基于BERT与密集复合网络的长文本语义匹配模型

陈岳林[1],高铸成[1],蔡晓东[2]

(1.桂林电子科技大学 机电工程学院,广西 桂林 541000;2.桂林电子科技大学 信息与通信学院,广西 桂林 541000)

摘 要:针对长文本语义匹配中词向量前后之间联系不易捕获以及主题信息可能不唯一,通常使得语义匹配效果不佳的问题,提出了一种基于BERT与密集复合网络的长文本语义匹配方法,通过BERT嵌入与复合网络的密集连接,显著提高了长语义匹配的准确率。首先,将句子对输入BERT预训练模型,通过迭代反馈得到精准的词向量表示,进而得到高质量的句子对语义信息。其次,设计了一种密集复合网络,先由双向长短期记忆网络(Bi-LSTM)获得句子对的全局语义信息,然后由TextCNN提取并整合局部语义信息得到每个句子的关键特征和句子对间的对应关系,并将BERT与Bi-LSTM的隐藏输出与TextCNN的池化输出融合。最后,汇总训练过程中网络之间的关联状态,可以有效防止网络退化和增强模型判断能力。实验结果表明,在社区问题回答(CQA)长文本数据集上,本文方法平均提升幅度达到45%。

关键词:深度学习;长文本语义匹配;BERT;密集复合网络;Bi-LSTM;TextCNN
中图分类号:TP391.1 **文献标志码**:A **文章编号**:1671-5497(2024)01-0232-08
DOI:10.13229/j.cnki.jdxbgxb.20220239

Long text semantic matching model based on BERT and dense composite network

CHEN Yue-lin[1],GAO Zhu-cheng[1],CAI Xiao-dong[2]

(1. *School of Mechanical and Electrical Engineering*, *Guilin University of Electronic Technology*, *Guilin* 541000, *China*;
2. *School of Information and Communication*, *Guilin University of Electronic Technology*, *Guilin* 541000, *China*)

Abstract: In the semantic matching of long texts, it is challenging to capture the before-and-after connections and topic information, which often results in poor semantic matching. This paper proposes a long text semantic matching method based on BERT and dense composite network. Through the dense connection of BERT embedding and composite network, the accuracy of long semantic matching is significantly improved. First, the sentence pair is input into the BERT pre-training model, and accurate word vector representation is obtained through iterative feedback, and then high-quality sentence pair semantic information is obtained. Secondly, a dense composite network is designed. Bi-LSTM first obtains the global semantic information of sentence pairs, and then TextCNN extracts and integrates local semantic information to obtain the key features of each sentence

收稿日期:2022-03-12.
基金项目:广西创新驱动发展专项项目(桂科AA20302001).
作者简介:陈岳林(1963-),男,教授.研究方向:自然语言处理.E-mail:370883566@qq.com
通信作者:蔡晓东(1971-),男,教授,博士.研究方向:图像处理,大数据.E-mail:caixiaodong@guet.edu.com

and the correspondence between sentence pairs, and the BERT Fusion with the hidden output of Bi-LSTM and the pooled output of TextCNN. Finally, summarizing the association state between networks during the training process can effectively prevent network degradation and enhance the model's judgment ability. The experimental results show that on the community question answering (CQA) long text dataset, the method in this paper has a significant effect, with an average improvement of 45%.

Key words: deep learning; long text semantic matching; BERT; dense composite network; Bi-LSTM; TextCNN

