文章编号:2096-1472(2024)04-0027-06
DOI:10.19644/j.cnki.issn2096-1472.2024.004.006

基于知识蒸馏模型的文本情感分析

李锦辉[1],刘 继[1,2]

(1.新疆财经大学统计与数据科学学院,新疆 乌鲁木齐 830012;
2.新疆财经大学新疆社会经济统计与大数据中心,新疆 乌鲁木齐 830012)
✉1187357069@qq.com; Liuji5000@126.com

摘 要:为了解决预训练语言模型训练时间过长、参数多且难以部署,以及非预训练语言模型分类效果较差的问题,提出了基于知识蒸馏模型的文本情感分析。以预训练深度学习模型(Bidirectional Encoder Representations from Transformers,BERT)作为教师模型,选择双向长短期记忆网络(Bidirectional Long Short-Term Memory,BiLSTM)作为学生模型;在知识蒸馏过程中,将教师模型的 Softmax 层的输出作为"知识"蒸馏给学生模型,并将蒸馏后的模型应用到公共事件网络舆情文本情感分析中。实验结果表明,该模型参数仅为 BERT 模型的 1/13,使 BiLSTM 模型的准确率提升了 2.2 百分点,优于其他同类别轻量级模型,提高了文本情感分析效率。

关键词:知识蒸馏;网络舆情;BERT 模型;BiLSTM 模型
中图分类号:TP391.1 **文献标志码**:A

Text Sentiment Analysis Based on Knowledge Distillation Model

LI Jinhui[1], LIU Ji[1,2]

(1.*School of Statistics & Data Science, Xinjiang University of Finance & Economics, Urumqi* 830012, *China*;
2.*Xinjiang Social & Economic Statistics & Big Data Application Research Center, Xinjiang University of Finance & Economics, Urumqi* 830012, *China*)
✉1187357069@qq.com; Liuji5000@126.com

Abstract: To address the issues of lengthy training time, large number of parameters, and deployment challenges of pre-trained language models, as well as the comparatively poor performance of non-pre-trained language models in sentiment analysis, this paper proposes a text sentiment analysis based on knowledge distillation model. This model utilizes the pre-trained deep learning model, Bidirectional Encoder Representations from Transformers (BERT), as the teacher model and the Bidirectional Long Short-Term Memory (BiLSTM) as the student model. In the process of knowledge distillation, the output of the Softmax layer of the teacher model is distilled as "knowledge" for the student model, which is then applied to text sentiment analysis of online public opinions of public events. Experimental results demonstrate that the proposed model, with parameters only 1/13 of the BERT model, improves the accuracy of the BiLSTM model by 2.2 percentage points; it outperforms other similar lightweight models and enhancing the efficiency of text sentiment analysis.

Key words: knowledge distillation; online public opinions; BERT model; BiLSTM model

0 引言(Introduction)

在大数据和互联网飞速发展的背景下,社交平台(如微博、小红书、Twitter)中涌现了大量情绪化数据,此时舆情文本情感分析在该领域起到了重要作用[1]。现有的文本情感分析模型主要分为预训练语言模型和非预训练语言模型两种。预训练语言模型效果好,但训练时间过长且参数多,不适用于低资

收稿日期:2023-08-11
基金项目:国家自然社科基金资助项目(大数据背景下网络舆情智能治理:共同体构建、协同演进与引导机制,编号:72164034)

www.rjgczz.com

源设备;非预训练语言模型简单易用,但分析效果较差。

针对上述问题,本文提出基于知识蒸馏模型的文本情感分析,将教师模型的输出层作为"知识"蒸馏给学生模型,以便在压缩教师模型的同时,也能提升学生模型的准确率。通过该方法,深入挖掘和分析公众对突发事件的情感倾向,帮助政府部门及时了解公众对突发事件的态度和情感需求,也可以根据情感分析结果对舆情发展趋势进行预判,以此提高网络舆情智能化管理效能。

1 相关研究(Related research)

