通过 XLNet 获取词向量,利用 LSTM 结合注意力机制进行文本情感分析,进一步提高了模型分类预测的准确性。

通过对已有方法的深入学习,本文将预训练语言模型与 BiLSTM、自注意力机制、层归一化相结合,提出了一种结合自编码和广义自回归预训练语言模型的 DPT-BRNN(Dual Pre-trained-Bi-directional Recurrent Neural Network)文本分类模型。该模型综合考虑了不同词向量表达的优劣,分别通过 XLNet、BERT 来构建输入文本的动态语义词向量,并将其作为新的特征向量分别输入 BiLSTM 提取文本隐藏特征,获取语义间上下文依赖关系,同时与自注意力机制和层归一化连接,使文本分类的权重分配更加合理,从而提高文本分类效果。

2 结合双预训练语言模型的中文文本分类模型的构建

目前现有的主流预训练语言模型分为自回归语言模型(Auto Regression Language Model, AR)和自编码语言模型(Auto Encoding Language Model, AE)。其中,AR 模型主要用于评估文本的概率分布,该模型为单向模型;AE 模型为双向模型,可以学习到文本的上下文深层语义信息。

BERT 作为 AE 模型的典型成功案例,XLNet 作为改进的 AR 模型,本文将这 2 个模型应用到 DPT-BRNN 文本分类模型中,其模型结构如图 1 所示。该模型由 2 个通道组成。第一个通道通过 XLNet 对文本数据进行语义特征提取,构建基于词向量的特征表示;第二个通道使用 BERT 预训练语言模型,获取文本整体信息的词向量,然后将新特征输入 BiLSTM 以获取上下文语义表示。两通道都引入自注意力机制层和层归一化捕获文本的特征信息,最终将双通道的文本特征信息进行融合,输入输出层达到文本分类的目的。

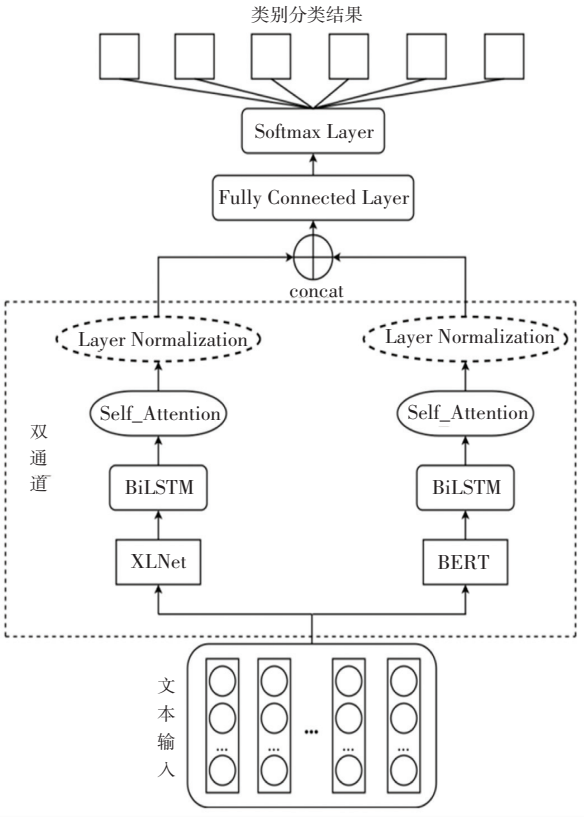

图 1 DPT-BRNN 文本分类模型

Fig. 1 DPT-BRNN text classification model

2.1 广义自回归语言模型

广义自回归语言模型(XLNet)提出的排列语言模型(Permutation Language Model, PLM)解决了传统自回归语言模型无法根据上下文预测结果的问题。假设当前输入句子的单词构成为 $[x_1, x_2, x_3, x_4, x_5]$,若要预测单词 x_3,使用 PLM 随机打乱单词排列顺序,依次运用自回归方法预测 x_3。PLM 部分概览图如图 2 所示。