0　引　言

文本语义匹配是自然语言处理(NLP)中一项至关重要的任务，在自动问题回答、自然语言推理、释义识别以及剽窃检测等领域有广泛应用。针对该任务提出了各种模型，比如传统的特征工程技术[1]，混合方法[2-4]，纯神经体系架构[5-7]，随着谷歌研究团队发布大型预训练模型BERT[8]，为该任务提供了新的研究思路。BERT作为一种上下文嵌入方式，在信息检索、序列标注、自动问答等方面取得了巨大成功，刷新了多个自然语言任务记录。近年来，国内学者也发布了先进模型，如陈源等[9]的自监督学习。但上述方法均只研究短文本语义匹配，随着人们在应用层面的要求不断提高，开始了对长文本语义匹配任务的不断探索。长文本的突出特点就是长，主要表现在长文本由若干短句组成，匹配难度相对于短文本直线上升。即使是BERT，在长文本语义匹配方面表现也不佳[10,11]。为了解决这个问题，Peinelt等[12]提出了tBERT，即在BERT基础上融入主题模型，将句子对中的相关主题词作为额外信息用于相似度判断，在一定程度上提升了准确率，但对于长文本，不同语境下主题不尽相同，这在一定程度上也加大了语义匹配的难度。最近，相关文献[13,14]提出对所有单词进行特征交互解决相似度度量问题，但在计算所有单词对之间的交互作用时，计算效率低下，也无法对与不同输入源相关的多个重要单词进行预测。

因此，长文本语义匹配任务具有两大挑战：使句子对在语义空间中具有良好的语义表示，以及根据语义表示得出准确的判断。文献[15-19]研究发现，BERT词嵌入具有各向异性，表现为：①词频影响了词向量空间分布，即根据词频区分不同的单词，然后计算不同词频的单词与原点的距离，发现词频越低，距离原点越远；②词频影响词向量空间稀疏性，即词频高的词汇分布的密集，词频低的词汇分布的稀疏。那么在词频低的词汇周围，就会存在很多意义不明的地方，词与词之间的语义空间将存在大量的不确定性，若在度量时所有单词的向量和正好落在定义不明处，则无法进行有效判断。除此之外，Transformer网络架构对长程依赖的捕获能力较差，原因与Transformer中self-attention的计算方式，以及position embedding有关。self-attention计算目标词与所有词之间的联系，无视词与词之间的距离，position embedding是一个硬编码，仅记录了每一个词的固定位置。所以Transformer网络架构无法很好捕获到长文本前后词之间的依赖关系。

因此本文受文献[20,21]的启发，提出了一种基于BERT与密集复合网络的长文本语义匹配模型(Long text semantic matching model based on BERT and dense composite network, BERT-DCN)，在BERT嵌入的基础上，将BERT嵌入表示输入双向长短期网络(Bi-LSTM与)TextCNN组合成的复合网络中，整合从BERT获取到的语义信息，最后将BERT与双向长短期网络(Bi-LSTM)的隐藏输出与TextCNN的池化输出融合，防止网络层数过多导致网络退化，最终进行语义匹配判断。实验结果表明，在社区问题回答(CQA)长文本数据集上，本文方法效果显著，平均提升幅度达到45%。

1　文本语义匹配模型设计

1.1　嵌入层

本文采用BERT预训练模型作为嵌入方式，将X与X'表示两个句子，令：

$$X=\{x_1,x_2,\cdots,x_m\} \tag{1}$$

$$X'=\{x'_1,x'_2,\cdots,x'_n\} \tag{2}$$

式中：$X=\{x_1,x_2,\cdots,x_m\}$与$X'=\{x'_1,x'_2,\cdots,x'_n\}$中的每个元素是组成$X$与$X'$的单词。将$(X,X')$作为嵌入层的输入，通过BERT模型后得到输出。[CLS]与[ESP]是BERT中的两个特殊符号，其中[CLS]作为输入的起始符，经BERT运算后

将作为句子对的隐藏层输出。[ESP]是分隔符,以区分两个不同的句子。这一步可理解为:

$$(T, T') = \mathrm{BERT}(X, X') \tag{3}$$

$$C = \mathrm{BERT}(\mathrm{CLS}) \tag{4}$$

式中:(T, T')是(X, X')经BERT输出后的嵌入表示;C表示BERT嵌入的隐藏输出。

1.2 复合网络编码层

本文利用Bi-LSTM的优势,对BERT的输出作为输入,送入Bi-LSTM进行运算,捕获前后词之间的语义依赖关系,增强语义信息。标准的Bi-LSTM网络主要包含遗忘门、输入门与输出门3个模块,以f_t、i_t、o_t表示,其具体算法如下:

