◆计算科学◆

基于 XLNet + BiGRU + Att (Label)的文本分类模型*

刘柏霆[1],管卫利[1,2**],李陶深[1,2]
(1.广西大学计算机与电子信息学院,广西南宁 530004;2.南宁学院数字经济学院,广西南宁 530299)

摘要:传统的词向量嵌入模型,如 Word2Vec、GloVe 等模型无法实现一词多义表达;传统的文本分类模型也未能很好地利用标签词的语义信息。基于此,提出一种基于 XLNet + BiGRU + Att (Label)的文本分类模型。首先用 XLNet 生成文本序列与标签序列的动态词向量表达;然后将文本向量输入到双向门控循环单元(BiGRU)中提取文本特征信息;最后将标签词与注意力机制结合,选出文本的倾向标签词,计算倾向标签词与文本向量的注意力得分,根据注意力得分更新文本向量。通过对比实验,本文模型比传统模型在文本分类任务中的准确率更高。使用 XLNet 作为词嵌入模型,在注意力计算时结合标签词能够提升模型的分类性能。

关键词:文本分类 XLNet BiGRU 标签词 注意力机制

中图分类号:TP391 **文献标识码**:A **文章编号**:1002-7378(2022)04-0412-08
DOI:10.13657/j.cnki.gxkxyxb.20221209.010

文本分类[1]是自然语言处理中的一项基本任务,被广泛应用在垃圾邮件识别、情感分析、文档主题分类等场景,提高其分类准确率是人们研究的重点。在文本分类任务中,词向量是深度学习模型中词语的基本表达形式,词向量语义的精确度直接影响文本最终的分类准确率。此外,注意力机制是深度学习模型中常用的特征选择工具,注意力打分的质量会决定模型的侧重点,从而影响分类的准确率。因此,本文拟从词向量与注意力机制出发,改进文本分类模型中的词嵌入模型与注意力机制,使模型的分类准确率提高,对于提高文本检测与分类任务的性能具有一定的现实意义。

对于现有的词嵌入模型,其生成的词向量分为静态与动态两种形式,静态的词向量训练模型有 Word2Vec (Word to Vector)[2]、GloVe (Global Vectors)[3]等。如方炯焜等[4]提出 GloVe-GRU 的模型结构,用 GloVe 作为词嵌入模型,生成 GloVe 的全局静态词向量并降低向量空间维度,这种词嵌入方式产生的静态词向量并不能处理一词多义的问题。动态的词向量嵌入模型是指模型生成的词向量是动态变化的,如果上下文改变,那么词向量也会跟着改变,同一个单词在语境中具有不同的词向量表达。常

收稿日期:2022-04-15 修回日期:2022-07-11
*国家自然科学基金项目(61762010)和广西科技计划项目(桂科 AD20297125)资助。

【作者简介】
刘柏霆(1999-),男,在读硕士研究生,主要从事自然语言处理研究。
【**通信作者】
管卫利(1977-),男,教授,硕士生导师,主要从事自然语言处理等研究,E-mail:2113391032@st.gxu.edu.cn。

【引用本文】
刘柏霆,管卫利,李陶深.基于 XLNet+BiGRU+Att(Label)的文本分类模型[J].广西科学院学报,2022,38(4):412-419.
LIU B T,GUAN W L,LI T S. Text Classification Model Based on XLNet+BiGRU+Att (Label) [J]. Journal of Guangxi Academy of Sciences,2022,38(4):412-419.

用的动态词向量嵌入模型有 ELMo(Embedding from Language Model)[5]、BERT[6]等。赵亚欧等[7]把 ELMo 与 Transformer 结合起来,将 ELMo 作为词嵌入模型,输出词向量到 Transformer[8]模型中做文本分类,但是 ELMo 内部采用长短期记忆网络(Long Short-Term Memory,LSTM)[9]结构来提取特征信息,LSTM 结构在提取特征的能力上较弱,相比于具有注意力机制的 Transformer 结构存在差距。除此之外,ELMo 模型内部只是将特征双向地拼接在一起,这种拼接方式未能很好地融合文本的语义信息。后来人们开始用 BERT[10]作为词向量预训练模型,如黄泽民等[11]用 BERT 作为词向量嵌入模型,将静态词向量输入到 BERT 中,利用 BERT 进一步优化静态词向量,赋予其动态词向量的特征,但是 BERT 也存在一些缺陷,如 BERT 在其训练阶段会随机遮掩掉一些词,而在下游任务的微调中却看不到这些被遮掩的词,导致两阶段差异,致使 BERT 的性能有一定程度的下降。XLNet[12]针对 BERT 现有的问题进行了改进,采用 XLNet 能够生成动态的词向量,同时也能生成比 BERT 更精准的词向量,因此本文采用 XLNet 作为词嵌入模型。