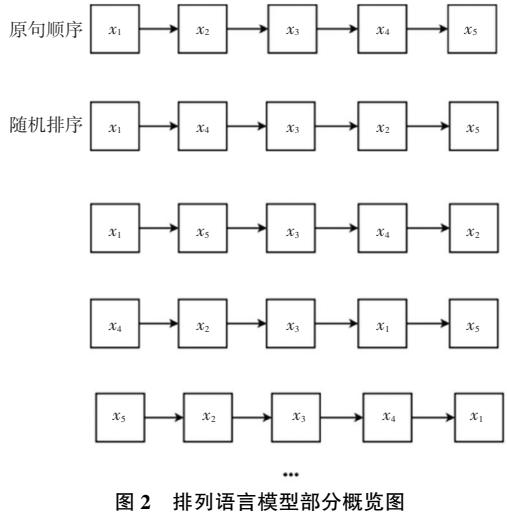

图 2 排列语言模型部分概览图

Fig. 2 Partial overview diagram of Permutation Language Model

为实现排列语言模型的基本思想。XLNet 采用基于目标感知表征的双流自注意力掩码机制，即内容流自注意力（Content Stream Attention）和查询流自注意力（Query Stream Attention）。同时，模型利用改进的注意力掩码（Attention Mask）使模型读取上下文信息。双流自注意力计算流程如图 3、图 4 所示。

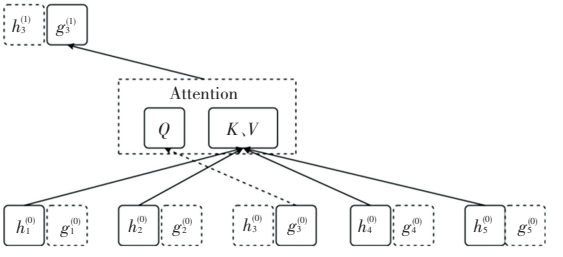

图 3　查询流自注意力计算

Fig. 3　Query stream attention calculation

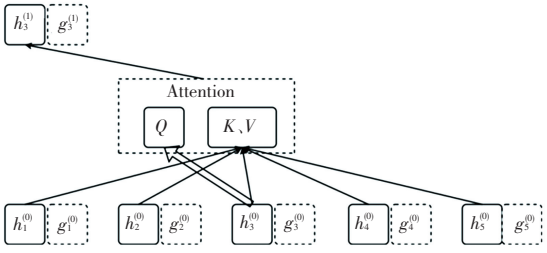

图 4　内容流自注意力计算

Fig. 4　Content stream attention calculation

图 3、图 4 中，g 表示各单词的位置信息，h 表示各单词的内容信息，$g_3^{(0)}$ 为被预测词、即 x_3 的位置信息。

XLNet 还应用到了 Transformer-XL 中的片段循环机制（Recurrence Mechanism）和相对位置编码（Relative Positional Encoding）。片段循环机制以分段的形式进行建模，文本序列引入循环机制后实现隐藏层信息循环传递的方式如图 5 所示。

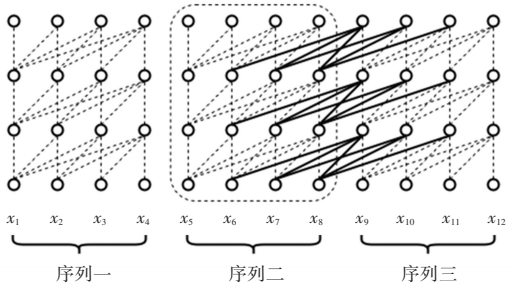

图 5　片段循环机制信息传递

Fig. 5　Recurrence mechanism information transmission

由图 5 可知，在对每个序列进行处理时，隐藏层从 2 个部分进行学习。一个是当前序列前面节点的输出，即图 5 中每个序列的虚线部分；另一个是当前序列之前序列节点的输出，即图 5 中实线部分。