$$f_t = \sigma(\boldsymbol{U}_{fh} h_{t-1} + \boldsymbol{W}_{fx} x_t + \boldsymbol{b}_f) \tag{5}$$

$$i_t = \sigma(\boldsymbol{U}_{ih} h_{t-1} + \boldsymbol{W}_{ix} x_t + \boldsymbol{b}_i) \tag{6}$$

$$o_t = \sigma(\boldsymbol{U}_{oh} h_{t-1} + \boldsymbol{W}_{ox} x_t + \boldsymbol{b}_o) \tag{7}$$

$$\tilde{c}_t = \tanh(\boldsymbol{U}_{\tilde{c}h} h_{t-1} + \boldsymbol{W}_{\tilde{c}h} x_t + \boldsymbol{b}_{\tilde{c}}) \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{9}$$

$$h_t = o_t \odot \tanh(c_t) \tag{10}$$

式中:x_t、o_t、h_t分别表示t时刻的输入、输出以及隐藏层输出;c_t表示当前神经元状态;$\boldsymbol{U}$、$\boldsymbol{W}$表示权重矩阵;$\boldsymbol{b}$表示偏置;σ表示Sigmoid函数;$\odot$表示元素点乘。这一步可以简单理解为:将前一步的BERT输出作为Bi-LSTM的输入,进一步编码得到更为完善的语义表示。即:

$$(S, S'), (\overrightarrow{H}, \overleftarrow{H}) = \mathrm{Bi\_LSTM}(T, T') \tag{11}$$

式中:(T, T')表示1.2节中BERT输出;(S, S')表示经Bi-LSTM后的句子对语义表示;$(\overrightarrow{H}, \overleftarrow{H})$表示Bi-LSTM的隐藏层输出。

在本文任务中,文本视为二维数据(L, D)。其中L代表序列长度,D代表词向量维度。因此TextCNN的输入即句子对的语义向量,然后通过卷积层进行卷积运算,卷积计算方面,可以通过设置多个卷积核(Filters)提取更深层次的语义特征,然后对提取的信息进行池化操作,降低特征维度,最后通过全连接层再进行分类。

TextCNN的输入即Bi-LSTM的输出(S, S'),通过卷积操作后获取到句子对的n-gram信息,更好地提取词之间的语义相关性,运算方法为:

$$\mathrm{Conv} = f((S, S') * \boldsymbol{W} + \boldsymbol{b}) \tag{12}$$

式中:*代表卷积运算;f是激活函数;$\boldsymbol{W}$是权重矩阵;$\boldsymbol{b}$是偏置。在此之后,对卷积结果进行最大池化,进一步提取主要特征:

$$\widetilde{\mathrm{Conv}} = \mathrm{max\,pool}(\mathrm{Conv}) \tag{13}$$

最后将结果导入全连接层,参与最后的分类。

1.3 密集连接预测层

通过BERT嵌入、Bi-LSTM编码、TextCNN卷积运算,得到了最终的特征$\widetilde{\mathrm{Conv}}$,但考虑到网络结构已经很复杂,可能会出现网络退化的问题。借鉴残差网络的思想,将最终TextCNN得到的$\widetilde{\mathrm{Conv}}$特征与初始BERT和Bi-LSTM的隐藏层输出进行融合拼接,使最终的特征信息具有鲁棒性。

$$y_{\mathrm{out}} = \mathrm{cat}[\widetilde{\mathrm{Conv}}; C; (\overrightarrow{H}, \overleftarrow{H})] \tag{14}$$

式中:y_{out}是融合之后的特征表示;cat表示向量拼接,最终经过Softmax层得到每个类的概率分布。

$$y_{\mathrm{pred}} = \mathrm{Soft\,max}(y_{\mathrm{out}} \boldsymbol{W} + \boldsymbol{b}) \tag{15}$$