注意力机制[13]是文本分类任务常用的方法,其能够有效地根据文本信息的重要程度为文本分配权重。现有的文本分类模型所使用的注意力机制没能利用好标签词所蕴含的语义信息,如杨兴锐等[14]在 BiLSTM-CNN 混合模型的基础上加入残差连接与注意力机制,对卷积输出的向量计算注意力得分。叶瀚等[15]融合注意力机制与句向量,将首句子向量作为查询向量 Query,将其余各句子向量作为待查向量 Key。对 BERT 模型输出的 CLS 向量进行注意力点乘计算,通过注意力分数得出的权重系数对各 CLS 向量序列求和再平均来达到压缩编码的目的。梁顺攀等[16]提出一种基于混合神经网络的文本分类方法,对传统的卷积神经网络进行改进,增强对文本局部特征的提取能力,采用普通的自注意力机制对最终的文本向量分配权重。以上模型在做注意力计算时,均未利用标签词提取特征。若能很好地利用标签词,在注意力权重分配上就会更有目的性。如在新闻文本分类任务中,以“军事”为标签词样本,文本里就会出现“坦克”“军演”等词语;在情感分类任务中,以“积极”为标签词样本,文本中就会出现“开心”“高兴”等词语。利用标签词对比文本中的内容,找出文本中与标签词语义相近的部分并给其分配较高的权重,在分类层中使模型更关注这些重点部分,最后就能够得到更准确的分类预测。

通过以上对现有文本分类模型的分析,本文提出一种基于 XLNet + BiGRU + Att (Label)的文本分类模型。首先,用 XLNet 代替 BERT、GloVe、Word2Vec 等模型作为预训练模型;然后,将标签词与注意力机制相结合,充分利用标签词的语义信息匹配出文本向量中的重点部分,以提高模型的分类性能。

1 方法描述

1.1 XLNet + BiGRU + Att (Label)模型

XLNet + BiGRU + Att (Label)模型可大致分为文本预处理层、XLNet 层、双向门控循环单元(BiGRU)层、注意力(Attention)层、全连接层(Linear) + Softmax 层。XLNet + BiGRU + Att (Label)模型结构如图 1 所示。模型的总体设计思想是采用 XLNet 生成文本序列与标签词序列的动态词向量表达;用 BiGRU 进一步提取出文本向量的全局特征;根据全局特征选出当前文本的倾向标签词;在注意力机制的基础上,根据倾向标签词针对性地对文本的全局特征向量进行重点划分。

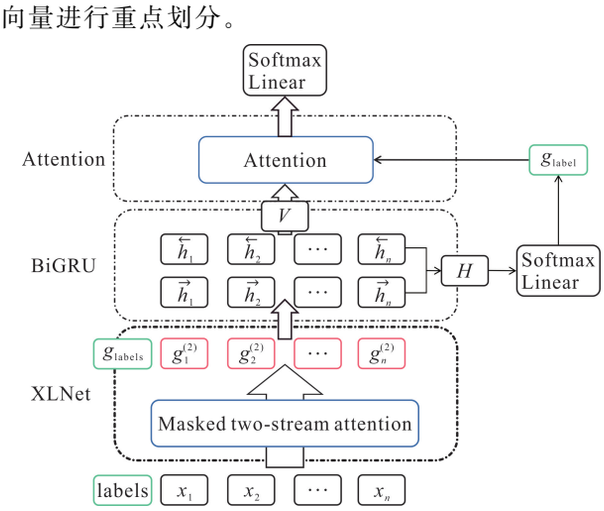

图 1 XLNet + BiGRU + Att (Label)模型结构

Fig. 1 Structrue of XLNet + BiGRU + Att (Label) model

模型处理的步骤大致如下。

步骤 1:进行数据预处理操作,将每个训练样本裁剪或填充为统一长度,对文本数据进行噪声去除并调用 Tokenizer 工具分词,经过处理后的文本连同标签词输入到 XLNet 模型中。