2.2　自编码语言模型

自编码语言模型（BERT）模型体系结构主要由多层 Transformer 模型结构组成，模型通过注意力机制将任意 2 个位置的单词距离转换为 1，有效地解决了 NLP 任务中出现的长期依赖问题。BERT 使用双向 Transformer 网络架构中的编码器（Encoder）搭建整个模型框架，同时，其模型框架由多层的 Transformer 模型结构组成，研究得出的网络结构模型如图 6 所示。

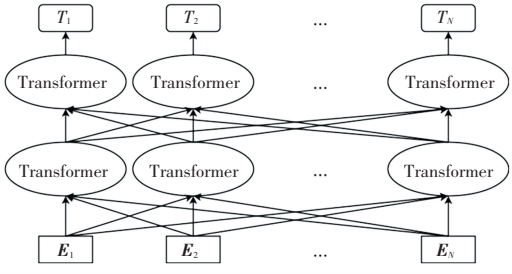

图 6　BERT 网络结构模型

Fig. 6　BERT network structure model

模型结构中的字向量 $\boldsymbol{E}_1$、$\boldsymbol{E}_2$、…、$\boldsymbol{E}_N$ 不仅包含当前文本的字符级向量（Token Embedding），而且也包括了每个字的位置向量（Position Embedding）和分段向量（Segment Embedding）。模型将 3 个向量相加求和之后，分别在文本的开头和结尾加上 CLS 和 SEP 的标记符号，然后输入双向 Transformer 编码器中，完成对每个字的双向编码表示。

2.3　BiLSTM + Attention

LSTM 由 t 时刻的输入向量 $\boldsymbol{x}_t$、单元状态 $\boldsymbol{c}_t$、临时单元状态 $\tilde{\boldsymbol{c}}_t$、隐层状态 $\boldsymbol{h}_t$、遗忘门的节点操作 $\boldsymbol{f}_t$、输入门 $\boldsymbol{i}_t$ 和输出门 $\boldsymbol{o}_t$ 组成。模型 t 时刻整个过程的计算方法公式具体如下：

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, x_t] + \boldsymbol{b}_f) \tag{1}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, x_t] + \boldsymbol{b}_i) \tag{2}$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c \cdot [h_{t-1}, x_t] + \boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \cdot \tilde{\boldsymbol{c}}_t \tag{4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, x_t] + \boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \cdot \tanh(\boldsymbol{c}_t) \tag{6}$$

虽然 LSTM 模型能够很好地捕捉较长距离的文本依赖关系，但仍无法学习文本从后向前的语义信息。因此，本文使用 BiLSTM 模型学习上下文双向的语义信息，同时在 BiLSTM 层之后加入自注意力机制层对词语权重进行重新分配。BiLSTM 结合自注意力机制的模型结构如图 7 所示。

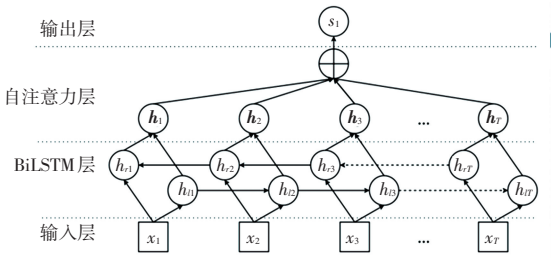

图 7 BiLSTM-Attention 模型图

Fig. 7 BiLSTM-Attention model diagram

由图 7 可知,将 BiLSTM 模型中的最后一个时序的输出向量 $\boldsymbol{h}_t$ 作为输入自注意力层的特征向量,其计算过程数学公式见如下:

$$\boldsymbol{h}_t = [h_{lt}, h_{rt}] \tag{7}$$

$$\boldsymbol{u}_t = \tanh(\boldsymbol{W}_s \boldsymbol{h}_t + \boldsymbol{b}_s) \tag{8}$$

$$\alpha_t = Soft\max(\boldsymbol{u}_t^{\mathrm{T}}, \boldsymbol{u}_s) \tag{9}$$

$$s_1 = \sum_t \alpha_t h_t \tag{10}$$

其中,$\boldsymbol{u}_t$ 为 $\boldsymbol{h}_t$ 的自注意力隐层表示;$\boldsymbol{W}_s$ 为权值矩阵;$\boldsymbol{b}_s$ 为偏置项;α_t 为 $\boldsymbol{u}_t$ 通过 *Softmax* 函数后得到的整个序列的归一化权值。