式中:$\boldsymbol{W}$和$\boldsymbol{b}$分别代表权重和偏置。

图1为模型整体结构图。

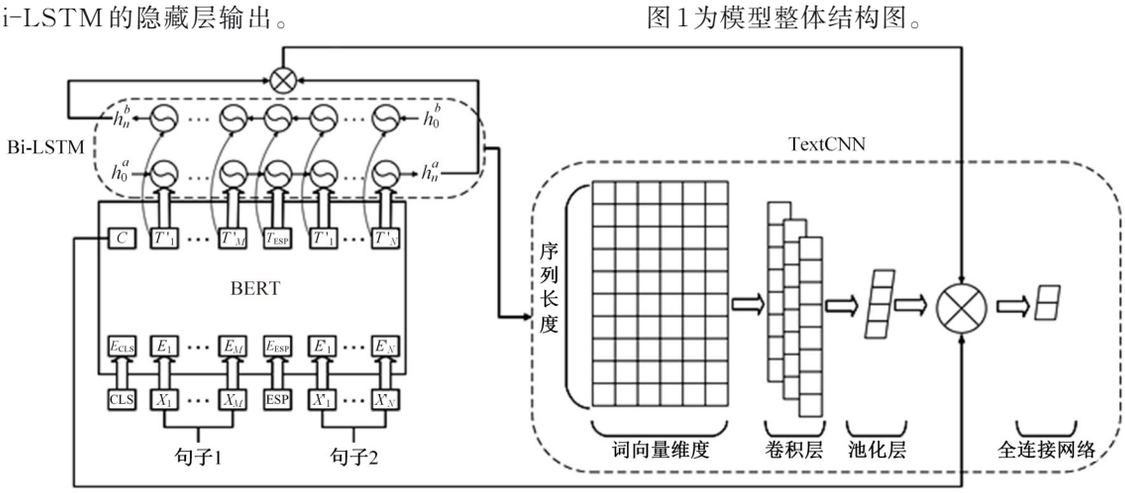

图1 模型整体结构图

Fig. 1 Overall structure of model

2　实验结果与分析

2.1　实验数据与参数设置

为了解决长句子语义匹配问题，采用具有代表性的CQA数据集，这是一个社区问答语义相似性检测数据集，分为A、B、C三个任务。这些任务中每条数据的序列长度基本超过50，较长的序列超过100。数据样例及数据集具体参数见表1和表2，标签0代表两句话语义不匹配，1代表两句话语义匹配。

表1　CQA数据集样例

Table 1　CQA dataset example

句子1	句子2	标签
playing basketball I am basketball player; who lives in umm ghuwlina (doha). I want to play basketball. if people play basketball can contact me to arrange games	New to Doha Hi folks, I'm moving to Doha on the 7th January, just looking for a bit of advice on what sort of things I could do for entertainment while I'm there. I'm interested in trying out new sports and golf, also wondered where the best places to go to watch the English football, also anywhere that would be likely to show Scotland internationals and the 6 nations rugby? Cheers	0
playing basketball I am basketball player; who lives in umm ghuwlina (doha). i want to play basketball. if people play basketball can contact me to arrange games	Basketball Ok so we're talked about this before but didn't seem to find a place to play. Good news guys! I found a court to play on :－) It's in Education City, anyone who doesn't know where that is no worries I'll give u directions or we can meet up elsewhere and u can follow me. Anyways, I just need to know who's in so I can book it cause it can get really busy soon so I need to do this in advance...so who's in? P.S. don't ask what I'm doing up at this hour! I don't even know!	1

表2　CQA数据集参数

Table 2　CQA dataset parameters

数据集	CQA		
	SemEval-A	SemEval-B	SemEval-C
训练集	20 340	3 169	31 690
验证集	3 720	700	7 000
测试集	2 930	880	8 800