步骤 2:XLNet 处理输入的文本和标签词序列,生成动态词向量的文本与标签词序列,输入的文本长度为 512,输出的向量维度为 768。将文本向量输入

BiGRU 中。

步骤 3：BiGRU 将根据文本的上下文提取语义特征，将 BiGRU 最后时刻的隐藏层堆叠作为当前文本的总结向量 H，即考虑了文本里每个单词语义的总结向量。将 BiGRU 的最后一层输出拆分为前向输出与反向输出，并将其相加作为文本的内容向量 V。

步骤 4：将总结向量 H 输入到 Linear＋Softmax 中，预测出当前文本的倾向标签词。将倾向标签词与内容向量 V 输入 Attention 层，计算出内容向量对应倾向标签词的注意力权重，根据注意力权重更新文本的内容向量。

步骤 5：根据更新后的内容向量 C 得出文本的预测结果。

1.2　词嵌入层(XLNet)

1.2.1　排列语言模型

自回归语言(Autoregressive Language，AR)[17]模型只能实现单向的预测，典型的代表就是生成式预训练模型(Generative Pre-Train Model，GPT)。自编码语言(Autoencoder Language，AE)模型虽然实现了对文本的双向预测，但是其引入的 MASK 机制导致模型在预训练与微调阶段存在不一致的问题。针对以上两个问题，Yang 等[12]在 XLNet 中引进了两个新的概念：排列语言模型(Permutation Language Model，PLM)与双流自注意力机制(Two-Stream Self-Attention)。

PLM 的编码思想是把自回归语言模型和自编码语言模型结合起来，在中间加入一个称之为“排列(Permutation)”的步骤，该步骤使模型能够对文本进行双向预测，具体实现方法是把文本序列打乱，然后将末尾若干个词给遮掩掉。Yang 等[12]通过对比实验发现掩盖掉的词数接近文本的 15%时效果最好，这恰恰对应了 BERT 的 15%掩盖率。在计算时，将词按照打乱后的排列顺序，采用自回归的方式逐个预测。如图 2 所示，假设原序列初始为[1,2,3,4]，将该序列打乱后的排序方式为[2,4,3,1]与[4,3,1,2]。图 2(a)中，如果预测 3，则根据排列顺序在 3 前面的 2 和 4 来预测，预测 3 的概率为 $p(3)=p(2)\times p(2|4)\times p(3|4,2)$。同理预测图 2(b)中 3 的概率为 $p(3)=p(3|4)$。由此可见，简单的打乱排序方式使得同一序列从同样的方向预测却同时考虑到了前文与后文，这样的处理方式保留了序列的上下文信息，也避免了采用[MASK]标记位，解决了以往模型只能单向预测的问题。PLM 的打乱方式不是真正的将文本序列的排序打乱，而是生成一个掩码矩阵，如图 3[12]中 Attention Masks 所示，在 Transformer 中的 Attention 计算时把不需要的信息掩盖掉，相当于在预测时有意地让一些词发挥作用或不发挥作用，来达到打乱顺序的目的。

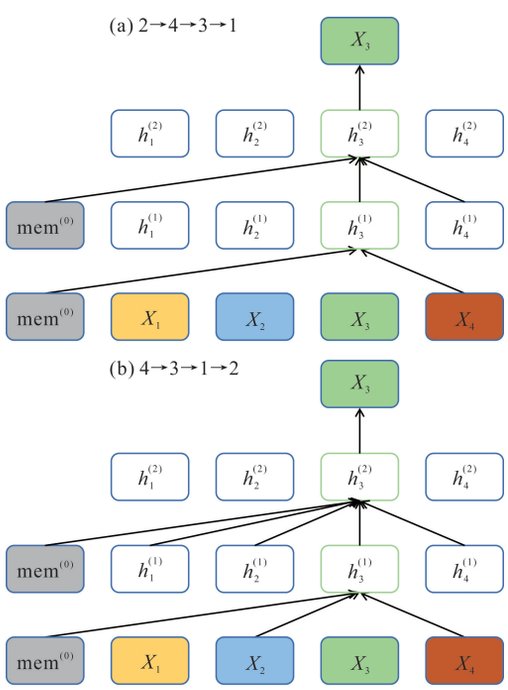

图 2　不同排列顺序的预测过程

Fig. 2　Prediction process of different order

对于长度为 T 的输入序列，其排列方式总共有 $T!$ 种。如输入长度为 5 的序列，那么就会有 5! 共 120 种排列方式，当序列过长时，就会导致模型的计算过于复杂。因此 Yang 等[12]通过式(1)对序列的各排列方式进行挑选。