2.4 归一化机制

为了提高模型的学习能力和表达能力,模型加入层归一化(Layer Normalization)机制对上层输出向量进行归一化处理。具体操作为根据以下公式将模型某一层所有神经元输入进行处理:

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} \alpha_i^l \tag{11}$$

$$\sigma^l = \sqrt{\frac{1}{H}\sum_{i=1}^{H} (a_i^l - \mu^l)^2} \tag{12}$$

$$\hat{\alpha}^l = \frac{\alpha^l - \mu^l}{\sigma^l} \tag{13}$$

$$y = g \cdot \hat{\alpha}^l + b \tag{14}$$

其中,μ^l、σ^l 分别表示模型各层神经元的均值和方差统计量;H 表示模型各层神经元的节点数;α_i^l 表示模型第 l 隐藏层的输出;y 表示模型上层输出经过归一化处理后的输出值;g 和 b 分别表示对输入向量的缩放和平移。

将归一化后的数据进行非线性激活函数运算,即:

$$h = f(y) \tag{15}$$

2.5 输出层

模型将文本序列经过 XLNet 通道和 BERT 通道后的特征向量进行融合,输入全连接层,然后利用 *Softmax* 分类器完成对文本内容的分类。*Softmax* 分类器输出类别分类概率的计算公式为:

$$P(y^{(i)} = j \mid x^{(i)}; \theta) = \frac{\exp(\theta_j^{\mathrm{T}} x^{(i)})}{\sum_{n=1}^{k} \exp(\theta_n^{\mathrm{T}} x^{(i)})} \tag{16}$$

其中,P 表示文本 x 分类至类别 j 的概率数值,θ 表示模型训练的参数。

3 实验结果与分析

3.1 实验环境与数据

本文的实验环境见表 1。

表 1 实验环境设置

Tab. 1 Setup of experimental environment

实验环境	详细信息
操作系统	Windows10
CPU	Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40 GHz
内存	28 G
显卡	NVIDIA GeForce RTX 3060
开发语言	Python 3.8.8
开发平台	Pytorch 1.9.1

本文选取 THUCNews 新闻数据集、搜狐新闻标题数据集[12]以及微博数据集,共 3 个文本数据集测试 DPT-BRNN 文本分类模型的分类效果。THUCNews 新闻数据集由清华大学提供并公开,包含财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐共 10 个类别新闻数据。搜狐新闻标题数据集由搜狗实验室提供,包括财经、健康、教育、军事、科技、汽车、时尚、体育、文化、娱乐共 10 个类别新闻数据。微博数据集来源于新浪微博评论,其中包含时政、社会、财经三个类别新闻数据。

本文将 3 个新闻文本数据集随机分为训练集、测试集和验证集,数据集概况见表 2。

表 2 数据集统计表

Tab. 2 Statistical table of datasets

数据集	总样本	训练集数	验证集数	测试集数	类目
THUNews	90 000	50 000	20 000	20 000	10
搜狐	65 000	50 000	5 000	10 000	10
微博	90 000	50 000	20 000	20 000	3

3.2　模型参数设置

本文 XLNet 采用哈工大讯飞联合实验室发布的 xlnet-base-cased[13] 模型，BERT 采用 Google 发布的预训练模型进行文本表示，Word2Vec 词向量使用 Skip-Gram 模型训练，得到维度为 300 维的词向量。模型的参数设置见表 3、表 4。