模型使用Pytorch深度学习框架搭建，在NVIDIA 2080TI GPU上训练，系统版本为Ubuntu16.04。嵌入层采用Pytorch版的BERT-base模型，Bi-LSTM中，设置隐藏层维度为600，TextCNN中，设置filter_size为[1-10]，filter_map为200。总体输入的batch为64，dropout设置为0.2，为了保证输入的序列长度的一致性，方便GPU运算，将序列长度的阈值设置为256，长度小于256的自动填充补齐，大于256的截取前256个字符。

2.2　评价指标

对实验结果的评价标准主要用到F_1分数，该指标又与精确率(Precision)、召回率(Recall)有关，具体的评价算法如下：

$$F_1=\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{16}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{17}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{18}$$

式中：TP为被模型预测为正类的正样本；TN为被模型预测为负类的负样本；FP为被模型预测为正类的负样本；FN为被模型预测为负类的正样本。

2.3　实验结果

首先针对CQA数据集进行测试，与各基线模型做对比，由于设置了随机种子数，每轮训练的结果都会有所不同，因此采用多轮训练求平均的方法评价其性能，每轮训练的结果如图2～4所示。

因此，A任务的平均F_1分数为0.750，B任务与C任务分别为0.587、0.612。与基线模型的对比如表3所示。

对于任务A，问答内容来自于同一个域，句子对之间文本重合度相对较高，有利于语义匹配识别，本文方法的性能相较于基线模型处于同一水平；任务B是对问题的释义，可以理解为句子2是

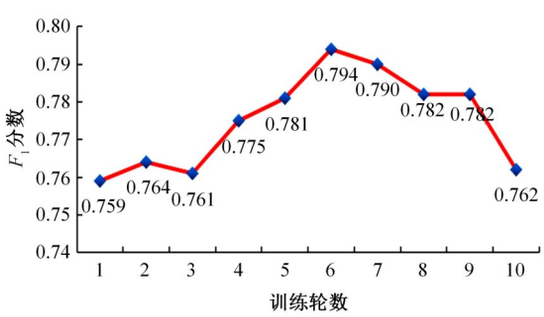

图2　A任务训练结果

Fig. 2　A task training result

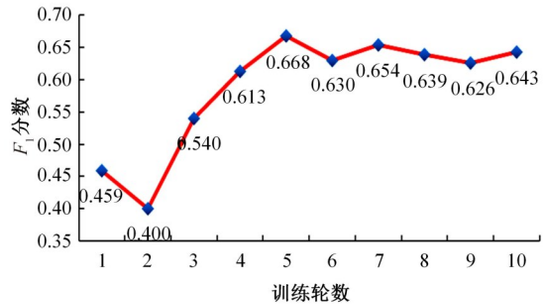

图3 B任务训练结果

Fig. 3 B task training result

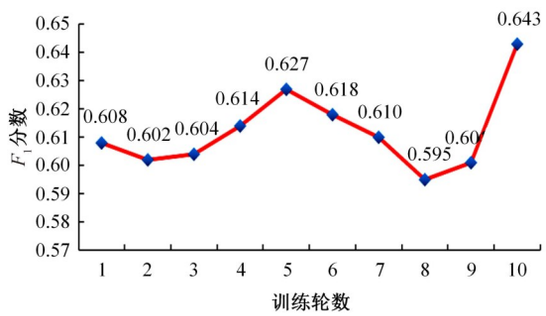

图4 C任务训练结果

Fig. 4 C task training result

表3 CQA数据集上实验结果对比

Table 3 Comparison of experimental results on CQA dataset

模型	F_1		
	SemEval-A	SemEval-B	SemEval-C
文献[1]	—	0.506	—
文献[2]	0.777	—	—
文献[4]	—	—	0.197
文献[7]	—	0.433	—
文献[12]	0.768	0.524	0.273
BERT-DCN	0.750	0.587	0.612

对句子1的转述,转述方式不同导致文本单词之间差异较大,提升0.063,提升幅度约为12%;对于任务C,它与任务A相似,不同点在于问答内容并非来自同一个域,其回答的内容是从外部域中引入,这增加了问题的难度,该任务中本文方法与最新方法相比提升0.339,提升幅度达到了124%。