$$\max_{\theta} E_{z\sim Z_T}\left[\sum_{t=1}^{T}\log p_\theta(x_{z_t}\mid X_{Z<t})\right], \tag{1}$$

其中，Z_T 为输入序列的所有排列方式；T 为输入序列的长度；z 为 Z_T 中的排列方式之一；z_t 为在 t 位置上 z 对应的值；$E_{z\sim Z_T}$ 为对所有排列方式求期望；$p_\theta(x_{z_t}|X_{Z<t})$为排列方式 z 上 t 位置词的预测。

基于式(1)对序列的所有排列方式求期望，根据期望值来选取模型中最优的排列序列、去除不合适的排列序列，以降低模型的计算复杂度。

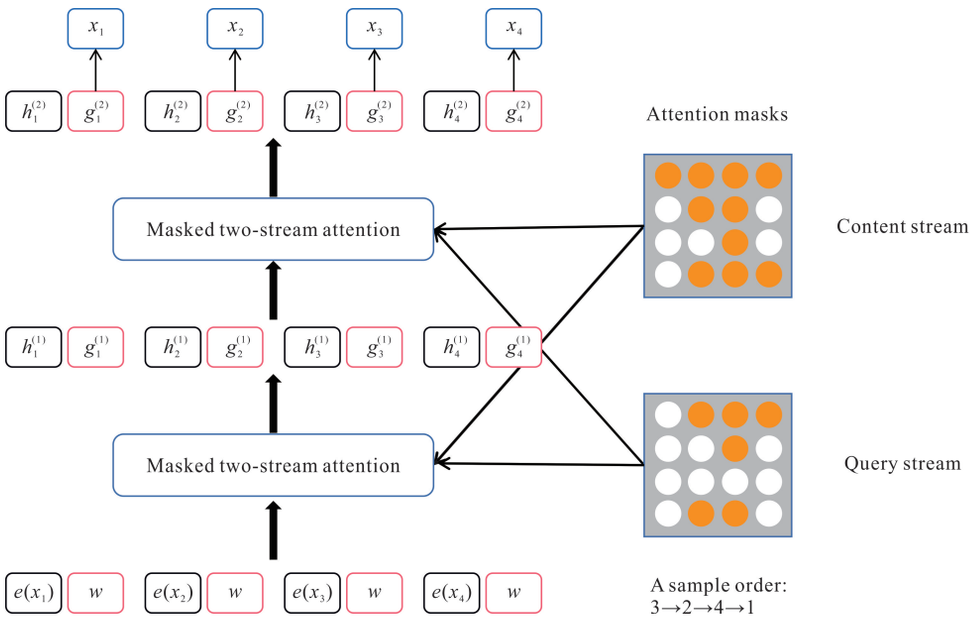

图 3　双流自注意力机制实现图[12]

Fig. 3　Implementation diagram of two-stream self attention[12]

1.2.2　双流自注意力机制

PLM 的计算过程存在一个问题,即由于输入词的顺序被打乱,模型无法知道所预测词对应原始序列的位置。传统的注意力机制把位置信息编码在 Token 里,如果用传统的注意力机制计算 PLM,则模型就无法看到被遮掩的词的位置信息。基于此,Yang 等[12]提出一种双流自注意力机制,将位置信息 g_θ 加入 AR 模型的目标函数中,如式(2)所示。

$$p_\theta(X_{z_t} = x \mid x_{z<t}) = \frac{\exp[e(x)^{\mathrm{T}} g_\theta(x_{z<t}, z_t)]}{\sum_{x'} \exp[e(x')^{\mathrm{T}} g_\theta(x_{z<t}, z_t)]}, \tag{2}$$