表 3　预训练模型参数

Tab. 3　Parameters of pre-trained models

参数名称	XLNet	BERT
Encoder 层数	12	12
隐藏层尺寸	768	768
自注意力机制头数	12	12

表 4　模型参数设置

Tab. 4　Setup of model parameters

参数名称	设置值
Adam 优化器学习率	1e-5
句子长度	34
Epoch	10
Batch_size	32
Dropout	0.5

3.3　评价指标

大多数分类模型的评估标准是：准确率（*Accuracy*）、精确率（*Precision*）、F_1 值（*F-measure*）以及召回率（*Recall*）。相关的混淆矩阵结构见表 5。

表 5　混淆矩阵

Tab. 5　Confusion matrix

预测值	实际值	
	Positive	Negative
Positive	*TP*	*FP*
Negative	*FN*	*TN*

其中，准确率表示正确预测的样本相对于所有分类样本的比重，计算公式如下：

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

精确率是指预测结果为正例的样本中，被正确预测为正样本的比例，计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{18}$$

召回率表示正确预测结果为正样本占全样本中实际正样本的比重，计算公式如下：

$$R = \frac{TP}{TP + FN} \tag{19}$$

F_1 是精确率和召回率的加权平均，计算公式如下：

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{20}$$

3.4　实验与分析

3.4.1　与预训练词向量模型的对比

与预训练词向量模型的对比实验结果见表 6。由实验结果可以看出，相较于使用 2 种单一的预训练词向量模型，同时使用 XLNet 和 BERT 的模型分类效果有所提升。本文模型在 3 个实验数据集上都取得了较好的分类效果，准确率分别达到了 83.68%、79.41% 和 98.72%；BERT-全连接模型在 3 个数据集上的准确率分别为 81.02%、78.02% 和 96.43%；XLNet-全连接模型在 3 个数据集上的准确率分别为 82.28%、78.40% 和 97.39%。本文模型设置双通道，在每个通道中分别使用 XLNet 和 BERT 预训练模型，同时又将预训练模型向量输入 BiLSTM、BiGRU 模型，并通过自注意力机制获取局部特征，归一化加速模型训练速度，最终将两通道输出信息进行拼接，从而使模型学习到充足的文本特征信息，并获得更加丰富的文本特征，同时又提高了模型的训练精度。因此，无论准确率、还是 F_1 值，本文模型均优于其它模型。其中，本文模型较 BERT-全连接模型准确率在 3 个数据集上分别提高了 2.66%、1.39% 和 2.29%，较 XLNet-全连接模型准确率在 3 个数据集上分别提高了 1.40%、1.01% 和 1.33%，证明了本文模型的分类有效性。

表 6　与预训练词向量模型的对比实验结果

Tab. 6　Experimental results compared with pre-trained word vector models　%

文本分类模型	THUNews 数据集		搜狐数据集		微博数据集	
	准确率	F_1 值	准确率	F_1 值	准确率	F_1 值
BERT	81.02	76.37	78.02	73.43	96.43	66.81
XLNet	82.28	78.22	78.40	74.31	97.39	70.50
DPT-BRNN	**83.68**	**79.77**	**79.41**	**75.00**	**98.72**	**82.50**

3.4.2 与其它词向量模型的对比

为了进一步验证 DPT-BRNN 模型的有效性,本文还对比多个在相同数据集上效果较好的基于 Word2Vec 词向量的先进模型,对比模型包括:TextCNN、BiLSTM、BiGRU、BiLSTM-CNN、BiLSTM-Attention[14] 和 MCCL[15]。模型对比结果见表 7。

表 7 与传统词向量模型的对比实验结果

Tab. 7 Comparative experimental results of traditional word vector models %

文本分类模型	THUNews 数据集		搜狐数据集		微博数据集	
	准确率	F_1 值	准确率	F_1 值	准确率	F_1 值
TextCNN	78.94	72.71	73.81	69.20	90.87	80.67
BiLSTM	79.36	72.77	74.72	69.46	90.55	81.53
BiGRU	79.10	77.62	74.46	69.02	90.69	81.71
BiLSTM-CNN	79.72	73.63	75.96	70.27	92.14	85.54
BiLSTM-Attention	80.04	78.89	76.45	72.04	92.05	85.02
MCCL	79.60	78.28	75.42	70.65	94.45	90.20
DPT-BRNN	**83.68**	**79.77**	**79.41**	**75.00**	**98.72**	**82.50**