分析表3可知,Peinelt等[12]提出的tBERT模型,仅以BERT作为基准嵌入与主题模型想融合,无法很好地解决长文本释义问题以及不同域之间的问答匹配,原因在于此类任务对大量关键词进行重新表达,加之文本序列过长,造成匹配难度增大。正好印证BERT词嵌入具有各向异性,嵌入空间在语义上不平滑,在某些领域定义不好以及BERT长序列依赖能力较差等问题。同时对于长文本,主题信息可能不唯一,也造成了匹配困难。本文提出的密集复合网络是一种针对性的方法解决此类问题,因此在该任务数据集上具有显著效果。

2.4 实验分析

2.4.1 消融研究

本文在BERT的基础上,通过以Bi-LSTM和TextCNN组成的复合网络,并借鉴残差网络的思想将各个网络的状态连接起来,参与匹配计算,最终得到了完整的语义匹配模型。本研究针对提升效果最大的Semeval-C数据集进行消融实验,以验证本文模型的有效性。将训练epoch固定为5,其余实验参数同2.1节,实验结果如表4所示。

表4 消融实验结果

Table 4 Ablation experiment results

模型	F_1
BERT-DCN	0.627
-Bi-LSTM	0.620
-TextCNN	0.593
-DCN	0.610

去掉Bi-LSTM网络后,模型退化为BERT-CNN,失去长依赖信息后,性能有所下降。去掉TextCNN,模型退化为BERT-Bi-LSTM,模型性能下降程度很大,证明TextCNN具有强大的分类判断能力。最后去掉完整DCN模块,即模型退化为基础BERT,此时性能同样不佳,但好于Bi-LSTM部分。因此可以得出结论,TextCNN在密集复合网络中起到了决定性的判断作用,虽然单一使用Bi-LSTM效果不佳,但与TextCNN融合后,能提升整体模型效果。

2.4.2 训练样本改进分析

在深度学习领域,算法和数据是影响结果的两个最重要因素,随着研究的深入,本文也对数据样本进行了相关探讨。在研究了先前测试中未匹配成功的案例,本文发现该数据集训练样本具有一定的缺陷,代表样例如表5所示。

该样例标签为1,表示语义匹配,但经多名研究人员评估,虽然两个句子都在讨论异族婚姻的主题,但句子1是询问某人异族婚姻的经历,而句子2谈论对异族婚姻的看法,因此本文认为两个句子不匹配。此类样例在数据集中存在一定比例,影响模型的判别性能,因此本文对数据集进行了清洗,对此类样例的标签进行重新标定,改动结果如表6所示。

本文经过对原始数据的标签进行重新标定,

表5 CQA数据集中有缺陷的样例

Table 5　Defective examples in the CQA dataset

句子1	句子2	标签
Mixed marriages with the world turning into a small village; mixed marriages have become more and more common. If you are in a mixed marriage relationship and living in Doha; Could you provide details on how you and your significant other met? where? how long have you been married for? and what are the goods and bads in a mixed marriage? Thank you (异族婚姻将世界变成一个小村庄;异族通婚变得越来越普遍。如果您处于异族婚姻关系并居住在多哈;您能否详细说明您和您的另一半是如何认识的?在哪里?你结婚多久了?异族婚姻有什么好处和坏处?谢谢)	Mix marriages do you think that 2 from 2 different cultures shall have a good marriage i mean will it last; (believes; customs; even language) not the same! i'm not talking about 2 from 2 different countries in Europe; or 2 in Asia; or 2 in Africa; ... i'm talking about the one's where completely no similarity? (异族婚姻你认为来自2种不同文化的2人应该有一段美好的婚姻吗?我的意思是这会持续下去吗?(相信;习俗;甚至语言)不一样!我不是在谈论来自欧洲2个不同国家的2个;或2个在亚洲;或2个在非洲;...我说的是完全没有相似之处的那个?)	1