文本情感分析又称情感倾向分析或意见挖掘,是从用户意见中获取信息的过程。目前,文本情感分析方法主要包括基于机器学习的模型和基于深度学习的模型两种。从机器学习的角度来看,基于机器学习的文本情感分析是通过使用有标注或者无标注的数据,利用传统统计机器学习算法抽取特征,然后进行情感倾向分析。例如,邓君等[2]提出 Word2Vec 和支持向量机(SVM)方法实现了对评论文本进行二分类。随着深度学习技术的发展,深度学习在处理文本信息领域取得了较大的进展。不同的网络搭建方法构成了不同的算法,典型的有卷积神经网络(CNN)、循环神经网络(RNN)和长短期记忆网络(LSTM)等。BEHERA 等[3]提出基于 Co-LSTM 的情感分析方法。BASIRI 等[4]将注意力机制与 CNN-RNN 融合用于文本情感分析。硬件技术的迭代推动了预训练语言模型的快速发展,在自然语言处理领域(NLP)取得重大突破。DEVLIN 等[5]首次提出自编码(AutoEncoder)预训练语言模型 BERT(Bidirectional Encoder Representations from Transformers),该模型提升了 11 项 NLP 任务的技术水平。为了解决单一模型存在的缺陷,有学者进行了模型融合。例如,马长林等[6]提出一种融合主题模型的情感分析算法。ALAYBA 等[7]将 CNN 和 LSTM 进行融合,提出基于 CNN-LSTM 的文本情感分类方法。刘继等[8]提出混合深度学习模型 MBERT-BiLSTM,该模型很好地解决了舆情正负样本失衡的问题。

随着深度学习的发展,自然语言处理领域取得了重大的突破。但在实际应用中,深度学习模型仍然存在诸多挑战。为了获得更高的准确率,模型通常会被设计得庞大而复杂,这就导致模型在训练和部署过程中需要消耗大量资源,因此很难部署在手机等边缘设备上。所以,设计一个具有高性能且满足低资源设备的模型尤为必要。当前,有 5 种方法可以获得高效的深度学习模型,直接设计轻量级网络、剪枝、量化、网络自动设计以及知识蒸馏(Knowledge Distillation,KD),其中知识蒸馏是由 HINTON 等[9]在其 *Distilling the Knowledge in a Neural Network* 论文中首次提出,它作为一种新型的模型压缩方法,目前已经成为深度学习研究领域的一个热点。

知识蒸馏采用教师-学生(Teacher-Student)的训练框架,该方法通常是把复杂的深层网络当作教师模型,浅层的小型网络当作学生模型。在文本情感分类任务中,轻量级网络(如 BiLSTM)的表现通常不佳,但可以利用知识蒸馏原理加强其分类能力。为了在压缩模型的同时能够进一步提升 BiLSTM 模型的文本情感分类能力,本文采用 BERT 作为教师模型,BiLSTM 作为学生模型,提出一种基于知识蒸馏的模型 Distill-BiLSTM 对网络舆情文本进行有效分析。

2 研究方法(Research method)

2.1 知识蒸馏

知识蒸馏是模型压缩中常用的方法之一。通常复杂度越高的模型其分类能力越好,但过大的模型可能存在冗余,训练时会消耗大量的计算时间,因此很难部署在低资源设备上。知识蒸馏以轻微损失模型的准确度为代价,压缩复杂模型。有学者从不同角度分析知识蒸馏的有效性。例如,FURLANELLO 等[10]指出,教师模型的最大 Softmax 概率值可以视为加权重要性,并通过实验表明,即使重新排列所有的非最大 Softmax 概率值也可以提高知识蒸馏的性能。YUAN 等[11]认为,知识蒸馏的成功不仅归功于类间相似性的信息,还归功于标签平滑正则化(Label Smoothing Regularization,LSR),在一些情况下,使用"软标签"的理论推理是正确的。

在蒸馏过程中,教师模型将其掌握的"知识"作为监督信号传递给学生模型,文献[8]将这种知识称为"暗知识"(Dark Knowledge)。学生模型在训练过程中接受这种"知识"提高其准确度,防止过拟合问题,使之接近教师的性能,实现知识的迁移,以此达到压缩模型的目的。蒸馏框架涉及两种标签:软标签(Soft Label)和硬标签(Hard Label)。教师模型经过温度蒸馏得到的概率输出称为软标签;通过 one-hot 方式进行编码的称为硬标签(真实标签)。在训练过程中,相较于硬标签,软标签往往携带更多的"知识"。