其中,x 表示当前要预测的词,X 表示预测词序列;$e(x)^{\mathrm{T}}$ 是当前输入的词向量的转置;$g_\theta(x_{z<t}, z_t)$ 是 Yang 等[12]在 XLNet 中提出的新概念,将 z_t 与在 t 位置前的词作为其输入值,在双流注意力机制中将 g_θ 作为查询向量 Q,其本身不包含内容信息,只包含位置信息。

“双流”即 Query stream 和 Content stream。Query stream 可以看到当前词的位置信息,不能看到其内容信息,而 Content stream 既可以看到当前词的内容信息也可以看到其位置信息。“双流”的更新公式如(3)和(4)所示。

$$g_{z_t}^{m} \leftarrow \mathrm{Attn}(Q = g_{z_t}^{m-1}, KV = h_{z<t}^{m-1}; \theta), \tag{3}$$

$$h_{z_t}^{m} \leftarrow \mathrm{Attn}(Q = h_{z_t}^{m-1}, KV = h_{z\leqslant t}^{m-1}; \theta), \tag{4}$$

其中,g 为查询隐状态;h 为内容隐状态;m 为 XLNet 的层数;Q 为查询向量 Query;K 为待查向量 Key;V 为内容向量 Value。Q、K、V 通过 Linear 得到其对应的矩阵。完整的双流自注意力机制实现原理如图 3 所示。输入序列是(x_1,x_2,x_3,x_4),采样顺序是 3→2→4→1。图 3 右边 Attention masks 为掩码矩阵,在注意力计算中有意让某些值“看不见”,以此来实现 MASK 的目的。如图 3 所示,模型从下往上计算,h 初始化为 $e(x_i)$,g 初始化为 w,然后根据掩码矩阵进行 Content stream 计算,预测 x_1 时可以看到全部词的信息,预测 x_2 时只能看到 x_2、x_3 的信息,依此类推。

1.3　双向门控循环单元(BiGRU)

GRU 在结构上与 LSTM 大致相似,GRU 由更新门和重置门构成,而 BiGRU 就是将两层 GRU 按照相反的方向堆叠起来,使模型能够从前后两个方向考虑文本的上下文信息。BiGRU 的网络结构如图 4 所示,x_t 为网络在 t 时刻的词向量输入;z_t 和 r_t 分别为更新门和重置门,h_{t-1} 为前一时间步的状态信息,h_t 为当前时间步的状态信息;σ 为 sigmoid 激活函数,tanh 为双曲正切激活函数,$\tilde{h}_t$ 为当前重置门的待定集合。

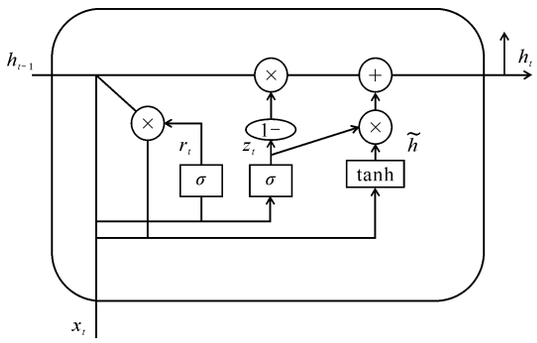

图 4　BiGRU 模型网络结构

Fig. 4　Network structure of BiGRU model

更新门决定上一时刻的状态信息中有多少是要留下的。重置门决定上一时刻的 h 有哪些被加入当前的待定集合 $\tilde{h}_t$ 中。更新门 z_t 和重置门 r_t 的计算公式如下：

$$z_t=\sigma(W_z\cdot[h_{t-1},x_t]), \tag{5}$$

$$r_t=\sigma(W_r\cdot[h_{t-1},x_t]), \tag{6}$$

其中，W_z 和 W_r 为可学习矩阵，x_t 为 t 时刻的输入。

从式(5)和(6)中可以看到，更新门与重置门不仅取决于上一时刻的状态信息，还取决于当前的输入 x。计算出重置门后，用重置门去挑选上一时刻的状态信息加入到待定集合中，当前时刻的待定集合 $\tilde{h}_t$ 计算公式如下：

$$\tilde{h}_t=\tanh(W_h\cdot[r_t\times h_{t-1},x_t]), \tag{7}$$

其中，tanh 为激活函数，W 为权重矩阵。由重置门决定 h_{t-1} 中有多少信息能留下，再结合当前输入 x 计算出本时间步的待定集合。当前时间步的状态信息计算公式如下：