表6　清洗后的数据

Table 6　Cleaned dataset

数据集(修改量/样本总量)	CQA		
	SemEval-A	SemEval-B	SemEval-C
训练集	382/20 340	339/3 169	630/31 690
验证集	129/3 720	32/700	180/7 000
测试集	91/2 930	24/880	55/8 800

对于A任务,总样本为26 990,标签重新标定的样本数为602,因此修改比例约为2.2%,同理B任务与C任务的修改比例分别为8.3%、2.6%。然后根据修改后的数据集,再次根据相同实验设置进行实验,结果如表7所示。

表7　修改样本后实验结果对比

Table 7　Comparison of experimental results after modifying the sample

数据集	F_1		
	SemEval-A	SemEval-B	SemEval-C
初始数据	0.750	0.587	0.612
修改后	0.768	0.607	0.622

根据实验结果,A、B、C三个任务相较于未改动前均有所提升,提升幅度约为2.4%、3.4%、1.6%。此次实验证明了以下两点:①本文所提出的模型确实具有良好的性能,通过对数据的清洗,可以进一步提升实验的结果;②对于语义类任务,数据的好坏尤其关键,语义从某种程度上具有一定主观性,对于不同的人评价语义可能有不同的结果,因此对于语义匹配任务,数据方面要切合自己本身的语义理解,才能得到有效的语义模型。

2.4.3　泛化能力分析

本文也挑选了Chinese-SNLI和LCQMC两个较为经典的文本语义匹配数据集验证本文模型的泛化能力,两个数据集的具体参数如表8所示。

表8　Chinese-SNLI与Quora数据集参数

Table 8　Chinese-SNLI and Quora dataset parameters

数据集	Chinese-SNLI	LCQMC
训练集	545 859	238 766
验证集	9 314	8 802
测试集	9 176	12 500

实验的参数设置沿用2.1小节,将epoch均设置为5进行实验,实验结果如表9和表10所示。

表9　Chinese-SNLI数据集实验结果

Table 9　Chinese-SNLI dataset experimental results

模型	Acc/%
Embed+add-attention*	75.1
BiLSTM+self-attention*	81.0
DiSAN*	81.5
BERT*	87.0
BERT-DCN	87.2

注:*表示引用已有的模型在实验数据集上进行实验。

表10　LCQMC数据集实验结果

Table 10　LCQMC dataset experimental results

模型	Acc/%
CNN*	72.8
CBOW*	73.7
BiMPM*	83.4
BERT*	87.4
BERT-DCN	87.7

注:*表示引用已有的模型在实验数据集上进行实验。

对于上述两个文本语义匹配数据集,本文采用准确率(Acc)作为评价指标。显然,本文方法相较于现有优秀模型仍有不错的效果,在数据集Chinese-SNLI和LCQMC上,提升0.2%和0.3%。

综合来看,本文方法在长文本语义匹配任务中效果突出,同时在经典语义匹配数据集上,也能展现出较好的效果。

3 结束语

本文提出了一种基于BERT与密集复合网络的文本语义匹配方法,该方法利用BERT作为词嵌入表示,经Bi-LSTM编码,捕获各词向量之间的前后依赖关系,弥补Transformer网络架构无法很好捕获到长句子前后词之间的依赖关系这一缺点,然后利用TextCNN捕捉句子中的局部关键特征,最后参与分类运算。实验结果表明,得益于BERT优越的词嵌入表示以及密集复合网络的特征提取能力,本文方法在对长文本语义匹配任务具有显著效果,运用于短文本语义匹配时性能也相当优越。

参考文献:

[1] Simone Filice1, Giovanni Da San Martinoao, Alessandro Moschitti, et al. SemEval-2017 task 3-learning pairwise patterns in community question answering [C]//Proceedings of the 11th International Workshop on Semantic Evaluation, Vancouver, Canada, 2017: 326-333.

[2] Wu Guo-shun, Sheng Yi-xuan, Lan Man, et al. Using traditional and deep learning methods to address community question answering task[C]//Proceedings of the 11th International Workshop on Semantic Evaluation, Vancouver, Canada, 2017: 365-369.