定义预测正确的类别概率称为绝对信息(Absolute Information);把非正确预测类别的概率称为相对信息(Relative Information)。模型在训练过程中经过 Softmax 层之后往往会把绝对信息赋予较大的值,把相对信息赋予较小的值,然而相对信息中包含着重要知识。为了平滑两种信息之间的差异性,引入温度系数 ρ 进行调节,通过控制 ρ 放大信息之间的相似性进而确定蒸馏程度。当温度越高,学生模型就越容易从相对信息中获取更多的知识。经过平滑后的概率分布 $q_i(Z_i,\rho)$就被称为软标签,其中 Z_i 表示模型对第 i 个类别的 logits 值,软标签概率的计算公式如下:

$$q_i(z_i;\rho)=\frac{\exp(z_i/\rho)}{\sum_{j=0}^{n}\exp(z_j/\rho)} \tag{1}$$

然而并不是温度越高越好,当温度过高时就会陷入一种平均主义。比如,一张手写数字 7 的图片对应的硬标签值为[1,0,0],将图片输入模型中得到 logits 值为[7,5,3],通过 Softmax 之后,得到软标签值为[0.83,0.12,0.05],模型的输出表明,手写数字 7 特别像数字 2,但特别不像数字 8;当温度为 3 时,通过公式(1)输出为[0.56,0.28,0.16],它们的相对大小就越接近;而当温度为 100 时,通过公式(1)输出为[0.34,0.33,0.33],就体现不出类别之间的差异性。

知识蒸馏损失函数由两个部分组成:一是 *KL* 散度损失函数,使用公式(2)计算;二是 *CE* 交叉熵损失函数,使用公式(3)计算。

www.rjgcz.com

$$KL(q(z^{T};\rho),q(z^{S};\rho))=\sum_{i=1}^{n}q(z_i^{T};\rho)\log\left(\frac{q(z_i^{T};\rho)}{q(z_i^{S};\rho)}\right) \quad (2)$$

$$CE(z^{hard},z^{S})=-\sum_{x}z^{hard}(x)\log(z^{S}(x)) \quad (3)$$

知识蒸馏的总损失函数是这两者的加权和，使用公式(4)表示。

$$L_{KD}=\alpha KL(q(z^{T};\rho),q(z^{S};\rho))+(1-\alpha)CE(z^{hard},z^{S}) \quad (4)$$

其中：z^{S} 为学生模型的 logits 融合输出，z^{hard} 为硬标签(真实标签)，z^{T} 为教师模型的 logits 融合输出，ρ 为温度系数，α 为平衡系数，知识蒸馏模型流程如图 1 所示。

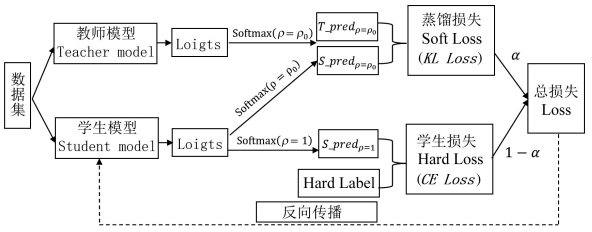

图 1　知识蒸馏模型流程

Fig. 1 Knowledge distillation model process

2.2　教师模型

本文采用大规模中文预训练 Bert-base-Chinese 作为教师模型，大量研究工作已经证明[12-13]，预训练语言模型可以提高许多自然语言处理任务(如文本情感分析、自然语言生成)的性能。预训练是指在大量未带有标签的文本上，以上一个词预测下一个词为目的进行模型训练，这样做的好处是可以使模型学习到每个词元的上下文表示，通过这种方式学到的向量称为词向量。通过训练得到的词向量和模型参数中包含许多在预训练阶段学习到的语义特征。基于预训练的语言模型只需要进行微调(Fine-tuning)，就可以应用到下游任务当中。

BERT 模型用于情感分类的过程如下：首先将单个句子以[CLS]+句子+[SEP]方式进行拼接，其次通过位置编码转换成词向量，将转换的词向量作为 Transformer 的输入进行训练，最后取出经过训练的词向量分类标识[CLS]所对应的向量，传给 Softmax 分类器就可以实现文本分类。BERT 教师模型框架如图 2 所示。