由实验结果可以看出,对比基于 Word2Vec 词向量的模型,本文模型无论准确率、还是 F_1 值都有所提高。但模型在微博数据集上的实验结果 F_1 值不如部分其他模型,一方面一定程度上可能是微博数据集数据不平衡所致,另一方面可能是词向量维度过高,部分模型出现过拟合所致。对比 TextCNN 模型,本文模型在 3 个数据集上准确率分别提高了 4.74%、5.60%和 7.85%。对比 BiLSTM 模型,本文模型在 3 个数据集上准确率分别提高了 4.32%、4.69%和 8.17%。对比 BiGRU 模型,本文模型在 3 个数据集上准确率分别提高了 4.58%、4.95%和 8.03%。BiLSTM 模型、BiGRU 模型在 3 个数据集上的训练结果大多较 TextCNN 模型优异,这是由于 BiLSTM 模型可以学习到句子的正向和逆向信息,从而能够更好地捕捉到上下文文本信息。对比 BiLSTM-CNN 模型,本文模型在 3 个数据集上准确率分别提高了 3.96%、3.45% 和 6.58%。BiLSTM-CNN 模型较 BiLSTM 模型在 3 个数据集上准确率分别提高了 0.36%、1.24% 和 1.59%。BiLSTM-CNN 模型较 TextCNN 模型在 3 个数据集上准确率分别提高了 0.78%、2.15%和 1.27%。对比 BiLSTM、TextCNN 和 BiLSTM_CNN 模型的实验结果,发现 TextCNN 与 BiLSTM 结合的网络结构能更有效地提取文本中的关键特征,提升分类准确率。对比 BiLSTM-Attention 模型,本文模型在 3 个数据集上准确率分别提高了 3.64%、2.96% 和 6.67%。BiLSTM-Attention 模型较 BiLSTM 模型在 3 个数据集上准确率分别提高了 0.68%、1.73%和 1.50%,引入注意力机制使得文本中每个词语的权重值得到重新分配,关键特征的语义信息更加明确,从而提升了 BiLSTM 模型读取上下文关键语义信息的能力。将本文所提模型与 MCCL 模型的分类结果进行比较,本文模型的准确率在 3 个数据集上分别提高了 4.08%、3.99%和 4.27%,这表明本文使用预训练语言模型提取词向量作为 BiGRU 与 BiLSTM 模型的输入,更充分地发挥了深度学习模型对文本特征的提取能力,并且同时在两通道分别引入自注意力机制层和归一化层对通道输出的特征分布进行调整,增强了模型的学习能力,有效地提升了模型分类的准确性。

4 结束语

本文结合预训练语言模型 XLNet、BERT、BiLSTM、自注意力机制以及层归一化,提出了双通道 DPT-BRNN 文本分类模型。模型首先将文本信息分别通过 XLNet、BERT 模型输出具有多样化信息的词向量,接着将词向量信息分别输入 BiLSTM,使其对文本序列信息进行捕捉学习,提取文本不同层次的上下文语义信息。然后,利用自注意力机制对文本深层次序列信息进行再提取,从而得到更加准确的文本关键语义信息,并将经过两通道的信息融合,得到更丰富的文本语义表示。通过将提出的文本分类模型与其他 8 种文本分类模型在 3 个数据集上进行对比分析,结果表明,本文所提出的文本分类模型在中文数据集上取得了比较好的分类结果,验证了本文模型分类的准确性和有效性。

在接下来的研究中,将通过扩充语料库来不断提升模型的分类性能,并着重从词语的语义扩展以及模型结构、参数方面进行优化改进,从而进一步提高模型学习效果的准确性,并降低训练过程的时间成本。

(下转第 14 页)