$$h_t=(1-z_t)\times h_{t-1}+z_t\times\tilde{h}_t。 \tag{8}$$

由于网络采用的是双向的 GRU，XLNet 模型的输出会从两个方向输入到 BiGRU 中，所以 BiGRU 会有两个方向相反的输出 h_t。将 BiGRU 正向的最后时刻输出与反向的最初时刻输出拼接起来，作为当前文本的总结向量 H，利用该总结向量即可计算出当前文本的倾向标签词。向量拼接的公式如下：

$$H=[\overrightarrow{h},\overleftarrow{h}]。 \tag{9}$$

将当前文本的 BiGRU 输出值拆分为前向输出和反向输出，再将二者相加作为文本的内容向量 V，输入注意力层。

1.4　结合标签词的注意力机制

结合标签词的注意力机制的目标是将标签词与注意力机制结合，选出文本的倾向标签词，计算倾向标签词与文本向量的注意力得分，根据注意力得分更新文本向量。具体的处理过程如下。

步骤 1　根据 BiGRU 的总结向量 H 挑选出当前文本的倾向标签词，挑选的公式为

$$\text{label\_index}=\text{argmax}(\text{softmax}(\text{Linear}(H))), \tag{10}$$

$$\text{label}=\text{lables}[\text{label\_index}], \tag{11}$$

其中，Linear 为全连接层，argmax 函数的作用是返回当前集合的最大值下标，softmax 为归一化函数。将文本的总结向量输入到 Linear，得到当前文本对应各标签词的分数；再将该分数集输入到 softmax 函数中，得到各标签词的概率，通过 argmax 函数返回最大概率的标签词下标。式(11)中的 labels 为标签词序列，序列长度为标签词个数，如果是五分类任务，那么标签词个数就为 5。将式(10)得出的标签词下标索引出其标签词向量，将该标签词向量作为当前文本的倾向标签词。

本文采用基于 Scaled dot-product attention[8] 的注意力机制，注意力 a 计算公式为

$$a=\text{softmax}(\frac{QK^{\mathrm{T}}}{d}), \tag{12}$$

其中，Q 为查询向量，K 为待查向量，d 为归一化参数。

步骤 2　将当前文本的倾向标签词作为查询向量 Q，将 BiGRU 层的输出向量作为待查向量 K。将输入的文本与倾向标签词进行点乘运算，计算出文本向量对应标签词语义的分值集合，将该分值集合通过 softmax 函数归一化得到文本向量内容的注意力得分，根据注意力得分更新文本的内容向量，更新的公式如下：

$$C=A\otimes V, \tag{13}$$

其中，A 为 batch_size 的注意力得分矩阵，V 为 batch_size 的内容向量，$\otimes$ 为矩阵相乘符号，C 为经过注意力权重更新后的内容向量。将 C 输入至网络的下一层进一步分类输出。结合标签词的注意力机制如图 5 所示。

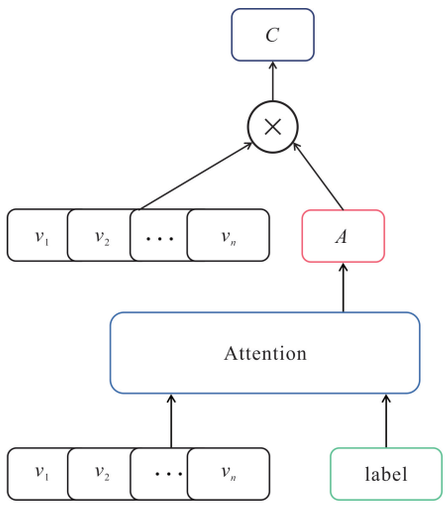

图 5　结合标签词的注意力机制

Fig. 5　Attention mechanism of combined label words

2　实验与分析

2.1　实验环境

本实验在 PC 设备上进行,系统为 Windows 10,GPU 为 RTX 2060 6 G,CPU 为 2.90 GHz,内存为 16 GB,硬盘为 500 GB 固态硬盘。实验开发平台为 Pycharm,所用框架为 Pytorch。