[3] Feng Wen-zheng, Wu Yu, Wu Wei, et al. Ranking system with neural matching features for community question answering[C]//Proceedings of the 11th International Workshop on Semantic Evaluation, Vancouver, Canada, 2017: 280-286.

[4] Yuta Koreeda, Takuya Hashito, Yoshiki Niwa, et al. Combination of neural similarity features and comment plausibility features[C]//Proceedings of the 11th International Workshop on Semantic Evaluation, Vancouver, Canada, 2017: 353-359.

[5] Wang Zhi-guo, Hamza Wael, Florian Radu. Bilateral multi-perspective matching for natural language sentences[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 2017: 4144-4150.

[6] Tan Chuan-qi, Wei Fu-ru, Wang Wen-hui, et al. Multiway attention networks for modeling sentence pairs[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018: 4411-4417.

[7] Jan Milan Deriu, Mark Cieliebak. Attention-based convolutional neural network for community question answering[C]//Proceedings of the 11th International Workshop on Semantic Evaluation, Vancouver, Canada, 2017: 334-338.

[8] Devlin Jacob, Chang Ming-wei, Lee Kenton, 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, Minneapolis, USA, 2019: 4171-4186.

[9] 陈源,丘心颖. 结合自监督学习的多任务文本语义匹配方法[J]. 北京大学学报:自然科学版, 2022, 58(1): 83-90.
Chen Yuan, Qiu Xin-ying. Multi-task semantic matching with self-supervised learning[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2022, 58(1): 83-90.

[10] Nils Reimers, Iryna Gurevych. Sentence BERT: Sentence embeddings using siamese BERTnetworks [C]//Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, China, 2019: 3982-3992.

[11] Li Bo-han, Zhou Hao, He Jun-xian, et al. On the sentence embeddings from pre-trained language models[C]//The Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 2020: 9119-9130.

[12] Peinelt N, Nguyen D, Liakata M. tBERT: topic models and BERT joining forces for semantic similarity detection[C]//Proceedings of the 58rd Annual Meeting of the Association for Computational Linguistics, Tokyo, Japan, 2020: 7047-7055.

[13] Chen Han-jie, Zheng Guang-tao, Ji Yang-feng. Generating hierarchical explanations on text classification via feature interaction detection[C]//Proceedings of the 58rd Annual Meeting of the Association for Computational Linguistics, Tokyo, Japan, 2020: 5578-5593.

[14] Tsang M, Cheng D, Liu H, et al. Feature interaction interpretability: a case for explaining ad-recommendation systems via neural interaction detection[J/OL]. [2020-12-11]. www.arXiv:2006.10966.

[15] Gao J, He D, Tan X, et al. Representation degeneration problem in training natural language generation models[J/OL]. [2019-12-10]. www.arXiv:1907.12009.

[16] Ethayarajh K. How contextual are contextualized word representations?[C]//Conference on Empirical Methods in Natural Language Processing and the 9th

International Joint Conference on Natural Language Processing, Hong Kong, China, 2019, 55-65.

[17] Yan Yuan-meng, Li Ru-mei, Wang Si-rui , et al. ConSERT: a contrastive framework for self-supervised sentence representation transfer[C]//Association for Computational Linguistics and International Joint Conference on Natural Language Processing, Online, 2021: 5065 - 5075.

[18] Schick T, Schütze H. Generating datasets with pretrained language models[J/OL][2021-10-21]. www.arXiv:2104.07540.

[19] Chen Han-jie, Song Feng, Jatin Ganhotra, et al. Explaining neural network predictions on sentence pairs via learning word-group masks[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, Mexico City, Mexico, 2021: 3917-3930.

[20] Balazs Jorge A, Matsuo Y. Gating mechanisms for combining character and word-level word representations: an empirical study[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, USA, 2019: 110-124.

[21] Choi H, Kim J, Joe S, et al. Evaluation of BERT and albert sentence embedding performance on downstream NLP tasks[C]//The 25th International Conference on Pattern Recognition, Online, 2021: 5482-5487.