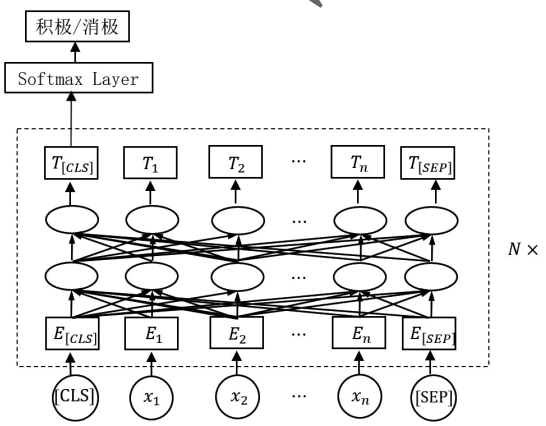

图 2　BERT 教师模型框架

Fig. 2 BERT teacher model framework

2.3　学生模型

本文采用 BiLSTM 作为学生模型，该模型是由一个前向 LSTM 和一个后向 LSTM 构成。在 BiLSTM 学生模型中，对于一个长度为 n 的输入序列 $X=[x_1,x_2,\cdots,x_n]$，其中每个 x_i 表示第 i 个词元的向量，词向量输入前向长短期记忆网络 $\overrightarrow{h}_f$ 的同时在后向长短期记忆网络 $\overleftarrow{h}_b$ 进行反向计算，其 t 时刻的传播公式如公式(5)、公式(6)所示：

$$\overrightarrow{h}_{ft}=\tanh(\overrightarrow{W}_h*\overrightarrow{x}_t+\overrightarrow{W}_h^{t-1}*\overrightarrow{h}_{t-1}+\overrightarrow{b}_h) \quad (5)$$

$$\overleftarrow{h}_{bt}=\tanh(\overleftarrow{W}_h*\overleftarrow{x}_t+\overleftarrow{W}_h^{t+1}*\overleftarrow{h}_{t+1}+\overleftarrow{b}_h) \quad (6)$$

其中，$\overrightarrow{W}_h$ 表示 $\overrightarrow{x}_t$ 的权重矩阵，$\overrightarrow{W}_h^{t-1}$ 表示在 $t-1$ 时刻正向隐藏层 h 的权重矩阵。$\overleftarrow{W}_h$ 表示 $\overleftarrow{x}_t$ 的权重矩阵，$\overleftarrow{W}_h^{t+1}$ 表示 $t+1$ 时刻反向隐藏层的权重矩阵。$\overrightarrow{b}_h$、$\overleftarrow{b}_h$ 分别表示前向隐藏层和后向隐藏层的偏置向量。

拼接前向 LSTM 最后一个时刻的隐藏状态 $\overrightarrow{h}_{fn}$ 和后向 LSTM1 时刻 $\overleftarrow{h}_{b1}$，$\boldsymbol{H}$ 表示 BiLSTM 模型的输出，拼接方式如公式(7)所示：

$$\boldsymbol{H}=Concat[\overrightarrow{h}_{fn};\overleftarrow{h}_{b1}] \quad (7)$$

利用拼接后的词向量 $\boldsymbol{H}$ 的输出作为输出层，然后利用 Softmax 分类器进行分类，这个过程为 BiLSTM 文本情感分类的完整过程。BiLSTM 学生模型框架如图 3 所示。

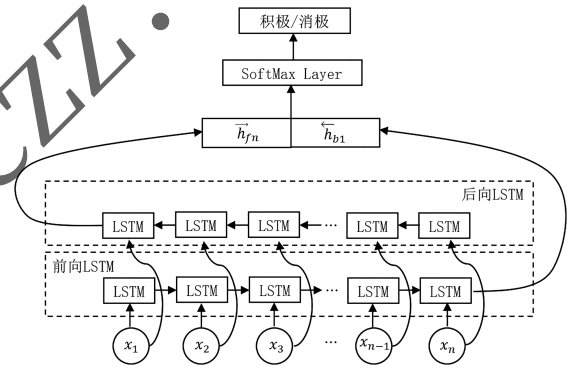

图 3　BiLSTM 学生模型框架

Fig. 3 BiLSTM student model framework

BiLSTM 模型可以同时兼顾前向信息和后向信息，在结合输入文本的语义信息和词性特征的同时，还能获取上下文相关的远期信息，从而有效地解决了由于距离过长而引起的梯度消失和梯度爆炸等问题。

2.4　基于 Distill-BiLSTM 的蒸馏模型

知识蒸馏采用教师-学生模型框架进行训练时，需要分别选择合适的模型作为教师模型和学生模型。通常，复杂但准确度高的作为教师模型；结构简单、容易部署的作为学生模型。文献[9]提出的 BERT 模型在当时 11 项自然语言处理任务中达到最高水平，该模型能够很好地利用文本语义特征充分获取样本信息；BiLSTM 模型结构简单，可以在考虑文本的语义信息和词性特征的基础上获取上下文相关的长时信息。所以，本文采用 Bert-base-Chinese 作为教师模型，BiLSTM 作为学生模型，提出基于 Distill-BiLSTM 的中文文本情感分类模型。此模型通过已经训练完成的 BERT 教师模型的软标签作为监督信号，指导 BiLSTM 学生模型进行训练，通过最小化蒸馏损失，使 BiLSTM 学生模型的性能接近 BERT 教师模型的性能，从而让