2.2　数据集

在英文数据集 bbc 以及 IMdb 数据集上进行实验。bbc 数据集总共有 2 225 条新闻样本,共分为 5 个类别,标签分别为“Business”“Technology”“Politics”“Entertainment”“Sport”,每个类别各有 445 条样本,在每个类别中挑选 245 条样本作为训练集,200 条样本作为测试集,设置每条样本长度为 510。IMdb 为二分类的影评数据集,标签为“Positive”和“Negative”。从 IMdb 数据集中随机选取 5 000 条样本作为训练集,选出 1 000 条样本作为测试集,其中正、负标签的样本个数相等,设置每条样本长度为 510。

2.3　参数设置

本文模型为 XLNet + BiGRU + Att (Label),设置对比实验模型为 Glove + BiGRU + Att、BERT + BiGRU + Att、XLNet + BiGRU + Att,其中 BERT 与 XLNet 均来自 Huggingface。各模型在 bbc 数据集和 IMdb 数据集中的参数设置见表 1。

表 1　模型参数设置

Table 1　Model parameter settings

模型 Model	批量大小 Batch size	迭代次数 Epoch	学习率 Leaning rate	词向量维度 Dimension of word vector
GloVe + BiGRU + Att	128	10/20	1e-4/1e-3	100
BERT + BiGRU + Att	16	10/6	1e-5	768
XLNet + BiGRU + Att	16	10/6	1e-5	768
XLNet + BiGRU + Att (Label)	16	10/6	1e-5	768

2.4　结果与分析

采用准确率作为评价指标,对比实验的结果见表 2。由表 2 可知,与 XLNet + BiGRU + Att 模型相比,本文模型在 bbc 数据集和 IMdb 数据集上的准确率分别提高 0.7%和 2.64%,说明结合标签词的注意力机制,能够根据标签词更有目的性地挑选出文本向量中的重要部分,给予重要文本内容较高的注意力得分。在 bbc 数据集和 IMdb 数据集上,本文模型的准确率比 BERT + BiGRU + Att 模型分别提高 1.04%与 4.02%,XLNet + BiGRU + Att 的准确率比 BERT + BiGRU + Att 分别提高 0.34%与 1.38%,说明 XLNet 作为词嵌入模型在性能上优于 BERT。这是由于 XLNet 中的编码方式能有效缓解 BERT 模型存在的缺陷,不仅实现了对文本的双向预测,还解决了模型在预训练与下游任务之间的差异性,因此能产生比 BERT 更精确的词向量。各模型的准确率均比 GloVe + BiGRU + Att 高,说明在文本分类任务中,动态词向量比静态词向量拥有更精确的语义表达。根据语境的不同生成不同的词向量表达,将 XLNet 作为预训练语言模型更适合文本分类任务。

表 2　各模型在 bbc 与 IMdb 数据集上的准确率(%)

Table 2　Accuracy of each model on bbc and IMdb dataset (%)

模型 Model	bbc 数据集 bbc dataset	IMdb 数据集 IMdb dataset
GloVe + BiGRU + Att	91.30	83.42
BERT + BiGRU + Att	95.12	87.48
XLNet + BiGRU + Att	95.46	88.86
XLNet + BiGRU + Att (Label)	96.16	91.50

3 结论

本文提出了一种基于 XLNet + BiGRU + Att (Label)的文本分类模型。研究结果表明,使用 XLNet 作为词嵌入模型能够解决静态词向量存在的问题,XLNet 利用 PLM 实现对词序列的双向预测,在注意力计算时引入双流机制实现位置编码的嵌入,克服了以往模型只能单向预测的缺点,因此生成的词向量也就更精确。将标签词与注意力机制结合起来,能够更精确地给文本向量分配权重,突出重点部分,从而提高模型分类的性能。本文的局限在于标签词融于注意力机制,对某些分类任务不适用,如医学疼痛等级分类、在线评论情感分级,在这些分类任务中标签词与文本语义没有关联,发挥不出本文模型的优势。下一步工作将继续研究在语义层面如何利用标签词进一步提升模型的分类性能。

参考文献

[1] 檀莹莹,王俊丽,张超波.基于图卷积神经网络的文本分类方法研究综述[J].计算机科学,2022,49(8):205-216.