www.rgzz.com

学生模型具有更好的泛化性能。BiLSTM 通过利用 BERT 模型的输出以及 BiLSTM 模型中反向传播误差的对应关系模拟 BERT 教师模型的知识。

此模型包括两个步骤：前向计算中，将数据通过位置编码转化为词向量方式输入 BERT 教师模型，将 logits 输出的概率分布作为学生模型的知识；BiLSTM 学生模型将数据通过位置编码转化为词向量方式输入，联合软标签和硬标签对学生模型进行知识蒸馏。本文采用 *KL* 散度作为蒸馏损失函数，并且 BiLSTM 学生模型通过结合软标签、硬标签避免学习到错误的知识。在蒸馏过程中，*KL* 散度损失函数和 *CE* 交叉熵分类损失函数用来计算模型损失。BiLSTM 学生模型通过反向传播计算误差，并通过计算梯度更新模型参数，从而更新 BiLSTM 学生模型的参数，Distill-BiLSTM 模型的损失函数如公式(4)所示。每次蒸馏结束后，将蒸馏损失和学生损失进行加权求和，并将其反馈给 BiLSTM 学生模型，然后利用 BiLSTM 学生模型对文本进行情感分类。

知识蒸馏模型如图 4 所示。

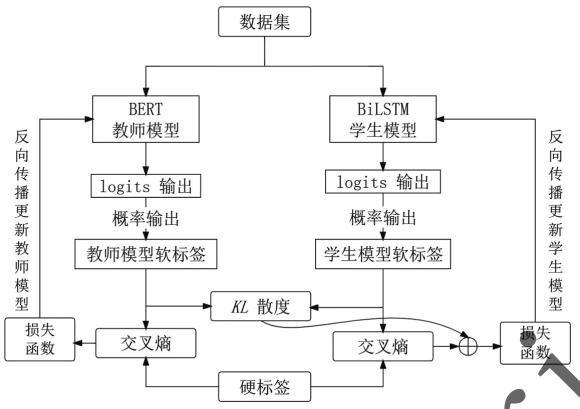

图 4 知识蒸馏模型

Fig. 4 Knowledge distillation model

3 相关实验(Related experiments)

3.1 实验环境配置

本实验在 Mat Pool 平台上操作，使用 Python3. 8 语言进行编程，使用 Pytorh 1. 11. 0 版本深度学习框架搭建模型。实验使用的 GPU 型号为 NVIDIA RTX A4000，显存大小为 16 GB，使用 CUDA 11. 3 进行 GPU 加速。

3.2 实验数据集

使用 Microsoft Edge 浏览器以“疫情”为关键词爬取微博客户端中的相关内容，总共采集 30 434 条数据，数据内容包含发布人名称、发布来源、发布时间、评论文本、点赞数、转发数、评论数。数据自身未带情感倾向标注，为确保数据的有效性，利用 Python 自带的自然语言处理库 SnowNLP 进行预分类。由于自带的 SnowNLP 库是有关电商评论的文本，对于分析本文与关键词“疫情”相关的数据会造成一定的偏差，进而影响模型的准确度。所以，第一步需要训练一个有关“疫情”的语料库，将爬取的数据随机抽取 1 000 条预先进行人工标注，将人工标注情感倾向为“正向”和“负向”分别存入 SnowNLP 库的“pos. txt”文件和“neg. txt”文件中，用以扩充样本的多样性，将模型运行后的“sentiment. marshal”替换原有的 SnowNLP 中的“sentiment. marshal”文件。第二步从爬取的数据中随机抽取不同于第一步选取的 1 000 条数据进行人工标注(标注方式同第一步)，用于测试数据。通过人工标注和机器算法进行对比，使得当模型的准确率达到 80%以上时用于数据分类。第三步是对爬取的数据运用第一步和第二步训练好的 SnowNLP 进行分类，将分类后的结果进行人工调整得到最终数据标签，数据预处理过程如图 5 所示。

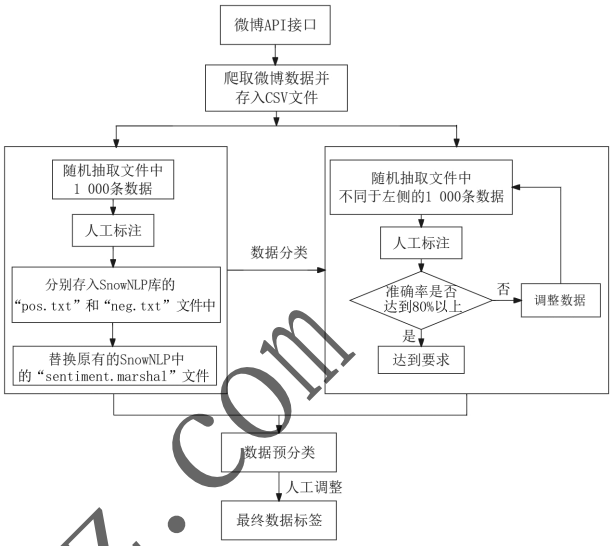

图 5 数据预处理流程

Fig. 5 Data preprocessing process

鉴于本文研究的是二分类问题，选定以 1 表示积极，0 表示消极。数据样本量共 30 434 条，以 8∶1∶1 的方式将数据集划分为训练集、验证集和测试集，即训练数据为 24 347 条，验证数据为 3 043 条，测试数据为 3 044 条，微博评论数据集合如表 1 所示。

表 1 微博评论数据集合

Tab.1 Weibo comment data collection

序号	内容	标签
1	希望如期而至的不只春天，还有疫情过后平安的所有人	1
2	等疫情结束去趟重庆吧明年，明年一定一定一定要去趟重庆去趟洪崖洞，吹吹嘉陵江的晚风	1
3	一酒店违反疫情防控规定被调查一酒店违反疫情防控规定被调查	0
4	这疫情搞得人从此有了囤物癖连咸菜都要囤上十包才放心	0

3.3 评价指标

选取准确率(*Acc*)、精确率(*P*)、召回率(*R*)和 *F*1 值作为评价指标。

$$Acc=\frac{TP+TN}{n} \tag{8}$$

$$P=\frac{TP}{TP+FP} \tag{9}$$

$$R=\frac{TP}{TP+FN} \tag{10}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{11}$$

www.rjgczz.com

其中:*TP* 表示预测为正例且实际为正例,*TN* 表示预测为负例且实际为负例,*FP* 表示预测为正例且实际为负例,*FN* 表示预测为负例且实际为正例,*n* 表示样本量。

3.4 参数设置

教师模型参数设定:采用 Bert-base-Chinese 作为教师模型进行训练,该模型共有 12 个 Transformer 的架构,隐藏层大小为 768,有 12 个多头自注意力,共 1.1 亿个参数,模型的优化方式采用 AdamW,初始学习率为 3e－5,词向量的最大长度为 256,训练轮数(*epoch*)为 10 且批量大小为 32,丢弃率(*Dropout*)默认为 0.1。

学生模型参数设定:采用 BiLSTM 作为学生模型,模型训练时的批量大小(*batch_size*)为 32,训练轮数(*epoch*)为 8,隐藏层维度(*num_hiddens*)为 256,学习率(*lr*)为 0.001,优化器(optimizer)选用 Adam,丢弃率(*Dropout*)选用 0.5。

蒸馏参数设定:利用知识蒸馏技术进行教师-学生架构训练,蒸馏过程涉及蒸馏温度 ρ 和平衡系数 α 两个参数,其中在蒸馏过程中起主导作用的是 α,它表示从教师模型蒸馏知识能力的大小。所以,本文选用不同的平衡系数,以模型结果的准确率为衡量标准,模型的准确率越高,表明从教师模型学习到的知识越多,蒸馏效果越好。

给定蒸馏温度 $\rho=3$,α 的取值为 0～0.9,当 $\alpha=1$ 时,表明知识蒸馏模型为原始的 BiLSTM 模型,所以在此部分不考虑 $\alpha=1$ 的情况。不同 α 值的设定如表 2 所示。

表2　不同 α 值的设定

Tab.2 Setting of different α values

α	准确率/%	α	准确率/%
0	78.94	0.5	78.12
0.1	78.98	0.6	78.06
0.2	80.72	0.7	77.89
0.3	78.42	0.8	77.33
0.4	78.22	0.9	77.04