[2] YANG Z T,ZHENG J. Research on Chinese text classification based on Word2vec [C]//2016 2nd IEEE International Conference on Computer and Communications (ICCC). Chengdu,China:IEEE,2016:1166-1170.

[3] PENNINGTON J,SOCHER R,MANNING C D. Glove:Global vectors for word representation [C]// Proceeding of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha,Qatar:Association for Computational Linguistics,2014:1532-1543.

[4] 方炯焜,陈平华,廖文雄.结合 GloVe 和 GRU 的文本分类模型[J].计算机工程与应用,2020,56(20):98-103.

[5] 李铮,陈莉,张爽.基于 ELMo 和 Bi-SAN 的中文文本情感分析[J].计算机应用研究,2021,38(8):2303-2307. DOI:10.19734/j.issn.1001-3695.2020.12.0543.

[6] 王宇晗,林民,李艳玲,等.基于 BERT 的嵌入式文本主题模型研究[J/OL].计算机工程与应用,2021[2022-04-12]. http://kns.cnki.net/kcms/detail/11.2127.TP.20211026.1535.010.html.

[7] 赵亚欧,张家重,李贻斌,等.基于 ELMo 和 Transformer 混合模型的情感分析[J].中文信息学报,2021,35(3):115-124.

[8] YANG X Y,YANG L,BI R,et al. A comprehensive verification of transformer in text classification [C]//SUN M S,HUANG X J,JI H,et al. CCL 2019:Chinese Computational Linguistics. Cham, Switzerland: Springer,2019:207-218.

[9] SHI M Y,WANG K X,LI C F. A C-LSTM with word embedding model for news text classification [C]//2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS). [S. l.]: IEEE, 2019: 253-257.

[10] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of deep bidirectional transformers for language understanding [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, USA: Association for Computational Linguistics,2019:4171-4186.

[11] 黄泽民,吴晓鸽,吴迎岗,等.结合 BERT 和 BiSRU-AT 的中文文本情感分类[J].计算机工程与科学,2021,43(9):1668-1675.

[12] YANG Z L,DAI Z H,YANG Y M,et al. XLNet:Generalized autoregressive pretraining for language understanding [C]//33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver,Canada:[s. n.],2019:5753-5763.

[13] 陈立潮,秦杰,陆望东,等.自注意力机制的短文本分类方法[J].计算机工程与设计,2022,43(3):728-734. DOI:10.16208/j.issn1000-7024.2022.03.018.

[14] 杨兴锐,赵寿为,张如学,等.结合自注意力和残差的 BiLSTM_CNN 文本分类模型[J].计算机工程与应用,2022,58(3):172-180.

[15] 叶瀚,孙海春,李欣,等.融合注意力机制与句向量压缩的长文本分类模型[J].数据分析与知识发现,2022,66(6):84-94.

[16] 梁顺攀,豆明明,于洪涛,等.基于混合神经网络的文本分类方法[J].计算机工程与设计,2022,43(2):573-579. DOI:10.16208/j.issn1000-7024.2022.02.038.

[17] 梁淑蓉,谢晓兰,陈基漓,等.基于 XLNet 的情感分析模型[J].科学技术与工程,2021,21(17):7200-7207.

Text Classification Model Based on XLNet + BiGRU + Att (Label)

LIU Boting[1], GUAN Weili[1,2**], LI Taoshen[1,2]
(1. School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, 530004, China; 2. College of Digital Economics, Nanning University, Nanning, Guangxi, 530299, China)

Abstract: Traditional word vector embedding models, such as Word2Vec and GloVe, cannot realize polysemy expression. Traditional text classification models also fail to make good use of the semantic information of label words. Based on this, a classification model based on XLNet + BiGRU + Att (Label) is proposed. Firstly, the dynamic word vector expression of text sequence and label sequence is generated by XLNet. Then, the text vector is input into the Bidirectional Gated Recurrent Unit (BiGRU) to extract the text feature information. At last, the label words are combined with the attention mechanism to select the tendency label words of the text, calculate the attention score of the tendency label words and the text vector, and update the text vector according to the attention score. Through comparative experiments, the model in this paper has higher accuracy than the traditional model in text classification tasks. Using XLNet as the word embedding model and combining label words in attention calculation can improve the classification performance of the model.
Key words: text classification; XLNet; BiGRU; label word; attention mechanism

责任编辑:梁　晓