为了更直观地反映出不同 α 值对模型的准确率的影响,将表 2 中的数据绘制成折线图,如图 6 所示。

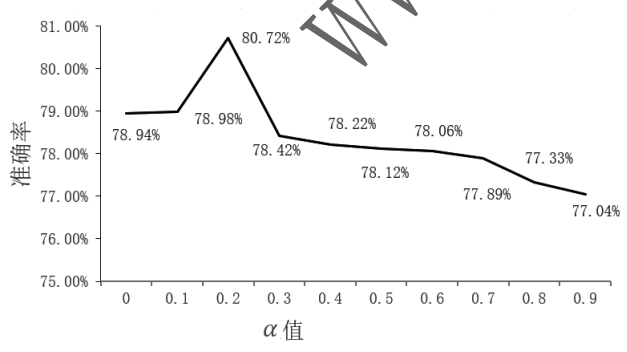

图 6　不同 α 值的折线图

Fig. 6 Line chart with different α values

从图 6 中可以观察到,当 $\alpha=0.2$ 时,模型的准确率达到 80.72%,为所有取值中最大。$\alpha=0.2$ 表明,模型的损失中,有 80%来自蒸馏损失,有 20%来自学生损失,此时学生模型从教师模型中学习到最多的知识。

4　实验结果(Experimental result)

准确率是指模型对整体数据分类结果的准确性进行评估,能够反映分类器对所有类别的分类准确性;精确率是指分类器预测为正类的数据中,真正为正类的占比,反映的是对正样本的区分能力;召回率是指真正为正类别的数据中,分类器预测为正类别的占比,反映的是对正样本的识别能力;*F*1 是综合考虑精确率与召回率的数值,能够更全面地反映模型的分类性能,其值越接近 1,表示分类器的性能越好。

(1)BERT:选用 Bert-base-Chinese 预训练模型,参数为模型原始设定的参数,将文本转化为词向量输入 BERT 模型进行分类。

(2)BiLSTM:由前向 LSTM 模型和后向 LSTM 模型构成,模型共两层,采用全连接层经过 Softmax 进行分类。

(3)GRU:调用 torch 中的 GRU,采用全连接层经过 Softmax 进行分类。

(4)LSTM:调用 torch 中的 LSTM,采用全连接层经过 Softmax 进行分类。

(5)TEXT-CNN:调用 BERT 模型词表经过全局最大池化层,最后一层的全连接的 Softmax 层输出每个类别的概率。

(6)Distill-BiLSTM:选用 BERT 作为教师模型、BiLSTM 作为学生模型进行知识蒸馏。

模型参数量对比如表 3 所示。

表3　模型参数量对比

Tab.3 Comparison of model parameter quantities

模型	参数量/MB
BERT	102.27
BiLSTM	8.04
LSTM	6.46
GRU	6.20
TEXT-CNN	5.54
Distill-BiLSTM	8.04

为了避免实验出现的偶然性,实验中共测试 5 次,再取平均值。模型结果比较如表 4 所示。

表4　模型结果比较

Tab.4 Comparison of model results

模型	*Acc*/%	*P*/%	*R*/%	*F*1/%
BERT	81.76	79.30	83.77	81.48
BiLSTM	78.52	66.67	71.43	68.97
LSTM	77.73	69.23	64.29	66.67
GRU	77.24	69.78	65.23	67.17
TEXT-CNN	76.15	70.93	67.19	67.9
Distill-BiLSTM	80.72	73.33	78.57	75.86

实验使用本文提出的方法将教师模型的“知识”蒸馏到学生模型后,将收集到的数据进行测试,对比模型的评价结果。根据表 3 和表 4,BERT 模型的参数量庞大,约为 102 MB,数据集的分类准确率达到了 81.76%,在所有模型分类能力中性能

www.rlgczz.com

最佳。本文提出的 Distill-BiLSTM 模型相比大规模预训练语言 BERT 模型，准确率只差 1.04 百分点，但是该模型的参数量约为 8 MB，约为 BERT 模型的 1/13；相比于轻量级 BiLSTM 模型，该模型与 BiLSTM 模型的参数量一致，但相较于 BiLSTM 模型准确率、精确率、召回率和 $F1$ 分别提升了 2.20 百分点、6.66 百分点、7.14 百分点和 6.89 百分点。此外，实验还比较了 LSTM、GRU、TEXT-CNN 以及 Distill-BiLSTM 几种模型的性能。结果表明，在同类别轻量级网络中，本文提出的 Distill-BiLSTM 模型的文本分类能力最佳。

不同模型的实验结果如图 7 所示。从图 7 可以观察到，Distill-BiLSTM 模型在同类别轻量级模型中的中文文本情感分类能力最佳。

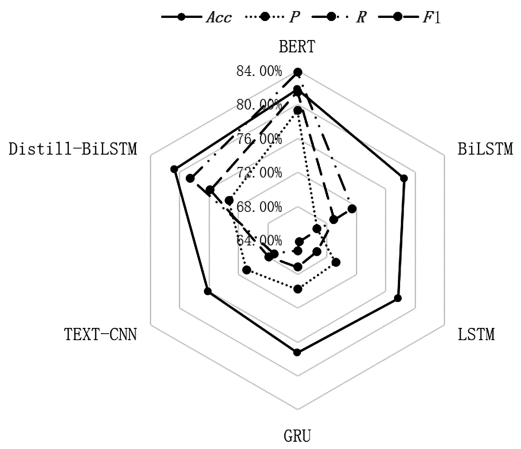

图 7 模型实验结果

Fig. 7 Model experiment results

5 结论(Conclusion)

本文针对大规模预训练语言 BERT 模型训练时间长、计算资源消耗大、难以部署等问题，提出了一种基于 Distill-BiLSTM 的中文文本情感分析模型。将 BERT 模型作为教师模型，使用 BiLSTM 模型作为学生模型，并运用知识蒸馏的思想将 BERT 模型的知识迁移到 BiLSTM 模型，从而实现文本情感分类。结果表明，Distill-BiLSTM 模型与 BERT 模型的分类能力相当，验证了本文提出方法的合理性及有效性。本文提出的模型旨在既能提升轻量级浅层 BiLSTM 模型的中文文本情感分类效果，又能降低 BERT 模型的复杂度和计算开销。在未来的研究中，可以尝试在教师模型中加入领域知识以进一步提高教师模型的分类能力，从而提高学生模型的分类精度。

参考文献(References)

[1] 王春东，张卉，莫秀良，等. 微博情感分析综述[J]. 计算机工程与科学，2022，44(1)：165-175.

[2] 邓君，孙绍丹，王阮，等. 基于 Word2Vec 和 SVM 的微博舆情情感演化分析[J]. 情报理论与实践，2020，43(8)：112-119.

[3] BEHERA R K，JENA M，RATH S K，et al. Co-LSTM：Convolutional LSTM model for sentiment analysis in social big data[J]. Information processing & management，2021，58(1)：102435.

[4] BASIRI M E，NEMATI S，ABDAR M，et al. ABCDM：an attention-based bidirectional CNN-RNN deep model for sentiment analysis[J]. Future generation computer systems，2021，115：279-294.

[5] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[C]//BURSTEIN J，DORAN C，SOLORIO T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，NAACL-HLT 2019. Minneapolis MN USA：ACL，2019：4171-4186.

[6] 马长林，王涛. 基于相关主题模型和多层知识表示的文本情感分析[J]. 郑州大学学报(理学版)，2021，53(4)：30-35.

[7] ALAYBA A M，PALADE V. Leveraging Arabic sentiment classification using an enhanced CNN-LSTM approach and effective Arabic text preparation[J]. Journal of King Saud university-computer and information sciences，2022，34(10)：9710-9722.

[8] 刘继，顾凤云. 基于 BERT 与 BiLSTM 混合方法的网络舆情非平衡文本情感分析[J]. 情报杂志，2022，41(4)：104-110.

[9] HINTON G E，VINYALS O，DEAN J. Distilling the knowledge in a neural network[J]. Computer science，2015，14(7)：38-39.

[10] FURLANELLO T，LIPTON Z C，TSCHANNEN M，et al. Born again neural networks[EB/OL]. (2018-06-29)[2023-03-25]. https://arxiv.org/pdf/1805.04770.pdf.

[11] YUAN L，TAY F E H，LI G，et al. Revisiting knowledge distillation via label smoothing regularization[EB/OL]. (2021-03-04)[2023-03-27]. https://arxiv.org/pdf/1909.11723.pdf.

[12] ZHAO A，YU Y. Knowledge-enabled BERT for aspect-based sentiment analysis[J]. Knowledge-based systems，2021，227(5)：107220.

[13] DABRE R，SHROTRIYA H，KUNCHUKUTTAN A，et al. IndicBART：A pre-trained model for indic natural language generation[EB/OL]. (2022-10-27)[2023-03-29]. https://arxiv.org/pdf/2109.02903.pdf.

作者简介：

李锦辉(1999-)，男，硕士生。研究领域：数据智能分析，知识蒸馏，网络舆情。

刘 继(1974-)，男，博士，教授。研究领域：数据智能分析，网络舆情。

www.rigczz.com